CLUSTERING COMPLEX MULTIMEDIA OBJECTS USING AN ENSEMBLE APPROACH

Ana Isabel Oviedo[1] and Oscar Ortega[2]
[1]*Computer Department, Universidad Pontificia Bolivariana, Medellin, Colombia*
[2]*System Engineering Department, Universidad de Antioquia, Medellin, Colombia*

Keywords: Complex multimedia objects, Clustering, Ensemble methods, Unsupervised learning.

Abstract: A complex multimedia object is an information unit composed by multiple media types like text, images, audio and video. Applications related with huge sets of such objects exceed the human capacity to synthesize useful information. The search for similarities and dissimilarities among objects is a task that has been done through clustering analysis, which tries to find groups in unlabeled data sets. Such analysis applied to complex multimedia object sets has a special restriction. The method must analyze the multiple media types present in the objects. This paper proposes a clustering ensemble that jointly assesses several media types present in this kind of objects. The proposed ensemble was applied to cluster webpages, constructing a text and image clustering prototypes. The Hubert's statistic was used to evaluate the ensemble performance, showing that the proposed method creates clustering structures more similar to the real classification than a joint-feature vector.

1 INTRODUCTION

A complex multimedia object (CMO) is an aggregation of heterogeneous data as a single unit. A CMO is composed by multiple media types like text, images, audio and video (Hunter and Choudhury, 2003) (Yang et al., 2008) (Zhuang et al., 2008). CMOs are present in several scenarios like web sites, music albums, electronic journals, electronic books, digitally recorded sound, digital moving images, digital television and social networks (Hunter and Choudhury, 2003) (Kriegel et al., 2008).

In applications with large amount of CMOs, the huge number and the complexity of the relationships among the objects exceed the human capacity to analyze and synthesize useful information and knowledge. The relationships among the objects can be expressed by similarities and dissimilarities that are searched in an automatic way using computers. The search for similarities and dissimilarities among CMOs is a task that has been done through clustering analysis, whose goal is to find natural groups in an unlabeled object set, such that objects in a group must be similar or related to one another, and must be different from the objects in other groups (Jain et al., 1999) (Romesburg, 2004) (Dy and Brodley, 2004) (Algergawy et al., 2008) (Jain, 2010). Humans are excellent seekers in two or three dimensional problems, but an automatic algorithm is necessary for higher dimensions (Jain, 2010).

Formally, the clustering task has an input set of n objects called $X = \{x_1, x_2, ..., x_n\}$, where each x_i is a feature vector of order d that represents the information of object i with $x_i = (f_1, f_2, ..., f_d) \in R^d$; each f_l is the lth feature. A clustering method attempts to distribute X into k groups given by $C = \{c_1, c_2, ..., c_k\}$, where k is the number of clusters with $k \leq n$ and c_j represents the jth cluster.

A clustering analysis can be performed in different ways. The literature review shows several clustering approaches such as hierarchical, partitioning, fuzzy, neural networks, probabilistic, graphs, evolutionary, kernels and spectral methods. Yet, when cluster analysis is applied to CMO sets, it has a special restriction: the method must analyze different media types. In the clustering approaches reviewed so far for this study, some methods combine multiple clustering structures when sets whose elements have the same media type are too complex to yield a unique clustering structure. Such methods are called alternative clustering, clustering aggregation, clustering ensemble and collaborative clustering.

In support of a clustering for CMOs, this paper proposes an ensemble approach which allows an independent clustering analysis of different media types present in this kind of objects and then the results are

Isabel Oviedo A. and Ortega O. (2012).
CLUSTERING COMPLEX MULTIMEDIA OBJECTS USING AN ENSEMBLE APPROACH.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 134-143
DOI: 10.5220/0003794501340143

combined with a voting function.

The paper outline is presented below. Section 2 presents the clustering lifecycle; section 3 presents the literature review; section 4 presents the proposed clustering ensemble; section 5 presents an evaluation of the ensemble and, finally, the conclusions are presented in section 6.

2 CLUSTERING LIFECYCLE

This section describes the clustering lifecycle for addressing the proposed clustering ensemble. The process of partitioning objects into clusters involves the following stages: feature subset selection, similarity measuring, clustering or grouping, cluster evaluation and result interpretation. The flowchart in figure 1 shows a feedback in the process depending on the cluster quality.

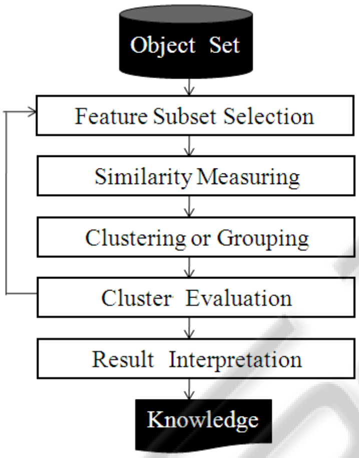

Figure 1: The clustering lifecycle.

