HAL open science

Mesures d'hétérogénéité sémantique des systèmes P2P non-structurés

Thomas Cerqueus, Sylvie Cazalens, Philippe Lamarre

To cite this version:

Thomas Cerqueus, Sylvie Cazalens, Philippe Lamarre. Mesures d'hétérogénéité sémantique des systèmes P2P non-structurés. Extraction et gestion des connaissances EGC'2010, Jan 2011, Brest, France. pp.341. hal-00625119

HAL Id: hal-00625119

https://hal.science/hal-00625119v1

Submitted on 20 Sep 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mesures d'hétérogénéité sémantique des systèmes P2P non-structurés

Thomas Cerqueus*, Sylvie Cazalens* et Philippe Lamarre*,**

*LINA - CNRS - UMR 6241, Université de Nantes, **INRIA
2, rue de la Houssinière - 44000 Nantes, France
{thomas.cerqueus, sylvie.cazalens, philippe.lamarre}@univ-nantes.fr

Résumé. L'autonomie des participants dans les systèmes P2P pour le partage de données peut conduire à une situation d'hétérogénéité sémantique dans le cas où les participants utilisent leurs propres ontologies pour représenter leurs données. Dans cet article nous commençons par définir des mesures de disparité entre participants en considérant leurs contextes sémantiques. En considérant la topologie du système et les disparités entre participants, nous proposons des mesures d'hétérogénéité sémantique d'un système P2P non-structuré.

1 Introduction

Dans les systèmes pair-à-pair conçus pour le partage de données, les participants sont autonomes. Dans le cas où ceux-ci utilisent des ontologies pour représenter leurs données, nous ne pouvons pas supposer qu'ils utilisent tous la même, car leurs objectifs et leurs besoins peuvent être différents. Dans de tels systèmes, il peut devenir difficile de faire en sorte que les pairs puissent partager des données et se comprendre. Dans cet article, nous focalisons sur la cause même du problème, à savoir le fait que des participants n'utilisent pas tous la même ontologie ou, même si tel était le cas, ne l'appréhendent pas exactement de la même façon. C'est ce que nous appelons l'*hétérogénéité sémantique* du système. Intuitivement, plus le système est hétérogène, plus il est difficile d'assurer l'interopérabilité en son sein. L'objectif de cet article est de définir des mesures d'hétérogénéité sémantique d'un système P2P.

La section 2 introduit un certain nombre de définitions. Puis la section 3 considère les disparités entre les contextes sémantiques de deux participants et propose deux mesures permettant de capturer différents aspects des disparités. La section 4 focalise sur la définition de l'hétérogénéité d'un système en exploitant les mesures définies dans la section 3. Là encore, plusieurs mesures sont proposées afin de prendre en compte différentes facettes de l'hétérogénéité. Nous concluons en section 5.

2 Définitions de base

Chaque participant d'une communauté ou d'un système peut vouloir utiliser sa propre ontologie. Et même en se référant à une ontologie commune, le contexte spécifique d'un participant (interprétation, usage, ...) peut amener des "distorsions" par rapport à un autre. Ainsi

certaines entités peuvent être perçues comme nettement plus similaires par l'un que par l'autre. Par exemple, un fleuriste aura tendance à considérer que le concept le plus proche de fleur est la rose (ou l'orchidée), alors que pour un guide de haute montagne, il pourrait s'agir de l'edelweiss, bien qu'ils connaissent tous deux ces fleurs. Une mesure de similarité entre concepts permet d'exprimer ces différences.

Ontologie Nous considérons qu'une ontologie est constituée de concepts, de relations, de propriétés, mais qu'elle ne contient pas d'individus (Staab et Studer, 2004). Les concepts sont reliés par des relations : subsomption, méronymie, relations de domaine, etc. Par ailleurs, à chaque concept peuvent être attribuées des propriétés. L'ensemble des concepts, des propriétés et des relations de l'ontologie o sont notés C_o, P_o et R_o, et l'union de ces trois ensembles est notée E_o. La figure 1 présente deux ontologies où les concepts sont représentés par des rectangles blancs, les propriétés par des rectangles gris, et les relations par des traits noirs (flèches ou losanges).

