Influence de l'hétérogénéité sémantique sur les performances d'un système de RI distribuée

Thomas Cerqueus, Sylvie Cazalens, Philippe Lamarre

Influence de l'hétérogénéité sémantique sur les performances d'un système de RI distribuée

Thomas Cerqueus* — **Sylvie Cazalens*** — **Philippe Lamarre****

* *LINA - UMR 6241*
Université de Nantes - 2, rue de la Houssinière. 44322 Nantes

** *LIRIS - CNRS - UMR 5205, F69621*
Université de Lyon - Campus de la Doua. 69622 Villeurbanne

{thomas.cerqueus, sylvie.cazalens}@univ-nantes.fr, philippe.lamarre@liris.cnrs.fr

RÉSUMÉ. *Nous considérons des systèmes pair-à-pair pour le partage de documents dans lesquels chaque pair utilise une ontologie pour représenter ses documents. Lorsque tous les pairs n'utilisent pas la même ontologie, le système est sémantiquement hétérogène, ce qui constitue à priori un frein à l'interopérabilité. Nous proposons un système dont l'organisation générique en couches logicielles sépare les algorithmes dédiés à la diminution de l'hétérogénéité de ceux utilisés pour la recherche d'information sémantique distribuée. Dans ce contexte, nous proposons une méthode de recherche d'information, puis nous focalisons sur l'impact qu'ont des algorithmes dédiés à la diminution de l'hétérogénéité sur l'interopérabilité (mesurée en termes de précision/rappel). Nos expérimentations considèrent des ontologies du domaine bio-médical et des documents issus de la base PubMed. Nous mesurons l'hétérogénéité sémantique du système et le taux de précision/rappel obtenus avant et après avoir appliqué deux algorithmes visant à diminuer l'hétérogénéité.*

ABSTRACT. *We consider peer-to-peer information sharing systems in which each peer uses an ontology to represent its documents. When the peers do not use the same ontology, the system is semantically heterogeneous, which prevents good interoperability. We propose a system with a generic layered software architecture that separates the algorithms dedicated to heterogeneity reduction from those which are used for semantic information retrieval. We propose an algorithm for distributed semantic retrieval in heterogeneous context. Then, we focus on the effects on interoperability (measured through precision and recall) of two algorithms for semantic heterogeneity reduction. Our experiments consider bio-medical ontologies and documents from the PubMed database. We measure semantic heterogeneity and precision/recall before and after using algorithms that aim to reduce heterogeneity.*

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

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

1. Introduction

Les systèmes pair-à-pair (P2P) se sont développés de manière importante ces dernières années pour échanger des documents (textuels, vidéos, etc). Ils constituent une solution intéressante pour définir des systèmes de partage d'information car, en plus de leurs propriétés de passage à l'échelle et de dynamicité, ils permettent l'autonomie des pairs et un contrôle décentralisé. Dans un tel système dédié à la recherche d'information, les pairs constituent des sources autonomes qui gèrent leurs propres documents. Ainsi, ils peuvent décider de ce qu'ils partagent, avec qui ils partagent, et comment ils représentent leurs documents, sans en référer à une entité centralisée. En particulier, ils peuvent choisir la façon dont ils indexent leurs documents. Nous nous plaçons dans le cadre d'une recherche d'information sémantique distribuée et nous supposons que chaque pair indexe ses documents par rapport à une ontologie. Lorsque les pairs utilisent tous la même ontologie, le système est sémantiquement homogène. Néanmoins ce cas de figure est peu probable car les pairs 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. Nous disons qu'il y a hétérogénéité sémantique lorsque les pairs n'utilisent pas la même ontologie. Ceci est un frein à l'interoperabilité sémantique (c.-à-d. à la capacité à communiquer et à échanger), car une requête émise par un pair risque d'être peu ou pas "comprise" par d'autres pairs.

Assurer un certain degré d'interoperabilité sémantique dans un système P2P où plusieurs ontologies sont utilisées semble indispensable. Plusieurs travaux ont étudié ce problème dans le domaine des bases de données. La plupart des solutions utilisent une traduction des concepts de la requête fournie par des alignements entre ontologies. Cette solution simple peut aussi être utilisée en recherche d'information, avec les mêmes inconvénients : si les pairs connaissent peu d'alignements, les traductions seront très pauvres et le taux de précision/rappel sera faible, indiquant un manque d'interopérabilité. Celle-ci peut être améliorée de deux façons différentes. Premièrement, il est possible de faire en sorte de réduire l'hétérogénéité sémantique. Par exemple si les pairs enrichissent l'ensemble des correspondances qu'ils connaissent, l'hétérogénéité est amoindrie, et ils sont plus aptes à traduire les requêtes : l'interopérabilité se trouve améliorée. Deuxièmement, il est possible de modifier la méthode de recherche et, par exemple, étendre la requête de façon à ce que sa traduction couvre plus de concepts connus par les autres pairs.

Un des fils directeurs de nos travaux est de bien distinguer la notion d'hétérogénéité de celle d'interopérabilité sémantique. Cela nous mène à également faire la distinction entre les algorithmes qui font décroître la première (et donc améliorent l'interopérabilité comme conséquence), et les algorithmes qui augmentent "directement" l'interopérabilité. Nous pensons que cela permet de mieux évaluer les apports des différents algortihmes et de mieux les comparer. Un premier problème a consisté à disposer de mesures pour chacune des notions. Si l'interoperabilité en RI peut se mesurer en termes de précision et rappel par exemple, nous manquions de mesures d'hétérogénéité. Dans (Cerqueus *et al.,* 2011c), nous avons proposé des mesures d'hétérogénéité sémantique, en montrant que cette notion comporte plusieurs facettes. Puis, nous avons défini des protocoles épidémiques (resp. CORDIS et GoOD-TA), spécifiquement conçus pour diminuer l'hétérogénéité sémantique, et montré dans quelle mesure ils le faisaient (Cerqueus *et al.,* 2011a)(Cerqueus *et al.,* 2011b).

