International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 2 | ISSN : 2456-3307

Stemming Process for Automatic Generation of Domain Module From E-Textbooks

Dhivya D, Kalaiarasi M, B. Uma Maheswari

Department of Computer Science and Engineering, Prince Dr. K. Vasudevan College of Engineering and Technology, Anna University, Chennai , Tamilnadu, India

ABSTRACT

To be effective, TSLSs (Technology- Supported Learning Systems) require an appropriate Domain Module.The Domain Module is considered the core of any TSLSs as it represents the knowledge about a subject matter be communicated to the learner.Building the Domain Module 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. Artificial intelligence techniques provide the means for the semiautomatic construction of the Domain Modules from electronic textbooks which may significantly contribute to reduce the development cost of the Domain Modules. DOM-Sortze is 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.

Keywords : Knowledge Acquisition, Domain Engineering, Ontology Design

I. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing business to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. The Domain Module is considered the core of any TSLSs as it represents the knowledge about a subject matter to be communicated to the learner.The most commonly used techniques in data mining are:

Artificial Neural Networks : Non-linear predictive models that learn through training and resemble biological neural networks in structure.

Decision Trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset.

Genetic Algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

Nearest Neighbor Method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k³1). Sometimes called the k-nearest neighbor technique.

Rule induction: The extraction of useful if-then rules from data based on statistical significance.

II. METHODS AND MATERIAL

1. Building the Domain Module

The approach here presented uses artificial intelligence methods and techniques such as natural language processing(NLP) and heuristic reasoning to achieve the semiautomatic generation of the Domain Module.

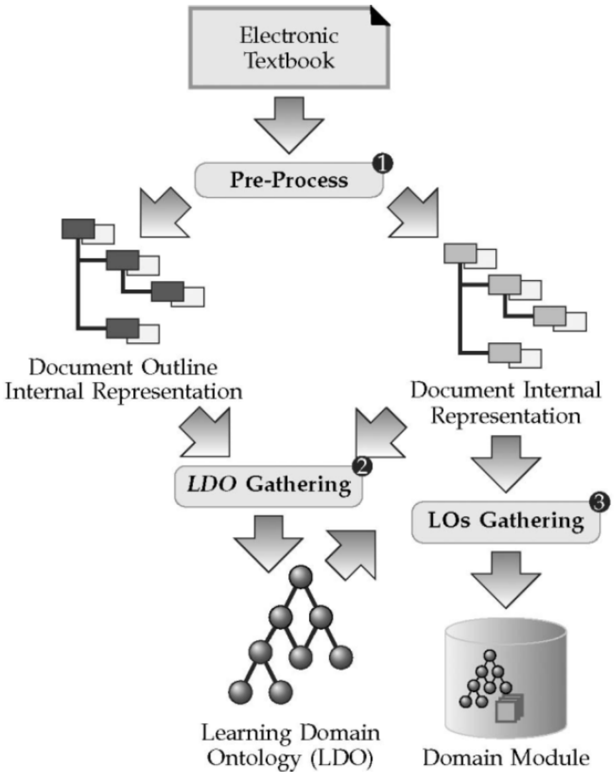

Electronic Textbook
Pre-Process
1
Document Outline Internal Representation
Document Internal Representation
LDO Gathering
2
LOs Gathering
3
Learning Domain Ontology (LDO)
Domain Module

Figure 1. Domain Module Building Process.

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 (see Fig. 1):

1. Textbook Preprocessing: First, the document must be prepared for the subsequent knowledge acquisition processes. This process is described in Section and the outcomes are then used to gather the two levels of knowledge encoded in the Domain Module.

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. The LDO will allow either the TSLS to plan the learning session or the students to guide themselves during the learn -ing process.

