Ontology based Automatic Module Generation from E-book

Keerthana.R
Assistant Professor
Jayam College of Engineering and Technology
Dharmapuri

Jayashree.N.R
Assistant Professor
Jayam College of Engineering and Technology
Dharmapuri

ABSTRACT

Technology Supported learning Systems have proved to be helpful in many learning situations. These systems require an appropriate representation of the knowledge to be learned, the Domain Module. The authoring of the Domain Module is cost and labor intensive. A novel DOM-Sortze is a system that uses natural language processing techniques, heuristic reasoning, and ontology for the semiautomatic construction of the Domain Module from electronic textbooks. To determine how it might help in the Domain Module authoring process, it has been tested with an electronic textbook. Its work presents novel DOM-Sortze and describes the experiment carried out. Novel DOM-Sortze comprises improving the generation of the LDO. It is planned to enhance the grammar for identifying pedagogical relationships. Novel DOM-Sortze is currently able to process images in the electronic document, it only considers their position in the text, and not where the image is referenced. Novel DOM-Sortze is being enhanced to support multilingual Domain Module generation. The LDO ontology supports the multilingual representation of the domain topics, and machine translation might be used to get approximate translations of the gathered LOs, used for searching and retrieving from the LOR or web pages.

Keywords

Domain module, LDO, DOM-Sortze, heuristic, Technical Supported Learning Systems, Ontology.

1. INTRODUCTION

The introduction of new technologies and scientific research are causing the revolution of information. This revolution can make some changes in the formal learning system. Today most of the learners choose the online education system because it is more flexible than others. Technology Supported Learning Systems (TSLSs), such as intelligent tutoring systems (ITSs), adaptive hypermedia systems (AHSs), and, especially, learning management systems (LMSs) such as Moodle1 or Black- board, are being widely used in many academic institutions and becoming essential for education. The effectiveness of Technology Supported Learning System (TSLS) is based on their module clustering. The proper domain modules give the proper understanding of data. The domain module contains the detailed description of the pedagogical representation of data. The Domain Module is considering as the core of any TSLSs it represents the knowledge about a subject matter that easily communicated to the learner. The Domain Module enables either the students to learn by themselves, in the case of exploratory learning systems, or to guide students through the learning process in instructive TSLSs.

The clustering of domain module is a hard task, because the modules extracted from large database normally it contain penta bytes of data. This work focuses on effective clustering of Electronic text book or E-book. Electronic textbook authors deal with similar problems while writing their documents, which are structured to facilitate comprehension and learning. Electronic textbooks might be used as the source to build the Domain Module. Introduce the system Novel Dom-Sortze, it overcome the difficulties of module clustering. Novel DOM-Sortze generates semiautomatic Domain Module from electronic textbooks. The Novel Dom-Sortze system including many techniques to develop effective domain modules, they are Natural Language Processing (NLP), clustering and Learning Domain Ontology. It aims to be domain-independent, so no domain-specific knowledge is used except the processed electronic textbook and the knowledge gathered from it. It also describes the Domain Module generation process, which has three main tasks: preparing the document for knowledge extraction, building the ontology that describes the domain and the generation of the learning objects (LOs).

The artificial intelligence methods and techniques such as Natural Language Processing (NLP) and heuristic reasoning to achieve the semiautomatic generation of the Domain Module. In this work, the Domain Module encodes knowledge at two different levels, the learning domain ontology (LDO) and the set of LOs. The Natural Language Processing (NLP) perform the task more user friendly that provides the interaction between user and learning system. Naïve Byes clustering algorithm is used for clustering of data, in this clustering based on their caption and inter relationship. The each and every Domain Module of the E-book depends on the pedagogical relationship of entities. These domain modules are checked internally and externally. Externally checking performs the caption checking of modules then that checks the inter relations of data entities after the checking ontology is applied to the domain modules. That formally describes types, relationship and properties of data then that stored in to the database it used for online learning system.

