From Multimedia Data Mining to Multimedia Big Data Mining

Gradinaru Bogdănel Constantin, Danubianu Mirela, Bărîla Adina Luminița
bogdan.gradinaru@usm.ro mirela.danubianu@usm.ro adina.barila@usm.ro
Department of Computers, Ștefan cel Mare University of Suceava, 13 University Street, Suceava, 720229, Romania

Summary
With the collection of huge volumes of text, image, audio, video or combinations of these, in a word multimedia data, the need to explore them in order to discover possible new, unexpected and possibly valuable information for decision making was born. Starting from the already existing data mining, but not as its extension, multimedia mining appeared as a distinct field with increased complexity and many characteristic aspects. Later, the concept of big data was extended to multimedia, resulting in multimedia big data, which in turn attracted the multimedia big data mining process. This paper aims to survey multimedia data mining, starting from the general concept and following the transition from multimedia data mining to multimedia big data mining, through an up-to-date synthesis of works in the field, which is a novelty, from our best of knowledge.
Keywords:
multimedia big data, multimedia data, multimedia data mining

1. Introduction

Since the emergence of high-performance data processing and storage systems, and with the spread of online services and mobile technologies, there has been a real revolution in multimedia data and their place in the field of information technology.

In many areas such as advertising and marketing, medicine, entertainment, education and training, surveillance, remote sensing, multimedia plays an important role. In addition, the most popular sites such as Yahoo, iCloud, YouTube and social networks Twitter, Facebook or Instagram combine text with images, video, and audio sequences, being valuable sources of multimedia data.

The volume of stored multimedia data is impressive, but the lack of methods and tools to analyze them and extract information, often hidden, leads to the so called "data rich, information poor" (DRIP) phenomenon, which assumes that although organizations are rich in data, they cannot produce valuable information that can give them a competitive advantage.

So, with such a rich content of multimedia data, there is a need for continuous development of the methods such as, processing, modeling, extracting, organizing and indexing this data for searching, retrieving, delivering, efficiently managing and sharing information from multimedia content, according to the requirements of every field.

The solution to these problems is Knowledge Discovery in Data (KDD) systems, dedicated in this case to multimedia data. The central step in this complex process of KDD aims to build patterns through multimedia data mining algorithms, briefly named multimedia mining.

Multimedia mining uses data mining methods and techniques adapted to the multimedia data characteristics: large volumes, variability, complexity, and heterogeneity. For this reason, it has a much greater complexity compared to data mining.

Subjectivism is the most important feature for understanding multimedia content. Multimedia mining is more than an extension of data mining, it is an interdisciplinary activity based on multimedia previewing, processing and retrieval, machine learning and artificial intelligence.

As we stated earlier, the explosive development at relatively low cost of modern technologies has led to people spending consistent time on the internet and social networks sharing information and creating multimedia data [1]. The technological advance materialized in more and more sophisticated devices used in medicine, education, intelligent process control, etc. it has also attracted the ability to generate and store impressive volumes of multimedia data. Given that, these data are characterized, in addition to volume and a wide variety of types and sources, high speed, often volatility can be said, without mistake, multimedia data meet the characteristics of Big Data [2].

This paper aims to provide a concise survey of the evolution of knowledge discovery in multimedia data, from multimedia data mining to multimedia big data mining. We started from the general concepts of knowledge discovery in data and data mining, which we customized in a multimedia context. Beyond this study, another element of originality of our work is the period considered, and according to our research, there are no other such works addressing aspects of this evolution until present.

Manuscript received November 5, 2022
Manuscript revised November 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.11.55

The paper is structured in five chapters as follows: Section 2 presents a short introduction in KDD and marks the position of data mining in this process, Section 3 refers to the concept of multimedia mining, and highlighting both the characteristics of multimedia data and multimedia Big data mining, presents the particularities induced by them to the modeling and pattern discovery activity, Section 4 shows an evolutive state-of-the art for multimedia data mining vs. multimedia Big Data mining and Section 5 ends with some conclusion and future work.

2. Knowledge discovery in databases and data mining

The need to make the right decisions in an extremely dynamic society has led to the need to have a correct, consistent and current information support, which involves, among other things, finding ways to extract valuable information, often hidden, from large data sets.

Defined as the process of ”identifying valid, novel, potentially useful, and understandable patterns in data”, Knowledge Discovery in Databases is a complex, iterative and interactive process. Although it targets databases as the primary source of data, it can be generalized to other sources as well [3].