Similarité intra-ontologie Une fonction mesurant la proximité entre deux concepts de la même ontologie est appelée similarité intra-ontologie. Formellement, nous considérons que la similarité intra-ontologie (notée sim_o) est normalisée : $sim_o : C_o \times C_o \rightarrow [0, 1]$. De nombreuses mesures ont été proposées, par exemple celles de Wu et Palmer (1994) et de Jiang et Conrath (1997), etc. En nous appuyant sur les travaux de Ventresque (2006), nous choisissons de ne pas faire l'hypothèse que la similarité intra-ontologie est une distance mathématique.

Contexte sémantique Nous introduisons le terme *contexte sémantique* d'un participant (noté Φ) pour désigner le couple formé d'une ontologie formalisant ses connaissances, et d'une fonction de similarité sur cette ontologie caractérisant la perception qu'il en a.

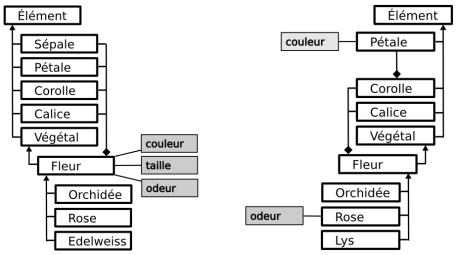


FIG. 1 – *Exemple de deux ontologies o_1 et o_2 ; la subsomption est notée ▲, la méronymie ♦.*

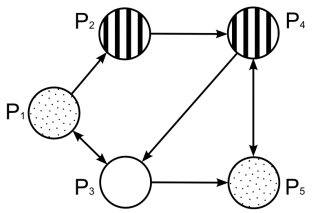


FIG. 2 – *Système P2P non-structuré de cinq participants ayant des contextes sémantiques différents.*

Similarité inter-ontologies Une fonction mesurant la proximité entre deux concepts de deux ontologies est appelée similarité inter-ontologies, et est notée SIM : $(E_o \times E_{o'}) \cup (E_{o'} \times E_o) \rightarrow [0, 1]$. Pour chaque paire d'entités $(e, e') \in E_o \times E_{o'}$, la fonction SIM détermine les valeurs de similarité de e' par rapport à e, et de e par rapport à e'. Elle possède les propriétés suivantes :

– $\text{SIM}(e, e') = 1$ si et seulement si e' est complètement similaire à e ;

– $\text{SIM}(e, e') = 0$ si et seulement si e' est complètement dissimilaire à e.

Nous ne faisons pas d'autres hypothèses sur les propriétés que respecte la fonction SIM.

Les valeurs de similarité peuvent être obtenues de plusieurs façons. Par exemple, des experts spécialistes du domaine représenté dans les ontologies peuvent déterminer ces valeurs. Les nombreux travaux réalisés dans le cadre des alignements d'ontologies peuvent aussi fournir de telles données (par ex. Euzenat et Shvaiko, 2007; Ehrig et Staab, 2004). En effet, les distances calculées lors du processus d'alignement permettent de déduire les valeurs de similarité SIM.

Nous proposons de considérer que deux entités $e \in E_o$ et $e' \in E_{o'}$ sont équivalentes si et seulement si $\text{SIM}(e, e') = 1$ et $\text{SIM}(e', e) = 1$. D'après cette définition, une entité $e \in E_o$ peut avoir plusieurs équivalents dans o', mais dans un souci de simplicité et de clarté, nous considérons qu'il n'y en a qu'un et qu'il est noté $eq^e_{o'}$. L'ensemble des entités de o ayant des équivalents dans o' est noté $E_o^{eq_{o'}}$.

3 Disparités sémantiques entre deux participants

Au niveau sémantique, les disparités peuvent apparaître car les domaines couverts, les granularités (c.-à-d. les niveaux de détails), et les perspectives (c.-à-d. les points de vue) peuvent être différents. Dans un premier temps, notre objectif est de proposer des mesures fournissant des informations sur les différents aspects et degrés de disparité entre les contextes sémantiques de deux participants. Du fait du manque d'espace, nous ne présentons ici que deux mesures, notées $\mathcal{D}$. Elles sont normalisées dans l'intervalle $[0, 1]$ et respectent les propriétés suivantes :

– minimalité : $\forall P, \mathcal{D}(P, P) = 0$;

– positivité : $\forall P, P', \mathcal{D}(P, P') \geqslant 0$.

