Semantic Web-Based Integration of Heterogeneous Web Resources

Elaheh Momeni
University of Vienna-Faculty of Computer Science
Liebiggasse 4/3-4, 1010 Vienna, Austria
elaheh.momeni.roochi@univie.ac.at

Abstract

Vast volumes of information from public Web portals are readily accessible from virtually any computer in the world. This can be seen as an enormous repository of information which brings significant business value for companies working in e-commerce activities. However, the main problems encountered when using this information are: (I) the information is published in various, non-machine-processable formats, (II) a lack of services that match and store information from various sources in a homogenous structure, and (III) the accessible datasets are rarely provided with e-commerce concepts in mind. These problems make them difficult to use by e-commerce applications. The main goal of this paper is to propose a methodology and analysis of components required for combining and integrating information into machine-processable dataset from different Web data sources, based on suitable e-commerce ontology. In order to demonstrate proposed methodology, the process of wrapping and matching the data from two public datasets will be discussed as an example.

Introduction

Web information, which is provided permanently by different Web portals, represents an enormous repository of globally distributed information. Accessing this repository can be a significant business value for companies which work in e-commerce. Therefore, e-commerce-oriented companies are very interested in finding ways to access relevant Web information and then to apply it in their everyday business in order to improve the output of their existing application infrastructure and increase their competitive advantage.

Within this framework, Semantic Web technology presents a common format for the combination and integration of data drawn from heterogeneous sources. It will enable machines to process data easily on a global scale. The essential aim of the Semantic Web vision is to make Web information practically processable by any computer in the world. This increase in effectiveness is constituted by locating, collecting and cross-relating content from more than one separate source. Furthermore, the goal of LOD-Linked Open Data (Bizer, Heath, and Berners-Lee 2009) is to enable people to share structured data on the Web as easily as they can share documents today, allowing the re-use and combination of all published data.

But the major problems of business applications based on Semantic Web technology and, in particular, the LOD movement are: (1) Information on the Web is published in many different formats (HTML, CSV, Text, etc). And, in addition, available information rarely is provided in machine-processable format. (2) A lack of services that match, link and store information from various sources in a unique processable structure. (3) Accessible datasets are rarely provided with e-commerce concepts (e.g., related information about offers, business entities, and price of one product/service).

These scenarios can be seen in the lack of a complete machine-processable, e-commerce-based repository for consumer electronics. For example, the specification of a Digital-Camera, the “Nikon-D70”, is listed by different sources, such as “ICEcat”[1] and “Wikipedia”[2]. By comparing these two specifications the following can be observed: First, we do not have any comprehensive access to all this information, for instance the property “Continuous shooting” is listed by “Wikipedia”, but it is not listed by the “ICEcat” specification. However, “ICEcat” lists many properties which are not listed by “Wikipedia”. Second, this information is published in heterogeneous formats (e.g., HTML, XML). Third, this information is available without any clear e-commerce background (for instance we do not have related information about offers, business entities and price of this Digital-Camera).

Therefore, the purpose of this paper is to propose a methodology for the creation of a complete machine-processable dataset from heterogeneous sources based on suitable e-commerce ontology. Furthermore, the requirements for generating such a dataset are discussed like an e-commerce model.

In the following, we will describe the detail of proposed methodology and the e-commerce model. In Section “Methodology Evaluation”, in order to demonstrate proposed methodology, the process of wrapping and matching the data from two public datasets will be discussed as an example. Section “Related Work” briefly overviews re-

[1] http://icecat.co.uk/

[2] www.wikipedia.org/

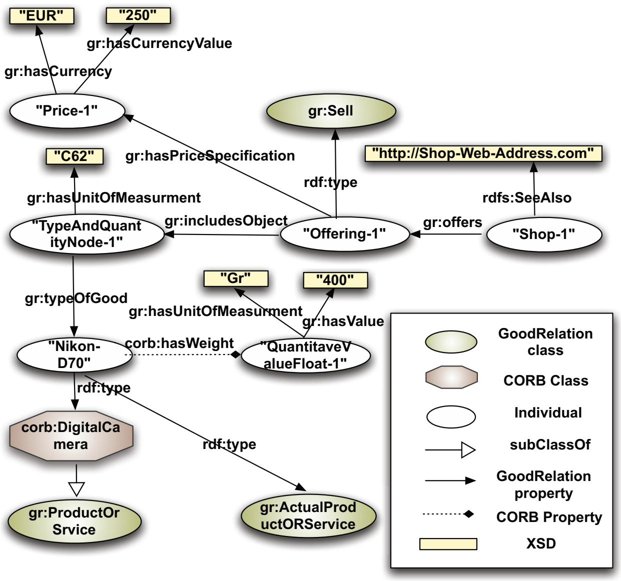


Figure 1: An Instantiated of the E-commerce Model

lated work and finally, "Summary" Section concludes the discussion and provides an outlook on possible area of future work.