L'objectif de cet article est d'étudier l'impact du taux d'hétérogénéité sur l'interopérabilité au sein d'un système de RI et l'apport des algorithmes CORDIS et GoOD-TA. Nous présentons d'abord le modèle de RI, en considérant des vecteurs sémantiques (section 2), ainsi que les mesures d'hétérogénéité (section 3). Puis nous focalisons sur la définition du système de RI. Nous proposons une architecture en plusieurs couches logicielles, en distingant celle dédiée à la réduction de l'hétérogénéité de celle dédiée à la recherche d'information (section 4). Puis nous définissons un protocole d'évaluation distribuée de requêtes top-k en contexte hétérogène (section 5). Enfin nous procédons à une série d'expérimentations (section 6). Nous utilisons des ontologies et des alignements du domaine bio-médical (Fridman Noy *et al.,* 2009), ainsi que le résultat d'annotations sémantiques de documents de la base PubMed[1]. Nous montrons l'apport de chacun des algorithmes dans l'augmentation de l'interoperabilité. L'originalité de ce travail se trouve dans le fait d'évaluer une méthode de recherche d'information en fonction de l'hétérogénéité sémantique du système P2P.

2. Modèle

2.1. *Ontologies légères et alignements*

Nous considérons qu'une ontologie o est composée d'un ensemble de concepts C_o, d'un ensemble de relations R_o entre ces concepts, et d'un ensemble de propriétés P_o (assignées à des concepts). L'union de ces trois ensembles d'entités est noté E_o. En pratique OWL permet de représenter des ontologies en définissant des *classes*, des *datatype properties* et des *object properties* (McGuinness *et al.,* 2004). Nous supposons également que chaque ontologie est identifiée par un unique URI.

L'alignement d'ontologies est un processus qui vise à identifier des correspondances entre les entités de deux ontologies (Euzenat *et al.,* 2007).

Définition 1 (Correspondance) *Une correspondance est un 4-uplet $\langle e, e', r, m\rangle$ tel que e et e' sont des entités de o et o', r est une relation entre e et e', et m est un degré de confiance.*

Dans ce travail, nous ne considérons que les relations d'équivalence ($\equiv$) entre concepts, et nous supposons qu'un concept a au plus un concept équivalent dans une autre ontologie. Un alignement entre les ontologies présentées sur la figure 1 pourrait contenir les correspondances $\langle Element_1, Element_2, \equiv, 1\rangle$, $\langle Fleur_1, Fleur_2, \equiv, 0.9\rangle$, et $\langle Petal_1, Petal_2, \equiv, 1\rangle$. Notons qu'un alignement n'est pas nécessairement parfait. En effet certaines correspondances peuvent manquer (par ex. $\langle odeur_1, fragrance_2, \equiv, 0.9\rangle$), et d'autres peuvent être incorrectes (par ex. $\langle Fleur_1, Lys_2, \equiv, 1\rangle$). Nous considérons dans un premier temps qu'un alignement ne contient pas de correspondance incorrecte.

1. www.ncbi.nlm.nih.gov/pubmed/

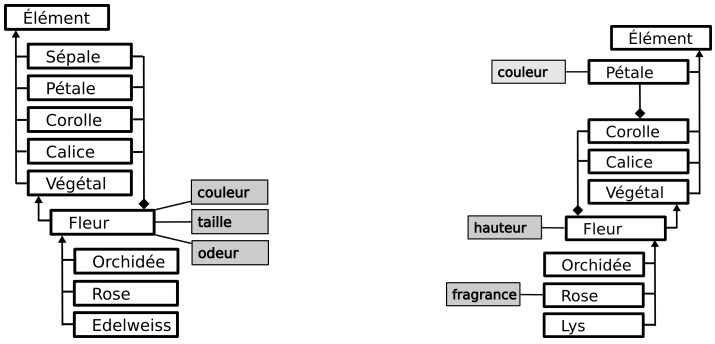

Figure 1. *Deux ontologies o_1 et o_2 composées de concepts, de propriétés et de relations.*

2.2. *Modèle de recherche d'information*

Modèle vectoriel sémantique

Nous considérons le modèle vectoriel sémantique. C'est une extension du modèle vectoriel (Salton *et al.,* 1975) dans lequel chaque dimension correspond à un concept d'une ontologie. Documents et requêtes sont chacun représentés par un vecteur à n dimensions où n est le nombre de concepts de l'ontologie (Ventresque, 2008). Un poids (valeur numérique comprise entre 0 et 1) est associé à chaque dimension en fonction de la représentativité du concept dans le document ou la requête. Par exemple, un vecteur exprimé dans l'espace de l'ontologie o_2 (cf. figure 1) peut être égal à : $[(Element, 0), (Petale, 0.8), (Corolle, 0), (Calice, 0), (Vegetal, 0), (Fleur, 1), (Orchidee, 0), (Rose, 0), (Lys, 0.3)]$. Le vecteur sémantique d'un document ou d'une requête peut être obtenu de différentes manières : indexation automatique, indexation manuelle, annotation. Par ailleurs nous ne faisons pas d'hypothèse sur la nature des documents considérés. Il peut sagir de documents textuels, de documents multimédias, de pages web, de diagnostics médicaux, etc. Le vecteur sémantique exprimé dans l'espace défini par o correspondant au document d est noté $\overrightarrow{d_o}$.

Mesure de pertinence

Pour mesurer la pertinence d'un document d par rapport à une requête q, nous mesurons le cosinus de l'angle formé par $\overrightarrow{d_o}$ et $\overrightarrow{q_o}$ (définis dans le même espace) :

$$\cos(\overrightarrow{d_o}, \overrightarrow{q_o}) = \frac{\overrightarrow{d_o} \cdot \overrightarrow{q_o}}{|\overrightarrow{d_o}| \times |\overrightarrow{q_o}|}$$

2.3. *Systèmes P2P non-structurés*

Chaque pair p d'un système P2P non-structuré est identifié de manière unique. Cet identifiant noté $id(p)$ peut être calculé à partir de l'adresse IP du pair et d'un numéro de port.

