MULTIMEDIA MINING

S. Kotsiantis[1], D. Kanellopoulos[2], P. Pintelas[3]
[1] Department of Mathematics, University of Patras, Patras 26 500, Greece, Email: sotos@math.upatras.gr
[2] Department of Electrical Engineering & Computer Technology, University of Patras, Patras 26 500, Greece, Email: dkanellop@teipat.gr
[3] Department of Mathematics, University of Patras, Patras 26 500, Greece, Email: pintelas@math.upatras.gr

Abstract: - Advances in multimedia acquisition and storage technology have led to tremendous growth in very large and detailed multimedia databases. If these multimedia files are analyzed, useful information to users can be revealed. Multimedia mining deals with the extraction of implicit knowledge, multimedia data relationships, or other patterns not explicitly stored in multimedia files. Multimedia mining is more than just an extension of data mining, as it is an interdisciplinary endeavor that draws upon expertise in computer vision, multimedia processing, multimedia retrieval, data mining, machine learning, database and artificial intelligence. This paper briefly describes the multimedia mining, while references cited cover the major theoretical issues.

Key-Words: - text mining; image mining; audio mining; video mining

1 Introduction

In digital data acquisition and storage technology, the rapid progress has led to the fast growing tremendous and amount of data stored in databases. Although valuable information may be hiding behind the data, the overwhelming data volume makes it difficult (if not impossible) for human beings to extract them without powerful tools. Multimedia mining systems that can automatically extract semantically meaningful information (knowledge) from multimedia files are increasingly in demand. For this reason, a large number of techniques have been proposed ranging from simple measures (e.g. color histogram for image, energy estimates for audio signal) to more sophisticated systems like speaker emotion recognition in audio [1], automatic summarization of TV programs [2].

Generally, multimedia database systems store and manage a large collection of multimedia objects, such as image, video, audio and hypertext data [3]. Thus, in multimedia documents, knowledge discovery deals with non-structured information. For this reason, we need tools for discovering relationships between objects or segments within multimedia document components, such as classifying images based on their content, extracting patterns in sound, categorizing speech and music, and recognizing and tracking objects in video streams.

In general, the multimedia files from a database must be first preprocessed to improve their quality. Subsequently, these multimedia files undergo various transformations and features extraction to generate the important features from the multimedia files. With the generated features, mining can be carried out using data mining techniques to discover significant patterns. These resulting patterns are then evaluated and interpreted in order to obtain the final application's knowledge.

The following section describes the application process for multimedia mining. In Section 3, the features extraction from multimedia data is analyzed. Section 4 presents data preprocessing that includes data cleaning, normalization, transformation and feature selection. In Section 5, the supervised and unsupervised models, which are used for multimedia mining are described. In Section 6, some applications of multimedia mining are referred, while some open problems are mentioned in the last section.

2 Process of application of multimedia mining

In Figure 1, we present the model of applying multimedia mining in different multimedia types. Data collection is the starting point of a learning system, as the quality of raw data determines the overall achievable performance. Then, the goal of data pre-processing is to discover important features from raw data. Data pre-processing includes data cleaning, normalization, transformation, feature selection, etc. Learning can be straightforward, if informative features can be identified at pre-processing stage. Detailed procedure depends highly on the nature of raw data and problem's domain. In some cases, prior knowledge can be extremely valuable. For many systems, this stage is still primarily conducted by domain experts.

The product of data pre-processing is the training set. Given a training set, a learning model has to be chosen to learn from it. It must be mentioned that the steps of multimedia mining are often iterative. The

analyst can also jump back and forth between major tasks in order to improve the results.