Nous ne faisons pas d'autres hypothèses sur les propriétés qu'elles possèdent.

Mesure basée sur les concepts Les concepts définis dans une ontologie reflètent le domaine de connaissance modélisé ainsi que le niveau d'expertise (car un même domaine peut être représenté avec plus ou moins de concepts). Pour mesurer la disparité entre deux participants P et P', nous considérons la similarité SIM de chaque concept de o avec chacun de ceux de o'. Nous choisissons de ne considérer que la valeur maximale. Si celle-ci est égale à 1, le concept ne génère pas de disparité. Dans le cas contraire, la disparité est d'autant plus forte que la dissimilarité est grande. Nous traduisons cela par une fonction $dispC$, fonction monotone décroissante définie sur $[0, 1]$ telle que $dispC(1) = 0$ et $dispC(0) = 1$. Nous définissons alors la disparité de P par rapport à P' (notion non symétrique) par :

$$\mathcal{D}_{concept}(P, P') = \frac{1}{|C_o|} \sum_{c \in C_o} dispC(\max_{c' \in C_{o'}} \text{SIM}(c, c'))$$

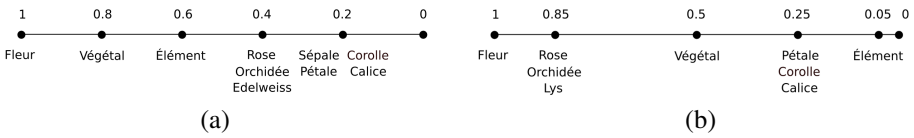
De la même manière, nous pouvons définir des mesures en se basant sur les propriétés et les relations des ontologies. Ainsi la notion de disparité est abordée avec des perspectives complémentaires.

Mesure basée sur le "désordre" Nous introduisons la notion de rang d'un concept c_1 par rapport à un concept c, notée $rang^{c_1}_{\Phi}(c)$ où : $rang^{c_1}_{\Phi}(c) = |\{s \in S^c_{\Phi} : s \geqslant sim_o(c_1, c)\}|$. Pour

un concept $c \in C_o$ ayant un équivalent dans o', l'ensemble S^c_Φ est défini par :
$S^c_\Phi = \{s \in [0,1] : \exists c' \in (C_o \cap E_o^{eq_{o'}})$ tel que $sim_o(c, c') = s\}$.

Sur les exemples présentés sur la figure 3, le concept *Rose*$_1$ est classé au quatrième rang selon la fonction de similarité sim_{o_1} par rapport au concept *Fleur*$_1$ (figure 3a) alors qu'il est classé au deuxième rang selon sim_{o_2} (figure 3b). Le fait que le classement soit différent peut avoir de l'importance. Par exemple, en recherche d'information, certaines méthodes utilisent les valeurs de similarité pour étendre les requêtes. Ainsi une requête constituée du concept *Fleur* pourrait être étendue avec le concept *Végétal* dans le contexte Φ_1 (car il est le concept le plus proche de *Fleur*) alors que dans le contexte Φ_2 l'expansion porterait sur les concepts *Rose*, *Lys* et *Orchidée*. Les réponses obtenues peuvent alors différer dans l'un et l'autre cas, certaines étant plus pertinentes. Nous proposons donc de mesurer le désordre autour de chaque concept $c \in (C_o \cap E_o^{eq_{o'}})$. Le désordre $des_{\Phi,\Phi'}(c)$ est défini par la distance *Rank Distance* (Dinu, 2003) où les différences de rangs sont normalisées dans $[0, 1]$:

$$des_{\Phi,\Phi'}(c) = \frac{1}{|C_o \cap E_o^{eq_{o'}}|} \sum_{c_0 \in (C_o \cap E_o^{eq_{o'}})} \frac{|rang^c_\Phi(c_0) - rang^{eq^c_{o'}}_{\Phi'}(eq^{c_0}_{o'})|}{\max(|S^c_\Phi|, |S^{eq^c_{o'}}_{\Phi'}|) - 1}$$



