Ontology Population from Web Product Information

Introduction

With the vast amount of information available on the Web, there is an increasing need to structure Web data in order to make it accessible to both users and machines. E-commerce is one of the areas in which growing data congestion on the Web has serious consequences. We propose a framework that is capable of **populating** a consumer electronic product **ontology** using **tabular** information from Web shops. Formalizing product information in this way enhances product comparison and recommendation applications. Our approach employs both lexical and syntactic matching for mapping properties and instantiating values.

A Consumer Electronics Ontology

Currently, there is a lack of detailed ontologies for consumer electronic products. Hence, we create the **OntoProduct** ontology:

- It is an extended version of the Consumer Electronics Ontology (**CEO**) with additional product attributes;
- It contains 270 consumer electronic product properties and 24 product classes;
- It is fully compatible with the well-known GoodRelations (**GR**) ontology for e-commerce;
- It utilizes the Units of Measurement Ontology (**MUO**) for linking units of measurements to quantitative values.

Ontology Population Framework

The population of ontologies is regularly preceded by a knowledge **extraction** phase, where, for instance, natural language Web pages on products are converted into raw **tabular** data. Our framework focuses on the **population** of ontologies with product information in the e-commerce domain, using extracted tabular data in the form of **key-value** pairs. For this, user-defined **annotations** for lexical and syntactic matching are employed, which facilitate the main tasks of our framework:

- **Classification**: optional step to determine product class; most Web stores nowadays have some kind of class or category data of each product already available.
- **Property Matching**: step for extracting ontology properties from tabular key-value pairs by lexically matching keys with ontology properties, and by matching regular expressions on values.
- **Value Instantiation**: phase for creating individuals in the ontology and associating them to their properties, while making use of parsers, content spotters, and instantiation tools.

Results & Conclusions

The property matching and value instantiation processes in the framework have been evaluated separately using a **golden standard**, under the assumption that the product class of a product description is known. The raw product data was obtained from two different Web sources, i.e., **Best Buy** and **Newegg.com**.

In our experiments, the proposed framework achieved a **solid performance** for the property matching and value instantiation processes. The former process resulted in a precision, recall, accuracy, and F1 score of 96.95%, 93.27%, 94.80%, and 95.07%, respectively, whereas the latter process resulted in scores of 77.12%, 76.09%, 62.07%, and 76.60%, respectively. Although some raw product keys in the test set were not present in the training set, many key-value pairs were still matched with ontology properties. **In practice**, this means that a semi-automatic approach would only require **training** the algorithm with a **few products** from each product class in order to achieve satisfactory performance on property matching for all the products in a Web shop.

After analyzing the results in more detail, we found that the regular expressions, in conjunction with the lexical representations, are often capable of correctly mapping key-value pairs to properties in the ontology. For example, the key “Product Dimensions” is correctly mapped to `ceo:hasWidth`, `ceo:hasHeight`, and `ceo:hasDepth`, demonstrating the usefulness of regular expressions in this context.

Acknowledgement

Damir Vandic is supported by an NWO Mosaic scholarship for project 017.007.142: Semantic Web Enhanced Product Search (SWEPS). Frederik Hogenboom is supported by the NWO Physical Sciences Free Competition project 612.001.009: Financial Events Recognition in News for Algorithmic Trading (FERNAT) and the Dutch national program COMMIT.

Contact

Damir Vandic
Postal: Erasmus University Rotterdam
P.O. Box 1738
NL-3000 DR Rotterdam
The Netherlands
Phone: +31 (0)10 408 8906
E-Mail: vandic@ese.eur.nl
Web: http://www.damirvandic.com

Erasmus University Rotterdam

Damir Vandic, Lennart J. Nederstigt, Steven S. Aanen, Flavius Frasincar, Frederik Hogenboom