Fouille de données dans les systèmes Pair-à-Pair pour améliorer la recherche de ressources

Florent Masseglia*, Pascal Poncelet**, Maguelonne Teisseire***

*INRIA Sophia Antipolis, Axis Project-Team, BP93 06802 Sophia Antipolis - France
Florent.Masseglia@sophia.inria.fr
**EMA-LGI2P/Site EERIE, Parc Scientifique Georges Besse, 30035 Nîmes Cedex, France
{Pascal.Poncelet}@ema.fr
***LIRMM UMR CNRS 5506, 161 Rue Ada, 34392 Montpellier cedex 5 - France
{teisseire}@lirmm.fr

Résumé. La quantité de sources d'information disponible sur Internet fait des systèmes d'échanges pair-à-pair (P2P) un genre nouveau d'architecture qui offre à une large communauté des applications pour partager des fichiers, des calculs, dialoguer ou communiquer en temps réel. Dans cet article, nous proposons une nouvelle approche pour améliorer la localisation d'une ressource sur un réseau P2P non structuré. En utilisant une nouvelle heuristique, nous proposons d'extraire des motifs qui apparaissent dans un grand nombre de noeuds du réseau. Cette connaissance est très utile pour proposer aux utilisateurs des fichiers souvent demandés (en requête ou en téléchargement) et éviter une trop grande consommation de la bande passante.

1 Introduction

La quantité de sources d'information disponible sur Internet fait des systèmes d'échanges pair-à-pair (P2P) un genre nouveau d'architecture qui offre à une large communauté des applications pour partager des fichiers, partager des calculs, dialoguer ou communiquer en temps réel, etc (Miller (2001), Ngan et al. (2003)). Les applications P2P fournissent également une bonne infrastructure pour les opérations sur de grandes masses de données ou avec de très nombreux calculs, comme la fouille de données. Dans ce cadre, nous considérons une nouvelle approche pour améliorer la localisation de ressources dans un environnement P2P non structuré selon deux aspects principaux pour extraire des comportements fréquents :

1. *L'ordre des séquences* entre les actions réalisées sur les nœuds (requête ou téléchargement) est pris en compte pour améliorer les résultats.
2. Les résultats des calculs distribués sont maintenus via un *"Pair centralisé"* pour réduire le nombre de communications entre les pairs connectés.

Connaître l'ordre des séquences des actions réalisées sur les pairs offre une connaissance importante. Par exemple, en examinant les actions réalisées, nous pouvons savoir que pour 77% des nœuds pour lesquels il y a une requête concernant "*Mandriva Linux*", le fichier "*Mandriva Linux 2005 CD1 i585-Limited-Edition-Mini.iso*" est choisi et téléchargé. Cette requête