(a) (b)

FIG. 3 – *Valeurs de similarités intra-ontologies (a) des concepts de l'ontologie o_1 par rapport au concept Fleur$_1$ et (b) des concepts de l'ontologie o_2 par rapport au Fleur$_2$.*

La moyenne des valeurs de $des_{\Phi,\Phi'}$ pour chaque concept de o ayant des équivalents dans o', définit une mesure globale, notée $\mathcal{D}_{desordre}(P, P')$. Une faible valeur signifie que globalement les concepts communs aux ontologies o et o' sont rangés de la même façon dans les deux contextes : on peut conclure que les perspectives sont proches. À l'inverse, une valeur forte indique que les perspectives des deux participants divergent fortement.

4 Hétérogénéité sémantique d'un système P2P

Dans cette section, nous définissons des mesures d'hétérogénéité d'un système P2P en utilisant l'une des mesures $\mathcal{D}$ présentées précédemment. Nous appréhendons l'hétérogénéité de plusieurs façons selon que nous considérons la diversité des contextes, les disparités entre les participants, ou la topologie du système. Nous différons en cela des travaux d'Euzenat et al. (2009) (qui ne considèrent pas la topologie du réseau) ou de Cudré-Mauroux et Aberer (2004) (qui ne considèrent pas les disparités).

Mesure basée sur le nombre de contextes sémantiques Intuitivement, si tous les participants partagent le même contexte sémantique, alors le système est parfaitement homogène. Par contre, plus le nombre de contextes sémantiques présents dans le système est élevé, plus

le système est hétérogène. Nous proposons la mesure $\mathcal{H}_{card}$ définie par : $\mathcal{H}_{card} = \frac{\phi-1}{|\mathcal{P}|-1}$, où ϕ représente le nombre de contextes différents utilisés dans le système, et $\mathcal{P}$ représente l'ensemble des participants. Cette mesure renvoie 0 si le système est homogène (un seul contexte utilisé), et 1 s'il est complètement hétérogène. Pour le système présenté sur la figure 2, nous avons : $\mathcal{H}_{card}(\mathcal{S}) = \frac{3-1}{5-1} = 0,5$.

Mesure basée sur les disparités La mesure $\mathcal{H}_{card}$ ne prend pas en compte les disparités entre les différents contextes sémantiques alors qu'elles influent fortement sur l'hétérogénéité. Comme nous ne faisons aucune hypothèse sur la manière dont les informations peuvent transiter de pair en pair, il est nécessaire de prendre en considération les disparités entre tous les couples de pairs. Cela nous conduit à définir la mesure $\mathcal{H}_{globale}$ par :

$$\mathcal{H}_{globale}(\mathcal{S}) = \frac{1}{|\mathcal{P}|^2 - |\mathcal{P}|} \sum_{\substack{P_i, P_j \in \mathcal{P} \\ P_i \neq P_j}} \mathcal{D}(P_i, P_j)$$

Une valeur forte d'hétérogénéité indique que les participants sont fortement hétérogènes deux-à-deux, tandis qu'une valeur faible peut s'expliquer de deux manières : soit le système contient de nombreux contextes sémantiques différents ($\mathcal{H}_{card}$ est élevé) entre lesquels les disparités sont faibles ; soit il contient peu de contextes différents ($\mathcal{H}_{card}$ est faible) entre lesquels les disparités sont importantes. Cela montre que ces deux mesures sont complémentaires.

Mesure centrée sur un participant Nous proposons ici de prendre en compte la topologie du système en mesurant à quel point le voisinage d'un participant P est hétérogène. En considérant l'ensemble des participants formant son voisinage de rayon n noté $\mathcal{V}_n^P$, nous proposons la mesure suivante :

$$\mathcal{H}_{part}^{P,n}(\mathcal{S}) = \frac{1}{|\mathcal{V}_n^P|} \sum_{P_i \in \mathcal{V}_n^P} \mathcal{D}(P, P_i)$$

La valeur est faible si les disparités entre P et chacun de ses voisins sont faibles ; dans le cas contraire elle est forte. La mesure $\mathcal{H}_{part}^{n}(\mathcal{S})$ peut être définie en considérant la moyenne de l'hétérogénéité autour de chaque participant. Cette mesure caractérise le système par rapport à son organisation.