Les liens entre pairs sont stockés par les pairs eux-mêmes : chaque pair p maintient une vue locale du système $vue(p)$ contenant les identifiants d'autres pairs. Ces derniers sont appelés les voisins de p (cet ensemble est noté $\mathcal{N}_p$).

Définition 2 (Système P2P non-structuré) *Un système P2P non-structuré est défini par un graphe $\mathcal{S} = \langle \mathcal{P}, \mathcal{N} \rangle$ où $\mathcal{P}$ est un ensemble de pairs, et $\mathcal{N}$ une relation définie par :* $\mathcal{N} = \{(p_i, p_j) \in \mathcal{P}^2 : p_j \in vue(p_i)\}$.

Nous définissons le voisinage d'un pair p dans un rayon n (noté $\mathcal{N}_p^n$) comme étant l'ensemble des pairs que p peut atteindre avec $l \leqslant n$ sauts.

Définition 3 (Fonction d'appariement pair-ontologie) *Étant donné un système P2P $\mathcal{S} = \langle \mathcal{P}, \mathcal{N} \rangle$ et un ensemble d'ontologies $\mathcal{O}$, nous considérons une fonction d'appariement pair-ontologie $\mu : \mathcal{P} \rightarrow \mathcal{O}$ qui attribue une ontologie à chaque pair.*

En pratique le fait qu'un pair p utilise une ontologie o (c.-à-d. $\mu(p) = o$) signifie que p représente ses documents et ses requêtes à l'aide des concepts de o. Afin de comprendre les requêtes qu'ils reçoivent, les pairs doivent connaître des correspondances. Nous notons κ_p l'ensemble des correspondances stockées par p.

3. Hétérogénéité sémantique

3.1. *Une notion multifacette*

L'hétérogénéité sémantique est une notion complexe qui provient du fait que tous les participants d'un système P2P ne s'accordent pas sur l'utilisation d'une ontologie unique. Dans (Cerqueus *et al.,* 2011c), nous avons identifié différentes facettes de l'hétérogénéité sémantique, et nous avons proposé plusieurs mesures permettant de la caractériser précisément.

Tout d'abord l'hétérogénéité sémantique dépend du nombre d'ontologies utilisées dans le système : plus il est élevé, plus le système est hétérogène. C'est ce que nous appelons la diversité sémantique. Elle peut mesurer ainsi :

$$\mathcal{H}_{Div}(\mathcal{S}) = \frac{|o_{\mathcal{S}}| - 1}{|\mathcal{P}| - 1} \quad [1]$$

où $o_{\mathcal{S}}$ est l'ensemble des ontologies utilisées dans le système $\mathcal{S}$, et $\mathcal{P}$ est l'ensemble des pairs de $\mathcal{S}$. L'hétérogénéité ne dépend pas seulement de la diversité sémantique car deux ontologies différentes peuvent aussi bien être très similaires ou très dissimilaires. Nous utilisons donc en plus la disparité entre deux pairs. Plusieurs définitions sont possibles. Ici nous considérons les correspondances existantes entres les ontologies des pairs. Globalement plus les pairs sont disparates les uns des autres, plus le système est hétérogène. Nous utilisons la mesure $\mathcal{H}_{Disp}$ définie par :

$$\mathcal{H}_{Disp}(\mathcal{S}) = \frac{1}{|\mathcal{P}|^2 - |\mathcal{P}|} \sum_{p \neq p' \in \mathcal{P}} d(p, p') \quad \text{où} \quad d(p, p') = 1 - \frac{|\kappa_{p'}^{\equiv}(o, o')|}{|C_o|} \quad [2]$$

où p (resp. p') utilise l'ontologie o (resp. o'), et $\kappa_{p'}(o, o')$ est l'ensemble des correspondances que p' connaît entre o et o'. Dans les systèmes P2P non-structurés, le voisinage d'un pair a une importance toute particulière car les requêtes qu'ils émet atteignent seulement les pairs qui le composent. 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 l'hétérogénéité centrée sur les pairs soit en considérant seulement les ontologies, soit en considérant les disparités entre les pairs et leurs voisinages. L'hétérogénéité centrée sur un pair p peut être mesurée par :

$$\mathcal{H}^n_{Dap}(\mathcal{S}, p) = \frac{1}{|\mathcal{N}^n_p|} \sum_{p_i \in \mathcal{N}^n_p} d(p, p_i) \quad [3]$$

où d est une mesure de disparité entre pairs (par ex. celle présentée dans [2]). Le paramètre n peut être fixé en fonction de l'application étudiée (par ex. $n = TTL$). En définitive plus l'hétérogénéité centrée sur chacun des pairs est élevée, plus l'hétérogénéité globale du système est élevée. Cette dernière peut être mesurée par :

$$\mathcal{H}^n_{DapAvg}(\mathcal{S}) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathcal{H}^n_{Dap}(\mathcal{S}, p) \quad [4]$$

Étant données ces différentes facettes, nous avons proposé des algorithmes permettant de réduire l'hétérogénéité de certaines d'entre elles.

3.2. *Réduction de l'hétérogénéité liée aux disparités entre pairs*

Nous avons proposé le protocole CORDIS pour réduire l'hétérogénéité sémantique de systèmes P2P non-structurés (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. Le but est de diminuer l'hétérogénéité du système liée aux disparités entre pairs. Disséminer des correspondances permet à des pairs d'en apprendre des nouvelles qui sont éventuellement utiles pour traduire les requêtes qu'ils reçoivent. CORDIS est un protocole épidémique : chaque pair initie de manière régulière un échange avec un autre pair. Chacun d'eux sélectionne et envoie des correspondances à l'autre.

3.3. *Réduction de l'hétérogénéité liée à la topologie du système*

Le protocole GoOD-TA est un protocole épidémique qui permet d'auto-organiser un système P2P non-structuré en fonction des connaissances sémantiques des pairs (Cerqueus *et al.,* 2011b). Les connaissances sémantiques d'un pair concernent l'ontologie utilisée par ce pair, ainsi que les correspondances qu'il connaît. L'objectif de GoOD-TA est de réduire l'hétérogénéité liée à la topologie 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.

