FEATURES MINING FOR MULTIMEDIA INDEXING AND RETRIEVAL

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

IRISA, University of Rennes I

ABSTRACT

The administration of very large collections of images accentuates the classical problems of indexing and efficiently querying. 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 the image collections and to improve the performance of the query processing. 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 the different clusters of features and exploit these rules to rewrite and to optimize the image query processing.

1. INTRODUCTION

Because of the growing demand for databases 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 content-information as well as application-specific semantics by mean of a variety of metadata (e.g., multimedia indexes, attributes-based annotations, and intentional descriptions), 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 the 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 assumption is that we can exploit metadata such as association rules extracted from mining visual features data and better organize the image collections according to these feature association rules. We propose a new method combining two data mining techniques: clustering and association rules mining that can be applied to all kind of very large still image databases in order to both optimize the data organization and the query processing.

The paper is organized as follows. In section 2, we present the new method including clustering and association rule mining. 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. It includes two steps: 1) the image indexing made off-line by clustering and association rule mining on global feature data and 2) the retrieval made on-line exploiting the association rule metadata in order to accelerate the access to the searched images.

2.1. Indexing images with clustering and association rule mining

The indexing of the image database has been implemented by three modules (see figure 1): the featuring module, the classifier and the rule generator. The data feature module calculates here five MPEG-7 features for the whole image set (but, this experiment can be extended to other global features). We worked with : two MPEG-7 color features (*ColorLayout; ScalableColor*), two MPEG-7 texture features (*HomogeneousTexture; EdgeHistogram*) and one form feature (*RegionShape*) [4].

For each feature, our system generates a XML file describing all the stored images (*I1 step* in Figure 1). The automatic classifier and the rule generator described in next subsections carry out the tasks of organizing the image collection (*I2 step*).

