Web Semantics: Science, Services and Agents on the World Wide Web 6 (2008) 171–190

Contents lists available at ScienceDirect

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

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

Syndication on the Web using a description logic approach☆

Christian Halaschek-Wiener[a,*], Vladimir Kolovski[b]

[a] Clados Management LLC, 1840 Gateway Dr., Suite 200, San Mateo, CA 94404, USA
[b] Oracle NEDC, Nashua, NH 03062, USA

ARTICLE INFO

Article history:
Received 30 November 2007
Received in revised form 11 June 2008
Accepted 29 June 2008
Available online 7 September 2008

Keywords:
Syndication systems
Publish/subscribe
Description logic
Incremental reasoning
Query answering

ABSTRACT

Syndication systems on the Web have attracted vast amounts of attention in recent years. As technologies have emerged and matured, there has been a transition to more expressive syndication approaches; that is, subscribers and publishers are provided with more expressive means of describing their interests and published content, enabling more accurate information filtering. In this paper, we formalize a syndication architecture that utilizes the Web Ontology Language (OWL) and description logic reasoning for selective content dissemination. This provides finer grained control for filtering and automated reasoning for discovering implicit subscription matches, both of which are not achievable in less expressive approaches. We then address one of the main limitations with such a syndication approach, namely matching newly published information with subscription requests in an efficient and practical manner. To this end, we investigate incremental query answering for a large subset of OWL and present an approach to reduce the portion of the ontology that must be considered for query answering in the event of updates. Lastly, an evaluation of the query approach is shown, demonstrating its effectiveness for syndication purposes.

1. Introduction

Web-based syndication systems have attracted a great amount of attention in recent years as the amount of streaming content on the Web has increased at dramatic rates. In typical syndication frameworks, users register their subscription requests with syndication brokers; similarly, content publishers register their feeds with syndication brokers, and it is then the broker's task to match newly published information with registered subscriptions. As technologies have emerged, there has been a transition to more expressive syndication approaches; that is publishers (and subscribers) are provided with more expressive means for describing their published content (respectively interests), allowing more accurate dissemination. This has been enabled by the maturation of technologies for sharing information on the Web and the standardization of representation languages for Web content. In particular, through the years there has been a transition from keyword based approaches to attribute-value pairs and more recently to XML. Given the lack of expressivity of XML (and XML Schema) as a knowledge modeling language, there has been interest in using the Resource Description Framework (RDF) and its accompanying schema language, RDF Schema, for syndication purposes. RDF has even been adopted as the standard representation format of RSS 1.0.[1]

Today's syndication approaches still provide relatively weak expressive power from a modeling perspective (i.e., XML and RDF are comparatively inexpressive languages) and provide very little automated reasoning support. However, if a more expressive syndication approach with a formal semantics can be provided, many benefits can be achieved; these include a rich semantics-based mechanism for expressing subscriptions and published content allowing increased selectivity and finer control for filtering, and automated reasoning for discovering subscription matches not found using traditional syntactic syndication approaches [45].

In this work, we consider using the Web Ontology Language (OWL) for representing published content. As the semantics of a large subset of OWL is aligned with description logics (DLs), reasoning techniques for DLs can then be leveraged for matching content with subscription requests. In such an approach, the previously mentioned benefits of using a formal representation language can therefore be achieved. An additional benefit of an OWL-based syndication approach is its native Web embedding and power as a data integration language. Further, such an approach can be seen as a natural extension of existing RSS 1.0 syndication systems, as OWL

☆ The results presented in this paper are the outcome of research performed while the author was a member of the MINDSWAP research group at the University of Maryland, College Park.

* Corresponding author.
E-mail addresses: christian@clados.com (C. Halaschek-Wiener), vladimir.kolovski@oracle.com (V. Kolovski).

[1] RSS 1.0 Specification: http://web.resource.org/rss/1.0/spec.

1570-8268/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.websem.2008.06.002

Table 1
Illustration of expressivity in OWL-based syndication

```
:RiskyCompany a owl:Class;
  owl:intersectionOf (
    [ a owl:Restriction; owl:onProperty :hasProduct;
      owl:someValuesFrom :AdverseEffectProduct ]
    :Company
  ) .
:AdverseEffectProduct a owl:Class;
  owl:intersectionOf (
    [ a owl:Restriction; owl:onProperty :causes;
      owl:someValuesFrom [ owl:unionOf ( :Infection :AllergicReaction ) ] ]
    :Product
  ) .
:causes a owl:ObjectProperty.
:onRecommendation a owl:ObjectProperty.
```

can be encoded in RDF.

To demonstrate the advantages of an OWL-based syndication approach, consider the following example: suppose we are disseminating news information in the financial domain. Also suppose that a stock trader is interested in articles that *could* discuss companies whose stocks are likely to become volatile; specifically, let us assume that the trader is interested in any *RiskyCompany* which the trader defines to be a *company* that has a *product* which *causes* an *infection* or *allergic reaction*. Using an XML-based approach syndication brokers can provide an XML schema that contains an element *RiskyCompany* and such companies can be declared to be this type of element. A limitation of such an approach is that publications (i.e., XML documents) must explicitly declare entities to be a *RiskyCompany*. This is because XML query languages such as XPath and XQuery only provide syntactic matching of the XML documents representing publications. If we consider an RDF-based approach, then the syndication broker can model the financial domain using RDF Schema. Therefore, additional matches can be obtained as one can logically infer that a company is a *RiskyCompany*. For example, if the domain of a property *hasProductWithAdverseEffect* is declared to be of type *RiskyCompany* and we are given that *BauschAndLomb hasProductWithAdverseEffect Renu*, then we would have a (inferred) match for the subscription; such logical inference (although simplistic) is not possible with an XML-based approach. However, in an RDF based approach, more complex logical definitions (and therefore finer-grained control) of *RiskyCompany* are not expressible.

