Paramétrage intelligent de l'alignement d'ontologies par l'intégrale de Choquet

Marouen Kachroudi*, Sami Zghal*,**, Sadok Ben Yahia*

* Université de Tunis El Manar - Faculté des Sciences de Tunis
Département des Sciences de l'Informatique - LIPAH
Campus Universitaire, 1060 Tunis, Tunisie
{marouen.kachroudi, sadok.benyahia}@fst.rnu.tn
** Université de Jendouba - Faculté des Sciences Juridiques
Economiques et de Gestion de Jendouba
Département Informatique
Avenue de l'UMA, 8189 Jendouba, Tunisie
sami.zghal@fsjegj.rnu.tn

Résumé. Le nombre croissant d'ontologies rend le processus d'alignement une composante essentielle du Web sémantique. Plusieurs outils ont été conçus dans le but de produire des alignements. La qualité des alignements fournis par ces outils est étroitement liée à certains paramètres qui régissent leurs traitements. Dans ce papier, nous proposons une nouvelle approche permettant l'adaptation automatique des paramètres d'alignement d'ontologies par l'utilisation de l'intégrale de Choquet, comme un opérateur d'agrégation. Les expérimentations montrent une nette amélioration des résultats par rapport à un paramétrage statique et figé.

1 Introduction

L'alignement d'ontologies se positionne comme une pierre angulaire du Web sémantique. Il facilite la réconciliation des ressources décrites par des ontologies différentes. Ce processus permet la production des correspondances entre les entités de deux ontologies. Dans ce contexte, une multitude de méthodes d'alignement ont émergé ces dernières années (Euzenat et al., 2011). Ces méthodes réussissent à produire une bonne qualité d'alignement en se basant sur une configuration adéquate des paramètres. Ces paramètres sont fixés en amont au processus d'alignement. Cependant, un tel paramétrage peut s'avérer parfois non adéquat puisqu'il ne prend pas en considération la nature intrinsèque des ontologies. A titre d'exemple, le système FALCON-AO (Hu et Qu, 2008) est un outil d'alignement comportant 21 paramètres différents, qui peuvent être fournis. Ces paramètres posent un problème quant à la détermination de leur combinaison optimale avant d'entamer la phase d'alignement.

En outre, il est à noter qu'un paramétrage particulier ne peut pas être universellement optimal. Dans ce qui suit, nous proposons une nouvelle approche pour le paramétrage automatique d'une méthode d'alignement. Cette approche repose sur l'exploitation de l'intégrale de Choquet dans le but de déterminer une configuration optimale des paramètres au cours du processus

d'alignement en fonction des ontologies à aligner. L'opération de paramétrage est caractérisée par un aspect totalement automatique et ne nécessitant pas l'intervention de l'utilisateur. Ainsi, les résultats encourageants, fournis après la phase d'évaluation, montrent que le processus de paramétrage automatique s'avère très intéressant.

2 L'intégrale de Choquet

L'intégrale de Choquet est considérée comme un opérateur d'agrégation (Kaci, 2011). Il permet l'amélioration de la puissance de l'analyse multicritères par la prise en compte de l'interaction entre les critères (Grabisch, 1996). En effet, cette notion permet de modéliser les phénomènes d'interaction entre les critères et la dépendance préférentielle. Elle utilise des mesures floues pour prendre en compte l'importance relative de chaque critère ainsi que les interactions mutuelles entre eux.

Définition 1 *La fonction de capacité (mesure floue) sur un ensemble $\mathcal{M}$ est une fonction μ : $2^{\mathcal{M}} \rightarrow [0,1]$ satisfaisant les conditions suivantes : la Normalisation ($\mu(\emptyset) = 0$ et $\mu(\mathcal{M}) = 1$) et la Monotonie ($\mu(A) \leq \mu(B)$ si $A \subseteq B \subseteq \mathcal{M}$).*

Une valeur de capacité, d'un ensemble de critères, est considérée comme son poids (importance). Les fonctions de capacité peuvent être considérées comme une extension des vecteurs de probabilités. Les fonctions de capacité permettent de modéliser l'interaction entre les critères au sein des systèmes nécessitant un grand nombre de paramètres.