The main purpose of research in this field is to develop computational methods in order to acquire basic and factual knowledge from huge amounts of data.

KDD is a complex, iterative and interactive process that relies on methods and techniques taken and adapted from different fields such as databases, statistics, artificial intelligence, machine learning and visualization.

Over time, several models of the KDD process have been developed. First of these models is the academic one, developed by Fayyad et al., and used in various domains, including engineering, medicine, e- business, or software development. Fayyad’s model contains nine stages: developing and understanding the application domain, creating a target data set, data cleaning and pre-processing, data reduction and projection, choosing the data mining task, choosing the data mining algorithm, data mining, interpreting mined patterns and consolidating discovered knowledge [4].

This was quickly followed by some industrial models. The most known, CRISP-DM, was developed by a large consortium of European companies and it has become the leading industrial model. It is thought of as a succession of six steps: business understanding, data understanding, data preparation, modeling, evaluation of the model and deployment [5].

Fig. 1 presents a conceptual model of the KDD process, referring to mainly the stages that involve the actual work with the data and their course from raw data to knowledge. The data selection phase involves choosing of a database or selecting of a subset of fields or records as the data source to be used for data mining. Adequate knowledge and understanding of the field are the foundation for identifying useful data. The quality and quantity of raw data are essential elements for a good performance of the whole process.

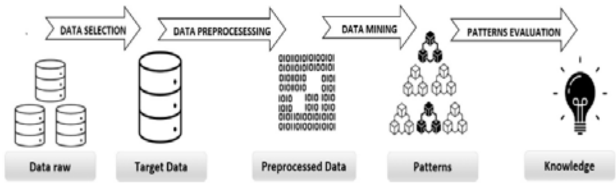

Fig. 1.KDD process conceptual model

The data selection phase involves choosing of a database or selecting of a subset of fields or records as the data source to be used for data mining. Adequate knowledge and understanding of the field are the foundation for identifying useful data. The quality and quantity of raw data are essential elements for a good performance of the whole process.

The preprocessing stage is dedicated, on the one hand, to finding the important characteristics from the raw data and, on the other hand, to bringing data in the suitable forms for the application of data mining algorithms, and it involves operations such as: integrating data from different sources, the choose of the most appropriate ways for representing or coding certain data or reducing the data sets dimensionalities.

Data cleaning eliminates data noise or solves the missing values problems. When there are large differences between the maximum and minimum values of the data, normalization has the role of reducing this values range.

Data transformation can take various forms. The most common are aimed at either dividing a field into several features or merging several fields. In both situations the result is the construction of a new feature that can have a stronger semantic value and can allow the generation of knowledge with increased semantic significance.

The processing of large data sets requires the development of efficient and robust algorithms. New technologies in the domain of databases must also be developed and discovered for the interactive, optimal and easy exploration of data but also for the discovery of knowledge. The rapid and steady growth of data is a challenge for their analysis and is based on models with a high level of complexity. But frequently, used data mining techniques are not effective for large data and for this reason we are constantly looking for solutions to reduce their size and dimensionality.

The dataset dimensionality is reduced by filtering or sampling, by feature selection and composition, or by feature extraction and, as result, the modelling process is faster and more efficient [8]. Feature selection aims to reduce a dataset dimensionality by identifying and removing as much irrelevant and redundant features as possible with respect to the task to be executed.

This could be a fully automatic process, and it often brings more benefits for data mining, such as: improved predictive accuracy, and reduced execution time for algorithms. There are two broad categories of feature selection methods – filters and wrappers [8].

The simplest approaches to feature selection are filters, known as open loop feature selection methods. Performing features selection through class separability criteria, they do not consider the effect of selected features on the performance of the whole process of knowledge discovery, and they usually provide a ranked list of features that are ordered according to a specific evaluation criterion such as: accuracy and consistency of data, information content or statistical dependencies between features [11].

They give also comparative information about the relevance of different features, but they don't indicate the desirable minimum set of the features [12].

Wrappers or closed loop feature selection methods consider the performance of the selected features set for the complete KDD process. Using as selection criteria the performance of prediction, wrappers assess the quality of selected features set through a comparation between the prediction algorithm performances when it is applied on the original set of features and on the reduced one [8].

For small data sets, the selection criteria of the features used do not require predictive evaluation. For this reason, filters do not consider changes in the performance of the knowledge discovery process.

