SEMANTIC INTEGRATION OF RECOVERED DATA FROM LEARNING PLATFORMS

CHOUKRI ALI

SARS team, ENSA Safi, University of Cadi Ayyad, Morocco

Abstract: With the explosion in the number of courses offered online and consequently the number of learners, the majority of schools faced two major problems: heterogeneity of used platforms and personalization of learning according to the learners' profiles.
Hence the need for the collection and integration of data into a formal consolidated system whose main basis is the integration layer which is the subject of this article

Keywords: Keywords: Information Integration System, Semantic Labeling, CRF, LSTM.

1. INTRODUCTION

Généralement, les données des plateformes éducatives sont stockées sous des différents formats et hébergées dans des sources différentes.
L'intégration consiste à regrouper des données résidant dans ces sources et à offrir une vue unifiée de ces données aux utilisateurs [1]. Elle permet de résoudre les difficultés liées à l'intégration structurelle et sémantique d'une part et les problèmes liés à l'hétérogénéité et à l'autonomie de la source d'autre part [2].
De nombreux travaux présentant des approches pour intégrer des sources de données hétérogènes à l'aide de la technologie sémantique sont proposées [3,4,5]. Cependant, aucun de ces travaux ne s'est pas intéressé à intégrer les métadonnées des cours et les données des apprenants issues de plateformes éducatives hétérogènes dans un système unifié en résolvant tous les conflits sémantiques.
L'objectif de notre initiative est de concevoir un système combinant des données contenues dans des plates-formes éducatives hétérogènes (e-learning, MOOC) avec une vue unifiée de ces sources.

2. INTEGRATION DE DONNEES

Un schéma intégré est conçu pour décrire la logique de la couche d'interface d'un système d'intégration de données. Les schémas locaux décrivent la logique des données dans les sources de données locales. Le mappage de schéma fait référence aux transformations entre les objets des sources locales et le schéma intégré. C'est un processus nécessaire pour toutes les approches d'intégration schématisées. Pour spécifier la correspondance entre la source de schéma et le schéma global, il existe de nombreuses alternatives de mappage.

- Global As View (GAV) est l'expression du schéma global en fonction du schéma local [6].
- LAV (Local As View) suppose l'existence d'un schéma global et définit le schéma local des sources de données à intégrer en tant que vues du schéma global [6].
- Les cartographies GLAV surmontent les limitations de GAV et de LAV. Dans la reformulation de requêtes dans l'approche GLAV, chaque règle de mappage est représentée par une requête conjonctive écrite dans le schéma global associé à une requête conjonctive écrite dans les schémas sources.

Dans la section qui suit, nous présentons brièvement une étude comparative, basée sur plusieurs critères et fonctionnalités, d'une gamme d'outils d'intégration de données.

Tableau 1. Comparaison des outils d'intégration d'informations.

Outil	Technique	mappage	Langage	ressources	Etiquette sémantique automatique
AutoMed [7]	Matching	BAV	AIQL	Relationnelle, XML, Fichiers plats	Non
AGORA [8]	Rewriting	LAV	Xquery	Relationnelle, XML	Non
KARMA [9]	Matching	GLAV	SPARQL	Relationnelle, XML,Spreadsheets, CSV, JSON	Oui
PICSEL [10]	View creating	LAV	CARIN	Services	Non
TSIMMIS [11]	View creating	GAV	MSL/LOREL	Semi-structured	Non

Suite aux résultats de cette étude, nous avons choisi d'implémenter le processus d'intégration à l'aide de *KARMA* par le fait qu'il utilise l'approche de cartographie *GLAV*, qui permet de surmonter les limitations de *GAV* et de *LAV* [6]. De plus, il est basé sur une ontologie pour résoudre des conflits sémantiques et a également la capacité d'apprendre et de reconnaître la mise en correspondance de données avec une ontologie basée sur un algorithme d'apprentissage.

3. ARCHITECTURE DE NOTRE SYSTEME D'INTEGRATION

Le profil des apprenants dans notre système est le pont qui relie l'apprentissage formel à l'apprentissage non formel. En effet, en enrichissant les profils des apprenants avec les informations issues de l'interaction des apprenants avec les MOOC, l'établissement pédagogique sera en mesure d'améliorer la qualité de l'apprentissage en adaptant leurs enseignements aux profils des apprenants et en recommandant des MOOC en fonction de leurs besoins.
Pour faciliter l'accès régulier aux sources de données, le système d'intégration proposé repose sur une approche d'intégration sémantique virtuelle. Les tâches du système d'intégration consistent à collecter et récupérer des données des apprenants et des données relatives aux cours de différentes plates-formes, puis modéliser les données des apprenants et des cours dans un format unifié pour faciliter la réponse à la demande de l'utilisateur et résoudre tous les conflits sémantiques.
La figure qui suit représente notre architecture qui est composée de trois couches : collecte, modélisation et mappage.