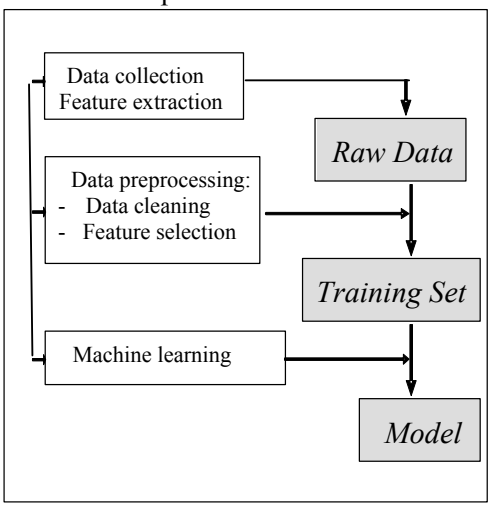

Figure 1. Multimedia mining process

Compared with data mining, multimedia mining reaches much higher complexity resulting from: a) the huge volume of data, b) the variability and heterogeneity of the multimedia data (e.g. diversity of sensors, time or conditions of acquisition etc) and c) the multimedia content's meaning is subjective.
The high dimensionality of the feature spaces and the size of the multimedia datasets make the feature extraction a challenging problem. In the following Section, we analyze the feature extraction process for multimedia data.

3 Feature extraction

There are two kinds of features: description-based and content-based. The former uses metadata, such as keywords, caption, size and time of creation. The later is based on the content of the object itself [3].

3.1 Feature extraction from text

Text categorization is a conventional classification problem applied to the textual domain. It solves the problem of assigning text content to predefined categories. In the learning stage, the labeled training data are first pre-processed to remove unwanted details and to "normalize" the data [4]. For example, in text documents punctuation symbols and non-alphanumeric characters are usually discarded, because they do not help in classification. Moreover, all characters are usually converted to lower case to simplify matters. The next step is to compute the features that are useful to distinguish one class from another. For a text document, this usually means identifying the keywords that summarize the contents of the document. How are these keywords learned? One way is to look for words that occur frequently in the document. These words tend to be what the document is about. Of course, words that occur too frequently, such as "the", "is", "in", "of" are no help at all, since they are prevalent in every document. These common English words may be removed using a "stop-list" of words during the pre-processing stage. From the remaining words, a good heuristic is to look for words that occur frequently in documents of the same class, but rarely in documents of other classes. In order to cope with documents of different lengths, relative frequency is preferred over absolute frequency [4].
Some authors used phrases, rather than individual words, as indexing terms [5], but the experimental results found to date have not been uniformly encouraging results. Another problem of text is the variant. Variant refers to the different forms of the same word, e.g. "go", "goes", "went", "gone", "going". This may be solved by stemming, which means replacing all variants of a word by a standard one [4].

3.2 Feature extraction from images

Image categorization classifies images into semantic databases that are manually precategorized. In the same semantic databases, images may have large variations with dissimilar visual descriptions (e.g. images of persons, images of industries etc.). In addition images from different semantic databases might share a common background (some flowers and sunset have similar colors).
In [6], the authors distinguish three types of feature vectors for image description: 1) pixel level features, 2) region level features, and 3) tile level features. Pixel level features store spectral and textural information about each pixel of the image. For example, the fraction of the endmembers, such as concrete or water, can describe the content of the pixels. Region level features describe groups of pixels. Following the segmentation process, each region is described by its boundary and a number of attributes, which present information about the content of the region in terms of the endmembers and texture, shape, size, fractal scale etc [6]. Tile level for image features present information about whole images using texture, percentages of endmembers, fractal scale and others.
Moreover, other researchers proposed an information-driven framework that aims to highlight the role of information at various levels of representation [7]. This framework adds one more level of information: the Pattern and Knowledge Level that integrates domain, related alphanumeric data and the semantic relationships discovered from the image data.

3.3 Feature extraction from Audio

Audio data play an important role in multimedia applications. Music information has two main branches: symbolic and audio information. Attack, duration, volume, velocity and instrument type of every single note are available information. Therefore, it is possible to easily access statistical measures such as tempo and mean key for each music item. Moreover, it is possible to attach to each item high-level descriptors, such as instrument kind and number. On the other hand, audio information deals with real world signals and any features need to be extracted through signal analysis.

The researchers of [8] used only perceptual features such as loudness, brightness, pitch etc. On the other hand, other researchers chose only perceptual features to represent sound clips [9]. Another researcher team used 12 cepstral features, as well [10].