4. Aperçu du système

Nous proposons un système dont l'architecture, générique et modulaire, est organisée en différentes couches logicielles. Dans sa configuration actuelle, il comporte les couches suivantes :

– Une couche basse composée de modules dédiés à la gestion du système P2P. C'est par exemple à ce niveau que se situe le module assurant le *peer sampling service*. Ce dernier est en mesure de proposer un échantillon de pairs du système, qui sert de base aux protocoles épidémiques.

– Une couche intermédiaire, dédiée à la réduction de l'hétérogénéité sémantique. Cette couche comprend deux modules correspondant respectivement aux protocoles CorDis et GoOD-TA. Elle supporte les mécanismes mis en œuvre pour faire décroitre l'hétérogénénité sémantique. Les protocoles de cette couche nécessitent des connaissances sur les ontologies et les correspondances données par les alignements.

– Une couche dédiée à la recherche d'information en environnement hétérogène. La méthode de RI proposée doit prendre en compte le fait que requêtes et documents sont susceptibles d'être représentés par rapport à des ontologies différentes. En cela, elle est chargée d'assurer l'interoperabilité sémantique, ce qui peut être fait de différentes manières. Dans la suite, nous proposons le protocole DiQuESH, qui effectue une recherche top-k, en supposant qu'avant d'évaluer la pertinence des documents, les requêtes sont traduites localement en utilisant les correspondances connues localement.

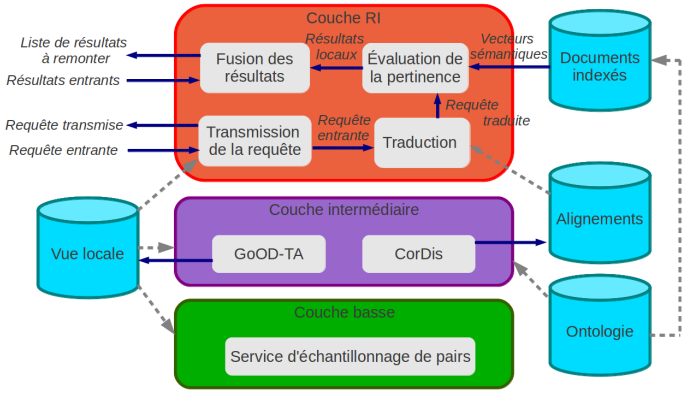

Figure 2. *Architecture d'un pair présentant les trois couches logicielles.*

La figure 2 présente l'architecture, elle-même modulaire, d'un pair quelconque du système. Les trois couches logicielles y sont représentées. Ces couches forment une ossature générique qui peut être complétée par d'autres modules ou couches. Par exemple, on peut ajouter une couche gérant le rapprochement des pairs selon leurs intérêts, ou les documents qu'ils stockent. Dans cette figure, les flèches en pointillé représentent des liens d'utilisation, tandis que celles avec des traits pleins représentent des entrées ou des sorties de modules. Par exemple la flèche pleine vers la base des alignements connus symbolise le fait que le protocole CORDIS permet au pair d'apprendre des correspondances. Dans la couche supérieure, les quatre modules forment le protocole DiQuESH décrit dans la section suivante.

5. Évaluation distribuée de requêtes

L'objectif de cette section est de définir un algorithme pour le traitement distribué des requêtes en environnement sémantiquement hétérogène. Nous présentons l'algorithme DiQuESH[2] avec les hypothèses suivantes.

Étant donnée une requête, chaque pair est capable d'affecter un score de pertinence à ses documents. Il contribue à l'interoperabilité en faisant au mieux pour répondre à la requête, même si l'ontologie utilisée pour représenter celle-ci diffère de la sienne. Les scores calculés par les différents pairs sont comparables et ont des valeurs comprises entre 0 et 1. De plus, nous supposons que les pairs ne sont pas malicieux : ils ne trichent pas sur les scores locaux de leurs documents. L'initiateur de la requête désire obtenir un ensemble de k meilleurs documents, pour un voisinage donné, c'est-à-dire qu'il n'existe pas hors de l'ensemble de résultats, un document qui aurait reçu un score plus élevé, en considérant le même voisinage.

Nous nous ramenons donc à une problématique d'évaluation distribuée de requêtes top-k. Nous utilisons le routage des requêtes et la remontée des résultats proposés dans l'algorithme générique Fully Distributed (Akbarinia *et al.,* 2006) où quatre phases se distinguent : transmission de la requête, traitement local, fusion et remontée des résultats, obtention des documents. Pour l'adapter au cas d'une recherche d'information en présence d'hétérogénéité sémantique nous spécifions le traitement local d'une requête à la fois pour la traduction et le calcul de la pertinence. Notons qu'avec cet algorithme, l'initiateur de la requête reçoit dans un premier temps une liste des k meilleurs résultats, un résultat étant un triplet (id_p, id_d, sc_d) où id_p est l'identifiant du pair possèdant le document, id_d est l'identifiant du document d chez le pair p et sc_d est le score de d obtenu chez p. Cette solution ne permet pas de déterminer en cours d'exécution si le même document est retourné plusieurs fois par différents pairs.

5.1. *Transmission de la requête*

Un pair p initiant une recherche envoie une requête à ses voisins directs ($\mathcal{N}_p$). Le message comprend le vecteur sémantique caractérisant la recherche $\overrightarrow{q_o}$, exprimé par rapport à l'ontologie o qu'utilise p, l'URI de o, le nombre k de documents souhaités, et la valeur de TTL (Time-

2. DiQuESH : **Di**stributed **Qu**ery **E**valuation in **S**emantically **H**eterogeneous context.

To-Live). Quand un pair p' reçoit une requête $\overrightarrow{q_o}$, il commence par la transmettre telle quelle à ses propres voisins en décrémentant seulement la valeur du TTL ; puis il la traite localement.

5.2. *Traitement local de la requête*

Cette phase, qui permet d'extraire la liste des k meilleurs résultats locaux, se décompose en deux étapes, traduction de la requête et évaluation de la pertinence. La première est omise si l'ontologie de la requête est la même que celle du pair.

