Processus de Décision Crédibiliste pour l'Alignement des Ontologies

Credibilistic Decision Process for Ontology Matching

Amira Essaid[12] Arnaud Martin[2] Grégory Smits[2] Boutheina Ben Yaghlane[3]

[1] LARODEC, Université de Tunis, ISG Tunis, Tunisie

[2] IRISA, Université de Rennes1, Lannion, France

[3] LARODEC, Université de Carthage, IHEC Carthage, Tunisie

[1] 41, Rue de la Liberté, cité Bouchoucha, 2000 Le Bardo, essaid_amira@yahoo.fr

[2] BP 3021, 22302 Lannion Cedex, { Arnaud.martin,gregory.smits }@univ-rennes1.fr

[3] Carthage Présidence 2016, boutheina.yaghlane@ihec.rnu.tn

Résumé :

L'appariement est une tache primordiale pour aboutir à la gestion de l'hétérogénéité sémantique d'ontologies distribuées et ainsi assurer leur interopérabilité entre les différents systèmes les utilisant. Cette mise en correspondance consiste principalement à détecter des relations sémantiques entre les entités de deux ontologies par application de méthodes d'alignement fondées sur des mesures de similarité. L'emploi de mesures de différentes natures fait inévitablement apparaître des conflits. Nous proposons dans cet article de gérer ce conflit à l'aide de la théorie des fonctions de croyance.

Mots-clés :

Appariement des ontologies, mesure de similarité, théorie des fonctions de croyance, conflit.

Abstract:

Ontology matching is one of the most important tasks to mitigate the effect of semantic heterogeneity and to assure the interoperability between the different systems that use them. The matching consists in detecting the semantic relations between entities of two ontologies and thus by the application of different techniques which are based on similarity measures. Using these measures can lead to conflicting aligments. We propose in this paper to manage the conflict using the Dempster-Shafer theory.

Keywords:

Ontology matching, similarity measure, belief function theory, conflict

1 Introduction

Pour faire face à l'émergence d'applications basées sur l'exploitation conjointe de sources de données distribuées dont le format est difficilement interprétable par des machines, la notion du *web sémantique* a été introduite [2]. Cette nouvelle génération du web tend à faciliter l'intégration et l'interopérabilité de sources de données entre différentes applications. Le web sémantique repose essentiellement sur l'utilisation d'*ontologies* décrivant la structure et la sémantique des données contenues dans les documents web. Toutefois, il n'existe pas d'ontologie partagée de référence pour chaque contexte applicatif, mais plusieurs ontologies développées indépendemment les unes des autres et couvrant souvent partiellement le contexte applicatif concerné. L'utilisation de ce genre de données repose donc sur une étape de résolution de l'hétérogénéité sémantique des sources à travers l'appariement des ontologies [5]. Cet appariement vise à trouver des correspondances entre les entités de deux ontologies. Ces entités peuvent être des concepts, des propriétés ou encore des instances. L'ensemble des correspondances appelé *alignement* peut être traité de différentes manières en fonction des besoins des applications. Une exploitation possible d'un tel alignement consiste à fusionner les deux ontologies pour obtenir une nouvelle ontologie. Le résultat de l'appariement peut aussi conduire à la génération de règles de raisonnement pour l'interprétation des ontologies appariées.La découverte manuelle des correspondances sémantiques entre deux ontologies est une tache coûteuse en temps, inefficace et pouvant conduire à des erreurs [7]. Dans [5], les auteurs recensent les différentes méthodes existantes et les principaux défis sous-jacents à cette tache d'appariement : la robustesse et l'évolutivité. La robustesse concerne le fait

que les erreurs mineures ne doivent pas avoir un impact sur le résultat de l'alignement et l'évolutivité quantifie la capacité de ces techniques à s'exécuter en un temps raisonnable, même en présence d'ontologies volumineuses [14]. Ces méthodes reposent essentiellement sur l'exploitation de mesures de similarité. Individuellement, aucune mesure de similarité permet d'obtenir un alignement parfait. Dans ce travail, nous émettons l'hypothèse qu'en exploitant la complémentarité de différentes mesures de similarité, un alignement de meilleure qualité serait obtenu. Cependant, combiner plusieurs méthodes conduit souvent à la nécessité de gérer des conflits d'alignements, conflits que nous proposons de gérer à l'aide de la théorie des fonctions de croyance [3], théorie reconnue comme un outil robuste dans la combinaison de jugements incertains.

