Building a Domain Model using Linked Data Principles

Michal HOLUB*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova, 842 16 Bratislava, Slovakia
holub@fiit.stuba.sk

The Linked Data principles are being used in many datasets published on the Web. The aim of our work is to use Linked Data in order to create models describing 1) the domain of software development, and 2) the domain of research in the field of software or web engineering. We use these models to represent the knowledge of IT professionals (analysts, programmers, testers), as well as the research interests of researchers in the respective fields. Using Linked Data allows us to connect entities in our models with the ones published on the Web and get more information about them.

Linked Data principles are being used in various datasets, which form a Linked Data Cloud. In the center of this cloud there are two large datasets: DBpedia [1] and YAGO [2]. Both use Wikipedia as their primary source of information, they extract it from infoboxes and categories. These datasets define as many entities as possible, so that other datasets can link to them.

We propose a method for automatic construction of a concept map serving as a basis of our domain models. For this purpose we use unstructured data from the Web, which we transform to concepts and links between them. Using a concept map we describe the knowledge of software developers as a set of technologies and principles they are familiar with. We also use a similar concept map to describe research areas, problems, principles, methods and models studied by researchers at our faculty.

Concepts are linked together using relationships like *is part of*, *is a*, *is written in*, or *uses*. Thanks to the relationships we can later do reasoning, e.g. deduce that when a programmer knows jUnit (a testing framework for Java) he also has to know a bit of Java (a programming language), because there is a relationship stating "jUnit uses Java".

As a source of information for the concept map population we use free online encyclopedia Wikipedia. We analyze the textual content of its articles to learn new concepts and their instances.

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2013 PeWe Workshop, April 5, 2013, pp. 73–74.

We use the existing concepts as a seed in the task of ontology learning. We search all Wikipedia's articles for occurrences of concepts from our concept map. Take "programming language" as an example of a concept. Article about it also links to a "List of programming languages", from which we can extract additional concepts, which are subclasses of "programming language".

When the ontology is ready, we populate it with instances, which we do as follows:

1. Select an article from Wikipedia containing a particular concept in its text.
2. Find the first sentence containing the verb *is* followed by one of the concept types.
3. Convert the title to a new concept instance (if it is not present) and create *is a* relationship between the instance and the concept.

Using this process we not only populate the domain model with particular technology, we also find all terms which can describe a technology used when developing software.

There can be other words following the verb *is* in the article not matching any concept from our map. These could express properties of the technology and we might enhance the ontology.

The ontology can be used in a system for gathering the knowledge of programmers. Let us assume the user adds "Java EE" to his skills. We identify it as a platform in the ontology. It is related to "programming language" by *uses* link. We can generate question "Which programming language used in Java EE are you familiar with?" This way we can get more skills from the user.

The domain models we create can be used as a basis in two adaptive systems. The first aims at capturing IT professionals' knowledge and skills, deduce further technologies they might know and enables users to search for a suitable candidate for a certain task or project. The second one allows users to bookmark, annotate and collaborate over research papers in digital libraries, as well as other Web documents. Here, we also use the model in order to answer queries in pseudo-natural language.

We evaluate the models and methods of their creation directly by comparing them to existing ones or by evaluating facts from them using domain experts. Moreover, we evaluate the models indirectly by incorporating them in adaptive personalized web-based systems and measure the improvement in the experience of users (i.e. they get better recommendations, search results, etc.).

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Auer S., Lehmann J.: What Have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content. In: *The Semantic Web: Research and Applications*, LNCS Vol. 4519. Springer, (2007), pp. 503-517.

[2] Suchanek F.M., Kasneci G., Weikum G. YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In: *Proc. of the 16th int. Conf. on World Wide Web*, ACM Press, (2007), pp. 697-706.