

Recherche dans de grandes bases d'images fixes : une nouvelle approche guidée par les règles d'association

Anicet Kouomou Choupo*, Annie Morin*, Laure Berti-Équille*

*IRISA Campus de Beaulieu 35042 Rennes-France
{akouomou, amorin, lberti}@irisa.fr
<http://www.irisa.fr/texmex/>

Résumé. Une base d'images fixes peut être décrite de plusieurs façons, notamment par des descripteurs visuels globaux de couleur, de texture, ou de forme. Les requêtes les plus fréquentes impliquent et combinent les résultats de plusieurs types de descripteurs : par exemple, "retrouver toutes les images ayant une couleur et une texture semblables à celles d'une image requête donnée". Pour retrouver plus efficacement et plus rapidement une image dans une grande base, nous exploitons des combinaisons appropriées de descripteurs et étudions l'intérêt des règles d'association entre clusters de descripteurs pour accélérer le temps de réponse à des requêtes sur de grandes bases d'images fixes.

Mots-clé : règles d'association, recherche par le contenu, image fixe, clustering, analyse des correspondances multiples.

1 Introduction

La technologie actuelle est favorable à l'accumulation des grandes quantités d'images fixes. Dans ce contexte, la recherche par le contenu pose deux principaux problèmes : l'organisation de la base et la description du contenu. Les techniques de recherche dépendent de la manière dont la base d'images fixes est organisée et indexée (Djeraba 2003, Berrani et al. 2002).

Une base d'images fixes peut être décrite de différentes façons, notamment par des descripteurs visuels globaux de couleur, de texture, ou de forme (au niveau pixel). Plusieurs descripteurs sont proposés dans la littérature (Manjunath et al. 2002). Chacun d'eux est défini selon l'information que l'on souhaite extraire de l'image. Il est légitime de se demander si l'on peut trouver des relations entre les descripteurs existants et si ces relations permettent une recherche plus rapide et plus efficace d'une image dans une base donnée.

L'objectif de ce travail exploratoire est d'étudier l'intérêt des règles d'association entre descripteurs pour accélérer le temps de réponse à des requêtes sur de grandes bases d'images fixes. Nous déterminons pour chaque descripteur des clusters d'images, puis nous générons sous forme de règles d'association les relations entre les différents clusters obtenus. Pour étudier la pertinence de ces associations entre clusters et pour valider notre approche, nous procédons en parallèle à une analyse des correspondances multiples sur le tableau croisant les images et les clusters. Enfin, il s'agit de sélectionner et d'exploiter les règles les plus pertinentes du point de vue des performances et de la qualité des résultats qui peuvent améliorer la recherche.

Cet article est organisé comme suit : dans la section 2 nous présentons sommairement le principe de description des images fixes ainsi qu'une synthèse de l'état de l'art des méthodes de recherche par le contenu. Notre méthode est présentée dans la section 3 et nous discutons

de nos premiers résultats. Une comparaison des résultats obtenus avec l'analyse des correspondances multiples fait l'objet du paragraphe 4. La section 5 conclut l'article et présente nos perspectives de travail et de recherche à court et moyen terme.

2 Description et recherche d'images fixes

2.1 Principe de description des images

Décrire une image consiste à en extraire le maximum d'informations pour répondre à une requête donnée. Notre étude est volontairement limitée à la description visuelle des images qui sont traitées comme des ensembles de pixels. De façon générale, étant donnée une image i , la description de i consiste à trouver un vecteur $d = f(i)$ qui résume les caractéristiques de l'image i . Le descripteur d de l'image i est un vecteur de réels ou d'entiers de dimension n et f la fonction de calcul du descripteur. Les descripteurs de nos images sont de trois types : la couleur (*ColorLayout*, *ScalableColor*), la texture (*HomogeneousTexture*, *EdgeHistogram*) et la forme (*RegionShape*); tous proposés dans le standard MPEG-7 (Manjunath et al. 2002).