Intuitivement, le voisinage d'un participant P lui est favorable s'il est composé de participants dont il est proche sémantiquement. Ainsi, les participants qui le comprennent le mieux sont accessibles facilement. Formellement, le voisinage d'un participant P lui est favorable si et seulement si $\forall i \in [\![1, n-1]\!]$ on a : $\mathcal{H}_{part}^{P,i}(\mathcal{S}) \leqslant \mathcal{H}_{part}^{P,i+1}(\mathcal{S})$. On peut alors comparer deux systèmes $\mathcal{S}_1$ et $\mathcal{S}_2$ contenant les mêmes participants. Un participant P est plus favorablement placé dans $\mathcal{S}_1$ que dans $\mathcal{S}_2$ si et seulement si $\forall i \in [\![1, n]\!]$ on a : $\mathcal{H}_{part}^{P,i}(\mathcal{S}_1) \leqslant \mathcal{H}_{part}^{P,i}(\mathcal{S}_2)$.

Nous considérons que le système $\mathcal{S}_1$ est mieux organisé que $\mathcal{S}_2$ si chaque participant P est plus favorablement placé dans $\mathcal{S}_1$ que dans $\mathcal{S}_2$. Cette condition étant stricte, nous pouvons aussi considérer que $\mathcal{S}_1$ est mieux organisé que $\mathcal{S}_2$ si au moins tous les participants favorablement placés dans $\mathcal{S}_2$ sont également favorablement placés dans $\mathcal{S}_1$, et que les autres ne sont pas moins favorablement placés. Le fait de pouvoir dire qu'un système est mieux organisé qu'un autre permet de qualifier la difficulté à assurer l'interopérabilité.

5 Conclusion

Dans cet article nous avons proposé un certain nombre de mesures de disparité sémantique entre deux participants en considérant leurs contextes sémantiques. Puis nous avons envisagé plusieurs définitions de l'hétérogénéité sémantique d'un système P2P en considérant la diversité des contextes, les disparités entre pairs, ou la topologie du système. Nous pensons que ces mesures sont complémentaires. À terme ce travail devrait permettre de caractériser finement différentes situations d'hétérogénéité en vue de mieux évaluer et comparer différentes classes d'algorithmes que ce soit pour l'évaluation de requêtes ou la clusterisation sémantique dans les systèmes P2P.

Références

Cudré-Mauroux, P. et K. Aberer (2004). A necessary condition for semantic interoperability in the large. In *3rd International Conference on Ontologies, Databases and Applications of Semantics (ODBASE)*, pp. 859–872.

Dinu, L. P. (2003). On the classification and aggregation of hierarchies with different constitutive elements. *Fundamenta Informaticae*, 39–50.

Ehrig, M. et S. Staab (2004). QOM - quick ontology mapping. In *3rd International Semantic Web Conference (ISWC)*, pp. 683–697.

Euzenat, J., C. Allocca, J. David, M. d'Aquin, C. Le Duc, et O. Svab-Zamazal (2009). Ontology distances for contextualisation. Deliverable, NeOn.

Euzenat, J. et P. Shvaiko (2007). *Ontology matching*. Springer-Verlag.

Jiang, J. J. et D. W. Conrath (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *10th Int. Conf. Research on Computational Linguistics*, pp. 19–33.

Staab, S. et S. Studer (2004). *Handbook on Ontologies*. Springer.

Ventresque, A. (2006). Une mesure de similarité sémantique utilisant des résultats de psychologie. In *3ème Conf. en Recherche d'Informations et ses Applications (CORIA)*, pp. 371–376.

Wu, Z. et M. Palmer (1994). Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 133–138.

Summary

In peer-to-peer data sharing systems with autonomous participants, interoperability within the system is made difficult because peers do not use the same ontology or do not understand it the same way. In this paper we firstly define measure of disparity between two participants according to their semantic contexts. Considering the topology of the system and the disparities between participants, we propose measures of semantic heterogeneity of an unstructured P2P system.