This paper presents Novel Dom-Sortze, a framework for the semiautomatic generation of the Domain Module from electronic textbooks. Novel Dom-Sortze aims to be domain independent, so no domain-specific knowledge is used except the processed electronic textbook and the knowledge gathered from it. Section 2 gives the work related to this paper. Section 3 describes the system model which entails two tasks: Algorithm such as Dom-Storze and Disadvantages (see Section 4 is about the proposed system. Finally, conclusions and future work implementation.

2. RELATED WORK

Learning Objects can be reused to support learning in different platforms or environments. A system that automatically builds Learning Objects from electronic textbooks using Natural Language Processing techniques, ontologies and heuristic reasoning is presented. ErauzOnt was able to gather definitions, examples and exercises for the topics of the Object Oriented Programming subject.[1]. Technology Supported Learning Systems have proved to be useful in learning situations, its development is hard task. Elkar DOM is a technique that provides a collaborative Domain Module authoring tool. [10].

Content reuse is one of the major concerns in the Technology *ErauzOnt* is a system that uses ontologies, Natural Language Processing techniques, and heuristic reasoning to generate Learning Objects from textbooks. It has been tested with several textbooks written in the Basque language in order to evaluate the automatic construction of Learning Objects[2].Retrieving and reusing Learning Objects (LOs) can lighten the workload of constructing new on-line courses or Technology Supported Learning Systems (TSLSs). This paper presents ErauzOnt, a framework for the automatic generation of new LOs from electronic documents using domain ontologies and NLP techniques. The semi-automatic development of ontologies is an important field of research. Introduce Wiktionary, which is collaborative online dictionary encoding information about words, word senses, and relations between them. One particular advantage of Wiktionary is its multilingual nature, which allows the construction of ontologies for different languages. For constructing their ontology Onto Wiktionary [3].

3. SYSTEM MODEL

Domain Module separation is a hard task which entails not only selecting the domain topics to be learned, but also defining the pedagogical relationships among the topics that determine how to plan the learning sessions. Textbook authors deal with similar problems while writing their documents, which are structured to facilitate comprehension and learning. Electronic textbooks might be used as the source to build the Domain Module, reproducing how average teachers behave while preparing their subjects: they choose a set of reference books that provide the main didactic resources (DRs)—definitions, examples, exercises—for the subject, and rely on them for scheduling their lectures.

3.1 Text Book Preprocessing

First, the document must be prepared for the subsequent knowledge acquisition processes. Electronic documents are available in many different formats, such as pdf, rtf, doc, or odf, a preprocess is carried out first to prepare the document. The content of electronic documents is organized using a hierarchical structure; documents contain chapters, which in turn are divided into sections, and so on. A tree-like internal representation of the document is built, so that the rest of the task can be performed with no dependence on the format the original document is stored in., and the outcomes are then used to gather the two levels of knowledge encoded in the Domain Module.

3.2 LDO Gathering

At this phase, the domain topics to be mastered, as well as the pedagogical relationships among them are identified and represented in the LDO. Ontology learning, i.e., gathering domain ontologies from different resources in an automatic or semiautomatic way has been addressed in many works Most of these projects aim at building or extending a domain ontology or populating lexical ontologies such as Wordnet. Ontology learning usually combines machine learning and NLP techniques to build domain ontologies or to enhance and populate some base ontologies.

3.3 Preprocessing Module

The system prepares the electronic document and gathers a standardized representation of it, to later run the knowledge acquisition processes. As electronic documents are available in many different formats, such as pdf, rtf, doc, or odf, a preprocess is carried out first to prepare the document. The content of electronic documents is organized using an ontology structure; documents contain chapters, which in turn are divided into sections, and so on. A tree-like internal representation of the document is built, so that the rest of the task can be performed with no dependence on the format the original document is stored in. In addition, the outline of the document, which might be located either at the beginning or the end of the document, can also be numbered or indented in different ways showing its structure. Thus, a homogenized internal representation of the outline is also gathered in the preprocessing.

