Organisation et routage sémantiques dans les systèmes pair-à-pair

Bruno Defude

GET-INT, CNRS UMR SAMOVAR

Département informatique
9 Rue Charles Fourier
91011 Evry CEDEX
Bruno.Defude@int-edu.eu

RÉSUMÉ. *Le pair-à-pair (P2P) s'est imposé ces dernières années comme la technologie majeure d'accès à des ressources multimédia sur l'Internet. On peut classifier ces systèmes selon leur modèle de recherche sous-jacent qui peut être soit non structuré (propagation aléatoire des requêtes dans le graphe des pairs), soit structuré (propagation des requêtes selon une structure d'organisation des pairs). L'approche non structurée bien que moins efficace sur le plan du routage des requêtes offre l'avantage de respecter au mieux l'autonomie des pairs et de pouvoir supporter des langages de requêtes plus expressifs. L'efficacité de la recherche dans les systèmes non structurés peut être améliorée en introduisant de la sémantique dans le processus de routage des requêtes. Cette sémantique est généralement construite à partir du contenu des pairs mais peut également faire intervenir leur comportement passé (historique des requêtes). Nous présentons dans cet exposé un panorama des principaux algorithmes d'organisation et de routage sémantiques. Nous montrons notamment les travaux que nous menons dans le cadre des projets Rare et Respire et qui visent à combiner l'information sur le contenu des pairs et sur leur intérêt.*

ABSTRACT. *Peer-to-peer systems (P2P) are becoming a leading technology for multimedia resource access on the Internet. These systems can be classified depending on their underlying search model, which can be unstructured (queries are flooding in the peer's network) or structured (queries are propagated according to a logical structure defined on peers). The unstructured model is less efficient but favours peer's autonomy and supports more expressive query languages. Search efficiency in unstructured systems may be improved using some semantic inside the query process. This semantic is generally defined on peer's content but may also used peer's behaviour (query history). In this talk, we present a general overview of algorithms for semantic routing and organisation. We will give a focus on the work we have done in projects Rare and Respire combining semantic about peer's content and behaviour.*

MOTS-CLÉS : *pair-à-pair, routage sémantique, recherche d'informations*

KEYWORDS: *peer to peer, semantic routing, information retrieval*

1. Introduction

Le pair-à-pair (P2P) s'est imposé ces dernières années comme la technologie majeure d'accès à des ressources multimédia sur l'Internet. D'après de récentes estimations, 60% du trafic de l'Internet viendrait de services P2P comme Kazaa, bittorrent.

On peut classifier ces systèmes selon leur modèle de recherche sous-jacent qui peut être soit non structuré (propagation aléatoire des requêtes dans le graphe des pairs), soit structuré (propagation des requêtes selon une structure d'organisation des pairs basée en général sur du hachage). L'approche non structurée bien que moins efficace sur le plan du routage des requêtes offre l'avantage de respecter au mieux l'autonomie des pairs et de pouvoir supporter des langages de requêtes plus expressifs.

De nombreux projets de recherche essayent d'améliorer l'efficacité de la recherche dans les systèmes non structurés en introduisant de la sémantique dans le processus de routage des requêtes. Cette sémantique est généralement construite à partir du contenu des pairs mais elle peut également faire intervenir le comportement passé des pairs (historique des requêtes).

Le reste d l'article est organisé comme suit. Nous décrivons succinctement dans la section 2 les grandes classes de systèmes pair-à-pair. Nous nous focalisons ensuite dans la section 3 sur les systèmes non structurés et nous montrons quelles sont les différentes dimensions à considérer pour prendre en compte de la sémantique dans ces systèmes. La section 4 décrit les travaux que nous menons dans le cadre des projets Rare (Rare 2007) et Respire (Respire 2007) et qui visent à combiner l'information sur le contenu des pairs et sur leur intérêt. Enfin, nous donnons quelques éléments d'ouverture dans la conclusion.

2. Les grandes classes de systèmes pair-à-pair

La première classe de systèmes pair-à-pair est le non structuré. Le principe de résolution de requêtes y est le suivant : un client soumet sa requête à un serveur quelconque du réseau (le plus proche de lui par exemple), celui-ci résout la requête localement et la propage récursivement à un certain nombre de ses voisins (aléatoirement choisis). La recherche s'arrête quand une certaine profondeur de recherche a été atteinte. Cette approche est bien sûr très inefficace et conduit à générer un grand nombre de messages. Elle fonctionne assez bien lorsque les ressources recherchées sont fréquemment les mêmes.