The first stage, feature subset selection, involves the attribute extraction that describes each object $x_i = (f_1, f_2, ..., f_d)$. The goal of this stage is to find the smallest feature subset that uncovers natural clusters (Dy and Brodley, 2004). The feature subset selection can be performed with two approaches: filter or wrapper algorithms (Dy and Brodley, 2004). The filter approach pre-selects the features before applying a clustering method. The wrapper approach incorporates feature selection into the clustering method.

The second stage, similarity measuring, is not easy to specify if the user does not have prior knowledge about the objects (Fred and Jain, 2005) (Francois et al., 2006). The goal of this stage is to specify how to measure the similarity between objects, which can be performed with two approaches: based on a probability distribution or based on a distance function. The most used approach is similarity measuring based on an Euclidean distance function; however, there are other distance measures as cosine, manhattan, chebyshev, mahalanobis, minkowski and hamming, which can produce diverse partitions for the same object set (Fred and Jain, 2005).

The third stage, grouping, involves the application of a clustering method using similarity measuring. The goal of this stage is to partition an object set $X = \{x_1, x_2, ..., x_n\}$ into k groups. The grouping stage can be performed with different approaches (Xu and Wunsch, 2005) like hierarchical, partitional, fuzzy, neural networks, probabilistic, graph, evolutionary, kernel and spectral methods.

The fourth stage, cluster evaluation, measures the quality of the partition obtained by the grouping stage (Fred and Jain, 2005) (Jain, 2010). There are two index approaches to evaluate cluster quality: external validation and internal validation. External validation evaluates the cluster quality based on a pre-specified clustering structure. Some external indexes are Rand, Fowlkes and Mallows, Hubert and Arabie, and Jaccard (Halkidi et al., 2002)(Hashimoto et al., 2009). Internal validation evaluates the cluster quality based on compactness and separability measures; the compactness expresses how similar objects are in the same cluster and the separability expresses how distinct objects are in different clusters. Some internal indexes are Dunn, Davies-Bouldin, Silhouette, Gath-Geva, Fukuyama-Sugeno and Xie-Beni (Halkidi et al., 2002)(Hashimoto et al., 2009).

The final stage, results interpretation, is one of the most important steps. Its goal is to provide meaningful information for users from the original objects with a compact description of each cluster. The interpretation of results can be performed in terms of cluster prototypes or of the most representative objects such as the centroid (Jain et al., 1999).

3 THE LITERATURE REVIEW

The literature review shows several clustering approaches (Xu and Wunsch, 2005) (Filippone et al., 2008) (Jain et al., 1999) (Jain, 2010), such as hierarchical, partitioning, fuzzy, neural networks, probabilistic, graphs, evolutionary, kernels and spectral methods. Yet, other kinds of methods are used when the dataset is too complex to yield a unique clustering structure. Such methods combine multiple clustering structures and have been applied to sets with the same media type.

The following literature review was organized in two parts: approaches for combining multiple clustering structures and clustering analysis applied to sets with the same media type.

3.1 Approaches for Combining Multiple Clustering Structures

Several approaches for combining multiple clustering structures have been formulated based on the idea that efficiency and accuracy can be increased by assessing different clustering structures (Strehl and Ghosh, 2003) (Gancarski and Wemmert, 2007). The clusterers combination is considered more difficult than classifiers combination in supervised learning, because the different clustering structures may not have the same number of groups and there is no information about the correspondence between the clusters of the different clustering structures (Strehl and Ghosh, 2003) (Gancarski and Wemmert, 2007). In the literature review, different approaches have been found for combining multiple clustering structures as altenative clustering, clustering aggregation, clustering ensemble and collaborative clustering.

3.1.1 Alternative Clustering

The first approach, called alternative clustering, generates different clustering structures and lets the user select the best structure according to his/her need. In (Caruana et al., 2006), the method is called meta-clustering: they organize together many base-level clusterings into a clustering of clusterings; thus the user navigates to the clustering(s) useful for his/her purposes. In (Bae and Bailey, 2006), the alternative clustering structures start from an existing structure and, in (Davidson and Qi, 2008), the use of constraints to characterize an existing clustering is proposed and then an alternative solution can be generated.

3.1.2 Clustering Aggregation

The second approach, called clustering aggregation, creates different clustering structures on the same dataset and the final result is obtained by selecting clusters among the structures. In (Law et al., 2004), a method called Multi-objective clustering is proposed, which applies several clustering algorithms corresponding to different objective functions and then the method picks the best set of objective functions for creating the final clustering structure. Another clustering aggregation approach is called Multi-run (Jiamthapthaksin et al., 2009), where the final clustering structure is a combination of high-quality clusters created from multiple runs; the goal is the parameter selection of a clustering algorithm.

3.1.3 Clustering Ensemble

The third approach, called clustering ensemble, combines multiple partitionings of the same object set without accessing the original features that determined the partitioning.The clustering structures can be generated in two ways: choice of objects representation or choice of clustering algorithms (Fred and Jain, 2005). This approach is focused on the merging process of the clustering structures using a hypergraph representation (Strehl and Ghosh, 2003), a co-association matrix with the similarity measure between patterns (Fred and Jain, 2005) or a probability distribution in the space of cluster labels (Topchy et al., 2005).