Proposed Methodology and E-commerce Model

With regard to the problems mentioned in the previous section, a methodology is required to define the workflow of a machine-processable, e-commerce-based RDF dataset creation process and the necessary building blocks. However, as a prerequisite to this process a model is required to introduce the necessary entities for the creation of such a dataset.

E-commerce-based Model

To provide a comprehensive e-commerce-based RDF dataset, an e-commerce ontology is required that allows the properties of a specific category of product/service to be described and the relationships between offerings made by means of those products/services, legal entities, and prices to be defined (Hepp 2005). For example, we must be able to define that a particular Web site describes an offer to sell a Digital-Camera of a certain model at a certain price, describing the specific properties of this Digital-Camera like the quantity value of its Weight or Dimensions. Furthermore, by surveying the e-commerce based ontology related work, GoodRelation ontology (Hepp 2008) is selected and extended to model our required entities. GoodRelations is defined as: *"a lightweight yet sophisticated vocabulary that allows manufacturers and shop operators to express the exact meaning of their offers made on the Web in a machine-procceasble way."* (Hepp 2008).

An individual ontology can be created for each different product/service category based on the GoodRelation ontology. Therefore, each RDF instance is created based on its category's ontology. Figure 1 shows and instantiated of the model for Digital-Camera category. For the "DigitalCamera" category a subclass of "ProductOrService" (a class of the GoodRelation ontology) is created. Furthermore, the specific properties of each category (e.g., "hasWeight") are described comprehensively. Describing the properties comprehensively plays an important role in matching the properties from different sources.

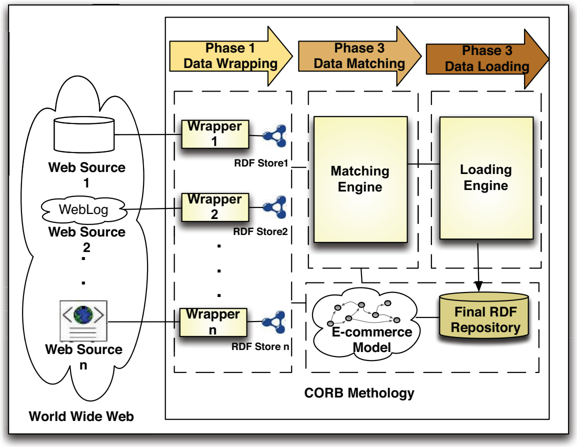


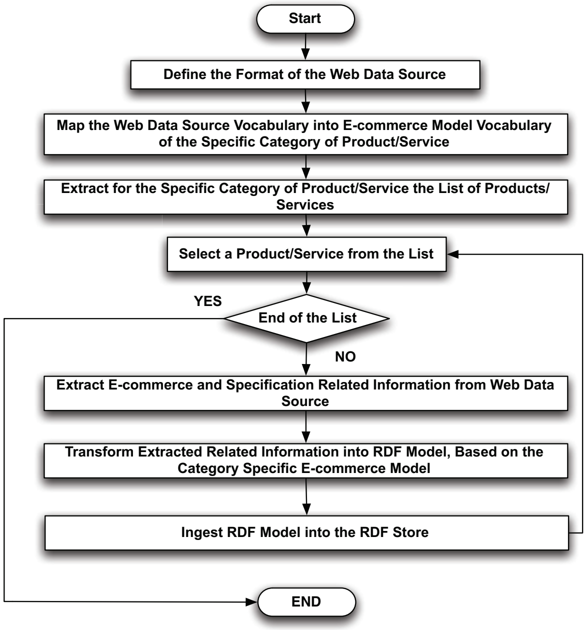
Figure 2: Overview of the CORB Methodology

Methodology

As we can see in Figure 2, CORB (Comprehensive Ontology-based Repository for Business-application) methodology proposes the following phases for the creation of such a comprehensive dataset:

Data Wrapping: In the first phase, CORB proposes wrapping data from various sources into a RDF format. Each wrapper has a different architecture and different components, according to its source. The main building blocks of each wrapper are:

- An extract engine: this component is responsible for searching for required data in a Web source, e.g., finding all the available data about a particular category of product/service, and loading this data to a defined destination. The internal structure of this engine is completely dependent on the structure of the Web data source. Many Web portals offer easy access to their data via an API. In other cases, this component can be selected from available Web data extraction solutions, which are overviewed in the "Related Works" Section. Furthermore, many available extract engines offer the extraction and loading of data into the various formats.
- A transform engine: this engine, by using transformation language (e.g., XSLT), is responsible for: mapping the Web Data Source Vocabulary to the specific extended e-commerce model Vocabulary (e.g., if a property, such as "hasWeight" for "DigitalCamera" category, is defined in the specific extended e-commerce model, and "ICEcat" defines this property as "Weight" via user interface