Dans la classe hiérarchique (Yang et al. 2003), le réseau est hiérarchisé en distinguant deux catégories de nœuds, les nœuds « puissants » (en terme de cpu, de bande passante) et les nœuds « faibles ». Les nœuds faibles sont regroupés et sont

associés à un nœud puissant. Seuls ces derniers fonctionnent en mode P2P, alors qu'à l'intérieur d'un groupe il s'agit d'un mode client-serveur classique.

Les systèmes structurés améliorent la fonction de recherche en organisant l'espace de stockage selon une structure d'ordre et en utilisant cette structure pour placer les ressources sur les nœuds (on parle également de Distributed Hash Table ou DHT). Plusieurs algorithmes comme Chord (Stoica et al. 2001), CAN (Ratsanamy et al. 2001), P-Grid (Aberer 2001), Pastry (Rowstron et al. 2001) ont été proposés qui diffèrent entre eux selon la structure d'ordre choisi (anneau, espace multi-dimensionnel, arbre binaire, …). Ces solutions amènent l'efficacité mais au prix de la perte d'autonomie des pairs puisque le placement des ressources est décidé par le système et non les utilisateurs.

3. Dimensions du routage et l'organisation sémantique des pairs

De nombreux travaux ces dernières années ont visés à améliorer la fonction de recherche dans les systèmes non structurés. L'idée de base est de remplacer le routage aléatoire par un routage guidé par la sémantique. Pour ce faire, plusieurs dimensions du problème sont à analyser :

- quelle sémantique : il s'agit de définir le type d'information à utiliser dans le processus de routage. Cela peut être l'information sur le contenu des pairs (les données/documents stockés), sur l'intérêt du pair (les requêtes déjà émises), sur les utilisateurs (profil d'un utilisateur ou de communautés d'utilisateurs) ;

- quelle représentation de la sémantique : cela va d'une simple information temporelle (les pairs ayant répondus récemment) à des modèles non structurés (simple liste plate de concepts) à des modèles structurés (ontologies) ;

- comment construire la sémantique : le processus de construction peut être manuel (par intervention de l'utilisateur) ou bien automatique (algorithmes d'apprentissage). On peut également trouver des approches mixtes où l'utilisateur intervient via des formes de feedbacks ;

- qu'est ce qui est partagé entre les pairs : a minima les pairs doivent partager des structures de données communes (représentation de la sémantique par exemple), mais cela peut aller au partage de fonctions (algorithmes d'apprentissage par exemple), voire au partage de connaissances (ontologie commune partagée) ;

- comment utiliser la sémantique : généralement la sémantique va être utilisée pour sélectionner le sous ensemble des pairs les plus « pertinents » pour une requête donnée. Cela peut aussi servir à organiser le réseau des pairs (classification des pairs selon leur contenu par exemple) ou bien à modifier les requêtes ;

- comment diffuser la sémantique : la connaissance construite localement sur un pair doit être diffusée aux autres pairs pour qu'ils puissent améliorer leur connaissance du réseau. Cette diffusion peut être globale (à tous les pairs) ou partielle (à quelques uns) et la fraîcheur est également importante (diffusion lors de

chaque modification ou bien périodique). Le coût de la diffusion en nombre de messages est par ailleurs crucial.

Les algorithmes proposés dans la littérature peuvent être analysés selon ces dimensions. Les systèmes de gestion de données pair-à-pair comme Piazza (Halevy et al. 2004), SomeWhere (Rousset et al. 2006), (Nejdl et al. 2003) travaillent sur des données structurées par un schéma (relationnel ou XML). Il s'agit de systèmes de médiation sans schéma global ou chaque pair dispose de son propre schéma local et de schéma de correspondance vers d'autres pairs. Il n'y a pas ici à proprement parler de processus de routage de requêtes, puisque les pairs vers lesquels propager une requête sont définis par les schémas de correspondance. Par contre, il faut ici de nouveaux algorithmes efficaces de réécriture et d'optimisation. (Ooi et al. 2003) proposent de sélectionner les pairs qui vont résoudre une requête relationnelle en faisant une comparaison entre le schéma de celle-ci et le schéma de la source locale.