However, some of the most frequently used features for audio classification are [11], [12]:

- Total Energy: The temporal energy of an audio frame is defined by the rms of the audio signal magnitude within each frame.
- Zero Crossing Rate (ZCR): ZCR is also a commonly used temporal feature. ZCR counts the number of times that an audio signal crosses its zero axis.
- Frequency Centroid (FC): It indicates the weighted average of all frequency components of a frame.
- Bandwidth (BW): Bandwidth is the weighted average of the squared differences between each frequency component and its frequency centroid.
- Pitch Period: It is a feature that measures the fundamental frequency of an audio signal.

3.4 Feature extraction from Video

In video mining, there are three types of videos: a) the produced (e.g. movies, news videos, and dramas), b) the raw (e.g. traffic videos, surveillance videos etc), and c) the medical video (e.g. ultra sound videos including echocardiogram).

Higher-level information from video includes:

- detecting trigger events (e.g. any vehicles entering a particular area, people exiting or entering a particular building)
- determining typical and anomalous patterns of activity, generating person-centric or object-centric views of an activity
- classifying activities into named categories (e.g. walking, riding a bicycle),
- clustering and determining interactions between entities [13].

The first stage for mining raw video data is grouping input frames to a set of basic units, which are relevant to the structure of the video. In produced videos, the most widely used basic unit is a shot, which is defined as a collection of frames recorded from a single camera operation. Shot detection methods can be classified into many categories: pixel based, statistics based, transform based, feature based and histogram based [14]. Color or grayscale histograms (such as in image mining) can also be used [15]. To segment video, color histograms, as well as motion and texture features can be used [16].

Generally, if the difference between the two consecutive frames is larger than a certain threshold value, then a shot boundary is considered between two corresponding frames. The difference can be determined by comparing the corresponding pixels of two images [17]. A set of other computational features derived from cinematic editing effects, motion and colors in videos are presented in [18].

4 Data Preprocess

In a multimedia database, there are numerous objects that have many different dimensions of interests. For example, only the color attribute can have 256 dimensions, with each counting the frequency of a given color in images. The image may still have other dimensions.

Selecting a subset of features is a method for reducing the problem size [21]. This reduces the dimensionality of the data and enables learning algorithms to operate faster and more effectively. The problem of feature interaction can also be addressed by constructing new features from the basic features set. This technique is called feature construction/transformation [22].

Sampling is also well accepted by the statistics community that argues “a powerful computationally intense procedure operating on a sub-sample of the data may in fact provide superior accuracy than a less sophisticated one using the entire data base” [19]. Moreover, discretization can significantly reduce the number of possible values of the continuous feature, as large number of possible feature values contributes to slow and ineffective process of machine learning [20]. Furthermore, normalization (“scaling down" transformation of the features) is also beneficial since there is often a large difference between the maximum and minimum values of the features.

5 Models for multimedia mining

Multimedia classification and clustering are the supervised and unsupervised classification of multimedia files into groups.

5.1 Classification models

Machine learning (ML) and meaningful information extraction can only be realized, when some objects have been identified and recognized by the machine. The object recognition problem can be referred as a supervised labeling problem. Starting with the supervised models, we mention the decision trees. An overview of existing works in decision trees is provided in [23]. Decision trees can be translated into a set of rules by creating a separate rule for each path from the root to a leaf in the tree. However, rules can also be directly induced from training data using a variety of rule-based algorithms. An excellent overview of existing works in rule based methods is given in [24].

Artificial Neural Networks (ANNs) are another method of inductive learning, based on computational models of biological neurons and networks. A recent overview of existing works in ANNs is given in [25]. A Bayesian network [26] is a graphical model for probabilistic relationships among a set of features.

Instance-based learning algorithms are lazy-learning algorithms [27] as they delay the induction or generalization process until classification is performed. During the training phase, the lazy-learning algorithms require less computation time than eager-learning algorithms (e.g. decision trees, neural and Bayes nets). However, during the classification process, they require more computation time.