Traduction de la requête

Pour traduire une requête $\overrightarrow{q_o}$, un pair p' utilisant l'ontologie o' exploite les correspondances qu'il connaît entre o et o'. Le résultat de la traduction est un vecteur sémantique $\overrightarrow{q_{o'}}$ exprimé dans l'espace vectoriel défini par o'. Ce processus est décrit dans l'algorithme 1.

Algorithme 1 : Traduction d'un vecteur sémantique $\overrightarrow{v_o}$ par le pair p'.

Input : un vecteur $\overrightarrow{v_o}$ exprimé dans l'espace de l'ontologie o.
Output : un vecteur $\overrightarrow{v_{o'}}$ exprimé dans l'espace de l'ontologie o'.

```
1 v_o' ← ∅ ;                          // Tous les poids de v_o' sont intialisés à 0
2 for c ∈ v_o do
3     if ∃c' ∈ C_o' : ⟨c, c', ≡, 1⟩ ∈ κ_p(o, o') then
4         v_o'[c'] ← v_o[c] ;          // v_o[c] désigne le poids de c dans v_o
```

Cet algorithme considère seulement les équivalences dont la valeur de confiance m est égale à 1 (algorithme 1, ligne 3). Nous pourrions aussi prendre en compte celles pour lesquelles m dépasse un certain seuil s, en modulant le poids de c' en fonction de m. Nous pourrions également considérer d'autres types de correspondances (subsomption, etc.) pour pondérer davantage de concepts dans l'espace de o', avec les limites connues d'un processus d'expansion.

Évaluation locale de pertinence

Traduire le vecteur de la requête dans l'ontologie du pair récepteur p' a pour effet de le projeter sur les seuls concepts que ce dernier parvient à traduire, c'est-à-dire qui figurent aussi dans l'ontologie o'. La pertinence des documents pourrait être calculée par rapport à cette projection, en oubliant totalement la requête initiale. Le risque est alors d'affecter à un document une valeur de pertinence inadaptée. Par exemple, supposons qu'un document soit représenté par un seul concept c et que la requête initiale le soit sur trois concepts (dont c). Supposons que la requête traduite ne contienne que c. Après cette traduction, si l'on considère directement le cosinus, la pertinence du document sera évaluée à 1. Or, si les trois concepts avaient été traduit, la valeur de pertinence aurait été plus faible. Ainsi, il semble que la pertinence calculée par rapport à une traduction approximative doive être pénalisée. Cela est d'autant plus important que l'algorithme compare ensuite les scores provenant de différents pairs. Ceux qui traduisent bien seraient alors pénalisés à tort.

Pour résoudre ce problème, nous proposons de pondérer le score obtenu par le document vis-à-vis de la requête traduite, en prenant en compte la déviation de cette dernière par rapport à la requête initiale. C'est un calcul qui peut être effectué par le pair récepteur, même s'il n'a pas tous les concepts de la requête dans son ontologie. La déviation correspond à l'erreur introduite lors de la traduction (incomplète) de la requête. Le score d'un document $\overrightarrow{d_{o'}}$ par rapport à la requête initiale $\overrightarrow{q_o}$ est donné par :

$$score(\overrightarrow{d_{o'}}, \overrightarrow{q_o}) = \cos(\overrightarrow{d_{o'}}, \overrightarrow{q_{o'}}) \times \cos(\overrightarrow{q_o}, \overrightarrow{\tilde{q}_o})$$

où $\overrightarrow{\tilde{q}_o}$ est le vecteur correspondant à la requête initiale $\overrightarrow{q_o}$ limité aux concepts qui ont été traduits et pris en compte dans $\overrightarrow{q_{o'}}$. Si d et q sont représentés dans le même espace o, alors $\overrightarrow{q_o} = \overrightarrow{\tilde{q}_o}$. Dans ce cas $\cos(\overrightarrow{q_o}, \overrightarrow{\tilde{q}_o}) = 1$, et donc le score de d par rapport à q est simplement $score(\overrightarrow{d_{o'}}, \overrightarrow{q_o}) = \cos(\overrightarrow{d_{o'}}, \overrightarrow{q_{o'}})$. Le premier avantage de cette méthode est que les scores des documents sont comparables car ils prennent en compte la déviation entre la requête initiale et la requête réellement traitée. Le second avantage est que la déviation est calculée de la même manière que la pertinence entre deux vecteurs sémantiques. Cette méthode est donc générique et peut être appliquée à d'autres mesures que celle du cosinus.

5.3. *Fusion et remontée des résultats - Obtention des documents*

Lors qu'un pair p a fini de traiter la requête localement, il attend les listes de résultats venant de ses voisins. Il effectue un tri-fusion de sa propre liste de résultats avec les listes de résultats qu'il a reçues. Il sélectionne alors les k résultats les plus pertinents et retourne cette liste au pair qui lui avait transmis la requête.

La phase d'obtention effective des documents par le pair initiateur de la requête est simple. Elle consiste à contacter les pairs possédant les documents les plus pertinents en spécifiant les identifiants des différents documents, les adresses des pairs et les identifiants de documents étant indiqués dans les triplets résulats remontés.

6. Expérimentations

L'objectif des expérimentations est de montrer quel est l'impact de l'hétérogénéité sémantique sur les résultats obtenus par une méthode de recherche d'information. C'est également l'occasion de montrer les bénéfices d'algorithmes de diminution de l'hétérogénéité tels que CORDIS et GoOD-TA. Le matériel à mettre en œuvre pour faire de telles expérimentations est considérable. Sans parler de la distribution, il faut non seulement disposer, comme dans les expérimentations de RI classique, d'un corpus de test (documents, requêtes et jugements de pertinence), mais aussi d'un nombre suffisant d'ontologies, par rapport auxquelles les documents doivent être indexé. Ne connaissant aucun corpus de ce type, nous avons utilisé les ontologies et les services de BioPortal.