Dans le cadre des systèmes travaillant sur de l'information faiblement structurée, plusieurs travaux ont proposés d'améliorer le routage en intégrant de la sémantique sur le contenu des pairs. (Voulgaris et al. 2004) propose d'utiliser la connaissance sur les pairs ayant récemment retournés des résultats. PlanetP (Cuenca-Acuna et al. 2003) construit une représentation du contenu des pairs par une technique de signature (filtres de Bloom) alors que (Crespo et al. 2002) supposent un vocabulaire commun d'indexation. Cette information est ensuite utilisée pour classer les pairs connus via une mesure de pertinence avec la requête, celle-ci étant propagée vers les pairs les plus pertinents. (Loser et al. 2007) étendent ces idées en sélectionnant dans la table de routage d'un pair, les pairs « experts » qui stockent des documents pertinents relativement à un thème et les pairs de « recommandation » qui connaissent des pairs « experts » d'un thème.

(Lumineau 2005) se situe entre les systèmes pour données structurées et non structurées. Il propose d'organiser le réseau de pairs en les classifiant à partir de leur schéma qui est une ontologie de concepts. Le routage se fait ensuite en sélectionnant les classes de pairs les plus proches de la requête. (Khambatti et al 2006) reprennent la même idée, mais utilisent l'intérêt des pairs pour construire les communautés.

(Nakauchi et al. 2001) propose un mécanisme d'expansion de requêtes basés sur des bases de connaissances locales construites à partir du corpus des documents stockés sur le pair et augmentées via des échanges entre pairs ou par feedback avec l'utilisateur.

Plusieurs travaux mixent approche P2P structurée et non structurée (Tang et al. 2003, Li et al. 2004, Lee et al. 2006). Les pairs stockent les documents en gardant leur autonomie de stockage mais par contre l'indexation qui est faite de ces documents est stockée dans une table de hachage distribuée, ce qui permet un accès efficace à l'index.

Des travaux récents (Michel et al. 2006) s'intéressent à l'amélioration de la représentation du contenu des pairs en enrichissant les listes de mots clés par l'analyse des co-occurrences de termes.

4. Correspondance entre contenu et intérêt

Nous avons repris l'approche de PlanetP (Chicky et al. 2006) en l'étendant avec une approche mêlant connaissances sur le contenu des pairs et sur leur intérêt. L'idée est de ne garder dans la table de routage d'un pair que la connaissance sur les pairs dont le contenu correspond à son intérêt ou bien dont l'intérêt est complémentaire du sien. Diverses approches peuvent être suivies pour représenter contenu et intérêt mais nous avons repris les filtres de Bloom comme index de contenu et un mécanisme analogue pour l'index d'intérêt. Celui-ci est calculé à partir des requêtes passées en introduisant une fonction d'oubli pour être capable de prendre en compte des changements d'intérêt. Le routage des requêtes se fait par sélection des pairs les plus proches (selon le contenu et/ou l'intérêt). La publication des index se fait soit par diffusion épidémique en permettant à un pair de ne garder que les index les plus intéressants pour lui, soit en « routant » l'index vers les pairs intéressés de manière analogue à une requête. La deuxième approche est beaucoup plus économe en messages mais ne garantit pas que tous les pairs intéressés reçoivent l'index.

5. Conclusion

De nombreux travaux proposent d'exploiter la sémantique sur le contenu des pairs pour organiser le réseau et/ou router plus efficacement les requêtes. Peu de travaux utilisent l'information sur les requêtes passées. Les points clés dans la définition de tels systèmes sont la qualité de l'indexation, la limitation de la connaissance à partager entre les pairs et surtout la minimisation du coût des mécanismes de maintenance des index. L'utilisation d'algorithmes d'apprentissage automatique semble prometteuse. L'évaluation des algorithmes proposés est par ailleurs une tâche difficile. Il n'y a pas encore de benchmark reconnu pour de tels systèmes qui réclament à la fois volumétrie des données, distribution des données et des requêtes sur les pairs et une certaine corrélation sémantique tant sur le contenu des pairs que sur les requêtes émises. Nous avons commencé à travailler à la définition d'un tel benchmark en utilisant les données de wikipedia et de dmoz.