The heart of the KDD process is the modelling or the data mining step. In this stage, the hidden patterns, relationships, and trends in the data are discovered, based on efficient algorithms. Pattern discovery may be approached by several techniques: classification, regression, clustering, association, time-series analysis, and visualization. Each of above mentioned can be implemented through various ways such as neural networks, fuzzy logic, pattern recognition or machine learning.

The last stage, strongly application dependent, aims to evaluate the quality of pattern discovered to decide if it is useful and it can be categorized as a new knowledge or not.

Although KDD is the whole process of extracting knowledge from data, there is still confusion about the difference between this concept and the data mining term. As shown in Fig. 1, and as it is evident from the study of the process models, data mining should be used only for a single stage of the KDD process.

Specifically, data mining involves the use of algorithms that produce a particular enumeration of patterns from large volumes of data, facilitating the discovery of seemingly unrelated data, relationships that could solve current or future problems. So, data mining is defined as the operation of extracting the interesting and previously unknown information and represents one phase in the complex process of Knowledge Discovery in Databases, aiming to solve two general categories of problems: prediction and description. It should be noted that each of the two problems is associated with its own methods, as shown in Fig. 2.

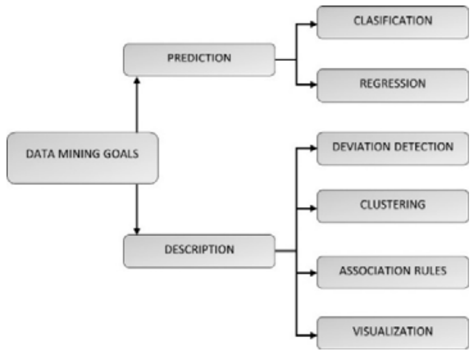

Fig. 2 Data mining objectives (adapted after [9])

Although the prediction is the one which brings maximum value, it is often preceded by the description.

Prediction involves patterns finding by classification or regression, which may anticipate future behaviors whereas description, based on clustering, deviation detection, discovering association rules and visualization techniques, aims to emphases the relevant characteristics of entities, events, or phenomena [9].

Finding a function or some rules that describe a class and allow to include a data item in one of several predefined classes is done by *classification*.

As part of supervised learning techniques category, it is performed through a two steps process. First, in the learning or the training phase, a classifier is built to describe a predetermined set of classes, through the so-called training set. This data set contains the class label of each case. Once built, the model is tested on data that are also labeled to determine its accuracy. If it is validated, the model will be further used to predict classes of unlabeled data [10].

Regression is also a prediction tool, but it is used to predict a value of a continuous variable based on the values of other variables used as predictors.

Clustering or cluster analysis identifies a finite number of categories or clusters to describe data. It brings together sets of entities into groups based on their similarities.

A cluster is a collection of objects that are very similar within the same cluster and are dissimilar to the objects in other clusters. Usually, the expression for similarity is a function of distance that can take different forms, appropriate to the types of data used to describe objects. It is not based on existence of predefined classes and class-labeled training sets, and for that reason it is an unsupervised learning technique [10].

Association rules find relationships or affinities between data which seem to have no semantic dependence. Initially, association rules were applied to data sets consisting of variable length transactions stored in transactional databases, but in time, new algorithms were developed and some algorithms have been adapted for various types of data, such as structured data stored in relational databases or unstructured data as text and image.

The last step in the KDD process is *pattern evaluation and validation*. If the model built in the data mining step has adequate performance, proven by measures such as accuracy, recall or other measures, appropriate to the context, it will be used further to obtain useful information and knowledge in the targeted domain.

If the model does not meet the quality conditions, the process can be restarted from any of the previous phases with the application of corrections to improve performance.

3. Multimedia Mininig

3.1 Multimedia Data

The term multimedia data refers, in a commonly accepted way, to those data coming from various media such as text, audio, or video used to provide multimodal information based on information technology [11].

They are characterized by several specific features, such as: huge volumes, temporality, spatiality, lack of structure and the need of particular logistics (storage space, fast transmission, content-based access) [14].

Multimedia databases are collections of related multimedia data. According to MPEG-7 Standard, a multimedia database contains the following data types [15] [16]: text, images (photographs, painting, maps or pictures), video (sequences of frames produced by various devices having the capability to record such sequences of images), audio provided by an aural recording device, animation sequences (animated images), graphic objects (drawing, illustration or sketches), composite multimedia – combinations of the above data types.

Briefly, media can be divided into two classes: dynamic, time-changing (holding audio and video data) also known as continuous media, and static, time-independent (holding text, image or graphics) belonging to discrete media, as it is presented in Fig. 3 [14].