3.1.4 Collaborative Clustering

The last approach, called collaborative clustering, has different methods that collaborate together during a refinement step of their results and share information throughout the clustering process, to converge towards a similar result until all the results have almost the same number of clusters, and all the clusters are statistically similar. At the end of this process, as the results have comparable structures, it is possible to define a correspondence function between the clusters and to apply a voting algorithm (Gancarski and Wemmert, 2007) (Forestier et al., 2008) (Forestier et al., 2010).

3.2 Clustering Analysis applied to Sets with the Same Media Type

The clustering methods have been so far applied to sets with the same media type, such as text, image, audio and video.

3.2.1 Text Clustering

The text clustering tries to find documents with many words in common, grouping these documents into the same cluster. Feature vector is the most widely used data structure for text representation; each document is a vector in a d-dimensional space, where d is the number of features in the entire document set and the vector entries represent the importance or weight given to each feature in a specific document (Meneses, 2006). There are different techniques, called weighting models, to calculate the features weight in text documents. Some techniques are boolean weighting, frequency weighting, TF x IDF weighting and TFC weighting. The most used is the standard function TF-IDF (Term Frequency Inverse

Document Frequency) (Sebastiani, 2002). TF is the frequency of the feature in a document and IDF is the inverse frequency of the feature in all documents:

$$idf(f_l) = log\left(\frac{n}{df(f_l)}\right), \quad (1)$$

where f_l is the lth feature, n is the total number of documents, and $df(f_l)$ is the number of documents that contains the lth feature.

3.2.2 Image Clustering

The image clustering tries to find an image mapping into clusters, such that images in the same cluster have essentially the same information. The images are commonly represented in feature vectors, graphs and trees. The visual features of an image can be classified in several types, such as color, texture and logical (Choubassi et al., 2007). The first type, color features, is used to describe the color distribution of the image constructing a frequency histogram in several color spaces such as RBG, YUV and HSV. The second type, texture features, tries to find visual patterns in images searching homogeneous regions. Some texture analysis techniques are energy, entropy, inverse difference moment, inertia and correlation. The last type, logical features, contains information about objects into images and their spatial relationships. Some of these features are curvature, shape, interest points and region positions.

3.2.3 Audio Clustering

The audio clustering tries to identify and group together all speech segments that were produced by the same speaker, background conditions or channel conditions (Lu et al., 2002) (Meinedo and Neto, 2003). The audio sources can be analyzed in three layers: acoustic characteristics, audio signatures and semantic models (Liu et al., 1998). The acoustic characteristics layer analyzes low level generic features such as loudness, pitch period and bandwidth of an audio signal. The audio signature layer is an intermediate-level associated with different sounding objects. The semantic models layer is a high level analysis that uses some prior known semantic rules about the structure of audio in different scene types. In the clustering process, the audio features can be extracted in short-term frame level and long-term clip level (Wang et al., 2000). A frame is defined as a group of neighboring samples with a stationary audio signal and short-term features such as volume and Fourier transform coefficients can be extracted. A clip is defined as a sequence of frames and clip-level features usually characterizing how frame-level features change over a clip (Wang et al., 2000). Some clip level features are volume based, ZCR based, pitch based, frequency based, etc.

3.2.4 Video Clustering

The video clustering has a challenge to simultaneously handle multimode videos with three elements: images, audio and motion (Hoi and Lyu, 2008). In the process of grouping video, the features are extracted from shots or frames (Zhong and Hongjiang, 1997). The shots capture continuous action in an uninterrupted segment of video frame sequences, with or without movement, so one shot is composed of one or more frames (Ngo et al., 2001) (Yeung et al., 1996). The features extracted from the shots are related to temporal aspects such as variance and movement, while the frames are drawn from issues related to static images such as color and texture. The most discernible difference between static images and video sequences comes from the movement and changes (Dimitrova and Golshani, 1995). However, most attempts of processing video do not analyze the movement for its difficult handling.

4 AN ENSEMBLE APPROACH FOR CMO CLUSTERING

Cluster analysis applied to a CMO set has a special restriction: the method must analyze the multiple media types present in this kind of objects. In the reviewed literature, the combination of multiple clustering structures can be used when the sets are too complex to give a unique clustering structure. Within such approaches, the clustering ensemble is used when the different clustering structures are independent and complementary. Considering that media types present in CMOs have independent and complementary information, this research proposes an ensemble of clustering structures generated from several media types, issuing the following research question: Will a clustering ensemble be able to find an underlying structure in a mix of several media types present in CMOs?

A clustering ensemble for CMOs permits a standalone exploration of the different media types present in this kind of objects. In this way, the proposed ensemble can take advantage of the research advances in text, image, audio and video clustering.The proposed ensemble approach for CMO clustering is presented in figure 2 with four kinds of components: a separator, several clusterers, a mapper and a combiner.