3.4 NLP Processing

Artificial intelligence methods and techniques such as natural language processing (NLP) and heuristic reasoning to achieve the semiautomatic generation of the Domain Module. In this work, the Domain Module encodes knowledge at two different levels, the learning domain ontology (LDO) and the set of LOs. LDO gathering. At this phase, the domain topics to be mastered, as well as the pedagogical relationships among them are identified and represented in the LDO. The LDO will allow either the TSLS to plan the learning session or the students to guide themselves during the learning process. NLP techniques to build domain ontologies or to enhance and populate some base ontologies. Different kinds of resources such as text corpora, document warehouses, machine readable dictionaries, or lexical ontologies are broadly used as sources of information for ontology learning.

3.5 Domain Module

In this work, the Domain Module encodes knowledge at two different levels, the learning domain ontology (LDO) and the set of LOs. The following steps are carried out to develop the Domain Module. Textbook preprocessing: First, the document must be prepared for the subsequent knowledge acquisition processes. The outcomes are then used to gather the two levels of knowledge encoded in the Domain Module. LDO gathering: At this phase, the domain topics to be mastered, as well as the pedagogical relationships among them are identified and represented in the LDO. The LDO will allow either the TSLS to plan the learning session or the students to guide themselves during the learning process. The LOs gathering. At this stage the LO—definitions, examples, exercises, and so on to be used during the learning process are identified and generated In this semiautomatic approach, the outcome of gathering the LDO and the LOs can be supervised by teachers and instructional designers both individually and collaboratively using lingo, a concept-map-based tool for the supervision of the Domain Module authoring process. Teachers could, this way, adapt the resulting Domain Module to their requirements or teaching preferences. Next, each of the steps is described in detail. The work here described has been applied on electronic documents written in the Basque language, but for the sake of readability, the examples will be shown in both Basque and English, although some information might be lost in translation.

3.6 Construct LDO

Ontology learning usually combines machine learning and NLP techniques to build domain ontologies or to enhance and populate some base ontology. Different kinds of resources such as text corpora, document warehouses, machine readable dictionaries, or lexical ontologies are broadly used as sources of information for ontology learning. The main sources of information for acquiring the LDO in a semiautomatic way as they are usually well structured and contain the main topics of the domain. Besides, they are considerably summarized, and therefore meaningful information can be extracted with a low-cost process. The reason is that authors of textbooks have previously analyzed the domain and decided how to organize the content according to pedagogical principles. Provided that the organization of the textbook is reflected in the outline, NLP techniques and a collection of heuristics are used to infer the implicit pedagogical relationships.

4. PROPOSED SYSTEM

This project presents Novel DOM-Sortze, a framework for the semiautomatic generation of the Domain Module from electronic textbooks. DOM-Sortze aims to be domain independent, so no domain-specific knowledge is used except the processed electronic textbook and the knowledge gathered from it. Novel DOM-Sortze is not aimed at building exhaustive domain ontology, but at providing aids to build ontology for didactic purposes. While most ontology learning approaches combine many resources or are restricted to certain particular domains, DOM-Sortze is domain-independent, and relies exclusively on the electronic textbook provided. In this experiment a textbook used in the mandatory secondary school was analyzed, and this might have limited the recall of the generation of the LDO.

