Web Semantics: Science, Services and Agents on the World Wide Web 7 (2009) 136–153

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

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

Automated ontology instantiation from tabular web sources—The AllRight system☆

Dietmar Jannach[a,*], Kostyantyn Shchekotykhin[b], Gerhard Friedrich[b]

[a] Technische Universität Dortmund, 44221 Dortmund, Germany
[b] University of Klagenfurt, 9020 Klagenfurt, Austria

ARTICLE INFO

Article history:
Received 21 January 2009
Accepted 19 April 2009
Available online 15 May 2009

Keywords:
Web mining
Web crawling
Ontology instantiation
Knowledge extraction

ABSTRACT

The process of populating an ontology-based system with high-quality and up-to-date instance information can be both time-consuming and prone to error. In many domains, however, one possible solution to this problem is to automate the instantiation process for a given ontology by searching (*mining*) the web for the required instance information.

The primary challenges facing such system include: (a) efficiently locating web pages that most probably contain the desired instance information, (b) extracting the instance information from a page, and (c) clustering documents that describe the same instance in order to exploit data redundancy on the web and thus improve the overall quality of the harvested data. In addition, these steps should require as little *seed knowledge* as possible.

In this paper, the AllRight ontology instantiation system is presented, which supports the full instantiation life-cycle and addresses the above-mentioned challenges through a combination of new and existing techniques. In particular the system was designed to deal with situations where the instance information is given in tabular form. The main innovative pillars of the system are a new high-recall *focused crawling* technique (xCrawl), a novel *table recognition* algorithm, innovative methods for document clustering and instance name recognition, as well as techniques for fact extraction, instance generation and query-based fact validation.

The successful evaluation of the system in different real-world application scenarios shows that the ontology instantiation process can be successfully automated using only a very limited amount of seed knowledge.

1. Introduction

The vision of the World Wide Web as a huge repository of *machine-processible* information may be realized in different ways. The first option is to rely on the large-scale use of semantic annotations that refer to entities defined in formal ontologies. Examples of such resource annotation systems include emerging new technologies such as RDF[1] or microformats that allow the augmentation of human-readable content (in HTML) with machine-processible information. However, even with appropriate tool support, it is unlikely that web documents will be sufficiently (semantically) annotated on a broad scale in the near future. Techniques for *automated tagging* have been recently proposed as a possible solution for overcoming this limitation as outlined for instance in [8] or [13]. These approaches have shown that they can partially solve the annotation problem.

The other option to exploit the Web of Data is to try to automatically "re-construct" the knowledge presented in (unstructured) web documents. Several *Web Mining* and *Information Extraction* (IE) techniques have been proposed to automate this task. Some rely on the redundancy of information on the web and use statistical methods [10,15], while others use natural language processing (NLP) techniques [29,30] to extract the required knowledge from unstructured documents. It has also been shown that domain *ontologies* can support knowledge extraction, which in turn means that the information extraction problem can be generalized to the problem of finding and inserting information that matches a given ontology into the system's knowledge base, a process also referred to as *ontology instantiation* or *ontology population*. The instantiated ontology can then in turn serve as a basis for the development and provision of so-called *knowledge services* in the Semantic Web [4].

The AllRightontology instantiation system and its innovations presented in this work were inspired by the requirements that arose

☆ This paper combines and significantly extends the work presented in [33–35].

* Corresponding author. Tel.: +49 231 755 7272.
E-mail address: dietmar.jannach@udo.edu (D. Jannach).

[1] http://www.w3.org/TR/rdfa-syntax/.

1570-8268/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.websem.2009.04.002

throughout the development of real-world recommendation and product-comparison services that we have implemented using the Advisor Suite framework [16,21]. These recommendation services were developed to support online customers' decision-making and purchase processes and are based on in-depth and accurate knowledge about the items being recommended. The central piece of information required in such interactive recommenders is thus all possible instances (facts) describing detailed product specifications. Keeping such specialized instances of an ontology-based application up-to-date can be a time-consuming and error-prone task, particularly in the fast-paced domain of electronic consumer goods (MP3 players, laptop computers, or digital cameras). The possibility of automating this knowledge-acquisition process, however soon became obvious, as in these domains the required item information is already available on the web, e.g. on manufacturer homepages or community portals.[2]

The goal was therefore to develop a Web Mining System (WMS) that is capable of finding and extracting this knowledge automatically. What was soon recognized, however, is that existing web mining techniques are not directly applicable, because, e.g. the instance information is in many cases presented in *tabular form*.