Nous tenons à préciser que l'approche exposée dans cet article est une description détaillée de trois étapes essentielles à l'obtention d'un ensemble d'alignements. Ces derniers feront l'objet, dans des travaux futurs, de construction d'une nouvelle ontologie incertaine à partir de deux ontologies appariées. Les trois étapes de notre approche sont :

1. Appariement des ontologies : Au cours de cette étape, nous utilisons trois méthodes d'appariement qui donneront pour chaque entité de l'ontologie source sa correspondante dans l'ontologie cible.
2. Modélisation dans le cadre de la théorie des fonctions de croyance : Les alignements obtenus au cours de l'étape précédente seront modélisés dans le cadre de la théorie des fonctions de croyance et les résultats d'alignement seront combinés par application de la règle conjonctive normalisée.
3. Prise de décision : En s'appuyant sur la formalisation du conflit obtenue lors de l'étape de combinaison, un processus de décision permettra d'identifier les alignements les plus crédibles.

Cet article est une amélioration de l'approche proposée dans [4] où nous nous sommes limités à la description des bases théoriques de notre approche à la différence de cet article où nous proposons une meilleure modélisation et nous exposons les résultats d'appariement obtenus sur des données réelles. Le reste de l'article est organisé comme suit : La section 2 présente les principes et techniques d'appariement. La section 3 est dédiée aux notions de base de la théorie des fonctions de croyance. Nous présentons dans la section 4 notre approche de modélisation des alignements dans le cadre de la théorie des fonctions de croyance pour ainsi conclure et donner les différentes perspectives dans la section 5.

2 Appariement des ontologies

2.1 Notions de base

Gruber définit une ontologie comme étant une spécification explicite d'une conceptualisation [6]. En effet, pour un domaine de discours, une ontologie est un modèle abstrait qui est défini dans un langage interprétable par une machine et qui met en évidence un ensemble de concepts et des relations entre ces concepts. Une ontologie est composée essentiellement :

– de *concepts* ou *classes* qui décrivent une collection d'objets pour un domaine particulier. Ces concepts sont organisés selon une hiérarchie taxinomique,
– d'*individus* qui représentent les instances de classe,
– de *relations* explicitant les liens établies entre les individus,
– d'*attributs* qui décrivent les propriétés des individus d'une classe,
– et d'*axiomes* permettant d'inférer de nouvelles connaissances.

L'appariement des ontologies apparaît comme une étape indispensable pour assurer la réconciliation entre ontologies hétérogènes et leur intéropérabilité sémantique vis-à-vis des différentes applications. Le processus d'appariement prend en entrée deux ontologies O_1 et O_2 [5] et produit un ensemble de corres-

pondances entre les entités des deux ontologies appariées. Une correspondance est un 5-uplet :

$$<\text{id}, e_1, e_2, \text{r}, \text{n}>$$

- *id* : identifiant de la correspondance.
- e_1 et e_2 : entités tel que e_1 appartient à une ontologie source O_1 et e_2 appartient à une ontologie cible O_2. Ces entités peuvent être des concepts, des propriétés ou encore des instances.
- *r* est une relation entre les entités (équivalence, disjonction, subsomption), où seule l'équivalence est étudiée dans notre approche.
- *n* est une mesure de confiance obtenue par application d'une mesure de similarité.

2.2 Les techniques d'appariement

Dans [5], les auteurs dressent un état de l'art complet des différentes méthodes d'appariement et notamment des mesures de similarité utilisables pour établir les correspondances entre les entités. Ces méthodes peuvent être qualifiées :

- de *terminologiques*, en comparant chaînes de caractères formées à partir des noms, des labels ou encore des commentaires des entités.
- de *structurelles*, où l'information structurelle des entités, i.e. les relations qui existent entre les entités (subsomption, rang, domaine, ...), est exploitée pour créer les correspondances. Ces méthodes reposent soit sur la structure interne des entités (cardinalité, transitivité, multiplicité, ...), soit sur leur structure externe, c'est-à-dire la position des entités dans la hiérarchie de l'ontologie.
- d'*extensionnelles*, lorsqu'elles exploitent les instances des concepts pour établir des similarités.
- de *sémantique*, lorsqu'elles s'appuient sur la sémantique de la théorie des modèles afin de justifier les résultats de l'alignement. Ces méthodes déductives sont souvent précédées d'une étape de prétraitement et exploitent les instances associées à une entité pour définir son contexte et l'interpréter.

