LIGHTWEIGHT SEMANTICS MODELING AND ACQUISITION FOR THE "WILD WEB"

Mária BIELIKOVÁ, Michal BARLA, Marián ŠIMKO

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
{bielik,barla,simko}@fiit.stuba.sk

Abstract. The current Web has many aspects. It is no longer only a place for content presentation. The Web is more and more a place where we actually spend time performing various tasks, a place where we look for interesting information based on discussions, opinions of others, as well as a place where we spend part of our recreation and leisure time. In addition, the Web provides an infrastructure for applications that offer various services. In this paper we concentrate on representation and acquisition of lightweight semantics for the "wild" Web, which is a must if we want to shift to a "smarter" Web and web applications, which cope with dynamic content and take into account user features to deliver personalized experience.

1. Introduction

A requirement for having an additional description of the Web content, for knowing its semantics and thus allowing machine processing, is almost as old as the Web is. A good example are classical search engines, which were relying on metadata tags manually inserted into static web pages in order to respond better to user queries.

While the semantics has been a primary focus for the Semantic Web initiative [4], it is crucial also for the typical "wild" Web of nowadays, which provides far more than a static content – it has become highly dynamic, it provides *functionality* on the top of the content and due to large amount of information it is becoming more and more *personalized* as we got used to *use* the Web through its services as our primary source of information and knowledge but also as a mean to connect with our friends or other people of our interest.

All this makes the requirement for semantics even more important – if we want to tailor functionality and underlying content to the needs of a particular visitor, we need to know not only about features of that visitor, but also about the content.

When providing semantics for web content, we interpret symbols and associate them with an appropriate sense in form of conceptual description. The complexity of descriptions may vary. Where heavyweight ontologies contain advanced structures such as axioms and enable complex reasoning, lightweight ontologies form only basic conceptual structures. Concept definition in lightweight ontologies is slightly simplified and "lightweight concepts" often represent rather a lexical reference to concepts than concepts themselves.

2. Lightweight semantics

We proposed lightweight semantics for modeling open and dynamic information spaces. Here our main concern is automatic or semiautomatic acquisition of semantics. So the question is not what we can do with the semantics when it is perfect (in sense of its formality and expressiveness), but how to acquire it for constantly changing and new content.

Our models consist of: a *designate* layer and a *metadata* layer (see Figure 1). Designate layer covers resource and user abstractions. We distinguish dual representation of resources: (1) resource instances are low-level representation of web resources (e.g., the content represented using XML, HTML), (2) resource designates constitute resource abstractions residing in a model. Differentiation between instances and designates supports the notion of reusability and extendibility in terms of content resource's lower level representation.

Metadata layer is formed by relevant domain terms – easily creatable descriptions that are related to particular topics present in the content. It is important to note that relevant domain terms do not represent concepts in strict ontological definition, cf. [5]. They rather represent lexical reference to the concepts, which form the models (unlike relevant domain terms, concepts are not explicit).

Elements in our models are interconnected via various types of relationships that represent various forms of relatedness between domain and user model elements.

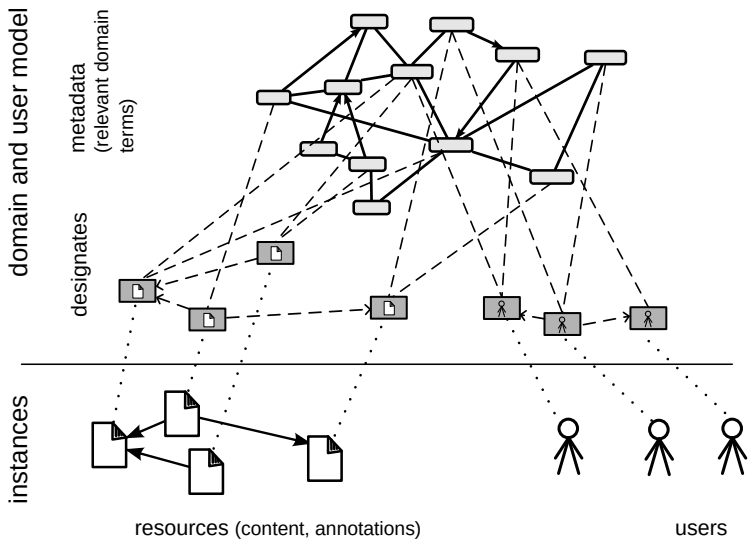

Figure 1. Domain and user model represented using the same lightweight approach.

