

Semantic Annotation based on Regular Expressions

Michal Laclavik¹, Emil Gatial¹, Zoltan Balogh¹, Ondrej Habala¹, Nguyen T. Giang¹, Ladislav Hluchy¹

¹ Institute of Informatics, Slovak Academy of Sciences, Dubravská cesta 9
845 07 Bratislava, Slovakia

laclavik.ui@savba.sk

Abstract. In this paper we describe a solution for creation of ontological metadata from text documents using a semantic annotation. A lot of solutions are known for creation of ontological metadata from text documents. Our solution is based on a regular expression. The solution uses a regular expression - ontology pairs to detect ontology concepts and annotate the document with them. The solution requires having ontology describing the problem domain. The paper explains two possible applications for using such technique.

1 Pattern Ontology Model and Annotation Algorithm

The instances of the Pattern class are used to define and identify relations between text and domain ontology, where the pattern property contains the regular expression which describes textual representation of the ontology element. The examined text is processed with the regular expression for every pattern and when it is found: the detected ontology element of hasClass or hasInstance represents text in the chosen problem domain. Moreover, when the hasClass property exists in the Pattern, the RDQL query is constructed and processed to find the individuals that match the condition:

- individual is the class of hasClass
- a property of individual contains the matched word

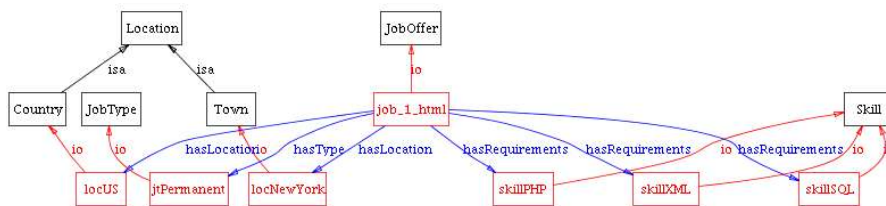


Fig. 1. Job Offer Individual with its properties detected by regular expressions showed in Figure 2 from Znalosti job offer application

Figure 1 shows results of the annotation based on patterns from Figure 2. We will describe an algorithm by an illustration example from the Znalosti project where a job offer text is analyzed and annotated by regular expressions patterns (Figure 2). Sys-