Une architecture d'acquisition et d'exploitation des connaissances pour les EIAH

Amal Zouaq, Roger Nkambou, Claude Frasson

Une architecture d'acquisition et d'exploitation des connaissances pour les EIAH

Amal Zouaq*, Roger Nkambou, Claude Frasson***

** Université de Montréal*
CP 6128, Succ. Centre-Ville, Montréal, QC, H3C3J7
zouaq@iro.umontreal.ca
frasson@iro.umontreal.ca
*** Université du Québec A Montréal (UQAM)*
CP 8888, Succ. Centre-Ville, Montréal, QC, H3C3P8
nkambou.roger@uqam.ca

RÉSUMÉ. *L'acquisition et l'exploitation des connaissances à des fins d'apprentissage doivent résulter d'une approche intégrée. Dans cet article, nous présentons un modèle conceptuel et une architecture technique à base d'ontologies pour l'extraction des connaissances et leur exploitation par un EIAH. Cette extraction est effectuée sur des documents du domaine et des objets d'apprentissage pour construire une ontologie du domaine. Elle s'appuie sur des techniques de traitement du langage naturel et de l'apprentissage machine et sur un mécanisme d'annotation manuel qui permet de faire ressortir les rôles pédagogiques dans les objets d'apprentissage. L'objectif est de constituer des objets de connaissances qui soient réutilisables par un EIAH. L'exploitation de ces objets se fait par un mécanisme d'agrégation automatique d'objets de connaissance et d'apprentissage et leur standardisation en SCORM et IMS-LD.*

MOTS-CLÉS: *acquisition des connaissances, ontologies, traitement de la langue naturelle, mémoire organisationnelle, objets de connaissance et d'apprentissage.*

KEY-WORDS: *Knowledge acquisition, ontologies, natural language processing, organizational memory, learning knowledge objects.*

1. Introduction

La difficulté de l'acquisition des connaissances du domaine est l'un des défis les plus sensibles des EIAH. En effet, cette opération repose généralement sur des experts humains du domaine et sur un processus d'explicitation de leurs connaissances. Outre le fait que cette pratique est ardue, elle nécessite de recommencer l'explicitation pour chaque domaine et est difficile à mettre à jour de manière à refléter les évolutions du domaine. Un processus d'acquisition (semi) automatique doit donc être mis en place. Il s'agit alors d'identifier les sources des connaissances du domaine. Les documents divers qui circulent dans une organisation (entreprise, université, communauté de pratique, etc.) peuvent représenter une telle source mais également les objets d'apprentissage stockés dans un entrepôt (*Learning Object Repository*). En effet, un objet d'apprentissage est une unité de connaissances intégrée, reliée à une discipline ou à un sujet. Il peut donc être considéré comme une source de connaissance fiable sur le sujet traité. Il importe donc de trouver une manière de déterminer la sémantique et les composants pédagogiques d'un objet d'apprentissage. Pour ce faire, il est possible d'utiliser des techniques de forage de données (*data mining*) telles que le traitement du langage naturel et les techniques d'apprentissage machine.

Dans cet article, nous décrivons une architecture technique et les outils qui permettent d'implémenter un modèle d'acquisition et d'exploitation des connaissances pour les EIAH. Nous abordons les fondements ontologiques du modèle en terme de structure, de sémantique, de pédagogie et de compétences. Plus particulièrement, nous insistons sur le mécanisme de composition automatique des objets de connaissances. Nous terminons par une synthèse de nos réalisations et une brève présentation des travaux futurs.

2. Architecture technique proposée

Nous visons un système intégré qui permette de modéliser aussi bien les besoins de formation, que nous exprimons sous forme de compétences et d'habiletés, que les représentations sémantiques et structurelles du document à analyser en terme de concepts et d'association entre ces concepts.

L'architecture est divisée en outils auteurs et en outils d'exploitation reliés aux EIAH. Les outils auteurs permettent d'extraire des connaissances dans les documents en entrée et de les stocker dans une structure appelée mémoire organisationnelle (MO). Ces objets de connaissances peuvent ensuite être retrouvés, via des outils de recherche, par un humain ou un programme afin d'être exploités par un EIAH à des fins d'apprentissage. La Figure 1 schématise notre architecture.