6.1. *Matériel*

Simulation de systèmes P2P

Pour ne pas avoir à développer et à deployer un véritable système P2P, nous avons utilisé le simulateur PeerSim (Montresor *et al.,* 2009). Ce simulateur est très utilisé dans la communauté P2P pour la validation. PeerSim permet d'instancier des systèmes P2P non-structurés servant de base à notre système.

Ontologies et alignements de BioPortal

Afin de simuler une situation réaliste, nous considérons des ontologies activement utilisées dans la communauté biomédical. Nous les avons obtenu au travers des services web BioPortal (Fridman Noy *et al.,* 2009). BioPortal est un dépôt ouvert de ressources biomédicales. En particulier il propose un accès à de nombreuses ontologies. Nous avons utilisé les servives BioPortal pour télécharger ces ontologies. Nous avons transformé toutes les ontologies au format OWL pour éviter les problèmes liés à l'hétérogénéité syntaxique (différence de format). Nous avons également téléchargé des alignements entre ces ontologies (toujours au travers de services web BioPortal). Ces alignements lient certaines des ontologies que nous avons téléchargées, alors que d'autres restent totalement déconnectées.

Documents

Nous avons extrait les annotations sémantiques de 4163 documents (appelés *ressources* sur BioPortal) à l'aide des services web BioPortal. Les documents correspondent à des articles scientifiques publiés dans des journaux du domaine biomédical. Ils sont issus de la base de données PubMed. Certains documents sont annotés avec des concepts de différentes ontologies. Les annotations sont utilisées pour créer les vecteurs sémantiques correspondant aux documents. Seules 39 ontologies sont utilisées pour représenter les documents.

Distribution des ontologies et des documents

Nous avons distribué aléatoirement dans le système P2P les 39 ontologies utilisées pour représenter les documents : chaque pair utilise une ontologie, et chaque ontologie est utilisée par le même nombre de pairs. Les documents sont également distribués aléatoirement : chaque pair stocke entre 50 et 100 documents indexés suivant l'ontologie qu'il utilise. Certains documents peuvent être répliqués dans le système. Des stratégies plus avancées ont été proposées pour placer les documents dans le système. Par exemple, Holz *et al.* (Holz *et al.,* 2007) suggèrent de regrouper les documents en fonction de leurs auteurs : chaque pair est responsable des documents d'un auteur, et les pairs responsables de deux auteurs ayant travailler ensemble sont voisins dans le système. Nous n'avons pas choisi cette stratégie car l'ensemble de documents que nous avons utilisé ne s'y prête pas : il y a presque autant d'auteurs que de documents, et très peu ont collaborer pour produire ces documents.

Requêtes et jugement de pertinence

Comme nous ne disposons pas de requêtes portant sur les documents que nous considérons, nous avons en avons généré un certain nombre. Chaque requête est un vecteur sé-

mantique composé de concepts d'une ontologie o : certains sont partagés entre o et d'autres ontologies, certains sont présents dans un ou plusieurs documents, et les autres sont choisis aléatoirement dans l'ontologie o.

Le jugement de pertinence nécessaire à l'évaluation de la méthode de recherche d'information est construit à partir de ces requêtes et de ces documents. Pour chaque requête nous identifions les k ($k = 10$) meilleurs documents en exécutant la méthode dans un système centralisé ayant accès à tous les documents et à toutes les correspondances entre ontologies : ce sont eux qui forment l'ensemble P des documents pertinents pour la requête. Nous prenons le soin de retirer les requêtes pour lesquelles aucun document pertinent n'est retourné. Finalement, nous disposons de 85 requêtes.

Métriques

Pour mesurer la qualité des résultats fournis par la méthode de RI pour une requête donnée, nous utilisons les mesures classiques de précision (Pr) et de rappel (Ra) :

$$Pr = \frac{|R \cap P|}{|R|} \text{ et } Ra = \frac{|R \cap P|}{|P|}$$

où P est l'ensemble des documents pertinents pour la requête, et R est l'ensemble des documents retournés par la méthode de RI.

Pour mesurer l'hétérogénéité des systèmes P2P, nous utilisons les mesures présentées dans la section 3 (formules [1], [2] et [4]).

6.2. *Paramètres de simulation*

Dans les expérimentations nous considérons des systèmes P2P de 1000 pairs. Dans un premier temps nous supposons qu'ils sont statiques, c'est-à-dire que les pairs présents dans un système ne le quittent pas, et qu'aucun nouveau pair ne le rejoint. Initialement tous les pairs sont directement connectés à quatre autres pairs. Nous considérons des requêtes top-k où $k = 10$: nous cherchons à obtenir les 10 documents les plus pertinents pour chaque requête. Par ailleurs nous fixons la valeur du TTL à 4. Par conséquent, chaque requête, qui est émise par un pair choisi au hasard dans le système, peut atteindre au plus 340 pairs.

Nous considérons quatre situations. La première correspond à un système P2P hétérogène dans lequel seul l'algorithme DiQuESH est mis en œuvre. La deuxième situation correspond au même système que dans la situation précédente, mais dans lequel l'algorithme CorDis est exécuté pendant 100 cycles. Les requêtes sont lancées après la fin de l'exécution de CorDis. La troisième situation est semblable à la deuxième mais l'algorithme exécuté est GoOD-TA. La dernière situation est une combinaison des deux situations précédentes : CorDis et GoOD-TA sont exécutés en parallèle.

L'algorithme CorDis permet de réduire l'hétérogénéité sémantique liée à la disparité entre pairs. La réduction est faible car trop peu de correspondances existent entre les ontologies qu'utilisent les pairs. L'algorithme GoOD-TA permet de réduire l'hétérogénéité sémantique liée à l'organisation du système à 0. Cela est possible car la diversité sémantique est

	$\mathcal{H}_{Div}$	$\mathcal{H}_{Disp}$	$\mathcal{H}_{DapAvg}$
DiQuESH seul	0,04	0,97	0,96
CorDis	0,04	0,96	0,96
GoOD-TA	0,04	0,97	0
CorDis + GoOD-TA	0,04	0,96	0

Tableau 1. *Valeurs d'hétérogénéité sémantique avant exécution des requêtes.*

très faible ($\mathcal{H}_{Div}(\mathcal{S}) = \frac{39-1}{1000-1} = 0,04$). En moyenne chaque ontologie est utilisée par 25 pairs donc ceux-ci se regroupent et l'hétérogénéité centrée sur chaque participant est nulle. Le tableau 1 présente les degrés d'hétérogénéité de chaque situation.

