Relationship Discovery from Educational Content

Petra VRABLECOVÁ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova, 842 16 Bratislava, Slovakia
petra.vrablecova@gmail.com

The domain model is an essential part of the adaptive learning system. It expresses the semantics of educational content in the form of metadata. We consider it to be a lightweight ontology, i.e., a set of terms and relations. Manual domain model building is a challenging task for teachers, hence there is an effort to automate it. We propose a method for automated acquisition of metadata from educational content, aimed at relationships discovery between terms.

There are many generic methods for automatic metadata acquisition from text, which are based on natural language processing. We decided to explore the statistical approach and its advantages, like no need for syntax knowledge and language independence. The uniform vocabulary of educational texts indicates better results of statistical methods. Only few works deal with automated domain model acquisition for adaptive systems and course authoring support. These consider another specific of educational content – its structure allowing also the usage of graph algorithms.

The content of adaptive learning system is a set of learning objects (LO) – mainly text documents, formed into a hierarchy (a tree or a book structure). They are linked through LO-LO relationships implying their relatedness. The purpose of our method is the automated creation of a lightweight ontology which will describe our set of LOs. It consists of a set of relevant domain terms (RDT) and relationships of different types between them. RDTs are assigned to LOs through an RDT-LO relationship that implies the semantic link between them (e.g., the term is a keyword).

Our method is preceded by the LO preprocessing, e.g., normalization, removal of stop words, lemmatization. Besides preprocessed texts a set of RDTs and RDT-LO relationships are needed as input. They can be extracted from LOs (e.g., by tf-idf). RDTs have to occur in the text of LOs, because the first step of our method is the application of Latent Semantic Analysis (LSA) on the preprocessed texts and RDTs. LSA produces relationships based on similarity of words surrounding RDTs in texts. The output is a net of related RDTs, i.e. set of RDT-RDT relationships. The second

* Supervisor: Marián Šimko, Institute of Informatics and Software Engineering

Spring 2013 PeWe Workshop, April 5, 2013, pp. 107–108.

step is the discovery of hierarchical RDT-RDT relationships which comprise the core of the domain model. We propose two methods for their determination.

The first one is based on term subsumption [2]. The sets of LOs, which have assigned RDTs from an LSA relationship identified in the first step, are compared. The sets also contain LOs of RDT's neighbors from the LSA net. If sets' intersection is not empty then the RDT belonging to the bigger set is marked as superordinate (Figure 1).

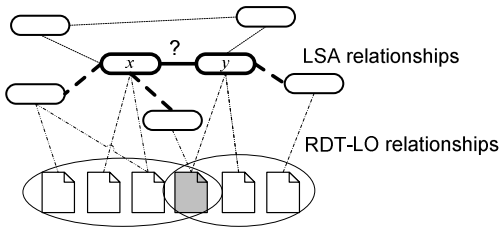

Figure 1. Determination of relationship type between terms x and y.

The second method is based on the Semantic Growback algorithm [1]. For two RDTs from an LSA relationship, are lists of top ranked RDTs created. The lists are produced by application of the PageRank algorithm with priors on the graph of LO-LO and RDT-LO relationships. If both RDTs are in both lists then the RDT which is on a higher position in both lists is marked as superordinate.

At the moment, our work is in the phase of evaluation. The goal of the evaluation is to find out whether the domain model built by our method is on the level of the manually built domain model. We perform tests on the Functional and Logic programming course. We experiment with various setups of the method and look for the optimal combinations. The preliminary results suggest that the most valuable contribution of the method is that it yields different kinds of relationships that cannot be discovered by applying linguistic approaches. The evaluation process also contains the integration of our method into an educational content management system.

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 55-60.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11.

References

[1] Diederich, J., Balke, W.: The Semantic GrowBag Algorithm: Automatically Deriving Categorization Systems. In: *Proc. of the 11th European Conf. on Research and Advanced Technology for Digital Libraries*, LNCS 4675. Springer, Berlin, 2007, pp. 1-13.

[2] Sanderson, M., Croft, B.: Deriving concept hierarchies from text. In: *Proc. of the 22nd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, ACM, 1999, pp. 206-213.