Définition 2 *Étant donnée une mesure floue μ, l'intégrale de Choquet C_μ sur un vecteur de critères $a = (a_1, a_2, ..., a_{n_c})$ est définie par : $C_\mu = \sum_{i=1}^{n_c} (a_{\sigma(i)} - a_{\sigma(i-1)}).\mu(\sigma(i), ..., \sigma(n_c))$ tels que : $a_{\sigma(0)} = 0 \leq a_{\sigma(1)} \leq ... \leq a_{\sigma(n_c)}$ représentent les indices permutés des critères.*

La plupart des méthodes d'agrégation multicritères se basent sur la somme pondérée, qui met en valeur l'importance de chaque critère indépendamment. L'intégrale de Choquet se base sur la notion d'interaction pour la résolution des problèmes multicritères. Elle permet de tenir compte de l'importance de chaque critère mais aussi de l'importance relative entre ces derniers. Pour cela, il faut distinguer la notion de l'importance globale de chaque critère et l'importance relative due à son interaction avec les autres. Outre les propriétés usuelles des opérateurs d'agrégation et la modélisation de l'importance relative des critères, la famille de l'intégrale de Choquet a la distinction de permettre la représentation de phénomènes d'interaction mutuelle qui peuvent exister. Cependant, pour interpréter le comportement de l'intégrale de Choquet, nous sommes amenés à calculer deux indices, à savoir : l'indice de Shapley et l'indice d'interaction entre les critères.

L'importance globale d'un critère i n'est pas déterminée uniquement par la mesure floue $\mu(i)$, mais elle prend en compte toutes les mesures $\mu(D)$ pour un sous-ensemble $D \subset N_c$ de toutes les coalitions d pour $i \in D$. En effet, nous pouvons avoir, $\mu(i)$ quasiment nulle suggérant que le critère i est sans importance. Cependant, nous pouvons avoir en joignant i à une coalition $D \subset N_c$, une valeur de $\mu(D \cup \{i\})$ qui soit plus grande que $\mu(D)$ suggérant ainsi l'importance du critère i dans la décision. Le calcul de l'importance globale se base ainsi sur la notion d'indice de Shapley, issue de la théorie des jeux coopératifs (Grabisch, 1996).

Pour tout critère i l'indice de Shapley est défini par : $I_i = \sum_{D \subset N_c \setminus \{i\}} \frac{(n_c - |D| - 2)!|D|!}{(n_c - 1)!}(\mu(D \cup \{i\}) - \mu(D))$, où $|D|$ représente le cardinale de D. L'indice de Shapley est représenté par le vecteur $< I_1, I_2, ..., I_{n_c} >$. Cet indice calcule la contribution moyenne du critère i dans toutes les coalitions. Une propriété fondamentale de l'indice de Shapley est que $\sum_{i=1}^{n_c} I_i = 1$. L'indice d'interaction entre les critères i et j est la moyenne de la quantité de synergie entre i et j en présence d'un groupe de critères D : $I_{ij} = \sum_{D \subset N_c \setminus \{i\}} \frac{(n_c - |D| - 2)!|D|!}{(n_c - 1)!}(\mu(D \cup \{i, j\}) - \mu(D \cup \{i\}) - \mu(D \cup \{j\}) + \mu(D))$. Trois situations peuvent être considérées selon l'importance des critères i et j pris ensemble : ***(a)*** si $\mu(\{i, j\}) > \mu(\{i\}) + \mu(\{j\})$ alors il y a une synergie de complémentarité entre ces deux critères ; ***(b)*** si $\mu(\{i, j\}) < \mu(\{i\}) + \mu(\{j\})$ alors il y a une redondance ou une synergie négative entre ces deux critères ; ***(c)*** si $\mu(\{i, j\}) = \mu(\{i\}) + \mu(\{j\})$ alors les critères sont indépendants.