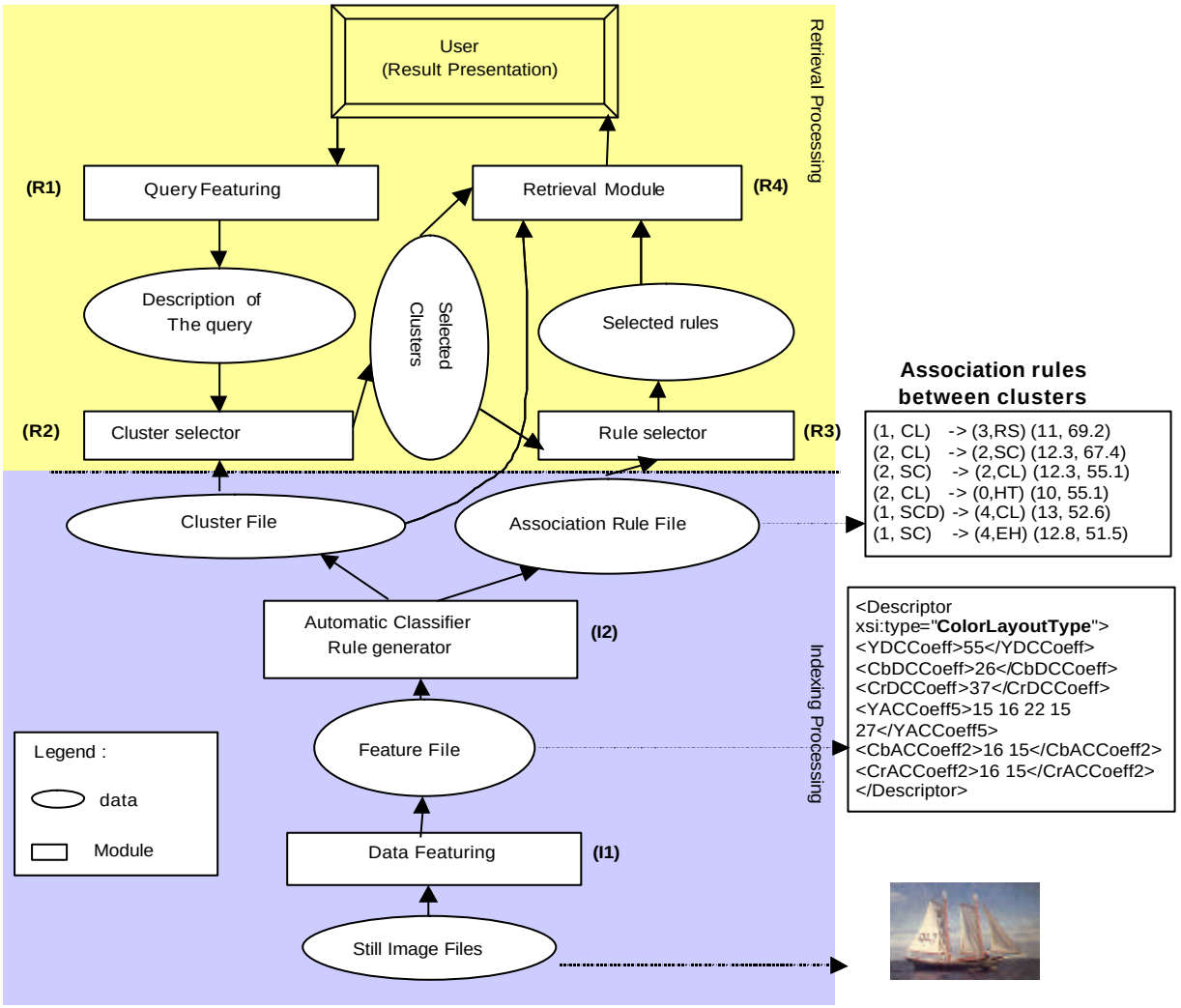
2.1.1. Clustering

To reduce the dimensionality, a clustering approach is used. It reduces considerably the number 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 to make classification. Each cluster is identified by a number (*cluster#*) and the name of the feature (*FeatureName*).

The rule generator then uses the files of clusters produced by the automatic classifier to extract relations, called association rules between the clusters.

2.1.2. Association rule mining

Association rule mining has been extensively 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 $D = \{T_1, .., T_n\}$ and a set of items $I = \{i_1, .., i_m\}$ 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 a transaction T in D, and A and B have no common items. For the association rule to be strong, the conditional probability of B given 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, where in the first step frequent item-sets are discovered (i.e. item-sets whose support is not less than a minimum support, called *minsup*) and in the second step association rules are derived from the frequent item-sets. In our approach, we used the *Apriori* algorithm [1] in order to discover association rules among the clusters of features extracted from the very large image database.

The generated association rules are the 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 the percentages indicating respectively the support and the confidence of the association rule. 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 has only one couple (see Figure 1 for the instances of the extracted rules).

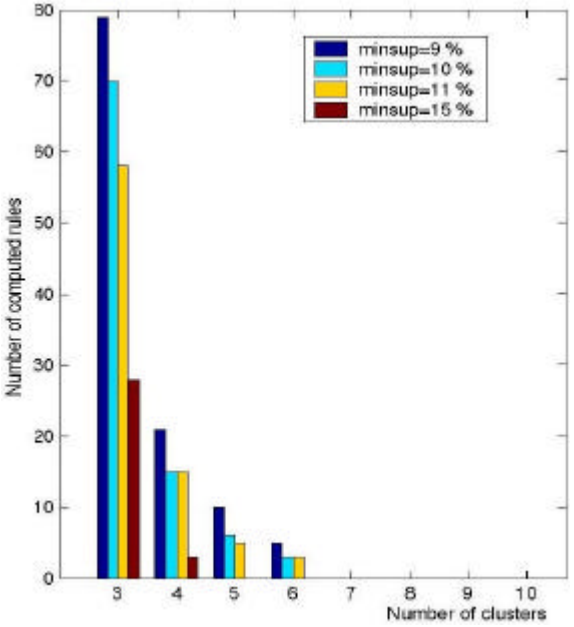
2.2. Retrieving images

While retrieving an image into the 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 assumption considered to be more general. The image submitted for the query is processed and all the global features managed by the indexing procedure (*R1 step*) are produced for this image by the featuring module. In our example, we considered two MPEG-7 features for colour, two for texture, and one for form.

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 step*). We use the association rules to reduce the number of clusters in which we make the sequential search. The rule selector chooses among the rules available those which describe the relations between the clusters provided by the cluster selector (*R3 step*). In other words, a rule is selected if all the clusters in the rule (in the left and the right parts) are elements of the whole set of clusters selected in the *R2* step. The selected clusters and rules are transmitted to the searching module (*R4 step*). If no rule is selected, then sequential search is made on all the selected clusters of image features. In the contrary case, sequential search will be done only in the clusters that don't appear in the right part of the rule. The results of sequential search in the clusters are then merged [5].

