Expert Systems with Applications 42 (2015) 1298–1313

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Automated product taxonomy mapping in an e-commerce environment

Steven S. Aanen, Damir Vandic, Flavius Frasincar *

Erasmus University Rotterdam, PO Box 1738, NL-3000 DR, Rotterdam, The Netherlands

ARTICLE INFO

Article history:
Available online 28 September 2014

Keywords:
Products
Semantic Web
Schema
Ontology
Matching
Mapping
Merging
E-commerce
Web shop

ABSTRACT

Over the last few years, we have experienced a steady growth in e-commerce. This growth introduces many problems for services that want to aggregate product information and offerings. One of the problems that aggregation services face is the matching of product categories from different Web shops. This paper proposes an algorithm to perform this task automatically, making it possible to aggregate product information from multiple Web sites, in order to deploy it for search, comparison, or recommender systems applications. The algorithm uses word sense disambiguation techniques to address varying denominations between different taxonomies. Path similarity is assessed between source and candidate target categories, based on lexical relatedness and structural information. The main focus of the proposed solution is to improve the disambiguation procedure in comparison to an existing state-of-the-art approach, while coping with product taxonomy-specific characteristics, like composite categories, and re-examining lexical similarity and similarity aggregation in this context. The performance evaluation based on data from three real-world Web shops demonstrates that the proposed algorithm improves the benchmarked approach by 62% on average F_1-measure.

1. Introduction

Recently, the Web has experienced a rapid growth, playing an increasingly important role in our society. The expectations are that the amount of information available on the Web will continue to grow exponentially; doubling in size roughly every five years (Zhang, Zhang, Yang, Cheng, & Zhou, 2008). This expresses the need to keep all this information structured. The vision of the Semantic Web from Berners-Lee, Hendler, and Lassila (2001) addresses this need, with the goal to make the Web more structured, interactive, useful, and containing meaningful data, understandable for both human and computer. This has lead to the usage of ontologies (Gruber, 1993): standardized representations of knowledge, in which concept relationships are explicitly defined. As a result, various research in the fields of ontology management and data annotation has been performed: the matching, construction and integration of ontologies Noy and Musen (2004a), Hepp, De Leenheer, De Moor, and Sure (2007), annotation of (Web) data Arlotta, Crescenzi, Mecca, and Merialdo (2003), Bizer et al. (2009), as well as different applications of the Semantic Web (Vandić, van Dam, & Frăsincar (2012), Benslimane, Dustdar, & Sheth (2008)). Unfortunately, the current Web has not yet evolved to the Semantic Web, since there is a lot of information that is not semantically annotated. Consequently, data has to be interpreted by humans, since machines do not understand them. Because machines do not understand the information embedded on Web pages, search engines are not always capable of finding the information that suits the user's needs the best.

Because machines currently do not understand true meaning of data, in the field of e-commerce, keyword-based search is often used. This type of search leads to a large fraction of customers that fail to find the product that fits their needs optimally (Horrigan, 2008). One of the reasons for this is that Web-wide parametric product search is not possible. Therefore, current product comparison tools are primarily based on pricing, instead of on product characteristics. The result is that customers will be forced to compare mostly on prices, as they can not scan thousands of products themselves. Although the price competition is — economically seen — not unwanted, it can well be that customers are prepared to buy more expensive products if those would fit their needs better. Selling more expensive products will increase revenue of online retailers, and thereby contribute to the economy. Thus for both online retailers and for customers, better product search, comparison and recommendation applications on the Web are desired.

To build product search, recommendation, or comparison tools, it is needed to deal with product categorization. In general, Web sites that deal with products, such as manufacturer pages or Web

* Corresponding author. Tel.: +31 (0)10 408 1340; fax: +31 (0)10 408 9162.
E-mail addresses: aanen@appophetweb.nl (S.S. Aanen), vandic@ese.eur.nl (D. Vandic), frasincar@ese.eur.nl (F. Frasincar).

http://dx.doi.org/10.1016/j.eswa.2014.09.032
0957-4174/

stores, have a hierarchy in which products are categorized. In this way, users are able to efficiently filter the kind of products that are desired, even though possibly many thousands of products are offered. These hierarchical categorizations are called taxonomies: tree-like structures in which concepts have supertype–subtype relationships. Taxonomies are related to schemas, in which richer concept relations, with also for example cardinality constraints, lead to a graph-like structure. To be able to aggregate information from multiple Web sites dealing with products, it is needed to merge their corresponding taxonomies in order to determine to which class the collected products belong to.

In e-commerce, taxonomies are often very heterogeneous, since no standardizations are being used, and hierarchies are often manually created. In the fields of ontology and taxonomy/schema matching, many different algorithms have been proposed to deal with the heterogeneity of information structures (Do, Melnik, & Rahm, 2002; Kalfoglou & Schorlemmer, 2003; Noy, 2004; Rahm & Bernstein, 2001; Shvaiko & Euzenat, 2005). However, since product taxonomies have some unique characteristics, such as composite categories (e.g.,'Electronics & Computers) and loose relationships (e.g., subcategory 'Hardware' under category 'PC Games', which is not a true subtype relation), specialized approaches are required.