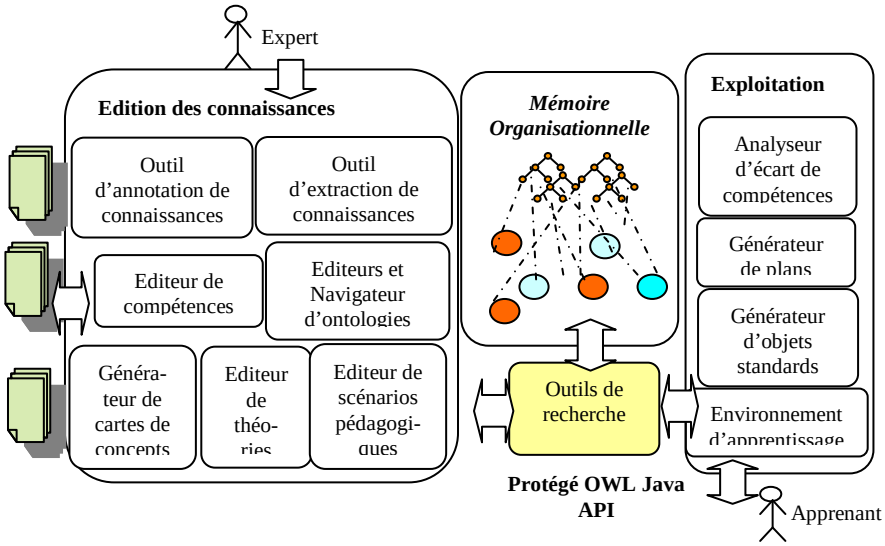

Figure 1. *Architecture d'acquisition et d'exploitation des connaissances*

Pour l'expression des schémas des ontologies, nous avons utilisé l'éditeur Protégé de l'université de Stanford (Protégé, 2006). La première dimension à prendre en compte est l'expression des besoins d'apprentissage. Elle est modélisée par l'ontologie des compétences.

2.0. Modélisation des compétences

Une approche d'apprentissage basée sur la définition des compétences permet d'assembler des objets d'apprentissage de manière ciblée. Dans notre cas, une compétence s'applique à un ensemble d'habiletés définies sur les concepts du domaine, et représente un objectif d'apprentissage. L'expression des habiletés est effectuée par l'utilisation de la taxonomie de Bloom (Bloom, 1956), qui a prouvé son utilité dans le domaine de l'apprentissage (Nkambou et al., 2003). Les compétences sont créées à l'aide d'un éditeur de compétences et stockées dans l'ontologie des compétences.

Une approche basée sur les compétences nécessite un remodelage des ressources d'apprentissage. Ceci est effectué par la modélisation de la structure et de la sémantique des documents.

2.1. Modélisation de la structure des documents textuels

L'extraction de la structure des documents est effectuée grâce à l'outil d'extraction de connaissances « *Knowledge Extractor* » qui analyse le document et retrouve automatiquement les paragraphes et les phrases de chaque paragraphe. Cette extraction se base sur l'architecture et l'environnement UIMA d'IBM (UIMA, 2006), qui permet l'analyse de différents formats de documents pour en extraire des

connaissances. Cette analyse s'appuie sur le développement de composants, les annotateurs, qui sont réutilisables et qui sont dévolus à des tâches spécifiques (dans notre cas, nous avons développé deux annotateurs en java).

2.2. Modélisation de la sémantique des documents textuels

2.2.0. Acquisition de l'ontologie du domaine

Une fois la structure déterminée, nous utilisons un algorithme d'apprentissage machine, Kea-3.0 (Frank et al., 1999), pour déterminer les mots-clés du document. Les phrases contenant ces mots-clés sont collectées et considérées comme des phrases-clés. Elles servent à lancer le processus d'analyse sémantique proprement dit. Nous utilisons l'analyseur statistique de l'université Stanford (Klein et Manning, 2003), pour retrouver les différentes catégories grammaticales des mots dans les phrases-clés. Plus précisément, nous employons un module de dépendances typées «*Typed Dependencies* » qui permet d'indiquer les liens entre les différents mots d'une phrase libellés de catégories grammaticales prédéfinies comme sujet, objet, etc. (De Marneffe et al., 2006). Ce processus nous permet d'obtenir, pour un document, ce que nous appelons une carte grammaticale de concepts.

