Etat de l'art sur ontologies et extraction de connaissances

Mouhamadou Saliou Diallo*, Moussa Lo*, Cheikh Talibouya Diop*, Fatou Kamara Sangaré*

*Université Gaston Berger de Saint-Louis(UGB)
UFR Des Sciences Appliquées et Technologie
Laboratoire d'Analyse Numérique et d'Informatique
BP. 234, Saint-Louis

ms.diallo@hotmail.fr

moussa.lo@ugb.edu.sn

cheikh_talibouya_diop@ugb.edu.sn

fatousk@gmail.com

RÉSUMÉ. L'extraction de connaissances à partir de données, communément appelé Data Mining a pour objectif de trouver des connaissances cachées parmi une masse de données importante. Depuis quelques temps, des travaux se sont intéressés à l'intégration des ontologies dans le processus d'extraction de connaissances. L'ingénierie ontologique permet de formaliser les concepts d'un domaine ainsi que les relations qu'ils entretiennent. Elle utilise les ontologies pour fournir une représentation formelle sur laquelle on peut effectuer des inférences. L'objectif de ce travail est de faire un état de l'art sur l'intégration de l'ingénierie ontologique et le processus de d'extraction de connaissance et de présenter ensuite nos perspectives de recherches.

ABSTRACT. The knowledge discovery in databases (KDD) or Data Mining has for objective to find knowledge hidden among an important mass of data. For some time, works were interested in the integration of the ontology in the process of knowledge discovery in databases. Ontology engineering allows formalizing the concepts of a domain as well as the relations which they maintain. It uses ontology to supply a formal representation on which we can make inferences. The objective of this work is a state of the art on the integration of ontology engineering and knowledge discovery process and to present then our future searches.

MOTS-CLÉS: Ontologies, Extraction de Connaissances.
KEYWORDS: Ontologies, Knowledge discovery in Data Base.

1. Introduction

Une ontologie est un ensemble de concepts et de relations, entre ces concepts, permettant de représenter les connaissances d'un domaine. Toutefois, la pertinence des informations contenues dans les ontologies repose sur des mises à jour régulières. Actuellement, ces mises à jour se font de façon manuelle [1], ce qui entraine une subjectivité des connaissances contenues dans les ontologies. Par conséquent, il est pertinent d'avoir des méthodes automatiques pour construire et mettre à jour les ontologies. L'utilisation des techniques de Data Mining est donc envisageable pour automatiser la construction et la mise à jour des ontologies. Le processus d'extraction de connaissance est à la fois itératif et interactif [2][3]et nécessite une bonne connaissances du domaine pour extraire des informations utiles. L'intégration des connaissances expertes dans le processus d'extraction de connaissance permet d'optimiser le processus et de valoriser le résultat [4] [5].Dans ce papier nous effectuons un état de l'art sur les différentes approches qui existent dans l'intégration de l'ingénierie ontologique et le processus d'extraction de connaissance.

La suite de ce papier est organisée de la manière suivante : dans la section 2, nous parlons des ontologies et les problèmes de construction et de mise à jour des ontologies. Dans la section3 nous présentons le processus d'extraction de connaissances, dans la section 4 nous présentons les différentes approches qui existent dans l'intégration de l'ingénierie ontologique et le processus d'extraction de connaissances et enfin dans la section 5, nous effectuons une conclusion et présentons nos perspectives de recherches.

2. Ontologies

Nées des besoins de représentation de connaissances, les ontologies visent à représenter les connaissances en étant à la fois interprétables par l'homme et la machine. Une ontologie est un ensemble de concepts et de relations permettant de représenter les connaissances d'un domaine [6] .Un domaine peut être par exemple: la médecine, la mécanique, l'urbanisation ou bien la manufacture. Par exemple dans le domaine de l'environnement, les êtres humains respirent l'air et consomme de l'eau. L'air et l'eau constituent donc des environnements spécifiques, la **Figure 1** illustre cette ontologie avec des rectangles qui représentent les concepts et les ellipses qui représentent les relations entre les concepts.