Start
Define the Format of the Web Data Source
Map the Web Data Source Vocabulary into E-commerce Model Vocabulary of the Specific Category of Product/Service
Extract for the Specific Category of Product/Service the List of Products/ Services
Select a Product/Service from the List
YES
End of the List
NO
Extract E-commerce and Specification Related Information from Web Data Source
Transform Extracted Related Information into RDF Model, Based on the Category Specific E-commerce Model
Ingest RDF Model into the RDF Store
END

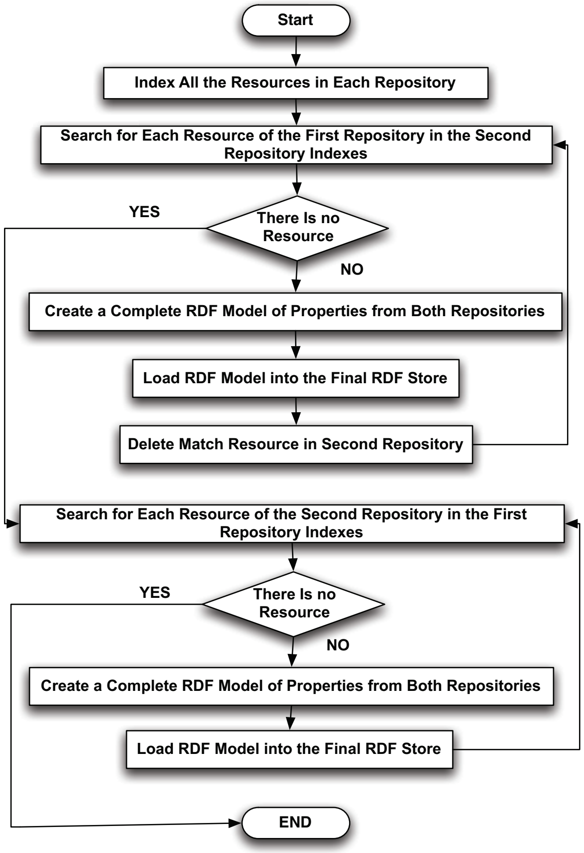
Figure 3: Internal Process of the Data Wrapping

interaction, as a result the mapping between properties descriptions is defined and saved in the transformation script), changing the data format to RDF, and standardizing it (e.g., creates a valid triple according to “Digital-Camera” model with a valid URI).

- An ingest engine: this engine loads all the created RDF instances from one source into a single RDF repository.

Figure 3 shows the internal process of this phase. In the case of wrapping information from HTML Web data sources, using available solutions instead of specially designed wrappers is more effective. But in most cases these tools wrap the data into XML or other formats, for which further steps are needed to transform them into a standard RDF format.

Data Matching: In the second phase, matching product/service resources from provided RDF-based repositories of different sources (provided in the previous phase) and completing the list of their properties and e-commerce relevant information is proposed. Therefore, product/service resources in each repository are first indexed (e.g., by using Apache Lucene). Second, for each resource in the first repository an index is searched for. The one to be selected is the one with the highest match rank in the second repository (This is based on the assumption that either the highest search rank is exactly the same product/service or there is no match at all). In order to ensure improved accuracy in the matching process between the product/service descriptions, it is clear that the result of the selection process in indexes needs to be carried out accurately in the index searching process. However, this issue is beyond the scope of this paper and will be discussed in a later work. Finally, for each product/service the list of the relevant properties and e-commerce entities (which are all available in the e-commerce model) are gathered from the appropriate resources of each RDF repository. Figure 4 shows the internal process of this phase.



Start
Index All the Resources in Each Repository
Search for Each Resource of the First Repository in the Second Repository Indexes
YES
There Is no Resource
NO
Create a Complete RDF Model of Properties from Both Repositories
Load RDF Model into the Final RDF Store
Delete Match Resource in Second Repository
Search for Each Resource of the Second Repository in the First Repository Indexes
YES
There Is no Resource
NO
Create a Complete RDF Model of Properties from Both Repositories
Load RDF Model into the Final RDF Store
END

Figure 4: Internal Process of the Data Matching

Data Loading: After finding the best matches, CORB proposes that all created triples be loaded into a single RDF repository in the third phase. Before loading the information, all the prepared information needs to be normalized and standardized.

