Recherche d'information sémantique dans les systèmes P2P hétérogènes

Thomas Cerqueus[1]

LINA - UMR 6241 – Université de Nantes
2, rue de la Houssinière - BP 92208
44322 Nantes cedex 3

thomas.cerqueus@univ-nantes.fr

RÉSUMÉ. *Nous considérons la recherche d'information sémantique dans les systèmes pair-à-pair. Ces derniers semblent être une solution intéressante pour le partage de données car ils garantissent le passage à l'échelle, et gère la dynamicité. Dans ce contexte, il est difficilement imaginable que tous les participants s'accordent sur l'utilisation d'une même représentation sémantique (schéma, ontologie, graphe conceptuel). Dans ce cas, le système est sémantiquement hétérogène. Cette situation limite l'interopérabilité entre participants. Dans cet article nous montrons quels sont les problématiques liées à l'hétérogénéité sémantique et nous présentons les solutions que nous proposons pour garantir un certain degré d'interopérabilité malgré l'hétérogénéité. L'originalité de ce travail se trouve dans le fait de tenter d'améliorer l'interopérabilité sémantique en s'attaquant directement à la source du problème : l'hétérogénéité.*

ABSTRACT. *We consider semantic information retrieval in P2P systems. These systems are an interesting solution for data sharing because they are scalable and handle dynamicity. In this context it is unlikely that all peers use the same semantic representation (schema, ontology, conceptual graph). In this case, the system is said to be semantically heterogeneous. This situation prevents good interoperability between participants. In this paper we present issues related to semantic heterogeneity, and we present the solutions that we propose to ensure a degree of interoperability in spite of heterogeneity. The specificity of this work lies in trying to improve interoperability by directly considering the problem of heterogeneity.*

MOTS-CLÉS : *Recherche d'information, systèmes P2P, hétérogénéité sémantique, interopérabilité.*

KEYWORDS: *Information retrieval, P2P systems, semantic heterogeneity, interoperability.*

1. Directeur : Philippe Lamarre – Co-encadrant : Sylvie Cazalens

1. Contexte général

Les systèmes pair-à-pair (P2P) constituent une solution intéresante pour le partage de données car ils supportent le passage à l'échelle et la dynamicité, et garantissent l'autonomie des participants. Dans notre travail nous considérons des systèmes P2P dans lesquels chaque pair utilise sa propre ontologie pour représenter les documents qu'il souhaite partager. Ces derniers sont indexés ou annotés par rapport à cette ontologie. Lorsqu'un pair veut rechercher des documents, il émet une requête et la propage dans le système. Comme le nombre de pairs dans le système peut être très important, il est peu probable qu'ils puissent tous se mettre d'accord sur l'utilisation d'une unique ontologie. En effet ils ont des objectifs, des contextes, des points de vue et des niveaux d'expertise différents, qui les amènent à modéliser leur domaine de manières différentes. Dans ce cas, nous disons que le système est sémantiquement hétérogène. La figure 1 présente un système P2P hétérogène. L'hétérogénéité sémantique est à priori un frein à l'interopérabilité car elle empêche les participants de se comprendre.

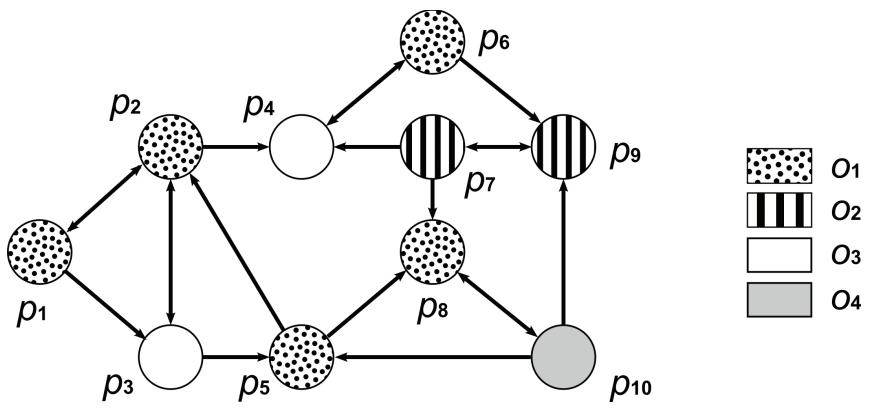

Figure 1. *Système P2P non-structuré sémantiquement hétérogène.*

L'objectif de notre travail est de définir un système dans lequel les pairs peuvent rechercher efficacement des documents (c.-à-d. interopérer) malgré le fait qu'ils utilisent des ontologies différentes pour représenter leurs documents.

