Mapping Product Taxonomies in E-commerce

Steven Aanen Lennart Nederstigt Damir Vandić Flavius Frăsincar

Econometric Institute, Erasmus University Rotterdam
P.O. Box 1738, 3000 DR Rotterdam, the Netherlands

The full version of this paper, entitled 'SCHEMA - AN ALGORITHM FOR AUTOMATED PRODUCT TAXONOMY MAPPING IN E-COMMERCE' appeared in: Extended Semantic Web Conference (ESWC 2012), volume 7295 of Lecture Notes in Computer Science, pages 300-314. Springer, 2012.

Abstract

In this paper we propose SCHEMA, an algorithm that automatically maps heterogeneous product taxonomies in the domain of e-commerce. SCHEMA employs a custom word sense disambiguation technique, based on the Lesk algorithm, in combination with the semantic lexicon WordNet. For finding candidate target categories and determining the path-similarity we propose a semantic category matching algorithm that takes into account the disambiguation process of a category. The mapping quality score is calculated using the Damerau-Levenshtein distance and a node-dissimilarity penalty. The performance of SCHEMA was tested on three real-life datasets and compared to PROMPT and the algorithm proposed by Park & Kim. The comparison shows that SCHEMA improves considerably recall and F_1-score, while maintaining similar precision.

1 Introduction

In recent years the Web has increased dramatically in both size and range, playing an increasingly important role in our society and world economy. For instance, the estimated revenue for e-commerce in the USA grew from $7.4 billion in 2000 to $34.7 billion in 2007 [2]. As a consequence, the aggregation of product data is becoming increasingly important. A common problem encountered in this task is the mapping of product taxonomies from different Web stores to an existing product taxonomy. By matching the product taxonomies from different Web stores, it becomes easier to compare products.

As a solution we propose the *Semantic Category Hierarchy for E-commerce Mapping Algorithm*, also called SCHEMA. This algorithm can be used to map heterogeneous product taxonomies from multiple sources to each other. The algorithm employs a word sense disambiguation technique that is based on the *Lesk algorithm* [3] to find synonyms of the correct sense for the source category name. Furthermore, it uses lexical similarity measures, such as the *Levenshtein distance*, along with structural information, to determine the best candidate category to map to. In order to evaluate SCHEMA, its performance is compared on recall and precision with PROMPT [4] and the algorithm proposed by Park & Kim [5].

2 SCHEMA

The SCHEMA algorithm takes as input a source taxonomy and a target taxonomy. For each category in the source taxonomy the algorithm produces a mapping to the target taxonomy. The algorithm can also provide a 'blank' mapping, in this case the best match in the target taxonomy did not exceed a certain quality threshold. SCHEMA executes the following three main steps for each category in the source taxonomy, which we will call the 'source category' from now on. The first step is to disambiguate the source category, which results in obtaining a set of synonyms of the correct sense. The second step involves selecting candidate categories from the target taxonomy using the set of synonyms obtained in the previous step. In the third step a comparison is performed with the source category to select the best-fitting candidate target category.

In the first step the algorithm first applies a splitting procedure on the parent of the source category, the source category itself, and the the source category's children. For example, the category "Music & Video" would be divided in the split terms 'Music' and 'Video'. The union of all split terms is considered as the context for the word sense disambiguation algorithm, which is based on the Lesk algorithm. The result of this step is the extended split term set, which is a set of synonym sets where each synonym set corresponds to a split term in the source category. In this step, as well as in subsequent steps, the Levenshtein distance is used to compare single terms.

The second step uses the extended split term set to select candidate target categories. This is done by also taking into account composite categories, e.g., 'Music & Video'. The issue with composite categories is that we do not want to map 'Music & Video' to 'Music', whereas mapping 'Music' to 'Music & Video' is fine. The way SCHEMA deals with this issue is that it uses each synonym set in the extended split term set to check for a match between two categories. The match checking is performed by the proposed category matching algorithm, which is based on the *longest common substring similarity*. A target category becomes a candidate if all the synonym sets of the source category are matching that target category.

The third step involves comparing candidate category paths to the source category path. For this purpose, we use the Damerau-Levenshtein distance [1]. This procedure computes a similarity between the source category and each candidate target category, taking into account the structure of the category paths. The mapping is selected by taking the candidate with the highest similarity. A threshold is used to avoid mapping to an unsuitable target taxonomy.

3 Evaluation and Conclusion

Three product taxonomies from real-life datasets were used for the evaluation. We used Amazon (2,500 categories), Overstock.com (1,000), and the Open Directory Project (44,000 categories). Using these three datasets, six different combinations of source and target taxonomies were performed. Using a sample of 500 for each data set, we manually mapped $6 \times 500 = 3000$ categories.

For the F_1-score, the evaluation shows that PROMPT has 20.75%, the Park & Kim algorithm 32.52%, and SCHEMA 55.10%. For the recall, these values are 16.69% for PROMPT, 25.19% for Park & Kim, and 80.73% for SCHEMA. For the precision, these values are 28.93% for PROMPT, 47.77% for Park & Kim, and 42.21% for SCHEMA. The results of our evaluation show that SCHEMA performs better than PROMPT and the algorithm of Park & Kim, on both recall and F_1-score, while maintaining a similar precision. This can be attributed to the ability of SCHEMA to cope with lexical variations in category names, as well as the ability to properly deal with composite categories.

In conclusion, the main objective for developing the SCHEMA algorithm was to facilitate the aggregation of product information from different Web sources by providing a product taxonomy mapping algorithm. In order to improve the recall, our algorithm employs word sense disambiguation and addresses the recurring issue of composite product categories. The performance of our algorithm was tested on three real-life datasets and compared with the performance of PROMPT and the algorithm of Park & Kim. This evaluation shows that SCHEMA achieves a considerably higher average recall than the other algorithms, while maintaining a similar precision.

References

[1] Fred J. Damerau. A Technique for Computer Detection and Correction of Spelling Errors. *Communications of the ACM*, 7(3):171–176, 1964.

[2] John B. Horrigan. Online Shopping. *Pew Internet & American Life Project Report*, 36, 2008.

[3] Michael Lesk. Automatic Sense Disambiguation using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *5th Annual International Conference on Systems Documentation (SIGDOC 1986)*, pages 24–26. ACM, 1986.

[4] Natalya F. Noy and Mark A. Musen. The PROMPT Suite: Interactive Tools for Ontology Merging and Mapping. *International Journal of Human-Computer Studies*, 59(6):983–1024, 2003.

[5] Sangun Park and Wooju Kim. Ontology Mapping between Heterogeneous Product Taxonomies in an Electronic Commerce Environment. *International Journal of Electronic Commerce*, 12(2):69–87, 2007.