Although tables are generally well-suited to the human reader as they are a compact and precise form for representing information, the above-mentioned information extraction techniques do not work well with them: methods that are based on Natural Language Processing (NLP) are for instance not applicable as tables in web documents typically do not contain full sentences but rather simple attribute-value pairs describing the features of an item; clustering methods that aim to identify several documents that describe the same instance (e.g. digital camera) based on term co-occurrences do not work either, because pages with *the same sets of keywords/tokens* (from the same portal) describe *different items*. Furthermore, these tables are not necessarily constructed from underlying databases, meaning that neither SPARQL[3] nor "Hidden-Web" techniques can be used. The AllRight system takes these particularities into account and presents several new techniques for dealing with tabular information.

The contributions of the AllRight to the state of the art can be summarized as follows. Overall, we show how automatic ontology instantiation can also be applied to domains in which data is given in tabular form (such as personal information, geographical data and so forth). In terms of technical innovation, the AllRight system includes a new crawling method for the fast location of tabular descriptions, a "visual" table identification technique, a novel way of applying clustering algorithms to deal with tabular descriptions, as well as a query-based fact validation method. Given the promising evaluation results with an accuracy of about 80% in different domains, we finally show how Semantic Web technology, ontologies and a combination of new and existing extraction techniques can help us to automatically *mine* the web for ontology instances in the context of real-world, industrial problem settings.

Note that although the design of the AllRight system was originally motivated by the requirements of a particular application scenario, it was devised as a more general, domain-independent ontology instantiation system in the sense of Alani et al. [4]: the input to the extraction system is thus a domain ontology that describes the structure of and the relations between the items to be searched for. The extraction process itself is domain-independent and is structured into a series of generic tasks such as web crawling, name extraction, validation, document clustering and fact extraction.

The paper is organized as follows. Based on a motivating example from the domain of digital cameras, we first sketch the overall ontology instantiation process flow, give then technical details of the individual steps required for information extraction, and finally present evaluation results for the individual subtasks. Throughout the paper, we compare the techniques presented with those outlined in related work and discuss commonalities and differences. The paper ends with a summary and an outlook on future work.

2. System overview

2.1. Process flow

Fig. 1 provides a high-level overview of the ontology instantiation process in the AllRight system in the context of the "recommender system" application scenario, where the problem consists of extracting detailed product data for a content-based recommender in an online store.

The first step is to develop a domain ontology that describes the basic characteristics of the items being searched for. In our example configuration, some parts of the knowledge are imported from the Advisor Suite system, which can, for instance, be used to model a list of interesting item characteristics. The system then generates keywords from this domain ontology (1) that are used for crawling the web for relevant web pages (2 and 3) with the new xCrawl method. The downloaded pages are then analyzed by the Identification Component (validator) (4) to determine if they contain the desired item descriptions. Again, this analysis is done on the basis of the knowledge from the ontology (5). Next, the validated set of documents is forwarded (6) to a module which generates clusters of pages describing the same product, extracts the specific facts for each product and feeds the new knowledge back to the ontology (7). Note that in principle fact extraction could be performed prior to duplicate removal [27]. However, such systems require more detailed seed knowledge and specialized and complex fact recognition techniques.

2.2. Core ontology and domain ontology

In order to remain applicable to a broad spectrum of application scenarios, the AllRightsystem differentiates between three "levels" of ontologies, see Fig. 2 for sample fragments. In the product data extraction scenario, the predefined *core ontology* forces the system to search for entities that have attributes of given types and certain units of measurement and that both the attributes and the units can be annotated with string-typed keywords. Note that in principle also other concepts and/or roles can be used (or re-used) in the core ontology. Changing the core ontology is however not easily possible in such a system because, e.g. all data access methods depend on the design of the core ontology. Still, our practical experiences show that the concepts and roles defined in the core ontology are sufficient to cover a wide range of applications areas in which data can be represented in the form of attributes and values.

As a part of the configuration process of the AllRight system for a specific domain, e.g. digital cameras or other consumer electronics, a *domain ontology* has to be modeled based on the core ontology. In the example, we thus state that "resolution" is a property of interest for entities of type "digital camera". The resolution is given as a real-valued number and is measured in megapixels. Appropriate keywords could be "effective pixels" and "million", respectively. Note that the task of defining the domain ontology corresponds to the definition of *seed knowledge*, which is required in every Web Mining System (WMS). Technically, the domain ontology is stored

[2] Our experiences showed that the quality and level of detail of commercially available data sets is not sufficient for our purposes.

[3] http://www.w3.org/TR/rdf-sparql-query/.