The Support Vector Machines (SVMs) is the newest technique that considers the notion of a "margin". Maximising the margin and thereby creating the largest possible distance between the separating hyperplane and the instances on either side of it, is proven to reduce an upper bound on the expected generalisation error [28].

5.2 Clustering Models

In unsupervised classification, the problem is to group a given collection of unlabeled multimedia files into meaningful clusters according to the multimedia content without a priori knowledge. Clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods. An excellent survey of clustering techniques can be found in [29].

Partitioning methods are divided into two major subcategories, the centroid and the medoids algorithms. The centroid algorithms represent each cluster by using the gravity centre of the instances. The medoid algorithms represent each cluster by means of the instances closest to the gravity centre. The hierarchical methods group data instances into a tree of clusters [30]. Density-based clustering algorithms try to find clusters based on density of data points in a region. The key idea of density-based clustering is that, for each instance of a cluster, the neighborhood of a given radius has to contain at least a minimum number of instances [31]. Grid-based clustering algorithms first quantize the clustering space into a finite number of cells (hyper-rectangles) and then perform the required operations on the quantized space. Cells that contain more than certain number of points are treated as dense and the dense cells are connected to form the clusters [32].

5.3 Association rules

The most association rules studies have been focusing on the corporate data typically in alphanumeric databases [33]. There are three measures of the association: support, confidence and interest. The support factor indicates the relative occurrence of both X and Y within the overall data set of transactions. It is defined as the ratio of the number of instances satisfying both X and Y over the total number of instances. The confidence factor is the probability of Y given X and is defined as the ratio of the number of instances satisfying both X and Y over the number of instances satisfying X. The support factor indicates the frequencies of the occurring patterns in the rule, and the confidence factor denotes the strength of implication of the rule. The interest factor is a measure of human interest in the rule. For example, a high interest means that if a transaction contains X, then it is much more likely to have Y than the other items.

Relatively little research has been conducted on mining multimedia data [34]. There are different types of associations: association between image content and non image content features. For example, if the upper part of the picture is at least 50% blue, it is likely to represent sky. Association mining in multimedia data can be transformed into problems of association mining in traditional transactional databases.

The image is can be modeled as a transaction, assigned with an ImageID, and the features of the images are the items contained in the transaction. Therefore, mining the frequently occurring patterns among different images becomes mining the frequent patterns in a set of transactions. In [35], the authors extend the concept of content-based multimedia association rules using feature localization. They

introduced the concept of progressive refinement in discovery of patterns in images.

6 Applications and systems

Satellite data is used in many different areas ranging from agriculture, forestry, and environmental studies. The applications using satellite data include measurements of crop and timber acreage, forecasting crop yields and forest harvest, monitoring urban growth, mapping of ice for shipping, mapping of pollution, recognition of certain rock types, and many others. For example, the CONQUEST system [36] combines satellite data with geophysical data to discover patterns in global climate change. The SKICAT system [37] integrates techniques for image processing and data classification in order to identify 'sky objects' captured in a very large satellite picture set. An example of video and audio data mining can be found in the Mining Cinematic Knowledge project [38], which created a movie mining system by examining the suitability of existing concepts in data mining to multimedia.

Moreover, the analysis and mining of traffic video sequences in order to discover information (such as vehicle identification, traffic flow, and the spatio-temporal relations of the vehicles at intersections) provide an economic approach for daily traffic operations. There are some multimedia data mining frameworks [39], [40] for traffic monitoring systems. Furthermore, various methods for the detection of faces in images and image sequences are reported in [41].

Detection of generic sport video documents seems almost impossible due to the large variety in sports. However, some authors presented a method that is capable of identifying mainstream sports videos [42].

7 Conclusion

This paper describes well known techniques for multimedia mining. In text mining there are two open problems: polysemy, synonymy. Polysemy refers to the fact that a word can have multiple meanings. Distinguishing between different meanings of a word (called word sense disambiguation) is not easy, often requiring the context in which the word appears. Synonymy means that different words can have the same or similar meaning.

In audio and video mining, a fundamental open problem also remains: The combination of information across multiple media (combining video and audio information into one comprehensive score).

