Compte-rendu de la journée
Passage à l'échelle de la recherche et de la fouille de contenus multimédia

(GDR ISIS et I3, 3 novembre 2010)

Laurent Amsaleg (IRISA), Michel Crucianu (CNAM), Frederic Precioso (ETIS, ENSEA)

10 janvier 2011

La journée, qui a regroupé 41 inscrits, s'est articulée autour de deux axes concernant respectivement :

1. le passage à l'échelle de la recherche par similarité sur la base de nouvelles représentations des données par hachage et codage (Joly, Jégou, Precioso) et de nouvelles techniques d'accès aux données et de distribution des calculs (Barton),
2. les problèmes d'échelle liés à des contextes spécifiques (Morin, Tavenard, Datcu).

Les travaux présentés, qui attaquent de front les problèmes d'échelle lorsque le volume des collections de données à indexer croît et qui visent des temps de réponse très courts quelles que soient les requêtes, sont tous basés sur une recherche approximée des plus proches voisins (Joly, Jégou, Precioso, Barton). Certains travaux se concentrent sur la conception de nouvelles familles de fonctions de hachage pour LSH pour optimiser le hachage par rapport à la distribution des données (Joly) ou permettre la recherche interactive itérative par le contenu (Precioso) dans les grandes bases d'images.

Ce problème d'indexation d'un grand ensemble de vecteurs descripteurs peut être vu comme un problème de codage source, où les vecteurs sont remplacés par leur version quantifiée. La méthode de quantification est alors cruciale pour obtenir une bonne précision en recherche par le contenu (Jégou) ou de nouvelles techniques de répartition sur un réseau (Barton) des indexes avec HBase et des calculs associés à ces indexes avec Hadoop.

L'après-midi les travaux ont concerné l'étude de représentations de données multimédia adaptées à des contextes spécifiques de fouille de données et de passage à l'échelle. Si nombre de techniques de la recherche par le contenu dans les bases d'images sont issues des travaux réalisés dans le cadre de la recherche d'information textuelle (par exemple la distance χ^2 ou le dictionnaire visuel), ces techniques n'ont été que peu explorées sous l'angle du passage à l'échelle. L'analyse des correspondances peut ainsi répondre à la problématique de l'indexation et de la recherche d'information dans des bases d'images et, en associant une structure de fichiers inversés à un découpage du calcul de l'analyse des correspondances, permettre la parallélisation des calculs sur GPU pour traiter rapidement de grands volumes de données (Morin).

Si la fouille de données multimédia sous-entend souvent « données images », d'autres modalités des données multimédia comme le son ou des modalités multiples comme les données satellitaires (images multi-spectrales), particulièrement volumineuses (1000 images = 100 Go), posent également des problèmes de passage à l'échelle. Ces contextes très spécifiques requièrent l'adaptation des méthodes actuelles.

Si peu de bases de données permettant l'évaluation du passage à l'échelle des algorithmes de recherche par le contenu (le bilan fait par M. Crucianu des discussions sur le sujet lors de *ACM Multimedia* 2010 est marqué par la rareté de ce type de base de données), notons tout de même quelques bases qui commencent à faire référence dans le domaine :

1. La base ImageNet (http://www.image-net.org qui est une base d'images associée à la base de données textuelles WordNet. Chaque nœud de la hiérarchie WordNet est ainsi décrit par plusieurs centaines voire milliers d'images (à l'heure actuelle 500 images en moyenne par nœud) pour atteindre 11 231 732 images ce jour. L'importance de la problématique du passage à l'échelle des algorithmes de recherche par le contenu se confirme par la création du

Challenge *Large Scale Visual Recognition Challenge 2010* (ILSVRC2010), en conjonction avec le Challenge PASCAL 2010. Ce Challenge concerne tout autant le passage à l'échelle quand la quantité des données à considérer est très grande (plus de 10 millions d'image) que lorsque le nombre de catégories à appréhender est très élevé (1000 catégories).

2. Les bases construites et traitées par l'INRIA qui sont accessibles depuis la page d'Hervé Jégou (http://www.irisa.fr/texmex/people/jegou/data.php) et qui concernent plusieurs contextes de passage du l'échelle : la base d'images *INRIA Holidays* pour l'évaluation de la recherche d'images par similarité, la base d'images *INRIA Copydays* pour l'évaluation de la recherche de copie, et la base *BIGANN* pour l'évaluation des algorithmes de recherche approximée des plus proches voisins.
3. La base d'images du MIT *Tiny Image Dataset* qui consiste en 79 302 017 imagettes couleur de 32x32 pixels (http://horatio.cs.nyu.edu/mit/tiny/data/index.html).

Accès aux présentations en PDF

Les organisateurs remercient l'ensemble des participants d'avoir transmis leurs présentations (et Alexis Joly pour avoir présenté des travaux en cours de publication).

"SALSAS: Sub-linear Active Learning Strategy with Approximate kNN Search for retrieval"
Frédéric Précioso (ENSEA), David Gorisse (ENSEA), Matthieu Cord (LIP6, UPMC)
http://cedric.cnam.fr/~crucianm/src/echelle/FPrecioso.pdf

"Random Maximum Margin Hashing"
Alexis Joly (Projet IMEDIA, INRIA Rocquencourt)

"Similarity Search at a Very Large Scale Using Hadoop and HBase"
Stanislav Barton (European Web Archive & Wisdom), Philippe Rigaux (CEDRIC& Wisdom)
http://cedric.cnam.fr/~crucianm/src/echelle/SBarton1.pdf
http://cedric.cnam.fr/~crucianm/src/echelle/SBarton2.pdf

"Approximate search as a source coding problem, with application to large scale image retrieval"
Hervé Jégou (INRIA)
http://cedric.cnam.fr/~crucianm/src/echelle/HJegou.pdf

"Utilisation de l’analyse factorielle des correspondances pour l’indexation et la recherche d’information dans une grande bases de données d’images"
Annie Morin (IRISA) Nguyen-Khang Pham (IRISA - Université de Cantho)
http://cedric.cnam.fr/~crucianm/src/echelle/AMorin.pdf

"L'alignement dynamique pour comparer des extraits audio ou vidéo : une bonne idée ?"
Romain Tavenard (Université de Rennes 1 / IRISA), Laurent Amsaleg (CNRS/ IRISA)
http://cedric.cnam.fr/~crucianm/src/echelle/RTavenard.pdf

"Scalability issues for Earth Observation Image Mining"
Mihai Datcu (German Aerospace Center DLR)