3. Automatic semantics acquisition for the “wild” web

We have two main options when considering acquisition of descriptions for the content of the “wild” Web: to base our approaches on available content and process it to extract required metadata, and/or to rely on users themselves to provide us with an evaluation of the content, its quality and usefulness for user’s current task. This evaluation is implicit, based on analysis of digital traces of user behavior within an information space.

Both of the mentioned approaches are useful if we want to employ lightweight modeling approaches for the purpose of information processing (including personalization, intelligent search) in the “wild” information space such as the Web is. We need an ability to acquire relevant domain terms from resources such as documents visited by the users and build the domain and user models on top of them. Because the Web is an open information space, we need to track down and process every page a user has visited in order to update her model appropriately. Apart from relevant domain terms, we need to acquire additional attributes describing the user’s access to the web resource – implicit feedback indicators such as time spent actively reading a page or amount of scrolling.

To achieve this, we developed an enhanced proxy server, which allows for realization of advanced operations on the top of requests flowing from user and responses coming back from the web servers all over the Internet [2]. The actual process of relevant domain terms extraction is based on both traditional natural language processing methods as well as on new online services such as OpenCalais (opencalais.com) or Alchemy (alchemyapi.com). The latter approach has a great potential as the services return not only relevant terms, but often provide also additional metadata including a binding to the concepts of LinkedData cloud. This allows us to disambiguate meaning of homonyms and establish relationships between relevant domain terms of our lightweight model. We believe that such a bottom-up approach for incorporating semantics is more viable than approaches relying on highly formalized ontologies.

Besides relevant domain term identification also discovery of relationships between relevant domain terms is an important step in semantics acquisition process. We are particularly focused on most elemental relationships, which we believe are sufficient for multitude of tasks related to intelligent/advanced information processing. In particular, we consider *relatedness* relationship as basic form of paradigmatic similarity between terms, and *is-a* relationships that form hierarchical skeleton of domain conceptualization. We already showed that such approach is, despite the pitfalls associated with natural language processing, feasible in the domain of web-based learning [3].

4. Conclusions

The Web became a dynamic and constantly changing place. The emergence of Read/Write Web opened the web content to millions of users to collaboratively edit and organize it. At the same time, we need description of the content in order to provide advanced functionality such as smart and personalized search.

We present lightweight semantics modeling as an approach to address current challenges of the information processing on the “wild Web”. We believe that lightweight semantics approach brings a benefit of a feasible automatic acquisition of semantic descriptions, while still being sufficient for majority of advanced information processing tasks. The major

features of our approach related to both domain and user modeling for advanced/intelligent information processing are:

- Separation between domain conceptualization and content – the content and its metadata are separated in order to allow proper reusability of content (with no need to change domain conceptualization) and flexible information processing (metadata-based rather than content-based).
- Extendibility together with a possibility of new types of content – domain model facilitates definition and creation of new types of resources. This allows us to consider various types of web content that can be involved in advanced processing providing more functionality. User experience is increased.
- Reusability across various applications – a distributed nature of the Web resulted in various applications processing the web content. In order to shift advanced processing beyond one application, domain and user models are reusable across the Web supported by proxy.
- Explicit support for collaboration – interactivity and collaboration improve user experience and increases users' competences, which makes travelling through the Web more convenient.

Acknowledgement: This work was partially supported by the grants VG1/0675/11/2011-2014, KEGA 345-032STU-4/2010, APVV-0208-10 and it is the partial result of the Research & Development Operational Programme for the Project Research of methods for acquisition, analysis and personalized conveying of information and knowledge, ITMS 26240220039, co-funded by the ERDF.

References

to other papers publishing the results that are summarized here

[1] Barla, M., Bieliková, M, Šimko, M.: Lightweight Semantics for the „Wild Web". In: *WWW/Internet 2011 : Proc. of the IADIS International Conference*. IADIS Press, (2011), pp. xxv-xxxii

[2] Barla, M., Bieliková, M. 2010. Ordinary Web Pages as a Source for Metadata Acquisition for Open Corpus User Modeling. In *Proc. of IADIS WWW/Internet 2010*. IADIS Press. pp. 227–233

[3] Šimko, M., Bieliková, M. 2009. Automatic Concept Relationships Discovery for an Adaptive E-course. In Barnes, T. et al. (Eds.). *Proc. of Educ. Data Mining 2009: 2nd Int. Conf. on Educ. Data Mining*. Cordoba, Spain, pp. 171–179.

Other references

[4] Berners-Lee, Tim; Mark Fischetti (1999). Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its inventor. Britain: Orion Business. ISBN 0-7528-2090-7.

[5] Cimiano, P. 2006. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer-Verlag, ISBN: 978-0-387-30632-2. 347p.