3.2 Multimedia Big Data

In Social networks, mobile applications, wearable devices, and economic and research dynamics have recently caused a dramatically change of the generating, collecting, storing and processing data processes and have led to the frequent use of the term Big Data.

Although the definition of the term is sometimes ambiguous, and in the literature, there is a wide variety of ways in which Big Data is viewed, the common element lies in the fact that it is considered to refer to those data collections whose management is not possible using traditional database (whether relational or non-relational) tools. At the same time, the analysis of these data requires adaptations, sometimes major, of KDD processes and implicitly of data mining tasks and algorithms. The impossibility of using traditional technologies is dictated by the characteristics of Big Data, known in the literature as a set of Vs [2].

The first, and best-known characteristics, first identified by Doug Laney in a 2001 Gartner report [17], relates to volume, variety and velocity, this meaning the speed with which data is produced and often become obsolete. Over time, other characteristics have been identified and defined, so that the number of Vs has reached, for specific cases, 42 [18]. The term "big" used in the name might suggest that the main attribute is a consistent volume of data. However, the size of the dataset must be judged according to context, as there are organizations whose activities involve working with gigabytes or terabytes, while others, such as social networks, collect and handle petabytes or exabytes of data. In all these cases, however, there may be complex data processing and analysis needs that are specific to Big Data applications [19].

More important than volume is the variety of data. This comes from the fact that data from different sources and often in different formats are combined to be processed together and further analyzed. Velocity is related to the speed at which data arrives and is processed, analyzed, transformed into information and provided to stakeholders. Assessing the velocity needs for Big Data is closely related to understanding the business process and user requirements.

It should be borne in mind that, in organizations, data can flow in two ways: it can be integrated from multiple sources, transformed and then batch loaded in data warehouses, from where it will be analyzed, or it can be generated, processed and analyzed in real-time streams. The only thing to bear in mind is that the right information must be provided at the right time for each individual case [19]. Another characteristic, this time proposed by IBM [20] is veracity, matched in this context with data quality.

As a consequence of these 4 V's and starting from the information that can be obtained through a proper Big Data analysis, a fifth V can be added - value. This is given by the fact that Big Data analysis helps to better understand the relationship between events in the environment and the behavior of the phenomena or entities studied, leading to higher quality decisions.

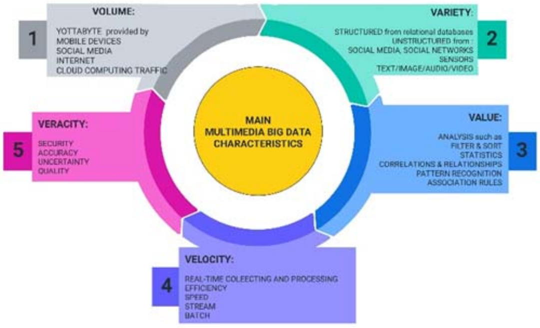

Fig. 4 gives a brief overview of these five Big Data features

Finally, it is worth noting that part of the characteristics of Big Data, volume and variety, also applies to multimedia data. Also, both Big Data and multimedia data work with all types of data, starting from structured, through semi-structured to unstructured.

If we consider large volume datasets, containing various data types such as combinations of text, images, video or audio but also having high velocity and veracity we can say that we are dealing with Big Data multimedia.

3.3 Multimedia (big) data mining

Generally, data mining is the process of finding surprising, interesting, and previous unknown patterns hidden in large amounts of data, which aims to improve decision making. Classically, data used in data mining are well-structured and, often, stored in large relational databases. Last years, the explosive growth of multimedia data volumes, led to the need of develop methods and tools to mine such data, in order to find valuable information to improve decision making. Multimedia data mining (MDM) is the process by which interesting information of implicit knowledge is found in multimedia databases.

Because multimedia databases are semi-structured or unstructured, multimedia data mining is more than an extension of data mining. Located to the confluence of many areas such as multimedia previewing, processing and retrieval, machine learning and artificial intelligence it has a major challenge in so-called semantic gap [21] which is the difficulty of deriving a high-level concept from low-level features (color histogram, homogeneous texture, contour-based shape, motion path) that are extracted from media data. A solution to this problem requires the creation of an ontology that covers different aspects of the concept and a hierarchy of elements that can deduce the probability of the concept from the probabilities of its components and the relationships between them.