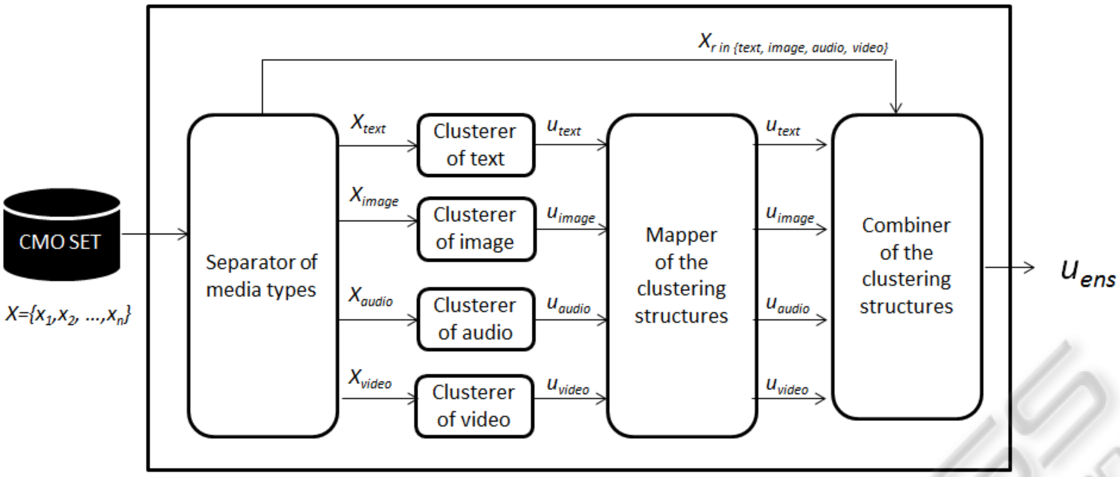

Figure 2: The proposed ensemble approach for clustering CMOs. The input is a CMO set and the output is a fuzzy matrix, expressing that objects can belong to more than one cluster with different membership degrees.

The input is a CMO set $X = \{x_1, x_2, ..., x_n\}$ and the output is the final clustering structure, which is a $n \times k$ fuzzy matrix u_{ens} expressing that objects can belong to more than one cluster with different membership degrees, where n is the number of CMO in the set and k is the number of clusters. The probability that object x_i belongs to cluster c_j is given by $u_{ens}[i, j]$, where $0 \leq u_{ens}[i, j] \leq 1$. In the fuzzy matrix the rows represent objects (i) and the columns represent clusters (j) (Zhang and Rueda, 2005)(Carvalho, 2007).

The proposed ensemble has a specific restriction: the clusterers must create clustering structures with equal number of groups k. The ensemble approach is described in the following subsections.

4.1 Separator of Media Types

In the literature review related with ensembles, the clustering structures are generated in two ways: choice of objects representation or choice of clustering algorithms (Fred and Jain, 2005). The proposed clustering ensemble creates the independent structures evaluating different media types, so the separator component preprocesses each CMO and generates a dataset X_r for each media type r in $\{text, image, audio, video\}$. Those datasets are the entries to the different clusterers and to the combinator component.

4.2 Clusterers for each Media Type

The ensemble has several clusterers that produce a clustering structure for a specific media type, where each structure is a fuzzy matrix. Let u_{text} be the fuzzy matrix generated by a clustering method with the text information of the CMO set. In the same way u_{image}, u_{audio} and u_{video} are generated. Each clustering structure must have the same number of groups k.

4.3 Mapper of the Clustering Structures

This component defines a correspondence function between the k clusters in the different clustering structures. For each clustering structure u_r, with r in $\{text, image, audio, video\}$, a vector of labels λ_r is defined, which has the cluster with highest membership value in u_r for each object. In (Forestier et al., 2010), the computation of a confusion matrix between each pair of vector λ_r is proposed for determining the cluster mapping. The confusion or matching matrix $M^{p,q}$ between the two vectors of labels λ_p and λ_q is a $k \times k$ matrix defined as:

$$M^{p,q} = \begin{pmatrix} \alpha_{1,1}^{p,q} & \ldots & \alpha_{1,k}^{p,q} \\ & \vdots & \\ \alpha_{k,1}^{p,q} & \ldots & \alpha_{k,k}^{p,q} \end{pmatrix} \qquad (2)$$

The confusion matrix represents the intersection $\alpha_{j,h}^{p,q}$ between the cluster c_j of the vector λ_p and the cluster c_h of the vector λ_q:

$$\alpha_{j,h}^{p,q} = \frac{|c_j \bigcap c_h|}{|c_j|} \qquad (3)$$

A cluster c_j is the corresponding cluster of c_h if it is the most similar to c_h. The similarity is computed observing the intersection $\alpha_{j,h}^{p,q}$ and the distribution $\rho_j^{p,q}$ of the cluster c_j in all the clusters of λ_q:

$$\rho_j^{p,q} = \sum_{t=1}^{k} (\alpha_{j,t}^{p,q})^2 \qquad (4)$$

Finally, the adequacy $\omega_{j,h}^{p,q}$ of a cluster c_j to a cluster c_h is:

$$\omega_{j,h}^{p,q} = \rho_j^{p,q} \times \alpha_{j,h}^{p,q} \tag{5}$$

Thus, the corresponding cluster of c_j in the vector of labels λ_q is the cluster c_h that maximizes the adequacy $\omega_{j,h}^{p,q}$. If there is a conflict between the corresponding clusters, the adequacy $\omega_{j,h}^{p,q}$ will resolve it by finding the next maximum. The output of the mapper component is the clustering structures u_r organized according to their corresponding clusters.