If we now consider an OWL-based approach, such functionality is clearly provided. For example, the knowledge broker can define a *R* iskyCompany as an OWL class whose necessary and sufficient conditions for inclusion are that it be a *company* that has some product which is an *AdverseEffectProduct*; similarly, an *AdverseEffectProduct* can be defined to be any *product* that causes some *infection* or *allergic reaction*. Using an OWL approach, this can easily be represented by the OWL descriptions in Table 1.[2]

Given this domain model, if we assume it is previously known that *BauschAndLomb* is a *company* that *has product Renu*, which is known to be a *Product*, and we receive the publication that *Renu causes* some *infection*, then standard DL reasoning services can be employed to automatically infer that *BauschAndLomb* is a *RiskyCompany* and thus there is a match for the subscription.

While OWL-based syndication approaches provide increased expressivity over XML and RDF, previous DL-based syndication approaches suffer from scalability issues due to the inherent complexity of DL reasoning [45,34,23]. This is an issue in domains such as the syndication of financial news because response times must be minimal as critical information must be delivered in near real time (e.g., for stock trading purposes). One of the main limitations in an OWL-based syndication approach is related to DL reasoning over changing data; this is primarily due to the static nature of existing DL reasoning techniques. In particular, the addition of information from newly published documents and data can be viewed as a change in the underlying knowledge base (KB). In current DL reasoning algorithms, reasoning on the updated KB is performed from scratch: consistency of the KB must be ensured, queries must be re-evaluated, etc. In this paper, we formalize an OWL-based syndication framework (originally presented in [24]), which leverages DL reasoning to determine subscription matches. We then address the scalability of the DL reasoning services necessary for the syndication framework, by investigating incremental query answering over OWL KBs; we specifically present a technique for reducing the portion of the KB that must be considered as candidate query bindings after an update. This effectively allows a smaller subset of the KB to be considered for possible subscription matches. The techniques we present are applicable to conjunctive retrieval queries that can be rolled-up into a distinguished variable (discussed in the next section) and containing only simple roles (i.e., no transitive roles or super-roles of a transitive role). The approach extends our previous work [24] to support arbitrary KBs expressed in the description logic $\mathcal{SHI}$ (a large subset of OWL). Lastly, an evaluation of the incremental reasoning techniques is provided, demonstrating their effectiveness for OWL-based syndication.

2. Preliminaries

In this section, we briefly provide an overview of OWL and description logics, query answering for DL KBs, and tableau algorithms for DL reasoning.

2.1. The Web Ontology Language

The W3C-approved Web Ontology Language (OWL) is the recommended standard for formally representing content on the Web. One of the main benefits of OWL is the support for formal reasoning, as the semantics of a variety of its sub-languages are firmly founded in description logics (a decidable fragment of First Order Logic). In particular, the sub-language OWL DL is a syntactic variant of the description logic $\mathcal{SHOIN}$ [28], with an OWL DL ontology corresponding to a $\mathcal{SHOIN}$KB. In this work, we address a subset of $\mathcal{SHOIN}$, namely $\mathcal{SHI}$; therefore, we briefly introduce the syntax and semantics of $\mathcal{SHI}$. Let **C**, **R**, **I** be non-empty and pair-wise disjoint sets of *atomic concepts*, *atomic roles*, and *individuals* respectively. The set of $\mathcal{SHI}$ roles (roles, for short) is the set $\mathbf{R} \cup \{R^- | R \in \mathbf{R}\}$, where R^- denotes the inverse of the atomic role R. Concepts are inductively using the following grammar:

$$C \leftarrow A | \neg C | C_1 \sqcap C_2 | C_1 \sqcup C_2 | \exists R.C. | \forall R.C.$$

where $A \in \mathbf{C}$, $a \in \mathbf{I}$, $C_{(i)}$ a $\mathcal{SHI}$ concept, R a role, and S a *simple* role (i.e., no transitive roles or super-roles of a transitive role).[3] We write $\top$ and $\bot$ to abbreviate $C \sqcup \neg C$ and $C \sqcap \neg C$, respectively. A *role inclusion axiom* is an expression of the form $R_1 \sqsubseteq R_2$, where R_1, R_2 are roles. A *transitivity axiom* is an expression of the form $Trans(R)$, where $R \in \mathbf{R}$. An RBox R is a finite set of role inclusion axioms and transitivity axioms. For C, D concepts, a *concept inclusion axiom* is an expression of the form $C \sqsubseteq D$. A TBox T is a finite set of concept inclusion axioms. An ABox A is a finite set of concept assertions of the form $C(a)$ (where C can be an arbitrary concept expression), role

[2] Note that this is expressed using standard Turtle syntax (as opposed to RDF/XML) and can be easily generated in today's OWL ontology editors.

[3] See [28] for a precise definition of simple roles.