Sujet de master recherche « Architectures logicielles distribuées » 2005–2006

Système pair à pair (P2P) pour la classification de documents multimédias

Encadrant principal : Julien COHEN
courriel : `Prenom.Nom@polytech.univ-nantes.fr`
tél. : 02 40 68 30 99

Co-encadrant(s) : Marc Gelgon

Ce stage consiste à concevoir un système pair à pair permettant de classifier de manière distribuée de très gros volumes de documents multimédias (plusieurs centaines de milliers d'images par exemple), en s'appuyant sur des techniques de classification déjà existantes.

Cadre du stage

Les documents multimédias (images, textes, films, son) peuvent être classés en catégories de manière à permettre une navigation intuitive entre les documents et à accélérer les temps de réponse à des requêtes sur ces documents (par exemple, afin de détecter en un temps acceptable le plagiat d'une image parmi un très grand nombre d'images).

Deux familles de critères peuvent être utilisées pour classer des documents : des informations extra-document (auteur, nom d'un lieu, description du contenu par mots clés, date) et des informations extraites automatiquement du document (couleurs dominantes, contraste, textures, formes, vitesse, etc).

Les deux activités principales autour de la classification sont la construction de la classification en fonction d'un ensemble de documents, et la consultation de celle-ci (trouver les photos de coucher de soleil dans cet ensemble, trouver toutes les photos ressemblant à une photo donnée).

Des travaux de recherche effectués dans l'équipe Atlas-Grim du LINA proposent une méthode efficace pour la construction et la consultation de telles classifications. Toutefois, en présence de très grands volumes de données, ces approches ne sont pas suffisantes (voir [1] pour un exemple réel).

Travail à réaliser

Ce stage consiste à porter les méthodes existantes vers un support d'exécution réparti de type pair à pair (P2P), c'est-à-dire un ensemble de machines gérées de manière décentralisée (auto-organisation, sans serveur ni contrôleur). Des travaux existent déjà dans ce domaine, comme [2], ou [3] [4].

La construction aussi bien que la consultation de la classification sur un système pair à pair devront être abordés. Pour la construction, on répartira l'ensemble des documents sur les machines disponibles, et une classification locale sera établie sur chacune. La problématique vient alors de la façon de faire évoluer une classification locale en fonction des informations obtenues sur les classifications des machines « voisines ».

Afin de rendre efficace l'étape de consultation de la classification, on peut ensuite établir une politique de répartition des classes entre les différentes machines (par exemple, une classe sera hébergée sur une unique machine), en tentant de garantir que le processus d'harmonisation des classifications peut se poursuivre.

Enfin, on étudiera le cas où les documents sont disponibles au fur et à mesure. En particulier, de nouvelles classes peuvent apparaître après la première étape de classification.

Un prototype devra être développé et servira à valider les choix effectués sur de très grands volumes de données.

Références

[1] George Tzanetakis, Jun Gao, and Peter Steenkiste. A scalable peer-to-peer system for music information retrieval. *Computer Music Journal*, 28(2) :24–33, June 2004.

[2] W. T. Müller, M. Eisenhardt, and A. Henrich. Efficient content-based P2P image retrieval using peer content descriptions. In Simone Santini and Raimondo Schettini, editors, *Proceedings of the SPIE, Internet Imaging V*, volume 5304, pages 57–68, December 2003.

[3] Cristina Schmidt and Manish Parashar. Enabling flexible queries with guarantees in p2p systems. *IEEE Internet Computing*, pages 19–26, May- 2004.

[4] M. Elena Renda and Jamie Callan. The robustness of content-based search in hierarchical peer to peer networks. In *Proceedings of the 13th Annual ACM Conference on Information and Knowledge Management (CIKM'04)*, Washington, D. C., November 2004.