# Multimedia Indexing and Retrieval with Features Association Rules Mining

Anicet Kouomou-Choupo, Laure Berti-Équille, Annie Morin
*IRISA, University of Rennes I*
*{akouomou,Laure.Berti-Equille,Annie.Morin}@irisa.fr*

## Abstract

*The administration of very large collections of images accentuates the classical problems of indexing and efficiently querying information. This paper describes a new method applied to very large still image databases that combines two data mining techniques: clustering and association rules mining in order to better organize image collections and to improve the performance of queries. The objective of our work is to exploit association rules discovered by mining global MPEG-7 features data and to adapt the query processing. In our experiment, we use five MPEG-7 features to describe several thousands of still images. For each feature, we initially determine several clusters of images by using a K-mean algorithm. Then, we generate association rules between different clusters of features and exploit these rules to rewrite the query and to optimize the query-by-content processing.*

## 1. Introduction

Because of the growing demand for database and information systems support in the area of modeling, managing, and processing digital media, there is a need to explicitly capture a fair amount of information content as well as application-specific semantics by mean of a variety of metadata (e.g., multimedia indexes, attributes-based annotations, and intentional descriptions), in order to allow appropriate access to, selection of, and processing of digital media involving very large raw data volumes. But most of the current practices in the context of multimedia data management are still quite *ad hoc*.

Content-based retrieval on raw data means that query capabilities are limited to the number of available matching algorithms. Performance is lacking when queries are executed on large data sets. Indirect retrieval and processing, however, that use abstract information or metadata seem to be a promising approach to enhance querying and processing. Our approach is to exploit metadata such as association rules extracted from mining visual global features data and to better organize and index image collections according to these feature association rules. We propose a new method combining two data mining techniques: clustering and association rules mining applied to all kind of very large still image databases in order to optimize the organization of data and the query processing.

The paper is organized as follows. In section 2, we present the new method including clustering and association rule mining we propose for better organizing very large image databases. In section 3, we present and discuss our experimental results. Finally, Section 4 concludes the paper and presents our future work.

## 2. Mining global features for query planning

The method we proposed is described in Figure 1 and includes two steps: 1) the off-line image indexing by clustering and association rule mining on global MPEG-7 feature data and 2) the on-line retrieval exploiting association rules metadata in order to accelerate access to queried images.

0-7803-8603-5/04/$20.00 ©2004 IEEE

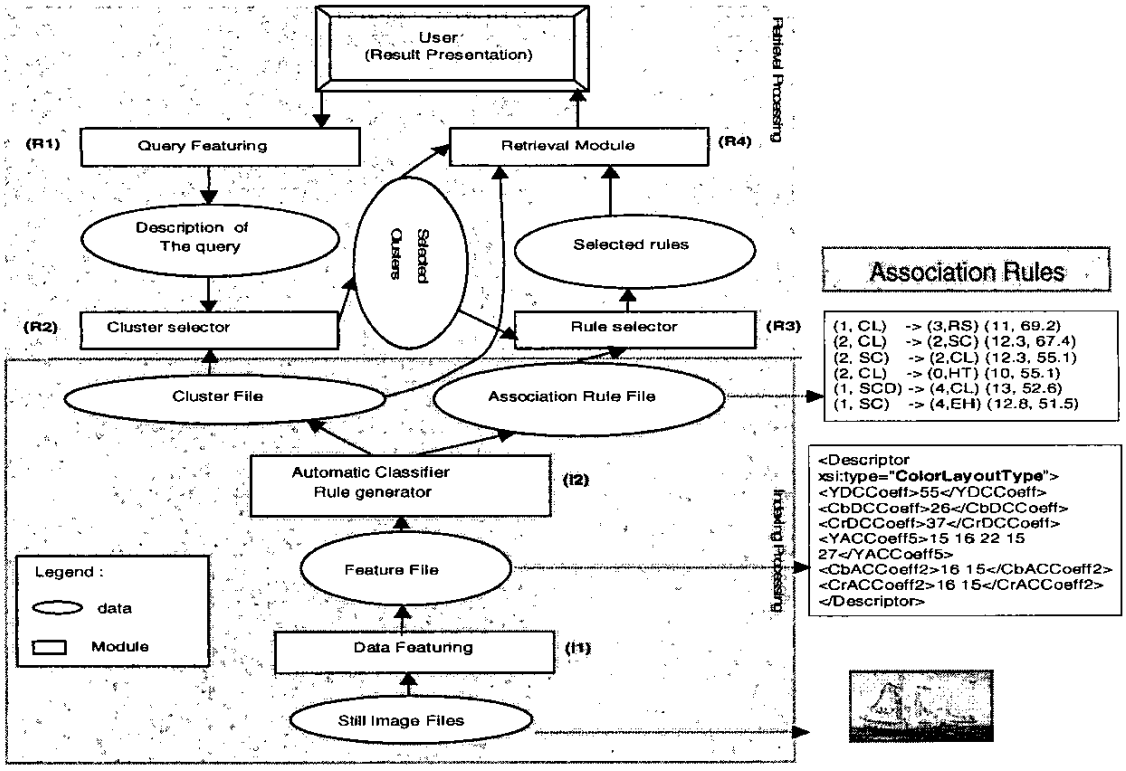
## 2.1. Indexing images with clustering and association rule mining

