Automated Educational Course Metadata Generation Based on Semantics Discovery

Marián Šimko and Mária Bieliková

Institute of Informatics and Software Engineering,
Faculty of Informatics and Information Technology, Slovak University of Technology,
Ilkovičova 3, 842 16 Bratislava, Slovakia
{simko,bielik}@fiit.stuba.sk

Abstract. The efficiency of an educational system is related to the ability to deliver personalized content to a student. The current educational systems use advanced mechanisms for adaptation by utilizing available knowledge about the domain area. However, describing a domain area in sufficient detail to allow accurate personalization is a tedious and time-consuming task. Only few works are related to the support of teachers by discovering the knowledge from educational material. In this paper we present a method for automated *metadata* generation addressing the educational knowledge discovery problem. We employ several techniques of data mining with regards to the e-learning environment. We evaluate the method on the functional programming course.

1 Introduction and Related Work

The domain model of an adaptive course represents an area that is the subject of learning. It consists of interlinked concepts – domain knowledge elements related to learning content [1]. The concepts are mutually interconnected forming a structure similar to a lightweight ontology. In educational systems concepts are also connected to learning objects, i.e. learning material portions containing concept instances. Let us consider a programming course containing a textbook chapter describing the Fibonacci sequence. With this learning object concepts like Fibonacci, recursion, cycle, etc. are associated. Concepts are not restricted to terms appearing within the text, nor topics of the textbook chapters. The concept space including relationships we also refer to as course *metadata* as it contains data about content being taught.

The bottleneck of the adaptive educational systems lies in the complexity of authoring. Such systems may contain thousands of presentation pages and hundreds of other fragments of learning material such as examples, explanations, animations and questions. These numbers are certainly sufficient for a study of a particular subject, but defining relationships between concepts in such a space is not only difficult but also impossible for a human being.

Our goal is to support the adaptive educational course authoring by the means of knowledge discovery techniques. In this paper we propose a method of automated metadata generation by revealing semantics hidden within the text.

We show that generated metadata are useful for e-learning needs, especially for recommendation. Furthermore, the teacher's effort is reduced since we are able to create promissing number of concepts and relationships automatically.

The work related to metadata generation in the area of adaptive e-learning by means of knowledge discovery is presented in [5]. Concept similarities are computed based on the comparison of concept's domain attributes. In contrast to our meaning, the concept in [5] also holds textual representation. This should be considered as intentional description, but then the reusability of such concepts is arguable. We are not aware of any other approaches to automated concept relationship generation in the adaptive e-learning field.

Finding relations between concepts is a subtask of the ontology learning field. Relations are induced based on linguistic analysis relying on preceding text annotation [2], incorporating formal concept analysis [3] or using existing resources such as WordNet [4]. The drawback of the approaches is the dependency on precise linguistic analysis. They rely on lexico-syntactical annotations, powerful POS taggers, existing domain ontologies, huge corpuses or external semantic resources. Such knowledge is often not available during e-course authoring. The solution for content authors should involve unsupervised approaches to unburden them from additional work. This need we address in the method we propose.

The task of structuring the concept space is also present in the area of topic maps. In this field the topics can be consider analogical to concepts. Authors in [6] generate relations between topics by analyzing the HTML structure of Wikipedia documents. Categorization methods are used in [8] where similar topics are discovered by latent semantic indexing (LSI) and K-means clustering. Unsupervised methods serve as guidance in topic ontology building. A similar approach is missing in the area of adaptive e-learning. Hence, our method is based on statistical unsupervised text processing and knowledge discovery.

2 Method Description

The goal of the proposed method is the automated creation of the domain model. The metadata are created automatically under the adaptive course author's (i.e., teacher's) supervision. Thus, his effort in the authoring process is reduced. Automated steps include concept extraction and relationship discovery.

2.1 Learning Objects Preprocessing

At the beginning we create the representation of learning objects relevant for further processing. We utilize a vector space model (VSM) based on the so-called *term relevance* which is the degree of importance of the term in the text (learning object). Beside the term frequency it comprises also other qualitative characteristics of the term. Learning objects preprocessing steps are as follows:

Vector representation composition. In this step we perform a lexical analysis of learning objects. Lexical units – tokens – are identified. We remove stop words having almost no semantic significance. Then we retrieve token's

lemmas – canonical forms. From lemmas we compose vectors containing term frequencies. In this moment we have the standard bag-of-words model.

Vectors adjustment. In this step we tune up the actual vectors weights (relevance) considering factors not related to the learning object content. The adjustment consists of two steps: (1) available index processing, (2) formatting processing. An *index* of domain keywords is often available in a learning environment (in textbooks, as course outcomes, etc.). We increase the relevance of such terms by multiplying it with coefficient empirically set to 5.0. *Formatting* processing covers the relevance adjustment according to formatting in source document. We utilize the rules similar to ones presented in [6] in this step.