2.2 Synthèse des méthodes de recherche existantes

La recherche d'images par le contenu se fait par plus proches voisins (PPV). L'indexation multidimensionnelle organise une base d'images afin d'accélérer la recherche. Un principe d'organisation consiste à former des clusters regroupant des descripteurs très voisins pour une distance donnée. On distingue, selon les critères d'exploitation des clusters, deux variantes de la recherche de PPV : la recherche exacte et la recherche approximative. Les performances de la recherche exacte se dégradent dans un espace de grande dimension (Bouet 2002).

La recherche approximative établit un compromis entre un temps raisonnable de recherche d'une part et l'exactitude des résultats ou la précision de la recherche d'autre part. Des modèles probabilistes peuvent être élaborés pour contrôler la précision des résultats (Berrani et al. 2002). Djeraba utilise une technique d'indexation voisine de la nôtre en ce sens qu'elle combine la construction des clusters et la détermination des règles d'association (Djeraba 2003). Sa base de travail est formée de plusieurs sous-groupes d'images classées thématiquement (animaux, plantes, ...) et l'idée consiste à caractériser ces groupes par des règles d'association. Lors de la recherche, une analyse de la requête à partir des règles d'association oriente le choix du sous-groupe d'images. Le travail de Djeraba a pour objectif l'amélioration de la qualité de recherche. Il ne s'intéresse pas au problème de fusion des résultats lorsque la recherche se fait suivant plusieurs descripteurs. De plus, le temps de la recherche n'est pas évalué.

D'autres travaux exploitent les règles d'association (Bouet 2002, Aouiche et al. 2003). Le premier tente de capturer le niveau sémantique par la découverte des relations entre les images d'une base de données pour réduire l'espace de recherche. Le second, destiné à l'auto-administration d'une base de données, fonde le calcul de l'index sur les ensembles d'attributs les plus fréquemment utilisés (dans les requêtes) à partir desquels il détermine des relations sous forme de motifs fréquents. Ce travail fait donc une indexation a posteriori de la base alors que nous faisons une indexation a priori. Nous ne supposons aucune organisation préalable de la base d'images et nous ne limitons ni le nombre ni le type de descripteurs utilisables.

3 Description de la nouvelle méthode

Nous proposons une méthode utilisant la classification automatique et la recherche de règles d'association. Notre approche se décompose en deux principales étapes fonctionnelles autour desquelles s'articule l'architecture de notre chaîne de traitement des images (voir Figure 1) : la procédure d'indexation et la procédure de recherche.

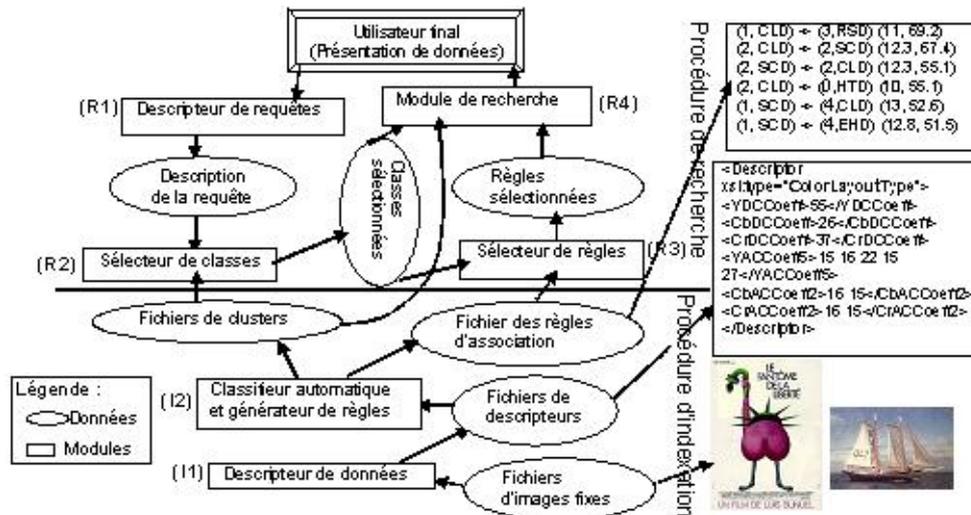


FIG. 1 – Processus d'indexation et de recherche d'images fixes.