3 La théorie des fonctions de croyance

3.1 Formalisme

La théorie des fonctions de croyance, également appelée théorie de Dempster-Shafer [3] [8], est fondée sur la manipulation de fonctions de masse. Une fonction de masse représente une évaluation quantitative des connaissances sur un problème donné. En effet, pour un problème donné, on définit un cadre de discernement Θ comme étant un ensemble fini de toutes les hypothèses possibles exhaustives et exclusives. Une fonction de masse est décrite sur l'ensemble de tous les sous-ensembles de Θ, noté 2^Θ. Cette fonction de masse ou masse élémentaire de croyance est décrite par :

$$m(\emptyset) = 0 \tag{1}$$

$$\sum_{A \subseteq \Theta} m(A) = 1 \tag{2}$$

Les éléments A tel que m(A)$>$ 0 sont appelés les éléments focaux.

3.2 Combinaison

En présence d'informations imparfaites (incertaines, imprécises et/ou incomplètes), la fusion se présente comme une solution pour obtenir une information plus pertinente et plus fiable. La théorie des fonctions de croyance est un outil intéressant et robuste de fusion de données. En effet, elle repose sur la possibilité de construire, pour un même cadre de discernement, une fonction de masse unique par combinaison des différentes fonctions de masses élémentaires issues de plusieurs sources d'informations distinctes et indépendantes et ceci en vue d'une prise de décision. Il existe un grand nombre de règles de combinaison [13], nous nous limitons dans cet article à la présentation de la règle conjonctive normalisée proposée par Dempster [3]. Pour deux fonctions de masse m_1 et m_2 et

pour tout $A \in 2^\Theta$, $A \neq \emptyset$, cette règle conjonctive normalisée est définie par :

$$m_{1\oplus 2}(A) = \frac{1}{1-k} \sum_{B\cap C=A} m_1(B) \times m_2(C) \tag{3}$$

où $k = \sum_{B\cap C=\emptyset} m_1(B) \times m_2(C)$ est souvent considérée comme une mesure de conflit entre les sources. Cependant cette normalisation par 1- k masque le conflit et a été introduite pour rester en monde fermé (toutes les hypothèses possibles du problème appartiennent au cadre de discernement Θ et $m(\emptyset) = 0$).

3.3 La prise de décision

Au regard de l'information obtenue suite à la combinaison des informations issues des différentes sources, on souhaitera le plus souvent désigner l'hypothèse la plus vraisemblable. D'une façon générale, comme présentée dans [12], les fonctions de décision (plausibilité, crédibilité, probabilité pignistique) prennent la décision sur les singletons du cadre de discernement.

– Maximum de crédibilité : la crédibilité (*bel*) représente le degré de croyance minimal apporté à un sous ensemble de 2^Θ. Elle mesure à quel point les informations données par une source soutiennent A. Cette fonction est définie pour tout $A \in 2^\Theta$ et à valeurs dans $[0, 1]$ par :

$$bel(A) = \sum_{B\subseteq A, B\neq\emptyset} m(B) \tag{4}$$

Le maximum de crédibilité consiste à retenir l'hypothèse la plus crédible. En d'autres termes, cette fonction de décision permet de retenir la meilleure hypothèse tout en donnant le minimum de chances à chacune des disjonctions. C'est un critère de décision pessimiste.

– Maximum de plausibilité (*pl*) : La plausibilité représente la croyance maximale affectée à un sous-ensemble de 2^Θ. Elle mesure à quel point les informations données par une source ne se contredisent pas. Cette fonction est définie pour tout $A \in 2^\Theta$ et à valeurs dans $[0, 1]$ par :

$$pl(A) = \sum_{A\cap B\neq\emptyset} m(B) \tag{5}$$

Le maximum de plausibilité consiste à retenir l'hypothèse la plus plausible. En effet, cette fonction de décision permet de retenir la meilleure hypothèse tout en donnant le maximum de chances à chacun des singletons. C'est un critère de décision optimiste.