Ce travail est partiellement financé par le projet ANR ARA MDSA Respire et le projet GET Rare.

6. Bibliographie

Aberer. K. P-Grid: A Self-Organizing Access Structure for P2P Information Systems, *Proceedings COOPIS Conference*, 2001

Chiky R., Hébrail G., Defude B. Définition et diffusion de signatures sémantiques dans les systèmes P2P, *Actes des journées Extraction et Gestion des Connaissances*, Lille, janvier 2006

Crespo A., Garcia-Molina H. Routing Indices for Peer-to-Peer Systems, *Proceedings of the 22nd International Conference on Distributed Computing Systems (ICDCS)*, 2002

Cuenca-Acuna F.M et al. Planetp: Using Gossiping to Build Content Addressable Peer-to-Peer Information Sharing Communities, *Proceedings IEEE International Symposium on High Performance Distributed Computing*, 2003

Halevy A. et al. The Piazza peer data management system. *IEEE TDKE 16(7)*, 2004.

Khambatti M., Dong Ryu K., Dasgupta P. Structuring Peer-to-Peer Networks using Interest-Based Communities, *Proceedings. P2PDBIS Workshop*, sept 2003

Lee D. L, Zhao D.J, Luo Q. Information Retrieval in a Peer-to-Peer Environment, *Proceedings of the International Workshop on Peer-to-Peer Information Management*, May 2006

Li M., Lee W.C. , Sivasubramaniam A. Semantic small world: An overlay network for peer-to-peer search, *Proceedings IEEE ICNP Conference*, 2004

Loser A., Staab S., Tempich C., Semantic Social Overlay Networks *IEEE Journal on Selected Areas in Communications, 25(1)*, 2007

Lumineau N. Organisation et localisation de données hétérogènes et réparties à travers un réseau pair-à-pair, Thèse de doctorat, Université Paris 6, décembre 2005

Michel S. et al. Discovering and Exploiting Keyword and Attribute-Value Co-occurrences to Improve P2P Routing Indices *Proceedings ACM CIKM Conference*, 2006

Nakauchi K. et al., Exploiting semantics in unstructured P2P systems, *IEE Trans. on Comm.*, july 2004

Nejdl W., Siberski W., Sintek M.. Design issues and challenges for RDF- and schema-based peer-to-peer systems. *ACM SIGMOD Record, 32(3)*, 2003.

Ooi B.C, Shu Y., Tan K-L.. Relational Data Sharing in Peer-based Data Management Systems, *ACM SIGMOD Record, 32(3)*, 2003

Rare. Routage par Apprentissage de Requêtes, http://www-inf.int-evry.fr/~defude/RARE, accédé en avril 2007

Ratsanamy S. et al. A Scalable Content Addressable Network, *Proceedings ACM SIGCOMM Conference*, 2001

Respire. Respire : Peer-to-Peer resources and services, querying and replication, http://respire.lip6.fr, accédé en avril 2007

Rousset M.C et al. SomeWhere in the Semantic Web, *Proceedings SOFSEM 2006 Conference*, Prague, janvier 2006

Rowstron A., Dreschel P. Pastry: Scalable, Distributed Object Location and Routing for Large-Scale Peer-to-Peer Systems, *Proceedings IFIP/ACM Conf. On Distributed Systems Platforms*, 2001

Stoica I. et al. Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications, *Proceedings ACM SIGCOMM Conference*, 2001

Tang C., Xu Z., Dwarkadas S., Peer-to-Peer Information Retrieval Using Self-Organizing Semantic Overlay Networks, *Proceedings. ACM SIGCOMM Conference*, 2003

Voulgaris S., Kermarrec A.M., Massoulié M., van Steen M. Exploiting semantic proximity in peer-to-peer content searching. *Proceedings 10th International Workshop on Future Trends in Distributed Computing Systems (FTDCS 2004)*, China, May 2004.

Yang, B. Garcia-Molina H. Designing a Super-Peer Network, *Proceedings ICDE Conference*, 2003