4.4 Combiner of the Clustering Structures

This component has a challenge: the clustering structures can have different amounts of data because some CMO could be incomplete. The proposed strategy is a voting function with two weights that represent the clustering structure quality and the size of the data space evaluated for each media type. The first weight, the clustering structure quality Q_r of the media type r in $\{text, image, audio, video\}$, is the value of an internal validity index for the clustering structure u_r. This weight is computed with the Xie-Beni index, which is the combination of compactness in the same cluster and separateness in different clusters:

$$Q_r = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n}(u_r[i,j])^2\|x_i - v_j\|^2}{n \cdot min_{ij}\|v_i - v_j\|^2} \tag{6}$$

where v_j is the center of cluster c_j (Xie and Beni, 1991). The second weight, $S_{r,i}$, is computed for each media type r of each object x_i. It represents the size of the data space evaluated for each media type r of the object x_i as the ratio between the amount of non-zero values in the feature vector of the media type r and the size of the complete object x_i:

$$S_{r,i} = \frac{|non\_zero(x_{r,i})|}{|x_i|} \tag{7}$$

Finally, the proposed voting function is a weighted average of the u_r fuzzy matrices:

$$u_{ens}[i,j] = \frac{1}{R}\sum_{r}((Q_r \times S_{r,i})u_r[i,j]) \tag{8}$$

where $R = 4$ is the number of fuzzy matrices and r in $\{text, image, audio, video\}$. The $u_{ens}[i,j]$ values represent the membership degree of the CMO represented by x_i to the cluster c_j. The $u_{ens}[i,j]$ values create the output of the ensemble that is a u_{ens} matrix.

5 EVALUATION

Sets with large amounts of CMOs arise in some applications, like webpage search engines, which index a large number of documents for information retrieval (Wong and Fu, 2000). Most of the webpage search engines divide the indexed documents into a number of classes. Due to the massive increase in the amount of web pages, the indexing must be developed by automatic systems through clustering analysis (Wong and Fu, 2000). Considering that a webpage is a CMO in itself, the proposed clustering ensemble is tested in a webpage set.

The goal of the evaluation is to determine which clustering approach creates structures closer to the true classification. The organization of this section is the following: the first subsection describes the clustering prototypes developed for the ensemble; the second subsection presents the experiment designs, and the last subsection presents results and a discussion about them.

5.1 Clustering Prototypes

For evaluating the proposed clustering ensemble in a CMO set, it is necessary to develop at least two clustering prototypes of different media types. In this paper are developed a text clustering prototype and an image clustering prototype.

The first prototype is developed with the text information extracted from the web pages and they are represented in feature vectors weighted with the TF x IDF function. The feature vectors for text representation have two specific problems: they are high dimensional and they possess a sparse condition (Dhillon and Modha, 2001) (Liu et al., 2003) (Feng et al., 2010). The vectors are high dimensional because they have been formed by a large number of features of the entire document set. Additionally, the vectors are very sparse because they contain few features of the total number of them in the entire document set, so the vectors have only a small number of non-zero or significant values.

For high dimensional and sparse vectors problems some authors recommend the application of a kernel method (Hashimoto et al., 2009), which formulates learning in a reproduction of the Hilbert space H of functions defined on the data domain, expanded in terms of a kernel trick. With a kernel trick, the objects to be clustered are mapped to a high dimensional feature space, computing a linear partition in the new space (Filippone et al., 2008). In applications where the dimensionality of each x_i exceeds n, a learning problem is computationally inefficient, particularly if objects are mapped into a Hilbert space (Hofmann et al., 2008). However, with the kernel trick, a set of nonlinearly separable objects can be transformed into a higher feature space dimension with the possibility to be linearly separable without knowledge about the mapping function (Xu and Wunsch, 2005).

So, the text clustering prototype is developed with a method called KFCM (kernel fuzzy c-means) (Yang et al., 2007), where the feature vectors are mapped into a high dimensional space by selecting a kernel function.Then they are separated into some clusters by the fuzzy c-means clustering algorithm.