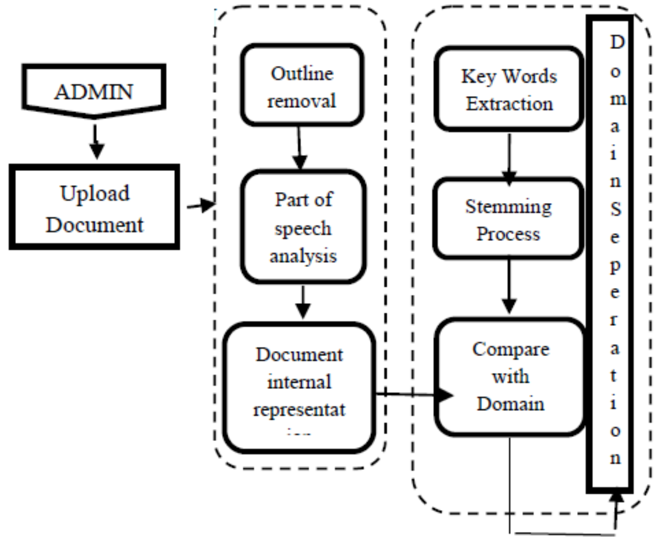
3. LOs Gathering: At this stage the LO includes definitions, examples, exercises to be used during the learning process are identified and generated. In this semiautomatic approach, the outcome of gathering the LDO and the LOs can be supervised by teachers and instructional designers both individually and collaboratively using Elkar-DOM.A concept-map-based tool for the supervision of the Domain Module authoring process.Retrieving and reusing Learning Objects (LOs) can lighten the workload of constructing new on-line courses or Technology Supported Learning Systems (TSLSs).The analysed textbooks tackle subjects such as cosmology, geology and anatomy. They are being used to build the domain module, i.e., a course, for Technology Supported Learning Systems (TSLSs) such as web-based intelligent tutoring systems or e-leaning systems.Machine Translation might be used to generate Los in other languages or at least to facilitate the retrieval of similar LOs in other languages. The metadata should be enhanced to contain references to “equivalent” LOs in other languages.As electronic documents are available in many different formats, such as pdf, rtf, doc, or odf, a pre-process is carried out first to prepare the document.

2. Related Works

Here, we discuss the related work and presents the contribution of finding a right way to develop a domain module to view books. DOMAIN GENERATION: The automatic or semiautomatic generation of the Domain Module for TSLSs has been rarely addressed presented a system for automatically building ITSs from machine readable representations of textbooks and proposed an environment to build ITSs from spreadsheets. The first requires the instructional designers to transcribe the textbook to a formal representation that can be processed, while the latter is limited to the mathematics domain. Nevertheless, there have been many attempts to automatically gather domain ontologies from diverse sources. It gathers a domain ontology from a set off text-based LOs with the aim of enhancing them with more knowledge. The authors report several experiments where the achieved recalls vary from 86.65 to 90.84 percent for classes, 75.28 to 84.33 percent for taxonomic relationships, and 80.08 to 93.12 percent conceptual relationships. TEXCOMON was also tested against TEXT-TO-ONTO[41], which obtained 73.06 percent success rates for classes,47.53 percent for hierarchical relationships, and 0.31to53.03

percent for conceptual relationships. Onto Learn [42] has been used to develop ontologies for tourism and economy. It uses online corpora, glossaries, and documents as the source for the ontology learning process, and has reported a 79.3 percent recall of the relationships for the tourism ontology and 45.5 percent for economy. The terminology recall was 55 percent. DOM-Sortze is not aimed at building an exhaustive domain ontology, but at providing aids to build an 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.

3. Architecture of Domain Module Generation



Admin login : Admin will maintain the database. The admin will login by giving their name and password. The admin will upload the book in to the specified DB. Outline Removal : The External document representations use standardized file formats such as JPEG, postscript etc..,It only accepts the pdf document. Files are readily available artifacts, which can be used to study document representations. Internal Representation : Hash value generated for the images, tables and formulas present in the document. Then, the Tokenisation process will take place. It is a process of extracting the phrases without grammer and the grammer gets omitted. Categorization : Stemming process will reduce the word to the root form, where lemmatization is concerned with linguistics i believe lemmatization is "go", "gone", "going", "goes", "been"and"went" where stemming a word would be reducing a word from "gone" to "go", so it can be matched to other stemmed words such as "going",as"going"stemmed would be "go" also, example:engineering,engineers,engineered,engin -eer these four words would not match up if they tested equality, however by stemming these words we can reduce them to a more basic form,

engineering-->engineer
engineers-->engineer
engineered-->engineer
engineer-->engineer

now we have stemmed words they will match for equality, so now if i try searching using the word for engineer, documents on engineering, engineers and engineered would be returned from a stemmed index/database. Stemming usually means to cut off characters from the end of the word,e.g. walked -> walk, walking -> walk. However, this does not necessarily produce a real word, e.g. a stemmer could also change house and houses to "hous". Also, cutting of characters isn't enough for irregular words, e.g. you cannot get from "went" to "go" by just cutting of characters. A lemmatizer solves these problems, i.e. it always produces real words, even for irregular forms. It usually needs a table of irregular forms for this, reducing words to a root form(stemming) - changing words into the basic form (lemmatization) . Upload Document ADMIN Outline removal Part of speech analysis Document internal representat ion Key Words Extraction Stemming Process Compare with Domain Domain Seperation Streaming process-Passing the particular value to its directed path. The directed path represents the specified domain. User login : The new user has to create the account by giving their details. The user has to login by giving their details. Once login completed the user provided with the domain separation page. The user can select the domain and refer the document.