In image mining an open problem remains: the combination of different types of image data. Documents from an OCR library and a video library need to be presented in a single ranked list.

References:

[1] Petrushin, V.A., Emotion Recognition in Speech Signal: Experimental Study, Development, and Application, *Proc. 6th International Conference on Spoken Language Processing (ICSLP 2000)*, Beijing, 2000. Vol. IV, pp 222–228.

[2]Maybury M.T. (Ed.) *Intelligent Multimedia Information Retrieval*, AAAI Press/MIT Press, Menlo Park, CA / Cambridge, MA, 1997.

[3]Yoshitaka A. and Ichikawa T., A survey on content-based retrieval for multimedia databases. *IEEE Trans.on Knowledge and Data Engineering*, Vol 11, 1999, pp. 81-93.

[4]Sebastiani Fabrizio, Machine learning in automated text categorization. *ACM Computing Surveys*, Vol 34, 2002, pp. 1-47.

[5]Schutze, H., Automatic word sense discrimination. Computational Ling., Vol 24, 1998, pp. 97–124.

[6]Zhang Ji, Wynne Hsu, Mong Li Lee. Image Mining: Issues, Frameworks and Techniques, in *Proc. of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001)*, San Francisco, CA, USA, 2001, pp. 13-20.

[7]Zhang J., W. Hsu and M. L. Lee. An Information-driven Framework for Image Mining, in *Proc. of 12th International Conference on Database and Expert Systems Applications*, Munich, 2001, pp. 232 - 242 .

[8]Wold E., T. Blum, D. Keislar, and J. Wheaton, Content-based classification, search and retrieval of audio, *IEEE Multimedia Magazine*, vol. 3, 1996, pp. 27-36.

[9]Liu Z., J. Huang, Y. Wang, and T. Chen, Audio Feature Extraction and Analysis for Scene Segmentation and Classification, *Journal of VLSI Signal Processing*, Vol. 20, 1998, pp.61-79.

[10]Foote J. et al, Content-based retrieval of music and audio, *Multimedia Storage Archiving Syst. II*, vol. 3229, 1997, pp. 138-147.

[11]Wang Y., Z. Liu, and J.-C. Huang, Multimedia Content Analysis, *IEEE Signal Processing Magazine*, Nov. 2000, pp. 12-36.

[12]Uitdenbogerd A. L., van Schyndel R. G., A review of factors affecting music recommender success, *3rd International Conference on Music Information Retrieval*, Paris, 2002, pp. 204-208.

[13]Rosenfeld A., D. Doermann, D. DeMenthon, Eds., *Video Mining*, Kluwer, 2003.

[14]Boresczky J. S. and L. A. Rowe, A comparison of video shot boundary detection techniques, *Storage & Retrieval for Image and Video*

Databases IV, Proc. SPIE 2670, 1996, pp. 170-179.

[15]Ardizzone E. and M. Cascia. Automatic video database indexing and retrieval. *Multimedia Tools and Applications*, Vol. 4, 1997, pp. 29–56.

[16]Yu H. and W.Wolf. A visual search system for video and image databases. In *Proc. IEEE Int'l Conf. On Multimedia Computing and Systems*, Ottawa, Canada, June 1997, pp. 517–524.

[17] Zhang, H.J., Low, C.Y., Smoliar, S.W. and Wu, J.H., Video parsing, retrieval and browsing: an integrated and content-based solution, *Proc. ACM Multimedia '95*, pp. 15-24.

[18]Truong B.T. and C. Dorai, Automatic genre identification for content-based video categorization, *Proc. of ICPR' 2000*, pp 230-33.

[19]Friedman, J.H., Data mining and statistics: What's the connection?, Computing Science and Statistics, Vol. 29, No. 1, 1998, pp. 3-9.

[20]Berka, P. & Bruha, I., Discretization and grouping: Preprocessing steps for data mining. *Lecture Notes in AI*, Springer, 1998, pp. 239-245.

[21]Liu, H. & Motoda H., *Feature Extraction, Construction and Selection: A Data Mining Perspective*, Kluwer, 1998.