2.2 Pseudoconcepts Extraction

After the preprocessing step the representation allowing concept candidates – *pseudoconcepts* extraction is prepared. This step consists of three substeps:

Relevant domain terms (RDT) selection. From the set of all learning object terms we select only those whose relevance exceeds a particular threshold k, empirically set to be equal to coefficient increasing the relevance of domain keyword. This way we find terms that represent certain semantic potential.

Relevant domain terms weight computation. Using extended tf-idf measure we compute the degree of RDT relatedness to learning objects:

$$w_{i,j} = \frac{rel_{i,j}}{\sum_k rel_{k,j}} \cdot idf_i \cdot \log \frac{|LO|}{|\{lo_j : t_i \in lo_j\}|} \tag{1}$$

where $w_{i,j}$ is relatedness of domain term t_i to learning object lo_j, $rel_{i,j}$ is relevance of domain term t_i in learning object lo_j, and LO is the set of learning objects in whole course.

Pseudoconcepts extraction and relationships creation. To let a RDT be promoted to a pseudoconcept we introduce the minimal relatedness threshold $r \in \langle 0; 1\rangle$. In our experiments it is a very small number (≈ 0.05), but it effectively filters out irrelevant domain terms. Between pseudoconcepts and learning objects we create relationships with the computed relevance weight $w_{i,j}$.

2.3 Relationship Discovery

Relationship discovery is the crucial step of our method. We apply several techniques of knowledge discovery on the actual domain model (consisting of pseudoconcepts connected with learning objects only) in order to obtain the degree of mutual pseudoconcepts relatedness. For each pseudoconcept we choose the most relevant neighbors – the neighbors most related to a given pseudoconcept.

Concept-to-concept similarity computation. For this step we proposed and experimented with three concept-to-concept similarity computation variants: vector approach, spreading activation and PageRank-based analysis. Each variant of similarity score computation provides a unique view of the actual

domain model state and employs a specific approach to knowledge discovery. Detailed description is beyond the scope of this paper and can be found in [9].

Most relevant neighbors selection. Finding the appropriate number of relevant neighbors is important for the generated domain model quality. In our experiments we select neighbors that accumulate $m\%$ of the sum of all neighbors' similarity scores to a given concept.

3 Experimental Evaluation

We evaluated the proposed method in the domain of programming learning on the functional programming course. We performed the adaptive e-course creation process using the CourseDesigner authoring tool which implements the automated metadata generation method. The subject of experiment was a half-term course consisting of 70 learning objects on the functional programming paradigm and programming techniques in the Lisp language. Learning objects were organized hierarchically and represented using the DocBook language.

The resulting course structure was compared to functional programming metadata created manually by a randomly chosen sample of 2007/08 course students. Manual creation of metadata comprised the assignment of weighted values to concept relationships. As assigning continuous values from interval $\langle 0; 1\rangle$ is a non-trivial task, weight values were from $\{0, 0.5, 1\}$ implying:

0 – concepts are not related to each other (no relation),
0.5 – concepts are partially (maybe) related to each other ("weak" relation),
1 – concepts are highly (certainly) related to each other ("strong" relation).

There were 366 relationships created, 216 were weak and 150 were strong.

Prior to the method application we assumed that the learning objects were loaded into a newly created course. The dictionary of domain keywords was also provided. During the concepts extraction step 76 concepts were extracted. The relationship discovery step was performed separately for each similarity computation variant. Between the 76 concepts 420, 442, and 316 relationships were retrieved respectively (see Fig. 1). To evaluate the obtained results we tracked the number of correct relationships retrieved by the method in relation to the total number of relationships retrieved (precision) and the number of correct relationships retrieved in relationship to the total number of relevant relationships defined manually (recall). To compare the results, we combined both into F-measure which is the weighted harmonic mean of precision and recall.

In order to gain more accurate evaluation, we extended the original recall measure to involve the manually constructed domain model relationship types:

$$R^* = \frac{retrieved \cap (relA \cup relB)}{relA \cup (relB \cup retrieved)} \tag{2}$$

where R^* is the extended recall measure, $retrieved$ is the set of all relationships retrieved by the method, $relA$ is the set of manually created "strong" relationships and $relB$ is the set of manually created "weak" relationships.