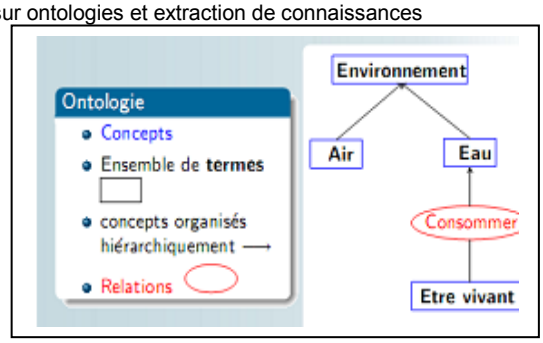

Figure 1.*Une ontologie dans le domaine de l'environnement [1].*

Toutefois, la pertinence des informations contenues dans les ontologies repose sur une mise à jour régulière. Mettre à jour une ontologie consiste à ajouter, dans l'ontologie initiale des concepts et des relations qui constituent les connaissances nouvellement acquises. Cette façon de mettre à jour les ontologies est fastidieuse et subjective [1]. Par conséquent, il est nécessaire de trouver des solutions pour construire et mettre à jour les ontologies de façon automatique.

3. Extraction de Connaissances

Avec l'augmentation de la capacité de stockage, nous avons assisté durant ces dernières années à une croissance importante des moyens de génération et de collection des données. Il s'est ainsi crée un besoin d'acquisition de nouvelles techniques et méthodes qui permettent d'extraire, des données, des informations utiles .C'est ainsi que l'on a commencé à parler de découvertes de connaissances à partir de données (KDD) ou encore Data Mining ou fouille de données [7]. Le Data Mining permet de résoudre plusieurs problématiques, il permet par exemple de déterminer les clients les plus fidèles d'une banque ou bien déterminer les tendances qui se dégagent dans les ventes de supermarché [7].Néanmoins la découverte des connaissances avec le Data Mining pose un certain nombre de problèmes notamment le temps de calcul des motifs et la découvertes de fausses informations. Il est donc nécessaire de trouver des solutions pour réduire le temps de calcul des motifs et pour diminuer la découverte de fausses informations.

Le processus d'extraction de connaissances à partir des données proposé par le CRISP-DM (Cross Industry Standard Process for Data Mining) est composé de six phases (voir Figure 2).Les phases ne sont pas strictement séquentiels, des allers-retours entre les différentes phases sont toujours requis. Les Flèches sur la Figure2 indiquent les dépendances les plus importantes, et les plus fréquentes entre les phases. Le cercle extérieur montre la nature cyclique du processus d'extraction de connaissance. Les phases qui composent le processus d'extraction de connaissance sont les suivantes :

-La phase de compréhension du domaine (Business Understanding en Anglais) est la phase de compréhension du domaine métier, des objectifs du projet et des exigences à respecter dans le domaine. Ces exigences sont ensuite formalisées en un problème de Data Mining.

-La phase de compréhension des données (Data Understanding en Anglais) est l'étape dans laquelle on essaie de déterminer les sous ensembles de données pouvant cacher des informations potentielles.

-La phase de préparation des données (Data Préparation en Anglais) est la phase la plus importante, elle permet de préparer les données qui seront remplis dans les outils de Data Mining.

-La phase de modélisation (Modelling en Anglais) c'est la phase de Data Mining proprement dite. Elle consiste à appliquer les algorithmes de Data Mining sur les données renvoyée par la phase de préparation des données.

-La phase d'évaluation (Evaluation en Anglais).Dans cette phase, les connaissances extraites sont évaluées et validées par les experts du domaine.

-La phase de déploiement (Deployment Understanding en Anglais) est la dernière phase du processus d'extraction de connaissances. Dans cette phase, les connaissances extraites sont présentées à l'utilisateur final(le décideur) qui n'est pas toujours un informaticien.

