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

Personnalisation et efficacité de l'accès à l'information multimédia

Encadrant principal : José Martinez
courriel : `José.Martinez@lina.univ-nantes.fr`
tél. : 02 40 68 32 36

Co-encadrant(s) :

Objectif du stage

Les systèmes de recherche d'information ont pour objectif l'assistance à l'accès à des masses d'information pas ou peu structurées [1]. Une étude de l'université de Berkeley [2] estime la quantité d'information produite en 2003 à cinq exaoctets, dont 92 % stockés de manière numérique. La loi de Moore [3], toujours valide pour prévoir l'évolution de la complexité des circuits intégrés, semble s'appliquer également à l'accroissement global de l'information produite : une augmentation régulière de 30 % par an a été constatée entre 1999 et 2002.

Cela amène à traiter le vaste problème du « passage à l'échelle ». L'accroissement du volume des collections a des répercussions immédiates sur l'étape d'évaluation d'une requête tant sur l'efficience (rapidité) que sur l'efficacité (qualité). L'*efficience* correspond à la complexité en temps de l'interrogation. Le temps de réponse à une requête doit toujours rester très court (quelques secondes) quelle que soit la taille du *corpus*. Il devient donc indispensable de limiter très rapidement et drastiquement le sous-ensemble des documents candidats, ceux qui seront examinés plus finement. L'impact sur l'*efficacité* est lié à la quantité de documents qu'il faut classer selon leur pertinence vis-à-vis de la requête [4].

Dans le cas des données multimédias, ces dernières sont déjà volumineuses (images) à très volumineuses (vidéos). Mais surtout, les métadonnées qui les indexent sont elles-mêmes volumineuses, notamment lorsqu'il s'agit d'information de bas niveau (couleur, texture, timbre, quantité de mouvement, trajectoires, etc.). Malheureusement, pour ces métadonnées là, le passage à l'échelle se heurte très rapidement à la « malédiction de la dimensionnalité » [5, 6]. Cela signifie qu'au-delà d'un seuil de 10 à 15 propriétés à indexer simultanément, les performances des index proposés (arbres-TV [7], -SR [8], -X [9], -A [10], etc.) sont aussi mauvaises que le parcours séquentiel total.

Des études préliminaires ont mis en évidence des voies de recherche possibles. En conclusion, les éléments de solution liés au traitement des requêtes sont de plusieurs ordres [11] :

- déporter autant que possible les traitements coûteux dans la phase d'indexation,
- remplacer les traitements coûteux qui persisteraient dans la phase d'appariement par des heuristiques,
- élaguer les traitements en tenant compte d'autres facteurs comme l'utilisateur, l'usage et le niveau d'abstraction des informations,
- revoir les modèles et notamment établir de nouvelles mesures, les descripteurs statistiques « standards » devenant trop peu discriminant dans des collections volumineuses hétérogènes, incomplètes et multimédias.

Nous nous intéressons plus particulièrement à la prise en compte du profil de l'utilisateur. Dans la phase d'indexation, il permettra de classer les documents suivant les centres d'intérêt des – groupes – d'utilisateurs. Dans la phase de recherche, il permettra de réduire rapidement, de manière auto-adaptative, l'ensemble des documents candidats en fonction du profil de l'utilisateur. De manière plus générale, la prise en compte du profil de l'utilisateur dans toutes les étapes du processus de la recherche d'information devrait améliorer l'efficience et l'efficacité dans les collections volumineuses.

Travail à réaliser

Le travail s'articulera autour de la modélisation et de l'exploitation d'un profil-utilisateur pour les données multimédias. Le but est d'exploiter le profil (i) pour pré-classer les données ou les sources de données susceptibles de satisfaire l'utilisateur, (ii) éviter autant que possible des accès effectifs aux systèmes sous-jacents ainsi que (iii) mettre à jour le profil en fonction des accès réalisés et des résultats obtenus.

Références

[1] G. B. Newby. The science of large scale information retrieval. Internet archives, 2000.

[2] P. Lyman, H. R. Varian, K. Swearingen, P. Charles, N. Good, L. L. Jordan, and J. Pal. How much information in 2003 ? http://www.sims.berkeley.edu/research/projects/how-much-info-2003, October 2003.

[3] G. E. Moore. Cramming more components into integrated circuits. *Electronics*, April 1965.

[4] D. Hawking, E . Voorhees, N. Craswell, and P. Bailey. Overview of the TREC-8 web track. In *Proceedings of the 8th Text Retrieval Conference (TREC-8)*, pages 131–150. NIST Special publication, 1999.

[5] R. Bellman. *Adaptive Control Processes : A Guided Tour*. Princeton University Press, 1961.

[6] S. A. Berrani, L. Amsaleg, and P. Gros. Recherche par similarités dans les bases de données multidimensionnelles : panorama des techniques d'indexation. *Ingénierie des Systèmes d'Information*, 7(5-6) :9–44, 2002.

[7] King-Ip Lin, H. V. Jagadish, and Christos Faloutsos. The TV-tree : An index structure for high-dimensional data. *VLDB Journal*, 3(4) :517–542, October 1994.

[8] Norio Katayama and Shin'ichi Satoh. The SR-tree : an index structure for high-dimensional nearest neighbor queries. In *ACM International Conference on Management of Data (SIGMOD)*, pages 369–380, Tucson, Arizona, May 1997.

[9] S. Berchtold, D. A. Keim, and H.-P. Kriegel. The X-tree : An index structure for high-dimensional data. In *22nd International Conference on Very Large Data Bases (VLDB)*, pages 28–39, Mumbai (Bombay), India, September 1996.

[10] Yasushi Sakurai, Masatoshi Yoshikawa, Shunsuke Uemura, and Haruhiko Kojima. The A-tree : An index structure for high-dimensional spaces using relative approximation. In *26th International Conference on Very Large Data Bases (VLDB)*, pages 516–526, Cairo, Egypt, September 2000.

[11] Mohand Boughanem, Sylvie Calabretto, Jean-Pierre Chevallet, José Martinez, and Lynda Lechani-Tamine. Rapport final de l'AS-91 du RTP-9 : "passage à l'échelle dans la taille des *corpus*". Rapport d'expertise auprès du CNRS, January 2004. http://www.irit.fr/ASVolume/.