3.1 L'indexation

L'indexation se fait hors-ligne. Elle est composée de deux modules : le descripteur de données, le classifieur et le générateur de règles. Le descripteur de données calcule tous les descripteurs à partir d'un ensemble de fichiers d'images stockés dans un répertoire. Nous travaillons avec les descripteurs MPEG-7 de couleur (*Color Layout*, *Scalable Color*), de texture (*Homogeneous Texture*, *Edge Histogram*) et de forme (*Region Shape*). Notre système génère un fichier décrivant toutes les images pour chacun des descripteurs utilisés (étape I1). Le classifieur automatique organise les fichiers de descripteurs en clusters pour chaque descripteur, construit un index d'accès à chaque groupe formé et le stocke dans un fichier. Chaque cluster est identifié par un numéro et la nature du descripteur pour lequel il est calculé. Le générateur de règles utilise les fichiers de clusters produits par le classifieur automatique pour extraire des relations entre clusters sous forme de règles d'association (étape I2). Une règle d'association ainsi générée est une implication de la forme :

$$(\langle \text{num cluster} \rangle, \langle \text{nom descripteur} \rangle) [(\langle \text{num cluster} \rangle, \langle \text{nom descripteur} \rangle) \dots] \rightarrow (\langle \text{num cluster} \rangle, \langle \text{nom descripteur} \rangle) (\langle \text{support} \rangle, \langle \text{confiance} \rangle)$$

ayant la sémantique suivante :

$\langle num\ cluster \rangle$ est le numéro de cluster pour le descripteur $\langle nom\ descripteur \rangle$; $\langle support \rangle$ et $\langle confiance \rangle$ sont des pourcentages indiquant respectivement le support et la confiance de la règle étudiée. La partie gauche d'une règle est composée d'un ou de plusieurs couples identifiant les clusters. Nous limitons la partie droite à un seul couple.

3.2 La recherche

La recherche se fait en-ligne. L'utilisateur soumet au système une image requête. Son but est de retrouver toutes les images semblables à la requête spécifiée. Deux cas peuvent se présenter : soit l'utilisateur choisit les descripteurs suivant lesquels la recherche doit être faite, soit l'utilisateur n'a aucune idée des descripteurs calculés. Nous penchons pour le second cas plus fréquent. Le descripteur de requête traite la requête soumise et produit un descripteur de chaque type dans la limite des descripteurs gérés par la procédure d'indexation (étape R1). Le sélecteur de classes utilise les fichiers de clusters et la description de la requête afin de déduire pour chaque descripteur, le cluster dans lequel la requête pourrait se trouver (étape R2).

Nous utilisons les règles d'association pour réduire le nombre de clusters (classes) dans lesquels nous faisons de la recherche séquentielle. Le sélecteur de règles choisit parmi les règles disponibles celles qui décrivent les relations entre les classes fournies par le sélecteur de classes (étape R3). Une règle est choisie si toutes les classes impliquées dans la règle sont des éléments de l'ensemble des classes sélectionnées à l'étape R2. Les classes et les règles sélectionnées sont transmises au module de recherche (étape R4). Si aucune règle n'est sélectionnée, alors la recherche séquentielle est faite dans toutes les classes sélectionnées ; sinon, la recherche séquentielle se fait uniquement dans les classes n'apparaissant pas en partie droite de règle. Les résultats de la recherche dans les classes sont ensuite fusionnés selon le principe décrit dans (Nepal 1999), puis présentés à l'utilisateur.