Image database indexing consists of three modules (see figure 1): the featuring module, the classifier and the rule generator. Data featuring module calculates MPEG-7 features for the whole image set. For our experiment, we work with two MPEG-7 color features (*ColorLayout*; *ScalableColor*), two MPEG-7 texture features (*HomogeneousTexture; EdgeHistogram*) and one form feature (*RegionShape*) [3]. Our image indexing process has two steps:

1. for each feature, our system generates a XML file describing all the stored images. Each image is then represented as a multidimensional numeric vector for used features (I1 indexation step in Figure 1)
2. the automatic classifier and the rule generator described in next subsections carry out tasks of organizing image collection (I2 indexation step).

**2.1.1. Clustering.** To reduce the dimensionality, a clustering approach is used. It reduces considerably the number of images subcollections by clustering similar images into similar groups. The clustering algorithm is a variant of k-medoids [2]. The automatic classifier organizes the file of features in clusters, and builds an access index. We use the K-mean algorithm for clustering. Each cluster contains images considered to be similar according to the distance associated to a given feature. It is then identified by a number (*cluster#*) and the name of the feature (*FeatureName*). Finally, an image is described by a set of clusters in which it belongs for all features. The rule generator then uses the files of clusters produced by the automatic classifier to extract relations (named association rules) between the clusters.

**2.1.2. Association rule mining.** Association rule mining has been intensively investigated in the data mining literature. Many efficient algorithms have been proposed, the most popular being *Apriori* [1]. Association rule mining typically aims at discovering associations between items in a transactional database. Given a set of transactions $D = \{T1, .., Tn\}$ and a set of items $I = \{i1, .., im\}$ such that any transaction $T$ in $D$ is a set of items in $I$, an association rule is an implication $A \rightarrow B$ where the antecedent $A$ and the consequent $B$



are subsets of $I$, and $A$ and $B$ have no common items. $A \subseteq I$ is called a *k-itemset* if the size of $A$ is $k$. For an association rule to be strong, the conditional probability of $B$ given $A$ has to be higher than a threshold called *minimum confidence*.

The support is defined such as: $s = |A \wedge B|/|D|$ and the confidence is defined such as: $c = |A \wedge B|/|A|$.

Association rules mining is normally a two-step process. In the first step, frequent item-sets (i.e. item-sets whose support is no less than a minimum support) are computed iteratively in the ascending order of their size using a level wise algorithm. In the second step, association rules in the form $A_2 \rightarrow A_1 - A_2 \mid A_2 \subset A_1$ with confidence greater or equal to *minimum confidence* are derived from each frequent item-set $A_1$ computed in the first step. In our approach, we used the *Apriori* algorithm [1] in order to discover association rules among the clusters of global MPEG-7 features extracted from the image database.

Generated association rules are implications such as:
(<*cluster#*>;<*FeatureName*>)
[(<*cluster#*>;<*FeatureName*>)...]
→ (<*cluster#*>;<*FeatureName*>) (<*s*>; <*c*>)

with the following semantics: <*cluster#*> is the cluster identifier for the feature <*FeatureName*>; <*s*> and <*c*> are percentages indicating respectively the support and the confidence of the rule previously defined. The left part of a rule is made up of one or several couples identifying the clusters such as (<*cluster#*>;<*FeatureName*>). The right part is limited to only one couple (see Figure 1 for six instances of the extracted rules).

### 2.2. Retrieving images

While retrieving an image from an image database, the goal of the user is to find all the images similar to the specified query. Two cases can arise: either the user selects the features whereby searching must be made, or the user does not have any idea of the features to use. We work with the second case we considered to be more general and the retrieval process consists of four main steps:

1. image submitted as query is processed and all the global features managed by the procedure of indexing (R1 retrieval step in Figure 1) are produced for this image by the featuring module. In our example, we considered five MPEG-7 features: two for the colour, two for the texture, and one for the form. Each of the 7727 images is described by five variables noted *CL* (*ColorLayout*) and *SC* (*ScalableColor*), *HT* (*HomogeneousTexture*) and *EH* (*EdgeHistogram*), *RS* (*RegionShape*)
2. the cluster selector uses the file of clusters and the description of the query to deduce for each feature, the cluster in which the query could be the closest (R2 retrieval step in Figure 1)
3. we use association rules to reduce the number of clusters in which we make sequential search. The rule selector chooses among rules available those which describe relations between clusters provided by the cluster selector (R3 retrieval step). In other words, a rule is selected if all the clusters implied in the rule (in the left and right part) are elements of the whole set of clusters selected in the R2 retrieval step
4. selected clusters and rules are transmitted to the searching module (R4 retrieval step). If no rule is selected, then sequential search is made in all the selected clusters of images. On the contrary case, sequential search will be done only in the clusters not appearing in the right part of the rule. Results of sequential search in the clusters are then merged according to the principle given in [4].