3 Agrégation de similarité par l'intégrale de Choquet

La forme la plus simple de l'agrégation est la moyenne arithmétique ou pondérée. Ce type d'opérateur n'est pas adapté pour l'agrégation des mesures de similarités vu qu'il exige que les mesures donnent des valeurs de façon indépendante. De toute évidence, cette condition n'est pas satisfaite. Ces mesures présentent une très grande interaction entre elles. Ainsi, l'utilisation de la moyenne pondérée peut conduire à un résultat biaisé car un groupe de mesures très similaires peut facilement submerger d'autres. Par conséquent, l'utilisation de l'intégrale de Choquet avec une fonction de capacité appropriée permet d'éviter ce problème. L'étape primordiale dans l'utilisation de l'intégrale de Choquet est de modéliser l'interaction entre les mesures via une fonction de capacité adéquate. Pour réaliser cette tâche, nous avons opté à l'utilisation de l'approche proposée par (Marichal et Roubens, 2000). Les auteurs ramènent le problème à un programme linéaire qui tient compte des contraintes préférentielles du décideur. Pour maximiser la valeur de la correspondance, $V_{corr} = \pi_1.SimTerm + \pi_2.SimTopo + \pi_3.SimSemant$ (composante terminologique $SimTerm$, composante topologique $SimTopo$ et composante sémantique $SimSemant$), suivant les propriétés de chaque ontologie, il faut tenir compte d'un certain nombre de contraintes :
- ***(i)*** si les valeurs de SIMTERM et SIMTOPO sont plus importantes que celle de SimSemant, alors il est préférable de les favoriser, sinon favoriser SIMSEMANT ;
- ***(ii)*** si la valeur de SIMTOPO est la plus petite, alors favoriser SIMTERM et SIMSEMANT ;
- ***(iii)*** si la valeur de SIMTERM est la plus petite, alors favoriser SIMTOPO et SIMSEMANT.

Chacune de ces contraintes modélise l'interaction positive ou négative qui peut exister entre les critères SIMTERM, SIMTOPO et SIMSEMANT. Les préférences sus décrites peuvent être représentées en utilisant un système linéaire. En outre, l'utilisateur est amené à assigner un poids à chaque critère qui reflète son importance relative dans le cadre de chaque contrainte.

$$\begin{cases} (1)\ 0.50\mu(SimTerm, SimTopo) > 0.35\mu(SimTerm) + 0.35\mu(SimTopo) \\ (2)\ 0.50\mu(SimTerm, SimSemant) > 0.35\mu(SimTerm) + 0.35\mu(SimSemant) \\ (3)\ 0.50\mu(SimTopo, SimSemant) > 0.35\mu(SimTopo) + 0.35\mu(SimSemant) \\ (4)\ 0.50\mu(SimSemant) > 0.35\mu(SimTerm, SimTopo) \\ (5)\ 0.50\mu(SimTerm) > 0.35\mu(SimSemant, SimTopo) \\ (6)\ 0.50\mu(SimTopo) > 0.35\mu(SimTerm, SimSemant). \end{cases}$$

En effet, la solution optimale à notre problème d'agrégation revient à résoudre à ce système linéaire sur notre ensemble $N_c = \{SimTerm, SimTopo, SimSemant\}$:

$$\begin{cases} (1)\ 0.50\mu(SimTerm, SimTopo) - 0.35\mu(SimTerm) - 0.35\mu(SimTopo) > 0 \\ (2)\ 0.50\mu(SimTerm, SimSemant) - 0.35\mu(SimTerm) - 0.35\mu(SimSemant) > 0 \\ (3)\ 0.50\mu(SimTopo, SimSemant) - 0.35\mu(SimTopo) - 0.35\mu(SimSemant) > 0 \\ (4)\ 0.50\mu(SimSemant) - 0.35\mu(SimTerm, SimTopo) > 0 \\ (5)\ 0.50\mu(SimTerm) - 0.35\mu(SimSemant, SimTopo) > 0 \\ (6)\ 0.50\mu(SimTopo) - 0.35\mu(SimTerm, SimSemant) > 0 \\ (7)\ \mu(SimTerm) + \mu(SimTopo) + \mu(SimSemant) = 1. \end{cases}$$