The second prototype, image clustering, is developed with color features and with the text information related to the images. The feature vector that represents the webpage x_i is constructed in two stages. First, the content of each image is downloaded and a frequency histogram is computed in the RGB space. Then, the histograms are averaged. In the second stage, the text information related to the images is indexed with the TF x IDF function. In order to fulfill this purpose, the ”img” labels of the webpage are extracted. The final vector is a feature concatenation of the two stages creating a high dimensional data; thus the KFCM method is used to create the fuzzy clusters.

5.2 Experiment Design

For using the proposed ensemble in a CMO clustering task, four tests are developed with different datasets. The tests use a database of web pages from the open directory project (http://www.dmoz.org/), which is a human-edited directory of the web (Osinski and Weiss, 2004).The tests were conducted varying the number of webpages n in $\{30,50,90,480\}$ for contrasting the following clustering approaches:

- The text clustering.
- The image clustering.
- The joint-feature vector clustering.
- The clustering ensemble with a voting function that averages the results of the text and image clustering.
- The proposed clustering ensemble with a voting function that averages the results of the text and image clustering using both weights: the clustering structure quality Q_r and the size $S_{r,i}$ of the data space evaluated for each media type.

For comparing the clustering quality, an external index validation is calculated. In this paper, the Hubert's Γ statistic has been chosen as the index validation, because it has detected the correct number of clusters in several experiments (Hubert and Arabie, 1985). Let $C^* = \{c_1^*, c_2^*, ..., c_k^*\}$ be a clustering structure obtained for an object set X, let $C = \{c_1, c_2, ..., c_k\}$ be the real clustering structure of the object set and let Z and Y be the matrices that represent such structures:

$Z(i,j) = \{1$, if x_i and x_j belong to the same cluster in C^*, and 0 otherwise$\}, \forall i,j = 1...n$

$Y(i,j) = \{1$, if x_i and x_j belong to the same cluster in C, and 0 otherwise$\}, \forall i,j = 1...n$

Hubert's statistic is defined as:

$$\Gamma(C^*,C) = (1/M)\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Z(i,j)Y(i,j), \quad (9)$$

where M is the maximum number of pairs in the object set ($M = n(n-1)/2$) and n is the total number of objects in the X set. High values of this index indicate a strong similarity between the real clusters C and the obtained clusters C^*.

5.3 Results and Discussion

Table 1 summarizes the results for the four tests described above showing Hubert's statistic for the contrasted approaches.

The values presented in Table 1 show that the obtained results from the image clustering approach are not close to the true classification of the objects. They also show that the performance of all the approaches declines when the number of objects in the tests increase, because the dimensionality of the feature vectors increases too.

Figure 3 shows the mean values of the test results. It can be seen in the figure that the weighted average ensemble has a better performance than the average ensemble, achieving a balance between the uneven performance of the text clustering approach and the image clustering approach.

An important finding is the contrast between the joint feature vector clustering approach and the weighted average ensemble approach. The procedure of Fisher's least significant difference (LSD) is used to determine whether the means of these two approaches are significantly different. The LSD procedure indicates that there is significant statistical difference between these approaches using a confidence level of 95 percent. So, Hubert's statistic indicates that the proposed ensemble creates clustering structures more similiar to the real classification than a joint feature vector for the evaluated sets.

An interesting discussion is related with the computational complexity of the proposed ensemble,

Table 1: Hubert's statistic for the contrasted approaches. High values of this statistic indicate better results. The best values are in bold font.

Test	Number of webpages	Contrasted Approaches				
		Text clustering	Image clustering	Joint vector clustering	Ensemble: average	Proposed ensemble: weighted average
1	30	0,7793	0,4850	0,5595	0,5794	**0,7908**
2	50	0,7494	0,4736	0,5517	0,5583	**0,7632**
3	90	**0,5436**	0,0369	0,0478	0,0369	**0,5436**
4	480	**0,3565**	0,0230	0,0200	0,0230	**0,3565**

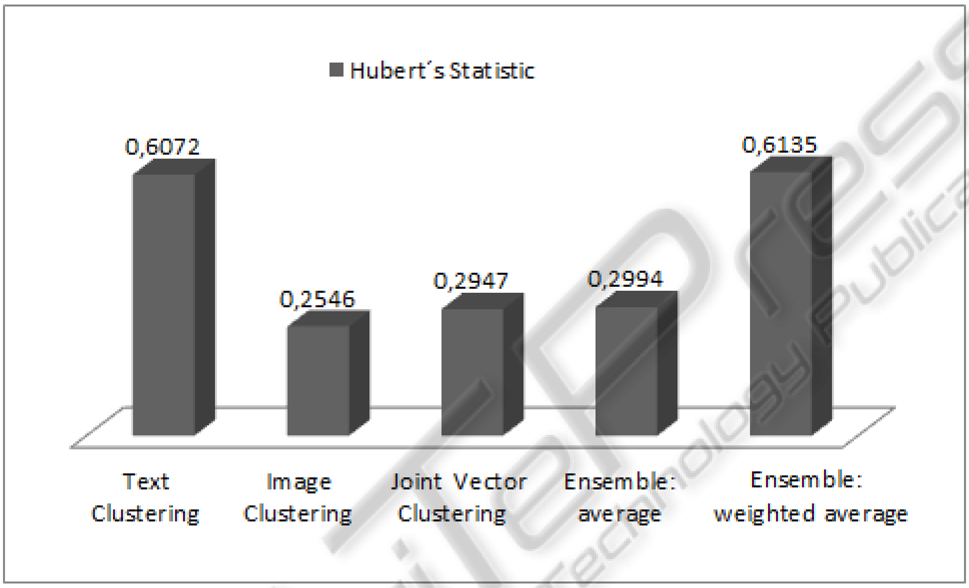

Figure 3: Mean values of Hubert's statistic. There is significant statistical difference between the joint vector clustering and the weighted average ensemble.