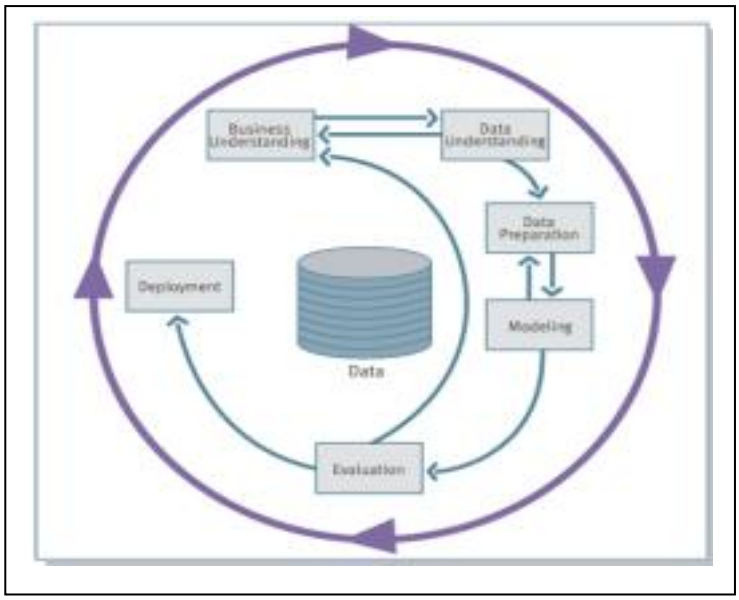

Figure 2. *Processus d'extraction de connaissances [8].*

4. Présentation des approches

Il existe actuellement trois approches dans l'intégration de l'ingénierie ontologique et le processus d'extraction de connaissances à savoir: Onto4KDD, KDD4Onto et Onto4KDD4Onto.

4.1. Onto4KDD

Le processus d'extraction de connaissances (voir Figure 2) nécessite une très grande connaissance du domaine pour déterminer les sous ensembles de données importants[2][9]. L'approche *Onto4KDD* utilise les ontologies pour optimiser le processus d'extraction de connaissance et pour valoriser les résultats. Dans les travaux de Bellandi et al [4], les auteurs proposent un Framework pour l'extraction de règles d'associations avec l'aide d'une ontologie. Ce Framework permet d'interroger les ontologies pour filtrer les instances qui seront utilisées dans l'extraction de règles d'associations, ce qui permet d'avoir des règles d'associations beaucoup plus intéressantes. Dans les travaux de Barb et al [10], les auteurs ont constaté que les règles d'associations extraites avec les techniques de Data Mining étaient incohérentes par rapport aux connaissances des experts dans le domaine geospatial. Cette incohérence est causée par l'absence, dans les règles d'associations extraites, de certains items très intéressants dans le domaine geospatial. En effet ces items étaient absents parce qu'ils étaient sous représentés. Pour résoudre ce problème, les auteurs ont utilisés une ontologie pour sur-échantillonner les sémantiques sous représentés. Ce sur-échantillonnage a permis ensuite d'avoir des règles d'association conformes avec la connaissance du domaine. Ce travail montre que l'intégration des ontologies dans le processus d'extraction de connaissance permet de résoudre des problèmes que les techniques de Data Mining ne peuvent pas résoudre à elles seules.

Nous constatons que ces travaux n'intègrent les ontologies que dans deux des six phases du processus d'extraction de connaissance, notamment dans la phase de préparation des données et dans la phase d'extraction proprement dite.

Tandis que dans les travaux de Brisson et Collard[5], les auteurs proposent une approche pour inclure la connaissance du domaine durant tout le long du processus d'extraction de connaissance mené par les experts. C'est-à-dire dans toutes les phases du processus d'extraction de connaissances proposé par le CRISP-DM (voir Figure 2) sauf dans la phase de déploiement.

4.2. KDD4Onto

Le processus de construction d'ontologie est manuel fastidieux et subjectif [1]. Afin d'accélérer le processus et de lui enlever toute forme de subjectivité, l'approche

KDD4Onto utilise les techniques de Data Mining pour construire et mettre à jour les ontologies de façon automatique.

Dans les travaux de Parekh et al[11], pour mettre à jour une ontologie, les auteurs proposent d'exploiter des glossaires et des dictionnaires pour déterminer les relations taxonomiques entre les termes trouvés ; ensuite les experts du domaine utilisent ces termes et leurs propres connaissances pour construire et mettre à jour les ontologies. Nous constatons que cette mise à jour est semi-automatique car elle ne se fait pas sans l'intervention de l'homme. Dans les travaux de Ben Ghezaiel et al[12],les auteurs proposent un enrichissement des ontologies de façon automatique par la combinaison de deux types de connaissances à savoir des connaissances issues d'une base générique minimal de règles associative et des connaissances issues de la structure ontologique initiale.

