arbido

2008/4 Informationswissenschaft: Die Instrumente der Zukunft

Vandergheynst Pierre,

Compression des données et archivage: le binôme du futur

Les ondes radio, les circuits téléphoniques et les câbles d'ordinateurs véhiculent quotidiennement des quantités astronomiques d'informations numériques. Or, comment les référencer si les professionnels de l'information documentaire, entre autres, veulent pouvoir les archiver et les usagers les utiliser? Un double défi attend les chercheurs: la compression et l'indexation.

Les faits

Lorsqu'on parle de «quantités astronomiques» d'informations numériques véhiculées chaque jour par les différents modes de transmission (TV, téléphones, internet, caméras de sur veillance, etc.), ce n'est pas une figure de style, loin s'en faut! Et le futur – proche! – va nous permettre de le vérifier à la puissance n.

Le livre blanc édité par l'IDC en mars 2007 (1) constate que la quantité d'informations numériques créée, saisie et transmise en 2006 était de 1,288 x 10 puissance 18 bytes. Ce qui correspond à 161 exabytes ou 161 billions de gigabytes; autrement dit environ 3 millions de fois l'information contenue dans tous les livres écrits depuis le début des temps. Mais le meilleur est encore à venir, puis que, toujours selon le rapport de l'IDC, le volume d'informations sera multiplié par 6 d'ici à 2010...

Se pose dès lors la question suivante: lorsque l'on sait que 95% de ces données ne sont pas structurées, comment les référencer? Or, la réponse à cette question est de toute première importance pour les professionnels de l'information documentaire qui seront appelés à utiliser les nouveaux outils que les scientifiques sont en train de mettre au point dans ce domaine.

Compresser, puis indexer

La solution comporte deux étapes: 1) il faut d'abord compresser, puis 2) indexer. La difficulté est de taille, puisqu'il s'agit de comprimer les données tout en les structurant «sémantiquement». On connaît déjà des formats de compression comme MPEG, ZIP, JPEG et, plus récent, JPEG2000 (voir encadré), mais ils ne sont encore que des embryons de solutions face au défi que représentent les volumes de données à valoriser.

Prenons par exemple les archives du Festival de Montreux, donc pour l'essentiel des données son et image. L'EPFL se charge actuellement de la numérisation de l'archivage de ce fonds. Mais comment accéder à l'information voulue dans des délais raisonnables? La réponse est sur toutes les lèvres: par recherche «sémantique».

Le défi de la recherche «sémantique»

Certes, mais ici aussi le défi est de taille. Les contenus sont de toute première importance dans ce contexte. Or, l'on sait que ces contenus comprennent du son, du texte, de l'image et de la vidéo. Il faut donc rechercher sur différents types de données. La recherche que l'on propose actuellement est indépendante d'un type de données à un autre. La solution réside donc dans l'intégration de ces données, afin qu'une recherche ciblée soit possible.

Autre exemple: les meetings virtuels, qui sont de plus en plus fréquents et qui seront certainement appelés à se multiplier à l'avenir si l'on considère l'explosion des coûts de déplacement due à la pénurie croissante des énergies non renouvelables. L'archivage de ces meetings (politiques, scientifiques, associatifs, sportifs, culturels) sera donc indispensable et nécessitera des solutions au niveau de la compression des données et de leur stockage qui n'existent pas encore. Le fameux «binôme du futur» sur lequel des milliers de chercheurs se penchent actuellement de par le monde ...

Conclusion

La tâche est donc titanesque pour les chercheurs et il faudra encore du temps avant que les professionnels de l'information documentaire puissent disposer d'outils leur permettant de fournir à leurs clients des prestations dignes de ce nom en matière de fonds audio visuels.

Références:

(1) *The Expanding Digital Universe. A Forecast of Worldwide Information Growth Through 2010,* sous la direction de John F. Gantz, mars 2007

La norme JPEG2000

JPEG2000 est un nouveau système de codage d’image utilisant l’état de l’art des techniques de compression et basé sur la transformée en ondelettes. Son architecture devrait être appropriée à un grand nombre d’applications depuis les appareils photos numériques jusqu’à l’imagerie médicale et d’autres secteurs clé. Le codage comporte des informations sur le contenu ainsi qu’une indexation primaire.

Pierre Vandergheynst

Professeur a? l’EPFL