3.3 Discussion

Notre méthode est en cours d'implémentation en C++ sous Linux. Les tests sont actuellement faits sur une base de 7727 images fixes. Pour chacun des 5 descripteurs MPEG-7 utilisés, nous regroupons les images en 5 clusters avec un algorithme de type k -means. La complexité temporelle du regroupement en clusters ne nous préoccupe pas ici car l'indexation se fait hors-ligne. Nous avons choisi 10% et 50% comme seuils respectifs de support et de confiance pour le calcul des règles d'association entre clusters avec l'algorithme *Apriori* (Agrawal et al. 1993). Dans ces conditions, le système produit 6 règles dont le support varie entre 10% et 13%. Ce support relativement faible s'explique : sa valeur est une fonction décroissante du nombre de clusters choisi par descripteur. Si l'on suppose une répartition uniforme des images dans chacun des 5 clusters pour chacun des descripteurs, alors le support est majoré par 20%.

L'usage des règles d'association réduit le nombre de descripteurs à explorer lors de la recherche. Dans ce cas, le temps de recherche est inférieur au temps de recherche séquentielle avec fusion des résultats. Le système conçu a été interrogé par 312 requêtes. Pour 112 d'entre elles, le système fait usage des règles d'association calculées, soit un taux d'utilisation de 35.9%. La recherche guidée par des règles d'association offre donc une perspective intéressante.

Le problème du calcul automatique du nombre de clusters reste entier. Nous adoptons une approche manuelle, car l'approche automatique demeure introuvable. Une tentative de résolution du problème est présentée dans (Fernandez et al. 2002), mais le temps de calcul

est plus grand. Notre système présente une autre limite si pour une requête donnée, aucune règle d'association n'est utilisée ; le temps de recherche peut être plus long que le temps de recherche séquentielle avec fusion. cette situation est due au temps de sélection des règles. Dans la suite de ce travail, nous confronterons notre méthode à l'approche classique de recherche séquentielle avec fusion des résultats sur les critères du temps et de la qualité de recherche.

4 Comparaison avec les méthodes statistiques

Pour compléter la recherche de règles, nous avons utilisé l'analyse des correspondances multiples (l'ACM). Chacune des 7727 images est décrite par cinq variables notées CLD (*Color Layout*) et SCD (*Scalable Color*), HTD (*Homogeneous Texture*) et EHD (*Edge Histogram*), RSD (*Region Shape*). Nous disposons pour chaque variable et chaque individu de la classe d'affectation de l'individu pour la variable en question. Ces valeurs vont de 1 à 5. Nous transformons ce tableau en tableau disjonctif complet. En effet, l'ACM est équivalente à l'analyse factorielle des correspondances simples sur un tableau obtenu en remplaçant dans le tableau d'origine chaque variable qualitative par l'ensemble des variables indicatrices des différentes modalités de cette variable. Nous travaillons donc sur un tableau de 7727 lignes et 25 indicatrices. Notre objectif est de comparer les résultats de l'ACM et ceux de la technique des règles d'association.

Les résultats les plus intéressants sont présentés à la figure 2, les nombres entre parenthèses indiquent la confiance de la règle. En plus des règles exhibées avec des techniques de règles d'association (voir Figure 1), on retrouve des règles plus compliquées mettant en jeu plusieurs variables. On constate que le descripteur de couleur CLD permet à lui seul d'induire d'autres valeurs de descripteurs. Nous n'avons conservé que les valeurs de confiance supérieures à 50% sauf pour la règle qui conduit à induire deux valeurs des modalités de SCD. Cette permanence d'associations impliquant CLD nous a conduit à estimer la topologie d'un réseau bayésien entre les cinq variables (voir figure 3). On note la position à l'origine du réseau de CLD. Il existe des dépendances conditionnelles entre les variables SCD, HTD, EHD, RSD et CLD.