which remains to be cubic $O(n^3)$ in the mapper component of the ensemble. This means that when the number of clusters significantly increases, the difficulty increases in a cubic order.

6 CONCLUSIONS

This paper addresses the clustering of complex multimedia objects with a special restriction: the method must analyze different media types. An ensemble was proposed for this purpose, which generates a clustering structure for each media type of a CMO set. The clustering structures generated must be reorganized in a mapping process that defines a correspondence function between the clusters of all the structures. Finally, a voting function with a weighted average of the clustering structures generates a final result. The proposal has a restriction: the ensemble does not consider different numbers of groups in the clustering structures. A new component should be included in the proposed ensemble for considering different numbers of groups,but this is part of a future research.

The proposed ensemble was applied to cluster webpages constructing a text clustering prototype and an image clustering prototype.Hubert's statistic was used to evaluate the ensemble performance using four datasets of web pages. Results showed that the proposed ensemble creates clustering structures more similar to the real classification than a joint feature vector.

A cubic computational complexity is a disadvantage of the ensamble due to the fact that, when the number of groups increases, the complexity increases in a cubic order, so this is an emergent research line. Other future works are: to improve the clustering prototypes, to create new clustering structures with other resources, and to use larger and new datasets.

ACKNOWLEDGEMENTS

The authors would like to acknowledge support from ARTICA (Alianza Regional para las TICs Aplicadas) and the Cocreation project which is supported by Universidad de Antioquia, Universidad Nacional Sede Medellin, Universidad Pontificia Bolivariana, Universidad EAFIT, Universidad de Medelln and UNE.

REFERENCES

Algergawy, A., Schallehn, E., and Saake, G. (2008). A schema matching-based approach to xml schema clustering. In *Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services*, pages 131–136, New York, NY, USA. ACM.

Bae, E. and Bailey, J. (2006). Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. In *IEEE International Conference on Data Mining*, pages 53–62.

Caruana, R., Elhawary, M., Nguyen, N., and Smith, C. (2006). Meta clustering. In *Proceedings of the Sixth International Conference on Data Mining*, ICDM 06, pages 107–118, Washington, DC, USA. IEEE Computer Society.

Carvalho, F. (2007). Fuzzy c-means clustering methods for symbolic interval data. *Pattern Recognition Letters*, 28(4):423–437.

Choubassi, M. E., Nefian, A., Kozintsev, I., Bouguet, J., and Wu, Y. (2007). Web image clustering. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages 15–20.

Davidson, I. and Qi, Z. (2008). Finding alternative clusterings using constraints. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 773–778, Washington, DC, USA. IEEE Computer Society.

Dhillon, I. S. and Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1-2):143–175.

Dimitrova, N. and Golshani, F. (1995). Motion recovery for video content classification. *ACM Trans. Inf. Syst.*, 13:408–439.

Dy, J. and Brodley, C. (2004). Feature selection for unsupervised learning. *Journal of Machine Learning Research*, 5:845–889.

Feng, Z., Bao, J., and Shen, J. (2010). Dynamic and adaptive self organizing maps applied to high dimensional large scale text clustering. In *Software Engineering and Service Sciences ICSESS*, pages 348–351. IEEE International Conference.

Filippone, M., Camastra, F., Masulli, F., and Rovetta, S. (2008). A survey of kernel and spectral methods for clustering. *Pattern Recognition*, 41:176–190.

Forestier, G., Wemmert, C., and Gancarski, P. (2010). Towards conflict resolution in collaborative clustering. In *Intelligent Systems (IS), 2010 5th IEEE International Conference*, pages 361–366.

Forestier, G., Wemmert, C., and Gançarski, P. (2008). Multisource images analysis using collaborative clustering. *EURASIP J. Adv. Signal Process*, 2008:133:1–133:11.

Francois, O., Ancelet, S., and Guillot, G. (2006). Bayesian clustering using hidden markov random fields in spatial population genetics. *Genetics*, 174:805–816.

Fred, A. and Jain, A. (2005). Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):835–850.

Gancarski, P. and Wemmert, C. (2007). Collaborative multistep mono-level multi-strategy classification. *Multimedia Tools Appl.*, 35:1–27.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2002). Cluster validity methods: part i. *ACM SIGMOD Record*, 31(2).