4. Algorithm

step1(): Get rid of plurals and -ed or -ing. e.g:
caresses -> caress
ponies -> poni
ties -> ti
caress -> caress
cats -> cat
feed -> feed
agreed -> agree
disabled -> disable

matting -> mat
mating -> mate
meeting -> meet
milling -> mill
messing -> mess
meetings -> meet

step2(): It turns terminal y to i when there is another vowel in the stem.

step3(): It maps double suffices to single ones. so -ization (= -ize plus-ation) maps to -ize etc. note that the string before the suffix must give m() > 0.

step4(): It deals with -ic-, -full, -ness etc. similar strategy to step3.

step5(): It takes off -ant, -ence etc., in context <c>vcvc<v>.

step6(): It removes a final -e if m() > 1.Stem the word placed into the Stemmer buffer through calls to add().Returns true if the stemming process resulted in a word different from the input. You can retrieve the result with get ResultLength()/getResultBuffer() or toString().

5. Gathering 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. 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.In the approach here presented, the LDO contains the main domain topics and the pedagogical relationships among them. Pedagogical relationships can be structural isA and part of or sequential prerequisite and next. The X isA !Y relationship declares that the topic X is a particular kind ofY . TheX part of !Y describes thatX is part of Y , i.e.,X is one ofthe topics to learn to fully master Y . The Y prerequisite !X relationship states that a topic X must be mastered before attempting to learn topic Y, while X next !Y expresses that it is recommended to learn topic Y right after mastering topic X.Ontology learning relies on the assumption that there is semantic knowledge underlying syntactic structures. For example, Text2Ont uses Hearst's patterns to gather taxonomic relationships, and nested term-based methods to identify the set of candidate domain topics. OntoLT, a Prote´ge´3 plug-in for the extraction of ontologies from text, uses a pattern that identifies taxonomic relationships on nested terms, relying on the appearance of a term and some modifiers (genus et differentiam).The LDO describes a certain domain for learning purposes ; it is an application ontology according to Guarino's classification and it also describes pedagogical knowledge. The LDO acquisition entails two main NLP- andheuristic-reasoningbased steps: outline analysis, whichresults in an initialversion of the LDO, and the documentbody analysis, which enhances the ontology with newtopics and relationships.During the LDO gathering process, an internal representation is used; in this representation, besides the learningtopics and the relationships, information about thegathering process itself-used heuristics,confidence on the heuristics, and so on is also included. Once the LDO has been gathered and reviewed by teachers or instructional designers using Elkar- DOM, the ontology is represented in OWL.

GATHERING LOS FROM DOCUMENTS The generation of LOs for the domain topics is achieved by identifying and gathering DRs, i.e., consistent fragments ofthe document related to one or more topics with a particular educational purpose. The identification and extraction of these pieces is carried out in an ontology-driven process that also uses NLP techniques. As the LO generating approach presented in this work aims to be domain independent, the only domain-specific knowledge used is the LDO that has been gathered from the electronic textbook in the previous phase. From now on, a DR will refer to a piece of the document meant to be used in the learning sessions (e.g., definition, exercise, . . .) while a LO refers to a reusable DR enriched with metadata. The LO generation process here described is carried out by ErauzOnt, which is part of the DOM-Sortze framework.Fig.2 describes the process for gathering the LOs from the electronic document, which entails the following tasks: generating DRs from the document, annotating the DRs to become LOs, and, finally, storing the generated LOs in a LOR for further use. The LDO, a DR grammar, discourse markers and a didactic ontology are used to gather DRs from the internal representation of the electronic textbook with the part-of-speech information. DR Generation is described in depth. The LDO, and the ALOCOM ontology are used to build the LOs from the gathered DRs and, finally, the LOs are stored in the LOR to facilitate their use and reuse.

6.1 LO Storage

Once the LOs and their preview files have been generated, they are inserted in the LOR to allow their retrieval and use in TSLSs. The LO publishing service is based on the SPI specification [36]. The LOR can be queried to find the appropriate LOs using SQI [37]. When the LO is composed, all its components are also appropriately labeled and stored in the LOR, as they might be useful in certain contexts.

III. RESULTS AND DISCUSSION

DOM-Sortze, the system presented throughout this paper, has been tested, with the intention of validating it, with a textbook provided. The main goal of this experiment was to evaluate how DOM-Sortze helps the teachers to build the Domain Module by measuring the knowledge, either in the LDO or the LOs, automatically gathered from the textbook. For the experiment, only text-based LOs were considered, so an adapted version of the electronic textbook in which the images were removed was used. The analyzed textbook has 30 pages, 29 if the outline is ignored and 8,495 words. The outline of the document has two levels: three main items, which have five sub items each, except the second item, which has four sub items. To evaluate the process of generation of the DomainModule using DOM-Sortze, a reference LDO and DR setwere needed to compare the obtained results. Therefore, three instructional designers collaborated to manually develop the LDO and identify fragments of the documents to be used as DRs for the domain topics. The instructional designers reached a consensus about the relevant domain topics and the pedagogical relationships among them to define the LDO.The generation of the LDO is evaluated based on the amount of automatically gathered knowledge, i.e., domain topics and pedagogical relationships and the correctness of the proposed topics and relationships. The details of this evaluation are presented. The evaluation of the LO generation considered the adequacy of the identified LOs, to which end the automatically gathered LOs were compared to the manually identified DRs.