2. Problématiques

L'hétérogénéité sémantique empêche les pairs de rechercher efficacement des documents car les requêtes qu'ils émettent peuvent être imcomprises (ou mal comprises) par les autres pairs du système. Par exemple lorsque le pair p_1 émet une requête, celle-ci risque d'être imcomprise par les pairs n'utilisant pas l'ontologie o_1 (cf. figure 1). Le processus d'alignements entre ontologies produit des correspondances qui permettent aux pairs de traduire les requêtes qu'ils reçoivent (Euzenat *et al.,* 2007). Néanmoins ils ne permettent pas de garantir une interopérabilité parfaite car certains concepts peuvent ne pas être partagés entre plusieurs ontologies.

Notre approche consiste à séparer les notions d'hétérogénéité et d'interopérabilité. Si dans notre cas, l'interopérabilité peut être mesurée avec les mesures de précision et de rappel (car nous considérons une application de RI), nous ne connaissons pas de mesures permettant de caractériser l'hétérogénéité. Par ailleurs nous ne connaissons pas d'algorithme conçu spécifiquement pour diminuer l'hétérogénéité sémantique. Nous pensons que réduire le degré d'hétérogénéité doit permettre d'améliorer l'interopérabilité. Pour résumer, les problématiques que nous abordons sont les suivantes : (i) comment caractériser et mesurer l'hétérogénéité sémantique d'un système P2P, (ii) comment réduire l'hétérogénéité sémantique de tels systèmes, et (iii) quels algorithmes proposer pour assurer l'interopérabilité. Les solutions que nous proposons sont présentées dans la section 3.

3. Propositions

3.1. *Mesures d'hétérogénéité sémantique*

L'hétérogénéité sémantique est une notion complexe qui provient du fait que tous les participants d'un système P2P n'utilisent pas la même ontologie. Une façon simple d'aborder l'hétérogénéité est de faire le ratio entre le nombre d'ontologies utilisées dans le système et le nombre de participants. Cette première approche (qui consiste à qualifier la diversité sémantique du système) permet de caractériser simplement l'hétérogénéité mais elle ne prend pas en compte tous ses aspects. En effet elle dépend également de la disparité entre les pairs, car deux ontologies différentes peuvent aussi bien être très similaires ou très dissimilaires. Globalement plus les pairs sont disparates les uns des autres, plus le système est hétérogène. La disparité entre deux pairs peut être mesurée de différentes manières (Euzenat *et al.,* 2007, David *et al.,* 2010). En général nous considérons qu'elle dépend des ontologies utilisées par les pairs, et des correspondances que les pairs connaissent entre leurs ontologies. Dans les systèmes P2P non-structurés, le voisinage d'un pair a une importance toute particulière car les requêtes qu'il émet atteignent seulement les pairs qui l'entourent. C'est pour cette raison que nous pensons que l'hétérogénéité sémantique dépend également de la manière dont les pairs sont organisés dans le système. Nous étudions alors l'hétérogénéité centrée sur chacun des participants en considérant les ontologies qu'ils utilisent, ou la disparité avec leurs voisinages : plus l'hétérogénéité centrée sur chacun des pairs est élevée, plus l'hétérogénéité globale du système est élevée. L'ensemble de ces intuitions a mené à la définition de cinq mesures permettant de capturer chacun des aspects de l'hétérogénéité (Cerqueus *et al.,* 2011a). Elles sont toutes normalisées entre 0 (système homogène) et 1 (système complétement hétérogène). Le fait de disposer de plusieurs mesures permet de caractériser finement l'hétérogénéité sémantique en fonction de chaque facette. L'équation [1] présente deux mesures d'hétérogénéité : celle mesurant la diversité, et celle mesurant la disparité globale du système (d est une mesure de disparité entre deux pairs). Pour des raisons de place nous ne présentons pas toutes les mesures (cf. (Cerqueus *et al.,* 2011a)).

$$\mathcal{H}_{Rich}(\mathcal{S}) = \frac{|o_{\mathcal{S}}| - 1}{|\mathcal{P}| - 1} \qquad \mathcal{H}_{Disp}(\mathcal{S}) = \frac{1}{|\mathcal{P}|^2 - |\mathcal{P}|} \sum_{p \neq p' \in \mathcal{P}} d(p, p') \qquad [1]$$

3.2. *Algorithmes de diminution de l'hétérogénéité sémantique*

Nous avons défini deux algorithmes de diminution de l'hétérogénéité sémantique : CORDIS et GoOD-TA. Tous deux sont des protocoles épidémiques (Kermarrec *et al.,* 2007). Dans un protocole épidémique, chaque pair initie régulièrement un échange avec un autre pair pour échanger des données. La nature des données échangées et la manière dont elles sont sélectionnées et traitées dépendent de l'objectif visé.