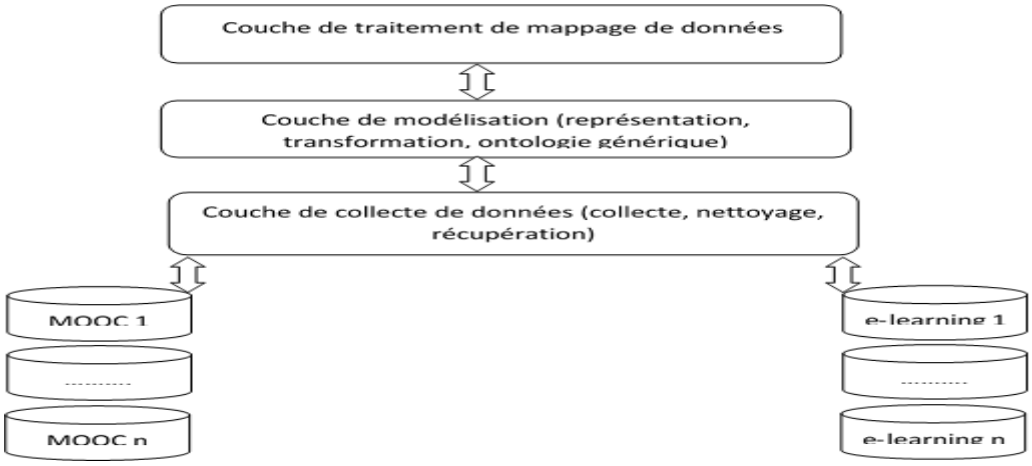

Fig.1. Architecture d'intégration de données éducatives hétérogènes.

3.1. Couche 1 : Collection de données

Trois plates-formes de MOOC sont considérées comme exemples, *OPEN edx, canvas* et le système de gestion de l'apprentissage *Moodle*.

3.2. Couche 2 : Modélisation de données

Le but ultime à cette étape est de convertir des données hétérogènes en un format unifié. Chaque source de données a sa propre structure et son propre vocabulaire. Notamment, les données de cours dans la plate-forme *OPEN edx* sont stockées dans la base de données *MongoDB* qui est une base de données *Nosql*, et les données de cours dans la plate-forme *Moodle* sont stockées dans la base de données *Mysql*. Il est également possible d'avoir un autre format tel que *Json* ou *XML*.
Pour surmonter ces problèmes d'hétérogénéité et de conflits, ce travail est basé sur une solution sémantique où l'ontologie joue un rôle important dans la fourniture de connaissances conceptuelles et vocabulaires sémantiques qui rendent le domaine disponible pour l'échange et la lecture d'informations dans le système.

3.3. Couche 3 : Mappage de données

Le processus de mappage dans l'outil *KARMA* comprend quatre étapes : l'attribution de types sémantiques, la spécification de relations, la génération de descriptions de source et la génération d'un document *RDF* [12].
Les entrées du processus de mappage sont les suivantes : une ontologie générique *OWL*, les sources de données que nous voulons mapper vers une ontologie générique et une base de données de types sémantiques que le système a appris à reconnaître en fonction des utilisations antérieures [12].
La sortie est un triplé *RDF* qui représente le contenu des sources alignées sur une ontologie générique.

3.4. Etiquetage sémantique pour une source de données relationnelles

L'outil *KARMA* utilise un modèle graphique probabiliste pour résoudre le problème de l'étiquetage sémantique. Il attribue des types sémantiques à chaque valeur d'un attribut, puis les combine pour déduire le type sémantique de l'attribut entier.
De plus, nous avons appliqué une méthode de similarité multilingue afin d'affecter à la métrique de similarité de l'algorithme d'apprentissage une meilleure précision.
Dans notre implémentation, nous avons utilisé un modèle de reconnaissance de types sémantiques basé sur une combinaison entre *CRF* et *LSTM* qui tire parti des modèles génératifs et discriminants pour améliorer la précision de la reconnaissance des types sémantiques.

5. RÉSULTATS ET DISCUSSION

Dans cette section, nous cherchons à comparer la précision de la reconnaissance de type sémantique entre le modèle *CRF* et le modèle *LSTM-CRF* [13].
Le modèle a été testé en utilisant quatre bases de données :

- Données des apprenants : *OPEN edx, moodle;*
- Données de cours : *OPEN edx* et bases de données *moodle*.

Deux expériences ont été effectuées :

- le modèle *CRF* a été appliqué au *KARMA* pour étiqueter chaque attribut de source en types sémantiques,
- le modèle *LSTM-CRF* pour étiqueter chaque attribut de source en type sémantique a également été appliqué.