Evaluation of the Gathered LDO

The evaluation of the construction of the LDO was carried out comparing the automatically identified domain topics and pedagogical relationships to the reference LDO, i.e., the LDO collaboratively defined by the instructional designers. The reference LDO entailed 83 domain topics, and 135 pedagogical relationships. Besides, the instructional designers classified the LDO elements in two levels according to their relevance. Level 1 entailed the most relevant domain topics as well as all the pedagogical relationships in which at least one of the topics. The rest were classified as Level 2elements. The elicitation of the LDO is carried out in two steps: first the outline is analyzed to gather the initial LDO and, then, the document body is processed to identify new topics and relationships.

Evaluation of the Gathered LOs

LO acquisition is more difficult to assess, as a LO might be the most appropriate in a particular context, while one of its components or a more complex LO (a composite LO that contains it) might fit better in other situations. To validate the generation of LOs, two aspects have been considered. On the one hand, the DR grammar was evaluated to determine its accuracy. On the other hand, the gathered LOs were checked against the manually identified set of DRs to get the percentage of automatically gathered DRs and their correctness.

IV. CONCLUSION

This method is used to develop a domain module based on classification done by admin who categorized the books uploaded in server and place them in a relevant domain. First the uploaded book get preprocessed and then stemming process is involved for comparison. When user login they will be provided with a relevant book that they were searching. The machine learning methods are planned to be used to infer new rules and it improves the identification of pedagogical relationships or the DRs in the electronic textbooks. 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.

V. REFERENCES

[1]. Houqiang Li,Jun Xu, Rui Cai, Tao Mei and Yong Rui, Fellow."Automatic Generation of Social Event Storyboard".2016.

[2]. C. Alexander, B. Fayock, and A. Winebarger. "Automatic event detection and characterization of solar events with iris", sdo/aia and hi-c. In AAS/Solar Physics Division Meeting, 2016.

[3]. P. Bailey,P.N. Bennett,F. Borisyuk,W.Chu, X. Cui, S.T. Dumaisand.R. W.White."Modeling the impact of short-and long-term behavior onsearch personalization". In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval,ACM 2012.

[4]. M. Bandara,K. Jayakodi, D. Meedeniya."An Automatic classifier for exam questions with wordnet and Cosine Similarity".2016.

[5]. Y.J. Chang, M.S. Huang,M.C. Hu and H.Y. Lo."Representative photo selection for restaurants in food blogs". In Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on,IEEE, 2015.

[6]. M. C. Chen and T.C. Chou. "Using incremental PLSI for threshold resilient online event analysis. Knowledge and Data Engineering", IEEE Transactions on, 20(3):289– 299, 2008.

[7]. H. L. Chieu and Y. K. Lee. "Query based event extraction along a timeline". In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2004.

[8]. H. Cui,W.Y. Ma, J.Y. Nie and J.R. Wen. "Probabilistic query expansion using query logs". In Proceedings of the 11th international conference on World Wide Web, pages 325–332. ACM, 2002.

[9]. S. Essid and C. F´evotte. "Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring". Multimedia, IEEE Transactions on, 15(2):415– 425, 2013.

[10]. A. Halevy, J. Madhavan, E. Sadikov and L. Wang. "Clustering query refinements by user intent". In Proceedings of the 19th international conference on World wide web, ACM, 2010.

[11]. Hao Lin,Hui Zhang,Junjie Wu, Yuvan Zuo."Topic Modellingof Short Text: A Pseudo Document view".2016.

[12]. D.D.Lee and H.S.Seung."Algorithms for non-negative matrix factorization". In Advances in neural information processing systems,2001.

[13]. S.Li,T. Mei,S. D.Roy and W.Zeng."Towards cross-domain learning for social video popularity prediction".Multimedia,IEEE Transactions on, 15(6):1255–1267, 2013.

[14]. A.Liu, W.Lin, and M. Narwaria. "Image quality assessment based on gradient similarity". Image Processing, IEEE Transactions on,21(4):1500–1512, 2012.

[15]. Qing yang and Rang Lu."Trend analysis of new Topics on Twitter".2012Multi-Attribute Decision Making Approach to Learning Management Systems Evaluation . *JOURNAL OF COMPUTERS, 2*(10), 28-37.

[16]. BÂRA, A., BOTHA, I., LUNGU, I., & OPREA, S. V. (2013). Decision Support System in National Power Companies A Practical Example (Part I) . *Database Systems Journal, IV*(1), 37-45.