La première contrainte donne une importance relative aux composantes terminologique et topologique ensemble. Pareillement, dans le cas d'une faible contribution de la composante topologique, la deuxième contrainte favorise la coalition entre la composante terminologique et celle sémantique. La troisième contrainte favorise la jointure entre SIMTOPO et SIMSEMANT au détriment de SIMTERM ou toute autre jointure. Les trois dernières inégalités favorisent une seule composante parmi les trois, à condition que l'interaction entre les deux composantes restantes soit négative. La résolution du système résultant de l'ensemble de ces contraintes a été menée en utilisant *Kappalab R package* [1](Grabisch et al., 2008).

4 Étude expérimentale

L'agrégation par l'intégrale de Choquet a été appliquée sur trois mesures de similarité résultant des trois modules d'alignement. La base Benchmark [2] de la campagne OAEI 2012 a été utilisée pour évaluer notre méthode de paramétrage. La base Benchmark comporte 111 tests. Chaque test permet d'évaluer la puissance de la méthode d'alignement sur un aspect particulier. Cette base comporte 4 sous cas, qui diffèrent selon la taille, tout en gardant les mêmes caractéristiques. Nous avons opté à l'utilisation de la sous base FINANCE. Les tests sont systématiquement dérivés d'une ontologie de référence et introduisent à un certain nombre de fluctuations. Ces modifications permettent d'évaluer le comportement de la méthode d'alignement face aux changements subis par les ontologies objet du processus d'alignement. En effet, ces ontologies montrent des altérations qui peuvent être catégorisées selon 6 niveaux, à savoir :
- ***(i)*** les noms d'entités peuvent être supprimés, remplacés par des synonymes ou traduits ;
- ***(ii)*** les commentaires peuvent être supprimés ou traduits ;
- ***(iii)*** les liens hiérarchiques peuvent être supprimés, étendus (*i.e.*, par rapport à l'ontologie 101) ou aplatis ;
- ***(iv)*** les instances peuvent être supprimées ;
- ***(v)*** les propriétés peuvent être supprimées ou ayant leurs restrictions de classes éliminées ;
- ***(vi)*** les classes peuvent être multipliées ou réduites.

1. http ://cran.r-project.org
2. http ://oaei.ontologymatching.org/2012/benchmarks

Une interaction positive (ou de moins en moins négative), reflète un indice de Shapley aussi important. La figure 1, montre que pour l'ontologie 101, SIMTERM a une importance qui dépasse SIMTOPO et SIMSEMANT. Avec l'absence des noms d'entités, la famille $20x$ marque une importance élevée pour SIMTOPO dépassant ainsi SIMTERM et SIMSEMANT. Les ontologies des familles $22x$ et $23x$ se distinguent par une morphologie structurelle pauvre (hiérarchie aplatie ou supprimée), ce qui défavorise la coalition de SIMTOPO avec les autres composantes. Les indices de Shapley pour les familles $24x$, $25x$ et $26x$ sont très proches, vu que toutes les composantes interagissent négativement à ce niveau. De point de vue mesure de similarité, les trois composantes à la fois sont incapables de fournir de bonnes valeurs, puisque les ontologies des familles $24x$, $25x$ et $26x$ sont doublement altérées, *i.e.*, terminologiquement et structurellement. Comme l'illustre la figure 2, et par rapport à un paramétrage figée, la moyenne de l'amélioration varie entre 6% (la famille $25x$) et 12% (la famille $20x$).