## 3. Experiment and Discussion

The method is implemented in C++. All the algorithms run on a PC under Linux. Its CPU is a Pentium 4 2.4 GHz, with 1024 Mb of main memory and 80 Gb of local disk. We worked with a base of 7727 still images. For each of the five MPEG-7 features, we gather the images into 5 clusters with the k-mean algorithm. We chose a minimum support of 10% and a minimum confidence of 50% for the calculation of association rules between clusters with the *Apriori* algorithm [1]. Under these conditions, the system produces 6 relevant rules whose support varies between 10% and 13%. This relatively weak support is explained. Indeed, the value of the support is a decreasing function of the number of clusters chosen by feature. If we suppose for example an uniform distribution of the images in each of the 5 clusters for each feature, then the support of the rules is majored by 20%. The use of association rules reduces the number of features to explore for on-line searching. In this case, the search time is lower than sequential search time including the results fusion. Our CBIR system has been queried by 500 queries. For 165 of them, the

system makes use of the generated association rules, that is to say the usage ratio of 33%.

Image retrieval guided by association rules offers an interesting perspective to explore for improving query by content performance. The searching time is reduced with the proposed method compared to the sequential searching time including the fusion of results (Table 1). Our objective is now to improve these performances by refining combination strategies of association rules.

Table 1. Experimental results of searching time for 500 queries over 7727 images

| | Average sequential searching time (s) | Average searching time with the method (s) | Time speed-up (s) |
|---|---|---|---|
| **Queries using association rules** | 46.50 | 44.77 | 1.73 |
| **Queries without using association rules** | 46.71 | 45.25 | 1.46 |

In order to validate our approach, we also compared it with statistical methods such as multiple correspondence analysis (MCA). Our objective was initially to check if we find results close to association rules and if we cannot find some others. Most interesting results appear in the following Table 2, the numbers between brackets indicate the confidence of the rule:

Table 2. Experimental results

| Strong relations between modalities of variables | Induced modalities |
|---|---|
| (2,CL) and (2,SC) | (0,HT) (54.4%) |
| (1,CL) or (3,CL) and (2,HT) | (0,SC) (59.7%) and/or (4,EH) (51.3%) |
| (0,CL) and (3,HT) | (0,SC) (47.9%) or (3,SC)(45.7%) |
| (4,CL) and (1,SC) | (4,EH) (52.6%) |
| (3,CL) and (0,SC) | (2,EH)(52.2%) and/or (3,RS) (54.4%) |

The variables are noted *CL* (*ColorLayout*) and *SC* (*ScalableColor*), *HT* (*HomogeneousTexture*) and *EH* (*EdgeHistogram*), *RS* (*RegionShape*).

This table is interesting because we find association rules showed in Figure 1 and generated by association rule mining and also more complicated rules with several variables obtained by multiple correspondence analysis. We especially remark that the feature of color *CL* is extremely significant and allows to induce other values of features. This permanence of associations implying *CL* led us to estimate the topology of a Bayesian network between the five variables (Figure 2). Indeed, we note that the position of the origin of the network is *CL*.

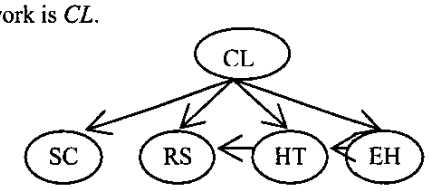


Figure 2. Bayesian network of features

For every kind of image databases, we can build off-line such Bayesian networks with global MPEG-7 features that can be used for improving the access time to queried images.

## 4. Conclusion

In this article, we describe a new method that can improve the global search time for a content-based information retrieval system. The interesting point is: this method can be applied to all kind of large image databases by exploiting association rules extracted from mining the clusters of global image features. Our research perspective is to adaptively combine and interchange features data in order to build optimized query plans (and also to rewrite queries by content) and to improve the performances and the quality of results.

## 5. References

[1] R. Agrawal, T. Imielinski, A. Swami, "Mining Association Rules Between Sets of Items in Large Databases", *ACM SIGMOD International Conference on Management of Data*, Volume 22, pp 207-216, Washington: ACM press 1993.

[2] S. A. Berrani, L. Amsaleg, P. Gros, "Approximate k-Nearest Neighbor Searches: A New Algorithm with Probabilistic Control of the Precision", *Tech. Rep. INRIA, No 4675*, 2002.

[3] B. S. Manjunath, P. Salembier, T. Sikora, *Introduction to MPEG-7*, John Wiley & Sons, 2002.

[4] S. Nepal, M. V. Ramakrishna, "Query Processing Issues in Image (Multimedia) Databases", *Proceedings of the 15th International Conference on Data Engineering (ICDE 99)*, pp 22-29, Australia 1999.