Le protocole CORDIS a pour objectif de réduire l'hétérogénéité sémantique de systèmes P2P non-structurés liée aux disparités entre pairs (Cerqueus *et al.,* 2011a). L'idée de ce protocole est de disséminer des correspondances dans le système afin de partager les correspondances connues de certains mais ignorées par d'autres. Plus les pairs connaissent de correspondances, plus ils sont en capacité de comprendre et traduire les requêtes émises par les autres pairs du système. Lorsque le pair p échange des correspondances avec p', il sélectionne des correspondances utiles pour p' et d'autres correspondances afin que p' puisse participer à la dissémination. Du point de vue de l'hétérogénéité, CORDIS permet de réduire les disparités entre les pairs. Les expérimentations que nous avons mené montrent qu'en allouant peu d'espace pour le stockage des correspondances, et en augmentant raisonnablement le trafic réseau, l'hétérogénéité sémantique liée à la disparité entre pairs peut être réduite de manière significative. Dans ces expérimentations, nous avons utilisé un ensemble d'ontologies et de correspondances, issues de la collection OntoFarm (Šváb *et al.,* 2005).

Dans (Cerqueus *et al.,* 2011b) nous avons défini le protocole GoOD-TA qui permet d'auto-organiser un système P2P non-structuré en fonction des connaissances sémantiques des pairs. Les connaissances sémantiques d'un pair concernent l'ontologie qu'il utilise, ainsi que les correspondances qu'il connaît. L'objectif de GoOD-TA est de réduire l'hétérogénéité liée à l'organisation des systèmes en faisant en sorte que les pairs proches sémantiquement soient proches dans le système. Dans ce protocole, chaque pair échange des descripteurs avec d'autres pairs. Le descripteur d'un pair est une représentation synthétique de ses connaissances sémantiques. Le fait de diffuser les descripteurs dans le système permet aux pairs de choisir explicitement leurs voisins. Ainsi le voisinage de chaque pair est composé de pairs capables de comprendre ses requêtes. De cette manière le pair a davantage de chance de trouver les données pertinentes par rapport aux requêtes qu'il émet. Nous avons mené des expérimentations, et montré que le protocole GoOD-TA permet de réduire de manière importante l'hétérogénéité sémantique liée à l'organisation des systèmes P2P non-structurés. Pour cela nous avons utilisé des ontologies utilisées dans le domaine bio-médical. Nous les avons téléchargées grâce aux services proposés par BioPortal (Fridman Noy *et al.,* 2009).

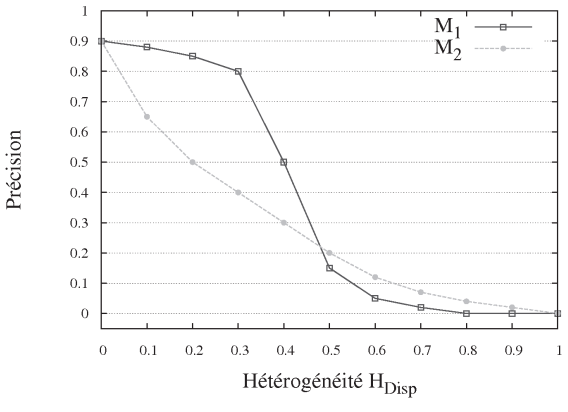

Figure 2. *Exemple de résultats permettant de comparer les performances de deux méthodes de RI en fonction d'une facette de l'hétérogénéité.*

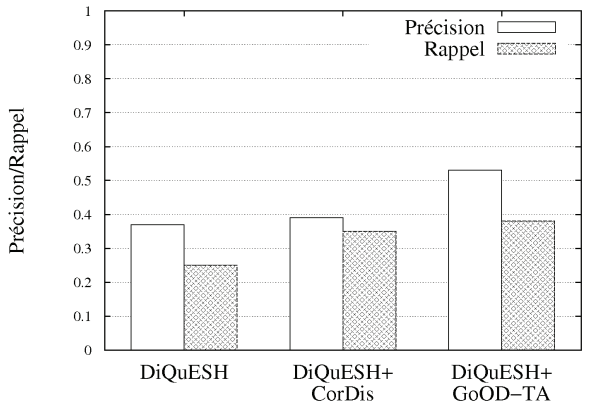

Figure 3. *Valeurs de précision/rappel obtenues par DiQuESH en fonction de l'algorithme appliqué : aucun,* CORDIS *ou GoOD-TA.*

3.3. *Étude de l'influence de l'hétérogénéité sur la recherche d'information*

Les mesures que nous avons proposées permettent d'étudier le comportement d'une méthode de RI en fonction de l'hétérogénéité. Elles peuvent permettre de tirer des conclusions du type : la méthode de recherche d'information M_1 est peu sensible à l'hétérogénéité liée aux disparités entre pairs (mesurée avec $\mathcal{H}_{Disp}$) lorsque celle-ci est inférieure à 0,3, et elle est complétement inefficace lorsque l'hétérogénéité est supérieure à 0,6. Comme nous considérons une méthode de RI, nous mesurons ses performances en termes de taux de précision et de rappel. Nous pouvons également comparer deux méthodes de RI (M_1 et M_2) dans différentes situations d'hétérogénéité. La figure 2 présente une courbe des valeurs de précision obtenues par deux méthodes M_1 et M_2 en fonction de la mesure d'hétérogénéité $\mathcal{H}_{Disp}$. Les courbes présentées servent à illustrer notre propos, et ne sont pas issues d'expérimentations.