Methodology Evaluation

In order to demonstrate CORB methodology, we created an RDF-based repository from 612 different Camcorder models from “ICEcat”[3] (free open catalogue is available in xml format) and “DBpedia”[4] (structured information from Wikipedia which is available in RDF format). We extended GoodRelation ontology by defining the Camcorder category specific properties. In the “Data Wrapping” phase: First, we wrote a script to extract the information on each individual camcorder from “ICEcat” as an XML format (e-commerce and specification related information for each product). Sec-

[3] http://icecat.co.uk//en/menu/services/index.htm

[4] http://dbpedia.org/About

ond, we provided an XSLT transformation script which mapped the Camcorder specification vocabulary of "ICEcat" into our Camcorder model vocabularies and transformed the information format into RDF format. Third, we ingested all this converted data into an "ICEcat" RDF repository. Due to the availability of "DBpedia" resources as RDF, we did not run the wrapping process for this source. However, we wrote a script, which mapped the Camcorder specification vocabulary of "DBpedia" into our Camcorder model vocabulary.

In the "Data Matching" phase, we first, indexed all the "DBpedia" resources which have "category:Camcorders" as "skos:subject" property. Second, we listed all the products of the "ICEcat" RDF repository and for each product in the list searched for the match resource in the "DBpedia" indexes. Third, we built a triple for each product, each property in our camcorder model, and the value of the property (which was found from one of two resources). In the "Data Loading" phase we loaded all the created triples in a single RDF repository.

Finally, in the final RDF repository we have: 612 comprehensive camcorder instances, which contain all the properties from both sources, 5527 shop instances (which are offering camcorder instances), 5527 offer instances, and 5527 price instances for each offer.

Related Work

Matching and mapping Web data hava a long research history. Volz et al. (Bizer et al. 2009), for instance, proposed a tool (Silk-A Link Discovery Framework for the Web of Data) for finding relationships between entities within different data sources. The Silk can be used by data publishers to establish RDF links from their data sources to other data sources on the Web. Haslhofer and Klas. (Haslhofer and Klas 2010) provide a comprehensive study on techniques for metadata mapping. Furthermore, we can name researches such as (Bernstein and Haas 2008).

Early related work on the issue of Web data extraction can be traced back to tools such as: Lixto (Baumgartner, Frölich, and Gottlob 2007) is a visual and interactive wrapper generator which uses an expressive and exible logic-based declarative language. Deep Web Makro Recording/Replaying allows for deep Web navigation ability and it is also able to pars Java Script and has additional offset-based data extraction. Kapow Technologies[5] which enables the conversion of disparate streams of Web data into strategic business assets. Sundewsoft[6]is an Eclipse-based wrapper generator using Internet Explorer, Web macro recording, and data patterns. It exports to XML, CSV or Excel.

Summary

In this paper, we have presented CORB methodology which attempts to satisfy three main requirements when e-commerce-based applications use vast amounts of information from different public Web portals. It recommends creating a comprehensive e-commerce RDF-based dataset from different Web sources in three phases: (1) Data Wrapping, (2) Data Matching, and (3) Data Loading.

In order to provide this information with e-commerce concepts in mind, we propose using an extension of the GoodRelation ontology for each category of product/service, and to provide information from different Web data sources in machine-processable format (e.g. RDF), we propose implementing a specially designed wrapper which wraps the data from different sources into a standard machine-processable format and maps the specification of each product/service category to our category-related e-commerce ontology. And then, to match and store information from various sources in a single processable structure (at the level of instance matching), we propose a matching algorithm for finding information about a particular product from various available Web sources. This component matches the missing properties of one product/service from different Web sources to create comprehensive RDF instances.

As already noted, the future goal is to set up an experiment to find an appropriate approach for an accurate searching process for a product/service from a repository within the indexes of another repository.

References

Baumgartner, R.; Frölich, O.; and Gottlob, G. 2007. The lixto systems applications in business intelligence and semantic web. In *Proceedings of the 4th European conference on The Semantic Web: Research and Applications*, ESWC '07, 16–26. Berlin, Heidelberg: Springer-Verlag.

Bernstein, P. A., and Haas, L. M. 2008. Information integration in the enterprise. *Commun. ACM* 51(9):72–79.

Bizer, C.; Volz, J.; Kobilarov, G.; and Gaedke, M. 2009. Silk - a link discovery framework for the web of data. In *18th International World Wide Web Conference*.

Bizer, C.; Heath, T.; and Berners-Lee, T. 2009. Linked data - the story so far. *International Journal on Semantic Web and Information Systems* 5(3):1–22.

Haslhofer, B., and Klas, W. 2010. A survey of techniques for achieving metadata interoperability. volume 42, 1–37. New York, NY, USA: ACM.

Hepp, M. 2005. A methodology for deriving owl ontologies from products and services categorization standards. In *ECIS*.

Hepp, M. 2008. Goodrelations: An ontology for describing products and services offers on the web. In Gangemi, A., and Euzenat, J., eds., *EKAW*, volume 5268 of *Lecture Notes in Computer Science*, 329–346. Springer.

[5] http://kapowtech.com

[6] http://www.sundewsoft.com