The knowledge discovery in multimedia data processes and implicitly the stages of multimedia data mining take specific forms for each of the multimedia data types listed in the previous section. As a result, we can consider multimedia data mining as a generic term that can be customized as follows: text mining, image mining, video mining or audio mining [22] [23].

Text mining aims to find useful information and knowledge hidden in text content through specific techniques such as text classification (categorization), information extraction, information retrieval, clustering and summarization.

Text classification known also as Natural Language Processing (NLP) aims to collect, process and analyze text documents to discover the appropriate topics or indexes for each document. It is a form of supervised learning in which texts expressed in natural language are associated, depending on their content, with a predefined number of topics.

Clustering process, or cluster analysis, looks to identify intrinsic structures in the unlabeled textual data and based on these structures, to create meaningful clusters. Often, it is used as a pre-processing step which prepare the context for other text mining tools.

Information extraction is the text mining technique which aims to identify and extract entities and attributes from semi-structured or unstructured texts and to find their relationships. Once this information is extracted it is stored in a database for future use.

Information retrieval uses specific set of words or phrases to extract relevant patterns from the text applying various algorithms able to track and monitor user behaviors. Briefly, text summarization is the technique of shortening long texts by automatically generate compressed versions that retain important information for the end-user. The final goal is to create coherent and fluent summaries based on the main elements defined in documents.

Image mining aims to extract implicit knowledge or patterns not explicitly stored in groups of images. Not only an extension of classical data mining to image domain, but a great challenge in image mining is also to find how the low-level representation by pixels of a raw image or of an image sequence, can be processed in an efficient way to allow the identification of high-level objects or relationships [24]. Image mining is concentrated on finding specific patterns from large collection of images, using techniques classification, clustering, association, and visualization.

Video mining performs automatically extract of content and structure of video, features of moving objects and spatial or temporal correlations of those features. Additionally, without assumptions about the contents, it can discover patterns in video structure, in objects activities or in video events from vast amounts of video data [25]. Video mining techniques are useful to implement tasks such as classification, video summarization or abnormal events detection, aiming to discover semantic patterns. Using audio mining techniques, the content of audio signals can be automatically analyzed.

The main area where audio mining is used is automatic speech recognition through which we try to identify the speech within the audio. Because image, video and audio mining algorithms encounter the semantic gap problem, sometimes, to reduce or avoid this issue, multimedia mining uses the multimodal data mining approaches [26] [27], which takes into account that in many cases, different types of multimedia data coexist. For example, image data may be combined with text, or video with audio data. The *multimodal data mining* approaches offer better capabilities of knowledge discovery from multimedia database, allowing to find high-level conceptual relationships.

The osmosis between different ways of expressing data in multimedia can be exploited to find new high-level conceptual relationships, which involves learning the associations between these different ways of expression. Once learned, these relationships can be used for

multimodal data mining. Big Data analytics is what delivers value because it produces the needed information for high-quality decisions.

Beyond the fact that they require specific technologies and architectures, at a conceptual level, Big Data analytics follow the same steps as knowledge discovery from multimedia data. These steps are shown in Fig. 5.

The first step in this process is data collection from the right sources, followed by their storage. Given the characteristics of multimedia data, such as spatiality, temporality, variety or volume, it is necessary to use appropriate storage systems. In practice different solutions are used, from relational or object-relational databases to NoSQL, Hadoop-specific NTFS files, or other Big Data storage variants.

Recalling that multimedia data comes in different types, often in combination, it is obvious that careful

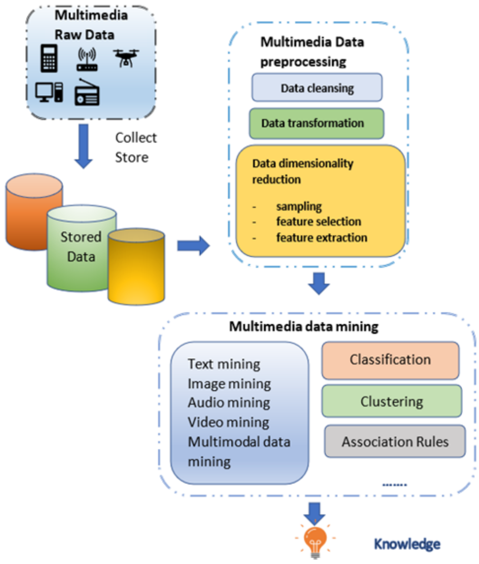

Fig. 5 Multimedia data mining process

pre-processing is needed before applying data mining algorithms. This step aims to transform raw data into good quality datasets suitable for pattern discovery algorithms.