Les mesures d'hétérogénéité permettent également de montrer quel est le bénéfice d'un algorithme de réduction de l'hétérogénéité sur les performances d'une méthode de RI. Dans (Cerqueus *et al.,* 2012) nous avons défini un algorithme de RI distribuée adapté aux systèmes P2P hétérogènes : DiQuESH. Il consiste à propager les requêtes dans un certain voisinage autour du pair initiateur, en laissant leurs traductions à la charge des pairs qui les reçoivent. Nous avons ensuite montrer l'impact des algorithmes CORDIS et GoOD-TA sur les performances de cet algorithme (en termes de précision/rappel). La figure 3 présente les taux de précision/rappel obtenus par DiQuESH dans trois situations : seul, après exécution de CORDIS, et après exécution de GoOD-TA. On voit clairement que l'exécution de l'un ou l'autre des algorithmes permet d'améliorer la qualité des résultats.

4. Perspectives

Dans ce travail nous avons présenté nos travaux récents et montré comment ils se combinent pour répondre à la problématique initiale. Les perspectives de ce travail sont les suivantes. Premièrement nous envisageons de combiner les algorithmes CORDIS et GoOD-TA pour tenter de diminuer davantage l'hétérogénéité. Nous pensons que la combinaison des deux doit permettre de la réduire plus rapidement et en limitant le coût (trafic réseau, stockage, etc.). De plus nous prévoyons de les combiner à d'autres algorithmes existants, par ex. (Nedos *et al.,* 2007). Deuxièmement nous comptons mener des expérimentations pour étudier précisément l'impact de l'hétérogénéité sur différentes méthodes de RI. Ce travail nécessitera de comparer les méthodes dans différentes situations d'hétérogénéité. La principale difficulté de cette tâche consiste à obtenir des données (ontologies, correspondances, documents, requêtes). Enfin nous envisageons d'élargir notre vision de l'hétérogénéité sémantique, pour considérer également les divergences qui peuvent exister au niveau des correspondances, car une correspondance entre ontologie peut être admise par un ensemble de participants, et pas par un autre. Nous pensons que ce travail et ces différentes perspectives permettent d'envisager la mise en place de systèmes à large échelle assurant l'interopérabilité entre les participants malgré l'hétérogénéité sémantique.

5. Bibliographie

Cerqueus T., Cazalens S., Lamarre P., « Gossiping correspondences to reduce semantic heterogeneity of unstructured P2P systems », *4th International Conference on Data Management in Grid and P2P Systems*, p. 37-48, 2011a.

Cerqueus T., Cazalens S., Lamarre P., Reducing Semantic Heterogeneity of Unstructured P2P Systems Through Gossip-Based Ontology-Driven Topology Adaptation, Technical Report n° hal-00643300, LINA, UMR 6241, 2011b.

Cerqueus T., Cazalens S., Lamarre P., « Influence de l'hétérogénéité sémantique sur les performances d'un système de RI distribuée », *Conférence en Recherche d'Infomations et Applications*, 2012.

David J., Euzenat J., Šváb-Zamazal O., « Ontology similarity in the alignment space », *9th International Semantic Web Conference*, 2010.

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

Fridman Noy N., Shah N. H., Whetzel P. L., Dai B., Dorf M., Griffith N., Jonquet C., Rubin D. L., Storey M.-A. D., Chute C. G., Musen M. A., « BioPortal : ontologies and integrated data resources at the click of a mouse », *Nucleic Acids Research*, vol. 37, p. 170-173, 2009.

Kermarrec A.-M., van Steen M., « Gossiping in Distributed Systems », *Operating Systems Review*, vol. 41, n° 5, p. 2-7, 2007.

Nedos A., Singh K., Cunningham R., Clarke S., « A Gossip Protocol to Support Service Discovery with Heterogeneous Ontologies in MANETs », *3rd IEEE International Conference on Wireless and Mobile Computing, Networking and Communications*, p. 53-61, 2007.

Šváb O., Svátek V., Berka P., Rak D., Tomášek P., « OntoFarm : Towards an Experimental Collection of Parallel Ontologies », *5th International Semantic Web Conference*, 2005. Poster.