Dans les travaux d'Elsayed et al [13], les auteurs proposent une construction, automatique, d'ontologies composées de deux phases : une phase de Data Mining et une phase de construction d'ontologie. La phase de Data Mining inclut la sélection des données, la préparation des données et l'extraction de connaissances à partir de ces données. Ces connaissances extraites sont utilisées, ensuite, dans la phase de construction des ontologies pour construire une ontologie. Ce travail permet ainsi de construire les ontologies sans L'intervention de l'homme.

4.3. Onto4KDD4Onto

Bien que quelques chercheurs travaillent dans l'une ou l'autre approche à savoir *KDD4Onto* ou *Onto4KDD*, rares sont ceux qui travaillent dans l'intégration des deux approches[14].

L'approche *Onto4KDD4Onto* est une approche hybride, qui intègre les approches *Onto4KDD* et *KDD4Onto*. Cette approche profite à la fois à la construction des ontologies et au processus d'extraction de connaissances[14].*Onto4KDD4Onto* permet ,à partir d'un ensemble de données, de construire une ontologie et d'extraire des connaissances sur cette ontologie pour construire une base de connaissance réutilisable[14][15].

Dans les travaux de Gottgtroy[14], l'auteur propose un cycle de vie hybride de découverte de connaissances dirigé par ontologie. Ce travail intègre les meilleures pratiques issues de l'ingénierie ontologique dans le cycle de vie proposé par le CRISP-DM [8]. Ce cycle de vie est composé de cinq phases dont les deux premières phases sont liées à l'ingénierie ontologique. Les deux secondes phases sont liées au processus d'extraction de connaissance et enfin la dernière phase concerne l'évaluation des connaissances extraites.

4.4. Bilan

Il existe actuellement trois approches dans l'intégration de l'ingénierie ontologique et le processus d'extraction de connaissances. Ces approches sont les suivants :

- **Onto4KDD** qui utilise les ontologies pour optimiser le processus d'extraction de Connaissance et pour valoriser les résultats,
- **KDD4Onto** qui utilise les algorithmes de Data Mining pour construire et mettre à Jour les ontologies de façon automatique.
- **Onto4KDD4Onto** qui est une approche hybride qui profite à la fois à la construction des ontologies et au processus d'extraction de connaissances.

L'intégration de l'ingénierie ontologique et le processus d'extraction de connaissances est utile à la fois à la construction des ontologies et au processus d'extraction de connaissances [11] [4] [14. En effet, chaque phase du processus d'extraction de connaissance peut bénéficier de la connaissance du domaine contenu dans les ontologies. D'autres parts les algorithmes de Data Mining peuvent être utilisés pour la construction et à la mise à jour des ontologies.

5. Conclusion et perspectives

Le processus d'extraction de connaissances est habituellement effectué par les experts qui utilisent leurs propres connaissances du domaine pour sélectionner les données les plus pertinentes dans le but de respecter les exigences du domaine[16].Par conséquent le succès du processus d'extraction de connaissances reste très profondément dépendant de l'expert du domaine qui effectue cette tâche. L'utilisation des ontologies dans le processus d'extraction de connaissance aide à simplifier et à structurer le processus d'extraction de connaissances offrant à un expert du domaine un modèle de référence pour les différentes phases du processus d'extraction de connaissance [4][5][9][10]. D'autre part, les algorithmes de Data Mining peuvent être utilisés pour accélérer le processus de construction et de mise à jour des ontologies [1][12][11][13].

Nos travaux futures concernent l'intégration des ontologies dans la dernière phase du processus d'extraction de connaissance proposé par le CRISP-DM[8].En effet, les travaux dans l'approche *Onto4KDD* ,qui utilise les ontologies pour améliorer le processus d'extraction de connaissance, intègrent les ontologies dans toutes les phases du processus d'extraction de connaissances(Voir Figure2) sauf dans la phase de déploiement [4][5][8][9],ce qui profite seulement aux analystes et pas à l'utilisateur final(le décideur). L'intégration des ontologies dans la dernière phase du processus

d'extraction de connaissance peut permettre de bien visualiser les connaissances extraites permettant ainsi au décideur d'avoir une bonne visibilité des connaissances extraites. Car les ontologies ont pour objectif de représenter les connaissances pour qu'elles soient exploitables par les hommes et les machines. Par conséquent nous proposons d'intégrer les ontologies dans la phase de déploiement du processus d'extraction de connaissances car elles peuvent permettre à l'utilisateur, dans cette phase, de bien visualiser les informations extraites.