association forte entre modalités de variables	modalités induites
(2,CLD) et (2,SCD)	(0,HTD) (54.4%)
(1,CLD) ou (3,CLD) et (2,HTD)	(0,SCD) (59.7%) et/ou (4,EHD) (51.3%)
(0,CLD) et (3,HTD)	(0,SCD) (47.9%) ou (3,SCD)(45.7%)
(4,CLD) et (1,SCD)	(4,EHD) (52.6%)
(3,CLD) et (0,SCD)	(2,EHD)(52.2%) et/ou (3,RSD) (54.4%)

FIG. 2 – Règles produites par l'ACM.

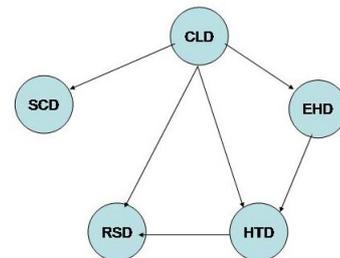


FIG. 3 – Estimation topologique des relations entre variables descripteurs.

Les méthodes statistiques confirment et complètent les règles d'association. Pour choisir le nombre de classes par descripteur, il faudrait examiner plus finement chacun d'eux. Cependant, un des problèmes est l'absence de sémantique et de similarité visuelle fournies par la classification. L'utilisation des règles d'association conjuguées aux méthodes statistiques peut accélérer

les recherches mais cela ne suffit pas pour une indexation et une recherche a posteriori.

5 Conclusion

Dans cet article, nous présentons un travail exploratoire et décrivons le principe de fonctionnement d'un système de recherche d'images par le contenu en cours de développement dont le but est d'améliorer les temps de recherche au sein d'une grande base d'images en exploitant les règles d'association extraites à partir des clusters de descripteurs d'images. A terme, l'idée est de pouvoir combiner voire interchanger des descripteurs d'images pour améliorer les performances et la qualité des résultats à une recherche par le contenu.

Références

- Agrawal R., Imielinski T., Swami A. (1993), Mining Association Rules Between Sets of Items in Large Databases, ACM SIGMOD, 1993, pp 207-216.
- Aouiche K., Darmont J., Gruenwald L. (2003), Extraction de motifs fréquents pour l'auto-administration des bases de données, RSTI, 2003, p 547.
- Berrani S.A., Amsaleg L., Gros P. (2002), Approximate k-Nearest Neighbor Searches: A New Algorithm with Probabilistic Control of the Precision, Rapport de Recherche INRIA, N° 4675, 2002.
- Bouet M., Khenchaf A. (2002), Traitement de l'information multimédia : Recherche du média image, RSTI - ISI. Bases de données et multimédia, 7(5-6), 2002, pp 65-90.
- Djeraba C. (2003), Association and Content-Based Retrieval, IEEE Transactions on Knowledge and Data Engineering, Vol. 15, N° 1, 2003, pp 118-135.
- Fernandez G., Meckaouche A., Peter P., Djeraba C. (2002), Intelligent Image Clustering, Lecture Notes in Computer Science, Vol. 2490, 2002, pp 406-419.
- Manjunath B.S., Salembier P., Sikora T. (2002), Introduction to MPEG-7, John Wiley & Sons, 2002.
- Nepal S., Ramakrishna M.V. (1999), Query Processing Issues in Image (Multimedia) Databases, Proceedings of the ICDE, 1999, pp 22-29.

Summary

Fixed image database can be described in several ways by global visual descriptors of color, texture, or form. Most frequent queries imply and combine results of several type of descriptors such as: "retrieve all images that have similar color and similar texture to the given example image". To retrieve more efficiently and more effectively an image of a large database, we exploit combinations of descriptors and we study the interest of association rules between clusters of descriptors to accelerate response time of queries on large fixed image databases. Multiple correspondence analysis is used to validate our approach.

Keywords: association rules, content based retrieval, fixed image, clustering, multiple correspondence analysis.