La prochaine étape consiste à trouver des liens sémantiques à partir des catégories grammaticales. Ceci s'effectue par le biais de patrons lexico syntaxiques et sémantiques. Ces patrons exploitent les relations grammaticales pour identifier les concepts du domaine (ex : sujet, objet direct, indirect, adjectif), les attributs, et des relations sémantiques représentées principalement par les verbes, les conjonctions, et les prépositions. Par exemple, dans la phrase : « A **runtime environment** *must be used to launch* the **individual content objects** *in* a **SCORM conformant package** », les mots en gras indiquent des concepts, et les expressions en italique indiquent des relations. On aboutit ainsi à des cartes de concepts sémantiques qui représentent le contenu du document. Les cartes de connaissances obtenues peuvent bien sûr être modifiées via un éditeur et doivent de toute façon être vérifiées par un expert humain. En effet, malgré les progrès réalisés, le traitement de la langue naturelle est loin d'être exempt d'erreurs.

2.2.1. Acquisition de l'Ontologie des rôles pédagogiques

L'annotation d'un document en terme pédagogique permet de retrouver des rôles pédagogiques reliés à des concepts du domaine. Par exemple, un document peut contenir une définition du concept x. L'ensemble des rôles pédagogiques est listé dans l'ontologie des rôles pédagogiques. Ils ne sont pas figés et peuvent évoluer en fonction du domaine à annoter.

2.3. Outils d'exploitation des connaissances

Les outils d'exploitation de l'architecture servent essentiellement à générer des objets de connaissances et d'apprentissage (LKOs) selon un scénario pédagogique

précis, tiré des théories de l'éducation. En effet, l'un des problèmes des objets d'apprentissage classiques est qu'ils ne disposent pas d'un cadre pédagogique explicite. Celui-ci est contenu implicitement dans la structure mise en place par un intervenant humain (Ullrich, 2004). Or cela peut constituer un obstacle à des programmes automatiques pour la recherche d'objets d'apprentissage pertinents, pour l'agrégation automatique d'objets d'apprentissage ou encore pour l'explicitation du contenu d'un objet d'apprentissage.

La génération de LKO doit donc s'appuyer sur des théories d'apprentissage. Bourdeau et al. (Bourdeau et al., 2004) évoquent d'ailleurs la nécessité de l'incorporation de structures conceptuelles communes pour modéliser les théories d'apprentissage et indiquent que ces structures doivent être encodées de manière déclarative afin de pouvoir désigner le système comme expert pédagogique conscient des théories qu'il peut mettre en œuvre (*theory-aware*).

Dans notre cas, nous considérons une théorie comme un ensemble d'étapes pédagogiques et nous la modélisons sous forme d'ontologie. Chaque étape pédagogique est liée à un ensemble de règles utilisant le formalisme SWRL (*Semantic Web Rule Language*). Ces règles représentent la partie déclarative de la théorie et sont couplées aux annotations pédagogiques ainsi qu'à des méthodes prédéfinies. Par exemple, pour «Fournir un résultat à l'apprenant », il est nécessaire de calculer d'abord le score de ce dernier. C'est pourquoi nous avons mis en place dans l'ontologie quatre méthodes génériques qui comprennent une méthode pour rechercher les pré-requis d'un concept, une méthode pour calculer le score de l'apprenant dans un exercice, une méthode pour retrouver les objectifs d'apprentissage et enfin, une méthode pour retrouver le contenu de la formation proprement dite.