6. Bibliographie

[1]Di Jorio, L. Abrouk, L. Fiot, C. Hérin, D. Teisseire, M. *«Enrichissement d'ontologie basé sur les motifs séquentiels »*. 23èmes Journées Bases de Données Avancées, BDA 2007, Marseille ,23-26 October 2007.

[2]Fayyad,U. Piatetsky,S.Hapiro,G. Smyth,P. *«The KDD Process for Extracting Useful Knowledge from Volumes of Data »*, Comm. ACM, vol. 39, 1996, p. 27-34.

[3]Chapman P., Clinton J., Kerber R., Khabaza T., Reinartz T., Shearer C., Wirthr. *«CRISP-DM 1.0: Step-by-Step Data Mining Guide »*.SPSS Inc, 2000.

[4]Bellandi A., Furletti B., Grossi V., Romei A. *«Ontology-driven Association Rules Extraction: a Case of Study»*. C&O: RR 2007, Proceedings of the International Workshop on Contexts and Ontologies: Representation and Reasoning.

[5]Brisson, L. Collard,M., *«An ontology driven data mining process»* proceedings of the Tenth International Conference on Enterprise Information Systems (ICEIS), June 12-16, Barcelona, Spain, 2008, pp. 54-61, ISBN 978-989-8111-37-1.

[6] Houacine, T. Azoune, S. *«Construction et exploitation d'une ontologie dans le domaine de lutte antiacridienne.».*Mémoire d'ingénieur à l'Institut National de Formation en Informatique (I.N.I) Oued-Smar, Alger (2008).

[7]Diop, Cheikh Talibouya. «Étude et mise en œuvre des aspects itératifs de l'extraction de règles d'association dans une base de données».Thèse unique, Laboratoire d'informatique (LI), université François Rabelais, Décembre 2003.

[8]CRISP-DM http://www.crisp-dm.org/.

[9] PINTO, F. M. SANTOS, M. F. *«Considering Application Domain Ontologies for Data Mining »* Journal WSEAS Transactions on Information Science and Applications archive Volume 6 Issue 9, September 2009.

[10] Barb, A.S.; Chi-Ren Shyu. *«Ontology Driven Content Mining and Semantic Queries for Satellite Image Databases»*. Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International.

[11]Parekh,V. Gwo, J. Finin,T.*«Mining Domain Specific Texts and Glossaries to Evaluate and Enrich Domain Ontologies».* International Conference of Information and Knowledge Engineering, Juin 2004.

[12]Ben Ghezaiel, L.Latiri, C.Ben Ahmed , M. Gouider-Khouja,N. *« Enrichissement d'ontologie par une base générique minimale de règles associatives Application aux maladies neurologiques : Les dystonies »*.COnférence en Recherche d'Infomations et Applications - CORIA 2010, 7th French Information Retrieval Conference, Sousse, Tunisia, March 18-20, 2010.

[13]Elsayed, A. El-Beltagy,S.R. Rafea, M. Hegazy, O. *« Applying data mining for ontology building»*. The 42nd Annual Conference on Statistics, Computer Science, and Operations Research (2007).

[14]Gottgtroy, P. *«Ontology Driven Knowledge Discovery Process: a proposal to integrate Ontology Engineering and KDD».*11th Pacific-Asia Conference on InformationSystems,PACIS2007,Proceedings,Paper72.
http://aisel.aisnet.org/pacis2007/72.

[15] Gottgtroy, P.,Nik Kasabov,P.,MacDonell, S.*«An ontology driven approach for knowledgediscovery in Biomedicine»*Capturing Intelligence Volume1,2006,Pages415-439 Fuzzy Logic and the Semantic Web.

[16]Tseng,Ming-Cheng. *«Evolutionary Mining of Association Rules with Ontological Information»* .These Universitaire, Institute of Information Engineering I-Shou University Republic of China, juin 2007.