6.3. *Résultats et discussion*

La figure 3 présente les valeurs moyennes de précision et de rappel obtenues dans chacune des situations étudiées pour l'ensemble des requêtes. Cette figure montre que CorDis permet d'augmenter de manière significative la valeur de rappel (de 0,25 à 0,35) tout en améliorant légérement la précision. Cela est dû au fait que les correspondances apprises par les pairs leurs permettent de mieux traduire les requêtes qu'ils reçoivent, et donc retrouver davantage de documents pertinents. Évidemment plus il y a de correspondances entre les ontologies, plus les pairs utilisant des ontologies différentes sont capables d'interopérer. Dans notre cas, les ontologies sont très peu liées. Par conséquent, l'algorithme CorDis réduit peu l'hétérogénéité, et la précision est modestement améliorée.

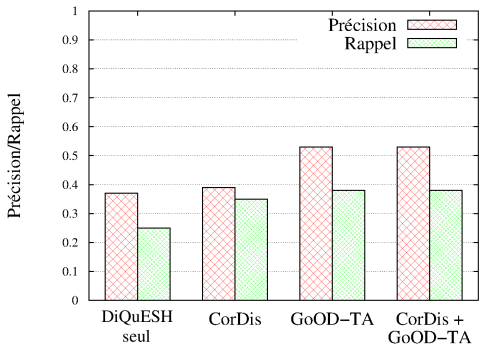

Figure 3. *Valeurs de précision/rappel obtenues dans les différentes situations étudiées.*

Par ailleurs l'algorithme GoOD-TA permet d'augmenter fortement la précision (de 0,25 à 0,39) et le rappel (de 0,38 à 0,53). Cela s'explique par le fait que chaque pair est placé à proximité des pairs susceptibles de comprendre et traiter les requêtes qu'il émet. En effet il est

préférable que les pairs capables de comprendre la requête sans la traduire soient accessibles, c'est-à-dire qu'ils soient dans le voisinage (à TTL pas) du pair qui l'émet. De plus il y a une forte corrélation entre l'ontologie utilisée par un pair, et ses centres d'intérêts (exprimés au travers des requêtes). Les documents susceptibles d'être pertinents par rapport à la requête sont à priori représentés dans un "vocabulaire" proche de celui de la requête. Par exemple, les requêtes émises par un spécialiste en biochimie métabolique seront vraisemblablement satisfaites par des spécialistes du même domaine (ou au moins d'un domaine proches). Il est donc probable qu'ils utilisent la même ontologie, ou des ontologies entre lesquelles il existe de nombreuses correspondances.

Enfin la figure 3 montre que la combinaison des deux algorithmes n'améliore pas la précision et le rappel par rapport à GoOD-TA seul. Cela s'explique par le fait que les pairs se regroupent en communautés dans lesquelles les pairs utilisent tous la même ontologie. Ce phénomène ne se produirait pas si la diversité sémantique (mesurée par $\mathcal{H}_{Div}$) était plus élevée, ou si les ontologies étaient plus proches sémantiquement.

7. Travaux liés

Les travaux présentés relèvent de la recherche d'information distribuée qui permet à des utilisateurs de gérer eux-mêmes leurs documents ou données sur un large réseau de nœuds. À la différence de certains travaux (Loupasakis *et al.,* 2011)(Markov, 2011), le système que nous présentons ne gère pas d'index global, qu'il soit ou non réparti, mais utilise un algorithme top-k. Récemment, des travaux en recherche d'information sémantique ont exploré l'apport des techonologies du web sémantique pour améliorer la recherche d'information (Rousset *et al.,* 2011). Cela nécessite en général d'être capable d'indexer les documents en fonction d'une ontologie (Baziz *et al.,* 2005). Dans (Dharanipragada *et al.,* 2010) les auteurs organisent les pairs en communautés. Chaque communauté maintient une base de connaissance qui contient des méta-données sur les concepts pour améliorer la recherche. Dans le système présenté, ce sont les pairs qui stockent les connaissances qui leur sont utiles. Dès lors que l'on utilise plusieurs ontologies, il est indispendable de savoir calculer des alignements entre elles (Euzenat *et al.,* 2007). Le problème se pose aussi entre schémas de bases de données. Les systèmes P2P où les pairs n'utilisent pas le même schéma ont fait l'objet de plusieurs études comme (Halevy *et al.,* 2003). Les correspondances sont utilisées pour traduire les requêtes SQL. Nous utilisons la même idée pour traduire les vecteurs sémantiques, mais en pénalisant les traductions approximatives et en maintenant une couche intermédiaire pour diminuer l'hétérogénéité sémantique. Toujours dans le domaine des bases de données, la notion d'interoparabilité est définie dans (Cudré-Mauroux *et al.,* 2004) : "deux pairs sont dits sémantiquement interopérables s'ils peuvent se transmettre des requêtes l'un l'autre, en suivant potentiellement plusieurs liens de traduction sémantique". Nous pensons que c'est une vue très particulière. Nous proposons de revenir au sens général d'interoperabilité comme la capacité à communiquer et à échanger des informations. À ce sens, elle doit être mesurée avec les métriques relatives à l'application choisie, comme la précision et le rappel dans notre cas. Si la notion d'hétérogénéité sémantique est souvent évoquée (Halevy, 2005)(Rousset *et al.,* 2011), il n'existe pas à notre connaissance de métriques associées, ni de travaux qui éta-

blissent un parallèle direct entre les deux notions. Dans (Cudré-Mauroux *et al.,* 2008), une "entropie des méta-données" (*metadata entropy*) est définie mathématiquement pour caractériser la rareté des méta-données. Elle reflète à la fois l'incomplétude et l'incertitude des données d'une seule source. Cette piste est peut-être exploitable pour une caractérisation distribuée. D'autres systèmes P2P sémantiques ont été proposés mais aucun ne semble considérer la notion d'hétérogénéité sémantique en tant que telle (Mena *et al.,* 2000) (Ng *et al.,* 2003). Notre travail se démarque de ceux-ci car nous focalisons sur la source du problème d'interopérabilité : l'hétérogénéité sémantique.