Although, some operations are common to all data types, such as solving missing values, normalizing, treating outliers or dimensionality reduction by selecting or extracting features using filters or wrappers [8], there are also particular preprocessing methods for each of the specific multimedia data types.

By applying the right type of multimedia mining, we can obtain new patterns, which in the presence of performances accepted by experts in the field, can be considered as knowledge that can be the basis for future quality decisions.

4. Multimedia data mining vs. multimedia data mining – state of art

The following is a brief analysis of how research conducted strictly in the field of multimedia mining has migrated to multimedia Big Data mining by considering Big Data technologies.

For this purpose, we made a survey of the available bibliography disseminating this research and we obtained for multimedia data mining a number of 46 papers and for multimedia Big Data mining 118 important (considered by us) papers.

In Fig. 6 and Fig .7 the yearly distribution of these papers is shown.

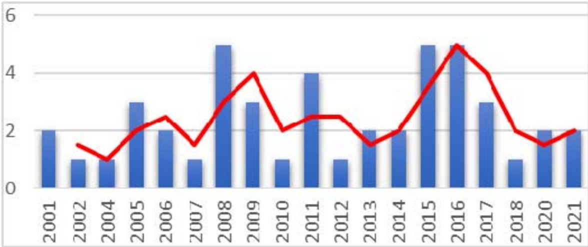

Fig. 6 The evolution of multimedia data mining research reflected in the early published papers

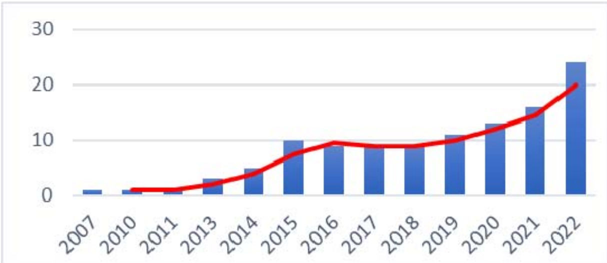

Fig. 7 The evolution of multimedia big data mining research reflected in the early published papers

It can easily be seen that before 2014-2015, when Big Data technology started to emerge and mature, research in the area of multimedia data mining was predominant. After this threshold, research on knowledge discovery from multimedia data gave way to the new direction namely multimedia Big Data mining.

5. Issues and challenges in multimedia mining

Even multimedia mining has brought considerable added value to our understanding of many aspects of behavior, feelings, disease or epidemic evolution, climate evolution and many other things, there is a flip side to the coin, as the nature of the data induces a number of issues and challenges.

These issues refer mainly to the limitations of the applied multimedia data mining approaches, including content-based retrieval and similarity search, generalization and multidimensional analysis, classification and prediction analysis, and mining associations in multimedia data.

Table 1 summarizes the issues raised by different types of multimedia mining, structured by the methods used for pattern discovery. All these translate into challenges that researchers and developers must face. Thus, during the maturation of multimedia data mining, and today, for multimedia Big Data mining solutions to solve specific problems of data acquisition and storage, efficient computing, scalability and security have been developed.

Table 1: Multimedia mining issues

Data Type	**Multimedia Data Mining Stage**		**Challenges**
Text	Preprocessing		- polysemy and synonymy
	Modeling	classification	-frequently, the classification models depend on only a few terms
		clustering	-finding the proper distance measures to deal with nonnumerical data
		association	-using semantics or concepts instead of words to reduce the problems with dimensionality, synonyms, homonyms [29].
Image	Preprocessing		-differences between global and local descriptors
	Modeling	classification	-finding a good model is computationally expensive.
		clustering	-grouping an untagged image collection into meaningful groups based on the content of the images [30]
		association	- scalability issue in terms of number of candidate patterns generated
Audio	Preprocessing		- segmentation of silence, music, speech and noise from the audio source.
	Modeling	classification	-methods based on models for segmentation or classification do not work in real time, and cannot be easily generalized [31]
		clustering	- automatic audio preprocessing uses large test databases, and lead to a decrease in recognition performance compared with hand labelled audio segmentation [32]
Video	Preprocessing		- feature extraction process depends very much on the structure of the video sequences, but also on the type of application.
	Modeling	classification	-shots of event contain significantly different features depending on camera techniques, object movements, location, and it is hard to build enough training data sets [33]
		clustering	- scalability for large video clusters
		association	- to reduce the semantic gap between the high-level concepts and low-level features [34]