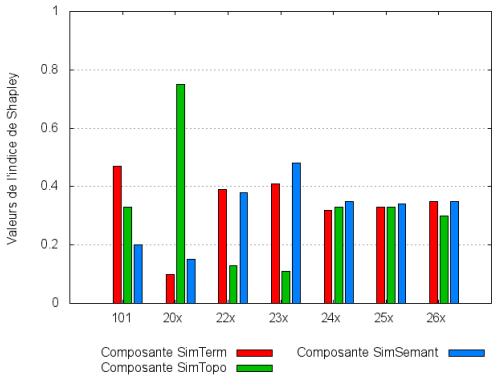

FIG. 1 – *Indices de Shapley par familles de tests.*

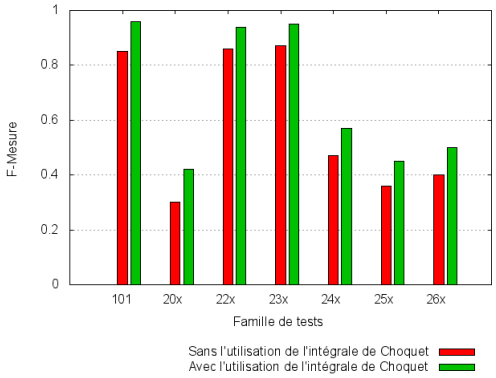

FIG. 2 – *Les valeurs de F-Mesure par famille de tests.*

5 Conclusion

Dans ce travail, nous avons proposé une nouvelle une méthode d'alignement d'un traitement "intelligent" pour agréger les valeurs de similarités fournies par différents modules. Une telle agrégation évitera à la méthode d'alignement de réagir d'une façon figée. La méthode proposée a été testée sur la base Benchmark (l'ontologie FINANCE) et a montré une nette amélioration sur les métriques d'évaluation. De ce fait, une expérimentation sur une plus large gamme d'ontologies réelles devrait nous permettre de perfectionner et d'optimiser notre outil et d'étudier son comportement sous l'impact d'autres contraintes.

Références

Euzenat, J., A. Ferrara, W. R. van Hage, L. Hollink, C. Meilicke, A. Nikolov, D. Ritze, F. Scharffe, P. Shvaiko, H. Stuckenschmidt, O. Sváb-Zamazal, et C. T. dos Santos (2011). Results of the ontology alignment evaluation initiative 2011. In *Proceedings of the* 6^{th} *International Workshop on Ontology Matching (OM-2011), Bonn, Germany, October 24, 2011*, Volume 814 of *CEUR-WS*.

Grabisch, M. (1996). The representation of importance and interaction of features by fuzzy measures. *Pattern Recognition Letters 17*(6), 567–575.

Grabisch, M., I. Kojadinovic, et P. Meyer (2008). A review of methods for capacity identification in choquet integral based multi-attribute utility theory : Applications of the kappalab r package. *European Journal of Operational Research 186*(2), 766–785.

Hu, W. et Y. Qu (2008). Falcon-ao : A practical ontology matching system. *Journal of Web Semantics 6*(3), 237–239.

Kaci, S. (2011). *Working with Preferences : Less Is More*. Cognitive Technologies. Springer.

Marichal, J.-L. et M. Roubens (2000). Determination of weights of interacting criteria from a reference set. *European Journal of Operational Research 124*(3), 641–650.

Summary

The increasing number of ontologies makes the process of ontology alignment an essential component of the Semantic Web. The quality of the provided alignments is closely related to some parameters that govern their treatment. In this paper, we introduce a new approach for the automatic adaptation of an ontology alignment alignment method parameters. This approach is based on the use of the Choquet integral as an advanced aggregation operator. It tends to dynamically assign the different weights to modules of an alignment method. Carried out experiments highlight a significant improvement vs that of a static configuration.