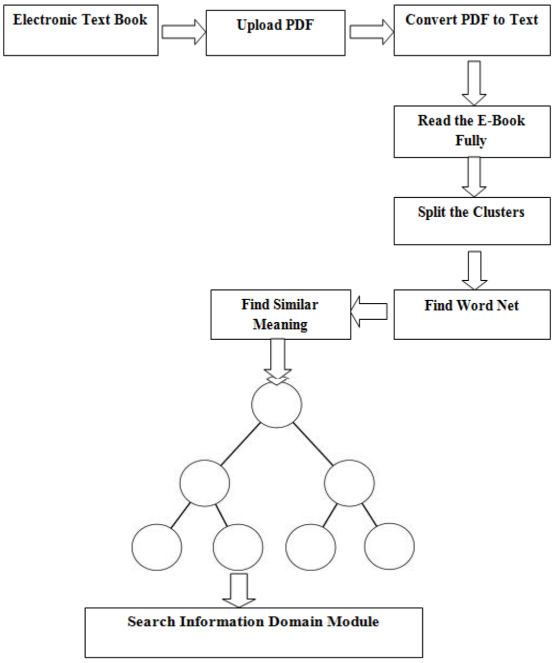

Figure:1 Proposed Architecture

4.1 Lingo Algorithm

When designing a web search clustering algorithm, special attention must be paid to ensuring that both content and description (labels) of the resulting groups are meaningful to humans. The automatically generated summaries of the original documents and hence are usually very small. Although SVD is capable of dealing with noisy data, without sufficient preprocessing, the majority of discovered abstract concepts would be related to meaningless frequent terms. The aim of the preprocessing phase is to prune from the input all characters and terms that can possibly affect the quality of group descriptions. Three steps are performed: text filtering removes HTML tags, entities and non-letter characters except for sentence boundaries. Next, each snippet’s language is identified and finally appropriate stemming and stop words removal end the preprocessing phase. We used stop words as potential indicators of a document’s language.

4.2 Performance and Result

Currently, Lingo is being enhanced to support new languages such as English. In fact, the acquisition of LOs has already been adapted and tested on a textbook on Object Oriented Programming obtaining similar results to those presented in this work. The analyzed book has Pages and words. The DR grammar for English achieved accuracy versus the percent achieved for Basque. The biggest differences were observed in the rules for definitions and problem-statements. The rules for definitions performed better in English, probably because the sentences employed in the book are shorter, and therefore less complex. The performance on problem-statements was worse for English. The identification of problem-statements in Basque is facilitated by an auxiliary verb used for imperative cases

The heuristic analysis works under the assumption that only one kind of structural relationship can exist between an outline item and all its sub items, as this fact was observed in almost all the analyzed outlines.

- Searching for e booking very complex
- Will be taken over time
- Produce duplicate result, and repeated same data

It is done only for two languages Basque and English. Not supported for Multi Lingual.

5. CONCLUSION

We presented Novel DOM-Sortze, a system for the semiautomatic generation of the Domain Module from electronic textbooks. The system employs NLP techniques, heuristic reasoning, and ontologies for the knowledge acquisition processes. DOM-Sortze has been tested using an electronic textbook and comparing the automatically generated elements with the Domain Module manually developed by instructional designers. The aim was to evaluate how DOM-Sortze contributes to Domain Module authoring. The electronic document used for the experiment was one of the books, written in the Basque language, used in the Nature Sciences subject in the first course of mandatory secondary education. As the experiment aimed to measure the knowledge acquisition from text, a version without images of the document was used as the source of data.

Novel DOM-Sortze is being enhanced to support new languages such as English. In fact, the acquisition of LOs has already been adapted and tested on a textbook on Object Oriented Programming obtaining similar results to those presented in this work. The analyzed book has 67 pages and 29,300 words. The DR grammar for English achieved 80.09 percent accuracy versus the 70 percent achieved for Basque. The biggest differences were observed in the rules for definitions and problem-statements. The rules for definitions performed better in English, probably because the sentences employed in the book are shorter, and therefore less complex. The performance on problem-statements was worse for English. The identification of problem-statement sin Basque is facilitated by an auxiliary verb used for imperative cases.

However, identifying imperative cases in English is harder. 75.93 percent of the LOs were gathered with 86.79 percent precision.

6. FUTURE IMPLEMENTATION

Novel DOM-Sortze comprises improving the generation of the LDO. It is planned to enhance the grammar for identifying pedagogical relationships to increase the recall of the relationships. Alternative ways to gather prerequisite relationships, which have a very poor recall, will be also tested. Besides, attributes of the domain topics, such as the domain relevance or the difficulty, which might be estimated using term hood measures are aimed to be automatically gathered. Although DOM-Sortze is currently able to process images in the electronic document, it only considers their position in the text, and not where the image is referenced, and therefore useful. Thus, the treatment of images must be improved. DOM-Sortze is being enhanced to support multilingual Domain Module generation. The LDO ontology supports the multilingual representation of the domain topics, and machine translation might be used to get approximate translations of the gathered LOs, used for searching and retrieving from the LOR or web pages. Besides, machine learning methods are planned to be used to infer new rules that might improve the identification of pedagogical relationships or the DRs in the electronic textbooks.