6. Conclusion

The explosive development of information and communication technologies, the emergence of relatively cheap and easy-to-use devices for large mass of beneficiaries has led to the possibility of generating huge amounts of data of the most diverse types, from structured data to unstructured text, image, video, audio or composite data which often hide valuable information that cannot be extracted with traditional methods. In the last decade, with the emergence of the Big Data concept, with characteristics that distinguish it from high-volume datasets, multimedia data mining research has migrated to multimedia Big Data mining. This paper has made a brief analysis of this phenomenon, starting from the presentation of what multimedia data means in contrast to multimedia Big Data, continuing with the data mining process for multimedia sets, and finally making a summary study of the evolution of research and the problems and challenges induced by both multimedia data and multimedia Big Data.

References

[1] Emily Adler. 2016. Social media engagement: The surprising facts about how much time people spend on the major social networks. Retrieved from http://www.businessinsider.com/social-media-engagement-statistics-2013-12
[2] Paraschiv A, Danubianu M, Big Data – present Opportunities and Challenges, Broad Research in Artificial Intelligence and Neuroscience – BRAIN, vol. 10(2), 2019
[3] Cios K., Pedrycz W., Swiniarski R., Kurgan L. (2007) Data Mining. A Knowledge Discovery Approach, *Springer Science-Business Media*, ISBN 978-0-387-33333-5, New York
[4] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-37.
[5] Wirth, R. and Hipp, J., CRISP-DM: Towards a standard process model for data mining. In Proceedings of *the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, pages 29-39, Manchester, UK, 2000
[6] Patel N., Sethi I., "Multimedia Data Mining: An Overview" 2007 DOI:10.1007/978-1-84628-799-2_2
[7] Addi Ait-Mlouk, Tarik Agouti, Fatima Gharnati "Multi-criteria decisional approach for extracting relevant association rules", 2016 *International Journal of Computational Science and Engineering* 15(3/4) DOI:10.1504/IJCSE.2017.10008087
[8] Danubianu, M., Pentiuc, S. G., & Danubianu, D. M. "Data dimensionality reduction for data mining: a combined filter-wrapper framework.", 2014 *International Journal of Computers Communications & Control*, 7(5), 824-831.
[9] Danubianu, M., Pentiuc, S. G. "Data dimensionality reduction framework for data mining", *Elektronika ir Elektrotechnika*, 19(4), 87-90, 2013
[10] Danubianu M., "A data preprocessing framework for students' outcome prediction by data mining techniques" In 2015 19th International Conference on System Theory, Control and Computing (ICSTCC) 2015 Oct 14 (pp. 836-841). IEEE
[11] Danubianu M., I. Tobolcea I., "Using data mining approach for personalizing the therapy of dyslalia," 2011 E-Health and Bioengineering Conference (EHB), 2011, pp. 1-4.
[12] Hall, M., Correlation-based feature selection for discrete and numeric class machine learning, Proc. Of International Conference on Machine Learning, 359-365, Morgan Kaufmann, 2000
[13] Douik A., Abdellaoui M., Cereal Grain Classification by Optimal Features and Intelligent Classifiers, *International Journal of Computers Communications & Control*, ISSN 1841-9836, 5(4):506-516, 2010.
[14] Yu, C., Williams, A., Lin, C. F., & Yu, W. C. (2008). Planning effective multimedia instruction. In Terry T. Kidd & Holim Song (Eds.), *Instructional systems and technology* (Vol. 1, pp. 216-231). Hershey, PA: Information Science Reference.
[15] Kalipsiz, O. (2000). Multimedia databases. Proceedings of IEEE International Conference, 111-115. Retrieved May4,2009, from: http://ieeexplore.ieee.org.proxy.library.msstate.edu/stamp/stamp.jsp?tp=&arnumber=859745&isnumber=18614
[16] Adjeroh, D.A. Nwosu, K.C. (1997). Multimedia database management: Requirements and issues. Multimedia, IEEE, 4(3), 24-33. Retrieved May 25, 2009, from http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=621580&isnumber=13526
[17] Manjunath BS, Salembier Ph, Sikora T (Eds.). Introduction to MPEG-7. Multimedia Con- tent Description Interface. Wiley, New York, 2002, 371 p.
[18] https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
[19] https://www.elderresearch.com/blog/42-v-of-big-data
[20] Danubianu M., Barila A. (2014) Big Data vs. Data Mining for Social Media Analytics, *International Conference of Social Media in Academia - Research and Teaching –SMART2014*
[21] https://www.ibmbigdatahub.com/infographic/four-vs-big-data
[22] Petrushin V, Introduction into multimedia Data Mining and Knowledge Discovery, Multimedia Data Mining and Knowledge Discovery, Springer – Verlag, 2007
[23] https://www.javatpoint.com/what-is-multimedia-data-mining
[24] Vijayakumar.V, A. Sakila, "Multimedia Minig Research – an Overview", *The International Journal of Computer Graphics & Animation (IJCGA)* Vol.5, No.1, January 2015
[25] Wynne H, Lee ML, Zhang J (2002) Image mining: trends and developments. J Intell Inf Syst 19(1):7–23

