Gestionnaire de déduplication

Fabrice Huet
fabrice.huet@unice.fr

30 janvier 2015

Nombre d'étudiants souhaités : 3-4

Description du sujet

La numérisation de tous les fichiers multimédia (photos, vidéos, musiques...) conduit à une accumulation de fichiers qu'il est difficile d'organiser. En effet, il n'existe pas de moyen automatique fiable pour classer les photos par thème ou par personne y apparaissant. Combiné au faible coût du stockage, il n'est pas rare d'avoir plusieurs copies d'un même fichier sur un disque dur. La déduplication est une opération qui consiste à détecter des fichiers identiques pour n'en garder qu'une copie. Elle peut se faire directement au niveau du système de fichiers (par exemple ZFS) ou grâce à un logiciel externe. De base, on peut considérer que 2 fichiers sont identiques si ils ont exactement le même contenu, à l'octet près. C'est très facile à détecter en comparant le checksum (MD5 ou SHA-1 par exemple) et relativement fiable. Mais pour certains fichiers, on peut aller plus loin. Par exemple une photo peut exister à différentes résolutions, ou avec de légères différences. Dans ce cas, le checksum sera différent alors que les fichiers seront similaires. Pour traiter ce cas, on peut par exemple utiliser du Locality Sensitive Hashing sur une signature de l'image.

Le but de ce projet est de construire un logiciel permettant de chercher des fichiers identiques ou similaires sur un système de fichiers. Ce logiciel utilisera une interface Web (HTML5 et JQuery) pour communiquer avec un back-end chargé de l'indexation et la recherche de contenus similaires (Java). Il existe une version initiale de ce logiciel dont il est possible de se servir. Elle fournit, entre autre, la recherche d'images similaires. Le logiciel devra au moins avoir les fonctionnalités suivantes :

- Gestion de centaines de milliers de fichiers
- Indexation des fichiers dans une BD de taille réduite
- Recherche de fichiers identiques ou similaires (images au moins)
- Recherche de repertoires contenant des fichiers identiques ou similaires
- Support du drag'n'drop pour chercher dans la base des images similaires à partir d'une image du web

Lieu

Université et réunions régulières à INRIA

Prérequis

Bonne connaissance de Java pour le back-end, HTML5+JS pour le front-end

Informations complémentaires