– Probabilité pignistique : Dans [9], Smets propose un compromis entre les deux règles de décision précédemment citées. En effet, la probabilité pignistique est une mesure qui permet d'équirépartir la masse placée sur chaque hypothèse différente d'une hypothèse singleton, sur les hypothèses qui la composent.

$$betP(X) = \sum_{A\in 2^\Theta, X\in A} \frac{m(A)}{|A|(1-m(\emptyset))} \tag{6}$$

où $|A|$ est la cardinalité de A. Prendre la décision en appliquant le maximum de la probabilité pignistique revient à choisir l'hypothèse singleton la plus probable.

Cependant, il est possible de prendre une décision sur les unions des singletons tel est le cas de la règle de décision proposée par Appriou [1]. En effet, elle permet de pondérer les règes de décision précédemment citées par une fonction qui rassemble toutes les préférences a priori relatives à la décision escomptée. Elle est décrite pour $\forall A \in 2^\Theta$ par :

$$A = \underset{X\in 2^\Theta}{argmax} \left(m_d(X) pl(X)\right) \tag{7}$$

où m_d est une masse définie par :

$$m_d(X) = K_d \lambda_X \left(\frac{1}{|X|^r}\right) \tag{8}$$

r est un paramètre appartenant à $[0, 1]$ permettant de choisir une décision allant du choix d'un singleton ($r = 1$) à l'indécision totale ($r = 0$). La valeur λ_X permet d'intégrer le manque de connaissance sur l'un des éléments X de 2^Θ. La constante K_d est un facteur de normalisation.

4 Processus de Décision Crédibiliste

Le processus de décision crédibiliste proposé dans cet article est illustré sur deux ontologies O_1 et O_2, relatives à l'organisation de conférences [1].

4.1 Etape 1 : Appariement des ontologies

La mise en correspondance de deux ontologies est effectuée à l'aide de techniques terminologiques d'appariement à savoir les distances de Levenshtein, Jaro et Hamming. Ce sont des méthodes qui, comme décrites précédemment, comparent les chaînes de caractères sans pour autant tenir compte des relations existantes entre les entités.

- **Distance de Levenshtein** : Elle est égale au nombre minimal d'opérations de suppression, d'insertion ou de substitution de caractères nécessaires pour la transformation d'une chaîne en une autre.
- **Distance de Jaro** : Elle mesure le nombre et l'ordre des caractères communs entre deux chaînes de caractères.
- **Distance de Hamming** : Elle mesure le nombre de positions au niveau desquelles les deux chaînes de caractères diffèrent.

La figure 1 montre un extrait des correspondances établies entre une ontologie source O_1 et une ontologie cible O_2 selon une mesure de similarité. Rappelons que dans un processus d'appariement, on désigne aléatoirement une ontologie de référence pour laquelle on essaie de chercher pour chacune de ses entités sa correspondante dans une ontologie cible. On observe que l'entité $ConferenceMember$ de l'ontologie O_1 est appariée aux entités $Conference$ et $Conference\_fees$ de l'ontologie O_2.

Par application des méthodes terminologiques d'appariement présentées précédemment, on