Based on the algorithm from Park and Kim (2007), which has been designed specifically for taxonomy matching in e-commerce, this paper will propose an improved approach, as there are several aspects in the Park & Kim algorithm that can be made better. More specifically, the focus of this paper will be on one of the major drawbacks of the existing algorithm: the word sense disambiguation process. The disambiguation is needed to find synonyms of the correct sense for category names. This is to account for the fact that different taxonomies make use of different words to characterize their classes, while having the same meaning. For example, 'Tools' can have completely different meaning depending on the parent category (e.g., 'gardening' vs. 'electronics'). Some Web shops might explicitly use 'Electronic Tools' or 'Gardening Tools' while others might just use 'Tools' and exploit the hierarchy to convey the intended meaning. The assumption is that when this process is improved, the overall recall and precision of the algorithm will rise as well. Apart from focusing on improving particularly this part of the algorithm, the goal is to also re-examine concepts such as composite category handling (neglected by Park & Kim), cope with depth variance between taxonomies, and propose new lexical similarity and similarity aggregation functions that better fit the e-commerce setting.

This paper is organized as following. First, we discuss in Section 2 related approaches for taxonomy/schema and ontology matching, as well as word sense disambiguation techniques and different lexical and semantic similarity measures. Similarity measures are needed to score candidate target categories for a given source category. Section 3 explains the implementation of the proposed algorithm, as well as the underlying ideas. In Section 4 the proposed algorithm will be evaluated against similar approaches using real-world data. Last, conclusions and possible future work are discussed in Section 5.

2. Related work

Product taxonomy mapping is part of the research fields of ontology and taxonomy/schema matching. Conceptual matching in general is used in various domains of information technology, like Semantic Web applications, ontology management, e-commerce, data warehousing, and database integration. Therefore, quite a lot of research has been done on the topic in the past decades Do et al. (2002), Shvaiko and Euzenat (2005). The main difference between ontology and taxonomy/schema matching can be found in the semantics.

Ontologies have the meaning of concepts and relations between them explicitly encoded in their data representation. Therefore, matching algorithms can choose to primarily use knowledge from within the ontology. Ontologies are logical systems, and can be seen as a logical set of axioms according to which data is annotated, as Shvaiko and Euzenat (2005) explain.

In taxonomy/schema matching however, data is often not annotated for meaning, besides the basic is-a relationships (and sometimes additional constraints as in database schemas), making it less structured than working with ontologies. Matching algorithms have to find out the meaning using external data or using the context of the data concepts within the schema. In other words, in ontology matching, computers work with data which they can understand. In schema matching, only hierarchical data is available of which a computer must first determine most relations and the meaning on its own.

Although there are some signs of initial research effort (Vandić et al. (2012)), in the field of e-commerce, the ideas of the Semantic Web are at their infancy in practice. Since no good applications exist at this moment, few Web stores have annotated their product pages with semantics, as specified by a product ontology like GoodRelations (Hepp, 2008). These semantics describe for example the relations between different products. Some exceptions do exist, but widely seen, the information on the Web — especially in product environments — is still not understood by computers.

For the above reason, taxonomy matching is more applicable than ontology matching in this field. However, both in ontology and in taxonomy/schema matching, the goal is to find relatedness between concepts, often using word sense disambiguation techniques and lexical or semantic similarity measures. Therefore, some ideas from the ontology matching domain can be applicable to taxonomy/schema matching as well. For this reason, we will discuss projects from both ontology matching and taxonomy/schema matching fields. In addition, since many of the approaches rely on relatedness of concepts and words, some measures for these will be discussed as well. As this research focuses on enhancing the word sense disambiguation process within the matching algorithm, we will also discuss some approaches for dealing with polysemy. Last, this section will give a brief overview of WordNet (Fellbaum, 1998), which is a semantic lexicon used by many matching algorithms and disambiguation techniques, including the proposed solution.

2.1. Ontology matching

This section discusses some approaches that deal with matching of ontologies. While this research focuses on taxonomy mapping, ontology alignment is a strongly related field of research which can give further insight in possible approaches of product taxonomy mapping.

As part of the *Protégé* environment for knowledge-based systems (Gennari et al., 2003), *PROMPT* was developed by Noy and Musen (2003). PROMPT is a framework for multiple ontology management, including various tools to deal with tasks that often occur in management of ontologies. One of these tools, *iPROMPT*, is an interactive approach for ontology merging. It guides the user through the merging process, by making suggestions on what should be merged, and identifying problems and inconsistencies. This information is based on the structure of the concepts, and relations between them within the ontology, as well as previous user actions. iPROMPT however only looks at the local context within the ontology, which is seen as a graph, for its decisions. In other words, it only takes direct relations of concepts into account. iPROMPT is very much user-dependent, and therefore not very suitable in the domain of automated product taxonomy matching. For this purpose, often large amounts of data have to be processed