[22]Markovitch S. & Rosenstein D., Feature Generation Using General Construction Functions, *Machine Learning*, Vol 49, 2002, pp. 59-98.

[23]Murthy, Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey, *Data Mining and Knowledge Discovery*, Vol. 2, 1998, pp. 345–389.

[24]Furnkranz, J., Separate-and-Conquer Rule Learning. *Artificial Intelligence Review*, Vol. 13, 1999, pp. 3- 54.

[25]Neocleous, C. & Schizas, C., Artificial Neural Network Learning: A Comparative Review, *LNAI* 2308, Springer, 2002, pp. 300–313.

[26]Jensen, F.. *An Introduction to Bayesian Networks*. Springer, 1996.

[27]Aha, D., *Lazy Learning*. Dordrecht: Kluwer Academic Publishers, 1997.

[28]Cristianini, N. & Shawe-Taylor, J., *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, 2000.

[29]Jain A.K., Murty M.N., Flyn, P.J. (1999), Data Clustering: A Review, *ACM Computing Surveys*, Vol. 31, No. 3, 1999, pp. 264 - 323.

[30]Zhang, T., Ramakrishnan, R., and Linvy, M., BIRCH: An efficient data clustering method for very large data sets. *Data Mining and Knowledge Discovery*, Vol. 1, 1997, pp. 141–182.

[31]Ester, M., Kriegel, H.-P., Sander, J., and Xu X., A density-based algorithm for discovering clusters in large spatial data sets with noise. *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining*. Portland, 1996, pp. 226–231.

[32]Wang W., Yang J. and Muntz.R. (1997), STING: A Statistical Information Grid Approach to Spatial Data Mining, *Proc. of the 23rd VLDB Conference,* Athens, Greece, 1997, pp. 186-195.

[33]Agrawal R., R. Srikant, Fast Algorithms for Mining Association Rules, *Proc. of the 20th Int'l Conference on Very Large Databases*, Santiago, Chile, 1994, pp. 487-499.

[34]Shyu Mei-Ling, Shu-Ching Chen and R. L. Kashyap, Generalized Affinity-Based Association Rule Mining for Multimedia Database Queries, *Knowledge and Information Systems,* Vol, 3, 2001, pp. 319-337.

[35]Za O., a Han, H. Zhu, In mining recurrent items in multimedia with progressive resolution refinement. *Proc. of the IEEE International Conference on Data Engineering*, 2000, pp. 461-470.

[36]Stolorz P., H. Nakamura, E. Mesrobian, R. Muntz, E. Shek, J. Santos, J Yi, K Ng, S. Chien, Mechoso C., and J. Farrara. Fast spatio-temporal data mining of large geophysical datasets. In *Proc. of Int'l Conf. on KDD*, 1995, pp. 300–305.

[37]Fayyad U., S. Djorgovski, and N. Weir. Automating the analysis and cataloging of sky surveys. *Advances in Knowledge Discovery with Data Mining*, 1996, pp. 471–493.

[38]Wijesekera D. and D. Barbara. Mining cinematic knowledge: Work in progress. In P*roc. of International Workshop on Multimedia Data Mining (MDM/KDD'2000)*, Boston, pp. 98–103.

[39]Dailey D., F. Cathey, and S. Pumrin. An algorithm to estimate mean traffic speed using uncalibrated cameras. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 1, 2000, pp. 98–107.

[40]Cucchiara R., M. Piccardi, and P. Mello. Image analysis and rule-based reasoning for a traffic monitoring system. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 1, June 2000, pp. 119–130.

[41]Snoek C.G.M. and M. Worring, Multimodal Video Indexing: A Review of the State-of-the-art *Multimedia Tools and Applications*, Vol. 25, No. 1, 2005, pp. 5-35.

[42]Kobla V., D. DeMenthon, and D. Doermann. Identification of sports videos using replay, text, and camera motion features. *In SPIE Conference on Storage and Retrieval for Media Databases*, volume 3972, 2000, pp. 332-343.