1. http ://oaei.ontologymatching.org/2013/conference/index.html

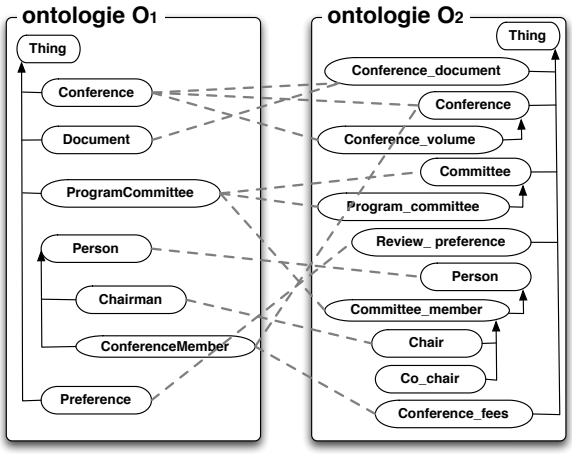

Figure 1 – Appariement entre deux ontologies pour une méthode donnée

aboutit à un ensemble de correspondances possibles. Pour limiter le nombre d'alignements considérés, un seuil de filtrage sur le degré de similarité a été défini empiriquement. Pour illustrer l'absence fréquent de consensus entre les méthodes d'appariement utilisées pour une même entité de O_1, le tableau 1 montre les appariements privilégiés pour chacune des trois méthodes considérées en prenant $ConferenceMember$ comme entité de départ de l'ontologie source O_1.

Tableau 1 – Résultats d'appariement de l'entité $e_1 = ConferenceMember$ de O_1 avec des entités de O_2

méthode	$e_2 \in O_2$	n
Levenshtein	$Conference\_fees$	0.687
Jaro	$Conference$	0.516
Hamming	$Conference$	0.625

4.2 Etape 2 : Modélisation dans le cadre de la théorie des fonctions de croyance

Les résultats d'appariement de l'étape précédente seront modélisés dans le cadre de la théorie des fonctions de croyance.

– *Cadre de discernement* : il sera composé de

toutes les entités de l'ontologie cible O_2 appariées à une entité de O_1 par au moins une des méthodes d'appariement considérées. Si l'on reprend l'illustration d'appariement de la figure 1, le cadre de discernement Θ est composé de toutes les entités de O_2 à l'exception de $Thing$ et Co_chair .

– *Sources d'information* : Chaque correspondance établie par une méthode d'appariement sera considérée comme une information dont la source est l'application d'une méthode d'appariement sur l'entité de O_1 concernée par la correspondance.
– *Les fonctions de masse* : Parmi l'ensemble des correspondances établies, on ne conserve que celle où l'entité source $e_1 \in O_1$ a un appariement de proposée pour toutes les méthodes d'appariement considérées. Une fois les correspondances retenues, on construit pour chaque source sa propre fonction de masse. La mesure de similarité obtenue pour une correspondance suite à l'application d'une méthode d'appariement est interprétée comme une masse. Vu que la somme des masses de croyance pour une source donnée doit être égale à 1, une masse sera affectée à l'ignorance totale.
  Reprenons les résultats du tableau 1. Ce tableau montre les informations de trois différentes sources que nous désignons par $S^{e_1}_{lev}$, $S^{e_1}_{jaro}$ et $S^{e_1}_{hamming}$, où $e_1 = ConferenceMember$. Selon $S^{e_1}_{lev}$, la masse associée à la correspondance avec $Conference\_fees$ de O_2, notée $m_{S^{e_1}_{lev}}(Conference\_fees) = 0.687$ est donc $m_{S^{e_1}_{lev}}(\Theta) = 1 - 0.687 = 0.313$. Pour cet exemple restreint aux correspondances privilégiées partant de e_1 pour les trois méthodes, les trois fonctions de masse sont les suivantes :
  – $m_{S^{e_1}_{lev}}(Conference\_fees) = 0.687$ et $m_{S^{e_1}_{lev}}(\Theta) = 0.313$,
  – $m_{S^{e_1}_{jaro}}(Conference) = 0.516$ et $m_{S^{e_1}_{jaro}}(\Theta) = 0.484$,
  – $m_{S^{e_1}_{hamming}}(Conference) = 0.625$ et $m_{S^{e_1}_{hamming}}(\Theta) = 0.375$.
– *Combinaison* : Au cours de cette étape, les fonctions de masse des trois sources seront combinées par application de la règle conjonctive normalisée. On obtient alors une masse pour chaque entité $e_2 \in O_2$ candidate à un appariement avec une entité $e_1 \in O_1$ donnée. Pour notre exemple où $e_1 = ConferenceMember$, on obtient les masses combinées suivantes :
  – $m^{e_1}_{comb}(Conference\_fees) = 0.2849$,
  – $m^{e_1}_{comb}(Conference) = 0.5853$,
  – $m^{e_1}_{comb}(\Theta) = 0.1298$.