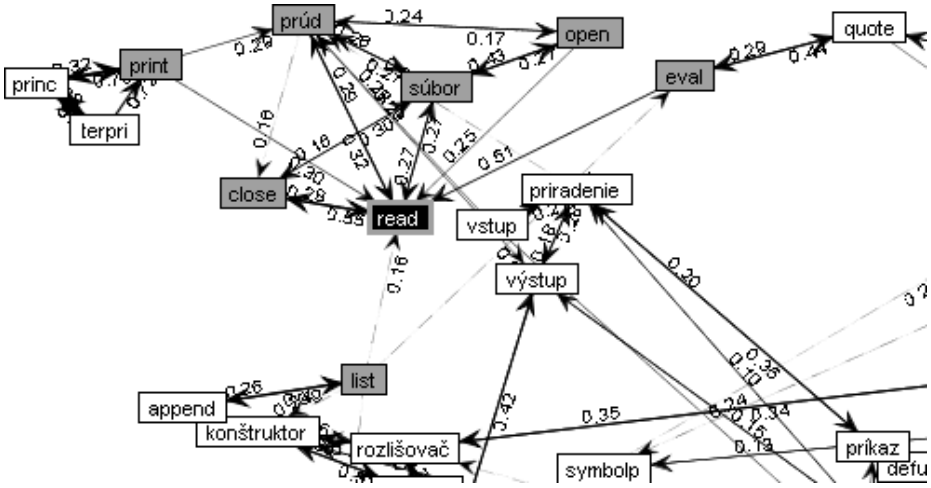

Fig. 1. Example of a domain model fragment after the relationship discovery step taken from CourseDesigner. The "read" concept is selected and its direct neighbors are colored grey. (The functional programming course is being taught in Slovak.)

The experiments yielded best results with PageRank-based analysis (F* = 0.652). The analysis of the generated relationships highlighted common NLP problems. None of the relationship discovery variants was able to significantly overcome natural language ambiguities. Less suitable results were obtained among concepts represented by terms occurring frequently, in more than one meaning or diffused over the whole course. A similar problem affected the concepts associated with a small, relatively independent group of learning objects as they were unable to create relevant connections with other semantically related concepts.

A legitimate question is what exactly does the F*-measure indicate in our experiment? We interpret it as the "completeness" of the generated metadata. Throughout the experiment, generated relationships not contained in the manually created relations were considered incorrect. Although manual relationship creators made their best effort to match real-world relations, relationships retrieved automatically need not to be irrelevant. They might represent bindings which were not explicitly realized even by the most concerned authors.

4 Conclusions

In this paper we presented an approach to the automated creation of a semantic layer over learning objects in an adaptive educational course. Our goal is to support adaptive educational course authoring and reduce the teacher's involvement. We proposed and evaluated the method for automated metadata generation based on the educational content processing. The method produces interconnected semantic elements – concepts.

Unfortunately, universal solutions enabling automatic metadata acquisition probably do not exist. To a certain degree complex domain ontologies may be used. However, they are currently not available as much as we want. Furthermore, it is questionable if we ever can produce course metadata (relationships between the concepts in particular) on a desired level of granularity.

The main contribution of this paper is the proposed method and the corresponding framework for preprocessing, discovering and finalizing the domain model structure which is crucial for successive reasoning on the semantically enriched content. As opposed to most current approaches which are limited to the content annotation when creating metadata, we go one step beyond by discovering both concepts and links ultimately creating a metadata layer above the learning objects. Without proper structure of the metadata it is not possible to reason and adapt navigation and presentation in large information spaces.

The proposed approach is not limited to learning objects represented by text. We can work with media content employing similarity measures for their interconnection and tagging for interconnections between the metadata layer and the content layer.

Acknowledgments. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 3/5187/07 and by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

1. Brusilovsky, P. Developing adaptive educational hypermedia systems: From design models to authoring tools. In Murray T., Blessing S., Ainsworth, S. (eds.): Authoring Tools for Advanced Technology Learning Environment. Kluwer, pp. 377-409.
2. Buitelaar, P., Olejnik, D., and Sintek, M. A protg plug-in for ontology extraction from text based on linguistic analysis. In Proc. of the 1st European Semantic Web Symposium (ESWS), 2004.
3. Cimiano, P., Hotho, A., Staab, S. Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. In Journal of AI Research, vol. 24,pp.305-339, 2005.
4. Cimiano, P., et al. Learning Taxonomic Relations from Heterogeneous Evidence. In Proc. of ECAI Workshop on Ontology Learning and Population, 2004.
5. Cristea, A. I., de Mooij, A. Designer Adaptation in Adaptive Hypermedia. In Proc. of Int. Conf. on Information Technology: Computers and Communications ITCC'03. Las Vegas, 2003. IEEE Computer Society.
6. Dicheva D., Dichev C. Helping Courseware Authors to Build Ontologies: the Case of TM4L. In 13th Int. Conf. on Artificial Intelligence in Education, 2007, pp. 77-84.
7. Diedrich, J., Balke, W-T. The Semantic GrowBag Algorithm: Automatically Deriving Categorization Systems. In Proc. of the 11th European Conf. on Research and Advanced Technology for Digital Libraries, ECDL 2007, pp 1-13.
8. Fortuna, B., Grobelnik, M., Mladenic, D. Semi-automatic Construction of Topic Ontology. In Semantics, Web and Mining, Joint Int. Workshop, EWMF 2005 and KDO 2005, Porto, Portugal, October 3-7, 2005.
9. Šimko, M., Bieliková, M. Automatic Concept Relationships Discovery for an Adaptive E-Course. In Proc. of 2nd Int. Conf. on Educational Data Mining, EDM 2009. Cordoba, Spain, 2009. Accepted.