Il est alors possible de générer des LKOs en fonction d'une compétence et selon une théorie d'apprentissage donnée. Au préalable, un outil de mesure de l'écart des compétences permet de mesurer, pour un apprenant donné, les habiletés déjà maîtrisées ou les pré-requis nécessaires. Ensuite, un outil, le générateur de plans pédagogiques, génère le LKO et permet à l'expert humain de vérifier le contenu du LKO et lui indique les objets de connaissances manquants.

3. Conclusion

Dans cet article, nous avons présenté un modèle conceptuel et une implémentation pour l'acquisition de connaissances et leur exploitation dans un EIAH. Pour ce faire, nous avons adapté et implanté des techniques d'extraction de connaissances à partir de textes utilisant le traitement de la langue naturelle et l'apprentissage machine ainsi qu'un mécanisme d'annotation. Cela nous a permis d'aboutir à un modèle ontologique indexant les documents et les objets d'apprentissage selon différentes dimensions: structure, contenu, compétences, et pédagogie. Les structures de données ainsi générées sont stockées dans une mémoire organisationnelle et servent de composants à un mécanisme d'agrégation

automatique d'objets de connaissance et d'apprentissage (LKOs). Ces LKOs se distinguent par un cadre pédagogique explicite provenant de théories de l'éducation. Nous avons effectué des tests de validation qui ont permis de s'assurer que les LKOs étaient exploitables sur des plateformes d'enseignement en ligne en tant qu'objets standards SCORM et IMS-LD. En effet, nous avons implanté un processus de standardisation des LKOs pour les rendre conformes à ces normes. Par ailleurs, les outils que nous utilisons pour le traitement de la langue naturelle (*Stanford Parser*) et la recherche de mots-clés dans les documents (*KEA-3.0*) sont des outils solides ayant démontré de bonnes performances dans leurs domaines respectifs. Nos travaux futurs s'orienteront vers l'élimination de bruits et la validation de l'ontologie du domaine ainsi que vers son exploitation de manière plus approfondie par un EIAH.

4. Bibliographie.

[Bloom 1956] Bloom, B.S. *Taxonomy of educational objectives: The classification of educational goals: Handbook I, cognitive domain*, Longman, New York, 1956.

[Bourdeau et al. 2004] Bourdeau, J., Mizoguchi, R., Psyché, V., and Nkambou, R. «Selecting Theories in an Ontology-Based ITS Authoring Environment », in *Proc. of Intelligent Tutoring Systems*, pp.150-161, Maceio, 2004.

[De Marneffe et al. 2006] De Marneffe, M-C., MacCartney, B. and Manning, C.D. « Generating Typed Dependency Parses from Phrase Structure Parses », in *Proc. of 5th Conference on Language Resources and Evaluation*, Genoa, 2006.

[Frank et al. 1999] Frank, E., Paynter, G.W., Witten, I..H., Gutwin, C., and Nevill-Manning, C.G. « Domain-specific Key Phrase Extraction », in *Proc. of the 16th International Joint Conference on Artificial Intelligenc*e, pp. 668-673, San Francisco, 1999.

[Klein & Manning 2003] Klein, D. and Manning, C.D. « Accurate Unlexicalized Parsing », in *Proc. of the 41st Meeting of the Association for Computational Linguistics*, pp. 423– 430, Sapporo, 2003.

[Nkambou et al. 2003] Nkambou, R., Frasson, C., and Gauthier, G. « CREAM-Tools: An Authoring Environment for Knowledge Engineering in Intelligent Tutoring Systems », in *Authoring Tools for Advanced Technology Learning Environments: Toward cost-effective, adaptative, interactive, and intelligent educational software*, pp. 93-138, Kluwer Publishers, 2003.

[Ullrich 2004] Ullrich, C. «Description of an instructional ontology and its application in web services for education», in *Proc. of Workshop on Applications of Semantic Web Technologies for E-learning*, pp. 17-23, Hiroshima, 2004.

5. Références sur le WEB.

[Protégé 2006]. Protégé Ontology Editor: http://protege.stanford.edu/

[UIMA 2006]. Unstructured Information Management Architecture: http://uima-framework.sourceforge.net/