4.3 Prise de décision

Une fois que nous avons tenu compte de toutes les informations provenant des différentes sources, l'étape de décision s'impose. Cette étape nous permettera de décider sur l'entité de l'ontologie cible à apparier avec chacune des entités de l'ontologie source. Au cours de cette étape de décision on va pouvoir décider par exemple si on doit apparier l'entité ConferenceMember à l'entité Conference_fees ou à l'entité Conference ou bien encore établir une indétermination. Dans ce cas, les correspondances concurrentes doivent être maintenues. Rappelons qu'à l'issue de l'étape précédente, nous avons obtenu les résultats suivants par application de la règle conjonctive normalisée :

– $m^{e_1}_{comb}(Conference\_fees) = 0.2849$,
– $m^{e_1}_{comb}(Conference) = 0.5853$,
– $m^{e_1}_{comb}(\Theta) = 0.1298$.

La règle de décision, à savoir le maximum de la probabilité pignistique, considère que le meilleur correspondant pour *ConferenceMember* de l'ontologie source est l'entité *Conference* de l'ontologie cible. Par application de la règle de décision proposée par Appriou, on obtient un résultat imprécis, c'est-à-dire une disjonction entre $Conference\_fees$ et *Conference* ($Conference\_fees \cup Conference$), ce qui se traduit par le fait qu'on pourra apparier *ConferenceMember* de l'ontologie source avec soit $Conference\_fees$ ou *Conference*.

5 Conclusion et perspectives

Dans cet article, nous avons proposé un processus de décision dans le cadre de l'appariement des ontologies. Cette approche se déroule principalement en trois étapes où nous étions amenés tout d'abord à chercher les alignements entre les entités de deux ontologies par application des méthodes terminologiques. Ensuite, nous avons procédé à la modélisation des résultats dans le cadre de la théorie des fonctions de croyance. La dernière étape consiste à prendre une décision quant au correspondant de chaque entité de l'ontologie source et ceci par application des règles de décision.

Comme perspectives à ce travail, nous envisageons tout d'abord d'utiliser d'autres méthodes d'appariement tenant compte des différents aspects sémantiques et structurels des entités à apparier et aussi de se fonder sur les résultats de décision pour construire une ontologie incertaine. Le passage à l'échelle du cadre de discernement important pour les ontologies nécessitera de développer des approches efficaces de décision sur l'espace puissance.

Références

[1] A. Appriou. Approche générique de la gestion de l'incertain dans les processus de fusion multisenseur. *Traitement du Signal*, 22, 307- 319, 2005.

[2] T. Berners-Lee, J. Hendler, O. Lassila. The semantic web. *Scientific American*, 284 : 34 - 43, 2001.

[3] A. Dempster. Upper and Lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38 : 325 - 339, 1967.

[4] A. Essaid, B. Ben Yaghlane, A. Martin. Gestion du conflit dans l'appariement des ontologies. *Atelier Graphes et Appariement d'Objets Complexes, Extraction et Gestion des Connaissances (EGC)*, 50 - 60, 2011.

[5] J. Euzenat, P. Shvaiko. *Ontology Matching*. Springer-Verlag, 2007.

[6] T.R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition Journal*, 5(2) : 199 - 220, 1993.

[7] N. F. Noy, M. A. Musen. Prompt : algorithm and tool for automated ontology merging and alignment. *In Proceeding of Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, 2000.

[8] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[9] P. Smets : Constructing the pignistic probability function in a context of uncertainty. *Uncertainty in Artificial Intelligence*, 5, 29 -39, 1990.

[10] P. Smets. Belief functions : The disjunctive rule of combination and the generalized bayesian theorem. *International Journal of Approximate Reasoning*, 9(1) : 1-35, 1993.

[11] P. Smets, R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66 : 191- 234, 1994.

[12] P. Smets. Decision making in the TBM : the necessity of the pignistic transformation. *International Journal of Approximate Reasonning* 38, 133 - 147, 2005.

[13] P. Smets. Analyzing the combination of conflicting belief functions. *Information Fusion*, 8(4) : 387 - 412, 2007.

[14] X. Su. Semantic enrichment for ontology mapping.*Norwegian University of Science and Technologie*, 2004.