L'objectif de ces tests est de comparer le taux de l'affectation et la reconnaissance de types sémantiques du modèle *LSTM-CRF* d'une part et *CRF* utilisé par les modèles *KARMA* d'autre part.

Tableau 2. Comparaison des résultats d'évaluation du modèle *CRF* de *Karma* et le modèle *LSTM–CRF*

Source	Nom table	reconnaissance correcte des types sémantiques (%)	
		Modèle CRF de Karma	Modèle LSTM-CRF
OPEN edx	Profils des apprenants	61,7	81,3
	Cours	57,8	80.0
moodle	Profils des apprenants	51,9	74,3
	Cours	56,4	80.6

L'outil *KARMA* basé sur le modèle *CRF* a pu déduire les types sémantiques pour les colonnes à une précision de 61,7% et nécessite une affectation manuelle pour le reste.

Le modèle *LSTM-CRF* a pu déduire les types sémantiques pour les colonnes à une précision de 81,3% et nécessite une affectation manuelle pour le reste.

4. CONCLUSIONS

La méthode *LSTM-CRF* est plus performante par rapport à *CRF* en terme de la précision de la reconnaissance des types sémantiques. L'utilisation du modèle *LSTM-CRF* est recommandée pour améliorer la précision des affectations semi-automatiques des types sémantiques et pour le mappage de sources de données vers un nœud dans l'ontologie.

RÉFÉRENCES

[1] Lenzerini M. (2002). Data Integration: A Theoretical Perspective. PODS ’02 Proceedings of the twenty-first ACM SIGMODSIGACT- SIGART symposium on Principles of database systems.
[2] Tatbul N, Karpenko O., Convey C., Yan J. (2001) . Data Integration Services. Brown University, Computer Science.
[3] Cheatham M., Pesquita C. (2017) Semantic Data Integration. In: Zomaya A., Sakr S. (eds) Handbook of Big Data Technologies. Springer, Cham
[4] SEBASTIAN KAGEMANN, SRIVIDYA K. BANSAL . (2015) . “MOOCLINK: BUILDING AND UTILIZING LINKED DATA FROM MASSIVE OPEN ONLINE COURSES”. IEEE 9TH INTERNATIONAL CONFERENCE ON SEMANTIC COMPUTING (ICSC), PP. 373-380, FEBRUARY, ANAHEIM, USA.
[5] N. Piedra, J. Chicaiza, J. López and E. Tovar. (2014). "An Architecture based on Linked Data technologies for the Integration and reuse of OER in MOOCs Context", Open Praxis, vol. 6, no. 2.
[6] Bayardo et al., InfoSleuth. (1997) . Semantic Integration of Information in Open and Dynamic Environments,.Proceedings of the 1997 ACM International Conference on Management of Data (SIGMOD), Tucson, Arizona, May 1997, http://www.mcc.com/projects/ 18 infosleuth.
[7] Boyd M., Kittivoravitkul S., Lazanitis C., McBrien P., Rizopoulos N. (2004). AutoMed: A BAV Data Integration System for Heterogeneous Data Sources. In: Persson A., Stirna J. (eds). Advanced Information Systems Engineering. CAiSE. Lecture Notes in Computer Science, vol 3084. Springer, Berlin, Heidelberg.
[8] I. Manolescu, D. Florescu, D. Kossmann. (2001). Answering XML queries over heterogeneous data sources. Proc. of the 27th Int. Conf. on Very Large Data Bases (VLDB 2001).
[9] Gupta, S., Szekely, P., Knoblock, C. A., Goel, A., Taheriyan, M., and Muslea, M. (2015). Karma: A System for Mapping Structured Sources into the Semantic Web. The Semantic Web: ESWC 2012 Satellite Events: ESWC 2012 Satellite Events, Heraklion, Crete, Greece, May 27-31, 2012.
[10] F. GOASDOUÉ, V. LATTÈS and M. ROUSSET .(2000). "THE USE OF CARIN LANGUAGE AND ALGORITHMS FOR INFORMATION INTEGRATION: THE PICSEL SYSTEM", (2000) International Journal of Cooperative Information Systems, vol. 09, no. 04, pp. 383-401.
[11] H. Garcia-Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, Y.Sagiv, J. Ullman, V. Vassalos, J. Widom. (997). The TSIMMIS approach to mediation: data models and languages. Journal of Intelligent InformationSystems 8 (2) 117–132.
[12] Szekely, P.A., Knoblock, C.A., Gupta, S., Taheriyan, M., & Wu, B. (2011). Exploiting semantics of web services for geospatial data fusion. GIS-SSO.
[13] Huang, Z.; Xu, W.; Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging, arXiv:1508.01991. [Google Scholar]