Hashimoto, W., Nakamura, T., and Miyamoto, S. (2009). Comparison and evaluation of different cluster validity measures including their kernelization. *Journal of Advanced Computational Intelligence*, 13(3).

Hofmann, T., Scholkopf, B., and Smola, A. (2008). Kernel methods in machine learning. *The Annals of Statistcs*, 36(3):1171–1220.

Hoi, S. and Lyu, M. (2008). A multimodal and multilevel ranking scheme for large-scale video retrieval. *Multimedia, IEEE Transactions on*, 10:607–619.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Hunter, J. and Choudhury, S. (2003). Implementing preservation strategies for complex multimedia objects. In *Seventh European Conference on Research and Advanced Technology for Digital Libraries, ECDL 2003*, pages 473–486. Springer.

Jain, A., Murty, M., and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666.

Jiamthapthaksin, R., Eick, C. F., and Rinsurongkawong, V. (2009). An architecture and algorithms for multi-run clustering. In *Computational Intelligence Symposium on Data Mining CIDM 09*, pages 306–313.

Kriegel, H.-P., Kunath, P., Pryakhin, A., and Schubert, M. (2008). Distribution-based similarity for multi-represented multimedia objects. In *Proceedings of the 14th international conference on Advances in multimedia modeling*, MMM 08, pages 155–164, Berlin, Heidelberg. Springer-Verlag.

Law, M. H. C., Topchy, A. P., and Jain, A. K. (2004). Multiobjective data clustering. In *Proceedings of the 2004 IEEE computer society conference on Computer vision and pattern recognition*, CVPR 04, pages 424–430, Washington, DC, USA. IEEE Computer Society.

Liu, T., Liu, S., Chen, Z., and Ma, W. (2003). An evaluation on feature selection for text clustering. In *Proceedings of the 20th International Conference on Machine Learning*, pages 448–495. AAAI Press.

Liu, Z., Wang, Y., and Chen, T. (1998). Audio feature extraction and analysis for scene segmentation and classification. In *Journal of VLSI Signal Processing System*, volume 20, pages 61–79.

Lu, L., Zhang, H.-J., Member, S., and Jiang, H. (2002). Content analysis for audio classification and segmentation. *IEEE Transactions on Speech and Audio Processing*, 10(4):504–516.

Meinedo, H. and Neto, J. (2003). Audio segmentation, classification and clustering in a broadcast news task. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*

Meneses, E. (2006). Vectors and graphs: Two representations to cluster web sites using hyperstructure. In *Latin American Web Congress*, pages 20–25.

Ngo, C.-W., Pong, T.-C., and Zhang, H.-J. (2001). On clustering and retrieval of video shots. In *Proceedings of the ninth ACM international conference on Multimedia*, MULTIMEDIA 01, pages 51–60, New York, NY, USA. ACM.

Osinski, S. and Weiss, D. (2004). Conceptual clustering using lingo algorithm: Evaluation on open directory project data. In *IIPWM04*, pages 369–377.

Romesburg, C. (2004). *Cluster Analysis for Researchers*. Lulu Press.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Strehl, A. and Ghosh, J. (2003). Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617.

Topchy, A., Jain, A. K., and Punch, W. (2005). Clustering ensembles: Models of consensus and weak partitions. *IEEE Transactions on pattern analysis and machine intelligence*, 27:1866–1881.

Wang, Y., Liu, Z., and Huang, J.-C. (2000). Multimedia content analysis using both audio and visual cues. *IEEE Signal Processing Magazine*, 17(6):12–36.

Wong, W. and Fu, A. (2000). Incremental document clustering for web page classification. In *In IEEE 2000 Int. Conf. on Info. Society in the 21st*, pages 5–8.

Xie, X. and Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(4):841–846.

Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Trans. Neural Networks*, 16(3):645–667.

Yang, A., Jiang, L., and Zhou, Y. (2007). A kfcm-based fuzzy classifier. In *Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery - Volume 02*, FSKD 07, pages 80–84, Washington, DC, USA. IEEE Computer Society.

Yang, Y., Zhuang, Y.-T., Wu, F., and Pan, Y.-H. (2008). Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. *IEEE Transactions on Multimedia*, 10(3):437–446.

Yeung, M., Yeo, B., and Liu, B. (1996). Extracting story units from long programs for video browsing and navigation. In *Proceedings of the 1996 International Conference on Multimedia Computing and Systems*, pages 296–305, Washington, DC, USA. IEEE Computer Society.

Zhang, Y. and Rueda, L. (2005). A geometric framework to visualize fuzzy-clustered data. In *Chilean Computer Science Society, SCCC.*

Zhong, D. and Hongjiang, D. Z. (1997). Clustering methods for video browsing and annotation. Technical report, In SPIE Conference on Storage and Retrieval for Image and Video Databases.

Zhuang, Y., Yi, Y., and Fei, W. (2008). Mining semantic correlation of heterogeneous multimedia data for cross- media retrieval. *IEEE Transactions on Multimedia*, 10(2):221–229.