8. Conclusion

Cet article suppose que les pairs du système utilisent des ontologies différentes pour représenter leurs documents. L'approche décrite distingue hétérogénéité sémantique et interoperabilité. Elles ne se mesurent pas de la même façon, même si l'une influence l'autre. Par exemple, s'agissant d'un système de RI, nous avons utilisé précision et rappel pour mesurer l'interoperabilité. Nous avons proposé un système comportant une couche dédiée à la réduction de l'hétérogénéité (protocoles CorDis et GoOD-TA), tandis que l'interoperabilité est assurée par le protocole DiQuESH que nous avons détaillé. Le calcul de la pertinence après traduction est un point clé. Nous avons mené des expérimentations pour lesquelles beaucoup de matériel était nécessaire : documents indexés, ontologies, alignements, jugements de pertinence... Il n'est pas facile de trouver un ensemble cohérent de tous ces éléments. Si l'utilisation de documents bio-médicaux et des ontologies BioPortal nous a facilité certaines tâches, elle nous a aussi limité dans nos expérimentations. Par exemple, le fait que seules 39 ontologies soient de fait utilisées pour indexer les documents ne permet pas de tester sur un grand nombre de pairs, la diversité sémantique devenant alors trop faible. De même, les caractéristiques des alignements entre ces ontologies font que les amélioratinos dûes au protocole CorDis sont faibles, même si elles restent parfaitement visibles. Par contre, l'apport de GoOD-TA en termes de diminution de l'hétérogénéité, et d'amélioration de l'interoperabilité est important. Ce type d'expérimentations nous permet aussi d'étudier la sensibilité d'un protocole assurant l'interoperabilité (DiQuESH) vis-à-vis de l'hétérogénéité : sans l'aide des algorithmes de la couche intermédiaire, le taux de précision/rappel est relativement faible.

Avec pour objectif d'améliorer l'interopérabilité sémantique pour la recherche d'information, il est possible de définir d'autres algorithmes visant à décroitre l'hétérogénéité ou à assurer directement une meilleure interoperabilité. Nous pensons que ces premiers résultats permettent d'envisager des études et comparaisons plus riches, permettant au final de construire des systèmes plus performants.

9. Bibliographie

Akbarinia R., Pacitti E., Valduriez P., « Reducing network traffic in unstructured P2P systems using Top-*k* queries », *Distributed and Parallel Databases*, vol. 19, n° 2-3, p. 67-86, 2006.

Baziz M., Boughanem M., Traboulsi S., « A concept-based approach for indexing documents in IR », *13ème congrès INFORSID*, p. 489-504, 2005.

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., « Semantic heterogeneity measures of unstructured P2P systems », *10th IEEE/WIC/ACM International Conference on Web intelligence*, p. 223-226, 2011c.

Cudré-Mauroux P., Aberer K., « A Necessary Condition for Semantic Interoperability in the Large », *3rd International Conference on Ontologies, Databases and Applications of Semantics*, p. 859-872, 2004.

Cudré-Mauroux P., Budura A., Hauswirth M., Aberer K., « PicShark : mitigating metadata scarcity through large-scale P2P collaboration », *The VLDB Journal*, vol. 17, n° 6, p. 1371-1384, 2008.

Dharanipragada J., Fausto G., Harisankar H., Uladzimir K., Two-layered architecture for peer-to-peer concept search, Technical Report n° DISI-10-02, University of Trento, 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, n° Web-Server-Issue, p. 170-173, 2009.

Halevy A., Ives Z., Mork P., Tatarinov I., « Piazza : data management infrastructure for semantic web applications », *12th International World Wide Web Conference*, p. 556-567, 2003.

Halevy A. Y., « Why your data won't mix », *ACM Queue*, vol. 3, n° 8, p. 50-58, 2005.

Holz F., Witschel H., Heinrich G., Heyer G., Teresniak S., « An Evaluation Framework for Semantic Search in P2P Networks », *Proceedings of the I2CS 2007*, 2007.

Loupasakis A., Ntarmos N., Triantafillou P., « eXO : Decentralized Autonomous Scalable Social Networking », *5th Biennial Conference on Innovative Data Systems Research*, p. 85-95, 2011.

Markov I., « Modeling document scores for distributed information retrieval », *34th international ACM SIGIR conference on Research and development in Information*, p. 1321-1322, 2011.

McGuinness D. L., van Harmelen F., OWL Web Ontology Language Overview, W3C recommendation, World Wide Web Consortium, 2004.

Mena E., Illarramendi A., Kashyap V., Sheth A. P., « OBSERVER : An Approach for Query Processing in Global Information Systems Based on Interoperation Across Pre-Existing Ontologies », *Distributed and Parallel Databases*, vol. 8, n° 2, p. 223-271, 2000.

Montresor A., Jelasity M., « PeerSim : A Scalable P2P Simulator. », *9th IEEE International Conference on Peer-to-Peer Computing*, p. 99-100, 2009. `http://peersim.sf.net`.

Ng W. S., Ooi B. C., Tan K.-L., Zhou A., « PeerDB : A P2P-based System for Distributed Data Sharing », *19th International Conference on Data Engineering*, p. 633-644, 2003.

Rousset C., Pinet F., Kang M. A., Corcho O., *Ontology for Interoperability*, vol. 1, Springer, chapter Ontology Fundamentals, p. 39-54, 2011.

Salton G., Wong A., Yang C. S., « A vector space model for automatic indexing », *Commun. ACM*, vol. 18, p. 613-620, 1975.

Ventresque A., Espaces vectoriels sémantiques : enrichissement et interprétation de requêtes dans un système d'information distribué et hétérogène, PhD thesis, Université de Nantes, 2008.