7. REFERENCES

[1] A. Conde, M. Larran˜aga, I. Calvo, J.A. Elorriaga, and A. Arruarte, “Automating the Authoring of Learning Material in Computer Engineering Education,”Proc. 42nd IEEE Frontiers in Education Conf. (FIE ’12),pp. 1376-1381, 2012.

[2] M. Larran˜aga, A. Conde, I. Calvo, A. Arruarte, and J.A. Elorriaga, “Evaluating the Automatic Extraction of Learning Objects from Electronic Textbooks Using Erauzont,” Proc. 11th Int’l Conf. Intelligent Tutoring Systems (ITS’12), pp. 655-656, 2012.

[3] Semi-AutomaticOntology Development: Processes and Resources, M.T. Pazienza and A. Stellato, eds., IGI Global, 2012.

[4] M. Larran˜aga, I. Calvo, J.A. Elorriaga, A. Arruarte, K. Verbert, and E. Duval, “ErauzOnt: A Framework for Gathering Learning Objects from Electronic Documents,”Proc. 11th IEEE Int’l Conf. Advanced Learning Technologies (ICALT ’11),pp. 656-658, 2011.

[5] P.-S.D. Chen, A.D. Lambert, and K.R. Guidry, “Engaging Online Learners: The Impact of Web-Based Learning Technology on College Student Engagement,”Computers and Education,vol. 54, no. 4, pp. 1222-1232, May 2010.

[6] A. Zouaq and R. Nkambou, “Evaluating the Generation of Domain Ontologies in the Knowledge Puzzle Project,”IEEE Trans. Knowledge and Data Eng.,vol. 21, no. 11, pp. 1559-1572, Nov. 2009.

[7] E. Agirre, O.L. de Lacalle, and A. Soroa, “Knowledge-Based WSD and Specific Domains: Performing Better Than Generic Supervised WSD,”Proc. 21st Int’l Joint Conf. Artifical Intelligence (IJCAI ’09),pp. 1501-1506, 2009.

[8] S. Ternier, D. Massart, F.V. Assche, N. Smith, B. Simon, and E. Duval, “A Simple Publishing Interface for Learning Object Repositories,”Proc. World Conf. Educational Multimedia, Hypermedia, and Telecomm. (ED-MEDIA ’08),pp. 1840-1845, 2008.

[9] K. Verbert, X. Ochoa, and E. Duval, “The ALOCOM Framework: Towards Scalable Content Reuse,”J. Digital Information, vol. 9, no. 1, 2008.

[10] M. Larran˜aga, J.A. Elorriaga, and A. Arruarte, “A Heuristic NLP Based Approach for Getting Didactic Resources from Electronic Documents,”Proc. European Conf. Technology Enhanced Learning (EC-TEL ’08),pp. 197-202, 2008.

[11] M. Larran˜aga, I. Niebla, U. Ruedat, J.A.Elorriaga “Towards Collaborative Domain Module Authoring,”Proc. Seventh IEEE Int’l Conf. Advanced Learning Technologies(ICALT ’07), pp. 814-818, July 2007.

[12] P.-S.D. Chen, A.D. Lambert, and K.R. Guidry, “Engaging Online Learners: The Impact of Web-Based Learning Technology on College Student Engagement,”Computers and Education, vol. 54, no. 4, pp. 1222-1232, May 2010.

[13] I. Aduriz, I. Aldezabal, I. Alegria, X. Artola, N. Ezeiza, and R.Urizar, “Euslem: A Lemmatiser/Tagger for Basque,” Proc. EURALEX, vol. 1, pp. 17-26, 1996.