3. EXPERIMENT AND DISCUSSION

The method has been implemented in C++ under Linux. We work for the moment with a base of 7727 still images. For each of the five MPEG-7 features, we gather the images into 5 clusters with a k-mean algorithm. The choice of the number of clusters depends on the number of significative rules discovered from the database and processed by the system. Figure 2 shows that : 1) when the number of clusters increases (more than 7 clusters), the number of significative rules decreases and 2) when the number of clusters decreases, the number of rules increases but rules are less significative. In our experiment, we choose a minimum support of 10% and a confidence minimum of 50% for computing the association rules between the clusters with the algorithm *Apriori* [1].



Under these conditions, the system produced 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, the 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 the association rules reduces the number of features to explore for the on-line searching. In this case, the search time is lower than sequential search time including the fusion of results.

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%.

The image retrieval guided by association rules offers an interesting perspective of research for improving the performances of the query processing for a CBIR system.
In order to validate our approach, we compared it with statistical methods such as the multiple correspondence analysis (MCA). Each of the 7727 images is described by the five variables noted *CL* (*Color Layout*) and *SC* (*ScalableColor*), *HT* (*HomogeneousTexture*) and *EH* (*EdgeHistogram*), *RS* (*ShapeArea*).
Our objective was initially to check if we could find results close to those of the association rules mining technique and if we could find some others. The most interesting results appear in the following table, the numbers between brackets indicate the confidence of the rule:

Strong relations between modalities of variables	**Induced modalities**
(2,CLD) and (2,SCD)	(0,HTD) (54.4%)
(1,CLD) or (3,CLD) and (2,HTD)	(0,SCD) (59.7%) and/or (4,EHD) (51.3%)
(0,CLD) and (3,HTD)	(0,SCD) (47.9%) or (3,SCD)(45.7%)
(4,CLD) and (1,SCD)	(4,EHD) (52.6%)
(3,CLD) and (0,SCD)	(2,EHD)(52.2%) and/or (3,RSD) (54.4%)

Table 1. Experimental results

This table is interesting because we found the association rules showed in Figure 1 and also, several more complex 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 the variable *CL* led us to estimate the topology of a bayesian network between the five variables (Figure 3). Indeed, we note that the position of the root of the network is *CL*. To present the underlying mathematical model of the MCA method is not the scope of the paper.

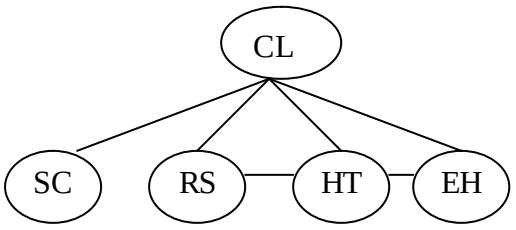


Figure 3. Bayesian network of features

Our objective is to propose a combination of techniques in order to build off-line, for every kind of image databases, a relevant bayesian networks of global features that can be used for improving the access time to searched images in the query-by-content processing.

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 very large image databases by exploiting the association rules extracted from mining the clusters of global image features. Our research perspective is to adaptively combine and interchange features in order to build optimized query plans (and also to rewrite the queries) and to improve the query performances and the quality of the results.

5. REFERENCES

[1] Agrawal R., Imielinski T., Swami A. (1993), Mining Association Rules Between Sets of Items in Large Databases, ACM SIGMOD International Conference on Management of Data, 1993, pp 207-216.

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

[3] Djeraba C. (2003), Association and Content-Based Retrival, IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No.1, 2003, pp 118-135.

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

[5] Nepal S., Ramakrishna M.V. (1999), Query Processing Issues in Image (Multimedia) Databases, Proceedings of the ICDE, 1999, pp 22-29.

[6] Obeid M., Jedynak B., Daoudi M. (2001), Image indexing and retrieval using intermediate features, Proceedings of the ninth ACM international conference on Multimedia, 2001, pp 531-533.

[7] Smeulders A.W.M., Worring M., Santini S., Gupta A., Jain R. (2000), Content-Based Image Retrieval at the End of the Early Years, Vol. 22, No.12, 2000, pp 1349-1380.