[26] Bhatt, C.A., Kankanhalli, M.S. Multimedia data mining: state of the art and challenges. Multimed Tools Appl 51, 35–76(2011). https://doi.org/10.1007/s11042-010-0645-5
[27] Barnard K, Duygulu P, Forsyth D, Freitas N, Blei D, Jordan M. Matching words and pictures. J. Machine Learning Reerach 3, 3 (2003), 1107–1135.
[28] Feng S.L., Manmatha R, and Lavrenko V. Multiple Bernoulli relevance models for image and video annotation. *In Proceedings International Conference on Computer Vision and Pattern Recognition*, 2004.Washington, DC.
[29] Nørvåg K, Øivind Eriksen T, Skogstad KI (2006) Mining association rules in temporal document collections. In: *International symposium on methodologies for intelligent systems (ISMIS)*, pp 745–754
[30] Zhang, Ji and Hsu, Wynne and Lee, Mong Li Image mining: issues, frameworks and techniques. In: *2nd ACM SIGKDD International Workshop on Multimedia Data Mining (MDM/KDD'01)*, 26 Aug 2001, San Francisco, CA, USA
[31] Ajmera J, McCowan I, Bourlard H (2002) Robust hmm-based speech/music segmentation. In: *IEEE international conference on acoustics, speech and signal processing*, pp 1746–1749
[32] Meinedo H, Neto J (2005) A stream-based audio segmentation, classification and clustering pre-processing system for broadcast news using ann models. In: *Interspeech—Eurospeech*
[33] Shirahama K, Sugihara C, Matsumura K, Matsuoka Y, Uehara K (2009) Mining event definitions from queries for video retrieval on the internet. In*: International conference on data mining workshops*, pp 176–183
[34] Lin L, Shyu ML (2009) Mining high-level features from video using associations and correlations. In: *International conference on semantic computing*, pp 137–144

BOGDĂNEL CONSTANTIN GRADINARU was born in March 7, 1979. He obtained his B.S and M.S in Computer Science at the Faculty of Electrical Engineering and Computer Science of "Stefan cel Mare" University of Suceava in 2002.
From 2002 to 2020 he was a design engineer in the field of surveying, cartography and real estate and building cadastre, and since 2020 he is a laboratory technician with higher education in the Department of Computer Science "Stefan cel Mare" University. He is also a 2nd year PhD student in the field of Computers and Information Technology,"Stefan cel Mare" University. His current research interests include various aspects of Multimedia
Data Mining which is increasingly turning into Big Data Mining and having a major impact in most popular fields.

MIRELA DANUBIANU was born in 1961. She received two B.S. and M.S. degrees in computer science, in 1985, and economics, in 2001, from the University of Craiova, and the Ph.D. degree in computer science from "Stefan cel Mare" University of Suceava, in 2006.
From 1985 to 1991, she was a Researcher with the Unconventional Devices Laboratory at ICMET Craiova, and from 1991 to 2001 she has been CIO at Dolj County Prefecture and from 2001 to 2014, she has been Lecturer with the Computers, Electronics and Automation Department, and since 2014 she is Associate Professor with Computers Department, "Stefan cel Mare" University. She is the author or coauthor of 5 books, 7 chapters and of more than 100 articles. Her research interests include databases theory and implementation, modern data architectures, data analytics, application of Data Science in economics, education, healthcare and project management.

Adina BARILA (b. June 4, 1968) has obtained the B.S. degree in Computer Science from "Alexandru Ioan Cuza" University of Iași in 1990, and PhD degree in Computer Science from "Stefan cel Mare" University of Suceava in 2015. She is Lecturer of the Computers Department at "Stefan cel Mare" University of Suceava. Her current research interests include databases theory and implementation, data analytics, application of Data Science in education and economics.