

ASSOCIATION RULES IMPLEMENTATION FOR AFFINITY ANALYSIS BETWEEN ELEMENTS COMPOSING MULTIMEDIA OBJECTS

¹ MENDOZA-PALECHOR, FABIO E. , ² CARRASCAL OVIEDO, ANA, ³ DE-LA-HOZ-FRANCO, EMIRO

¹Asstt Prof., Department of Computer Science and Electronic, Universidad de la Costa, Colombia

²Asstt Prof., Facultad de Ingeniería en Tecnologías de la Información y la Comunicación, Universidad Pontificia Bolivariana, Colombia

³Asstt Prof., Department of Computer Science and Electronic, Universidad de la Costa, Colombia

E-mail: ¹fmendoza1@cuc.edu.co , ²ana.oviedo@upb.edu.co , ³edelahoz@cuc.edu.co

ABSTRACT

The multimedia objects are a constantly growing resource in the world wide web, consequently it has generated as a necessity the design of methods and tools that allow to obtain new knowledge from the information analyzed. Association rules are a technique of Data Mining, whose purpose is to search for correlations between elements of a collection of data (data) as support for decision making from the identification and analysis of these correlations. Using algorithms such as: A priori, Frequent Parent Growth, QFP Algorithm, CBA, CMAR, CPAR, among others. On the other hand, multimedia applications today require the processing of unstructured data provided by multimedia objects, which are made up of text, images, audio and videos. For the storage, processing and management of multimedia objects, solutions have been generated that allow efficient search of data of interest to the end user, considering that the semantics of a multimedia object must be expressed by all the elements that composed of. In this article an analysis of the state of the art in relation to the implementation of the Association Rules in the processing of Multimedia objects is made, in addition the analysis of the consulted literature allows to generate questions about the possibility of generating a method of association rules for the analysis of these objects.

Keywords: *Association Rules, Multimedia Object, Data Mining, Data-Set, Correlations.*

1. INTRODUCTION

Currently is possible to highlight the increase of multimedia resources available through the web, being used by different media or devices, rising certain concerns about controlling their appearance as stated by [1], in the consulted literature is easy to find two scenarios for research, the first is focused on search of new knowledge and application of the data mining methods in these multimedia objects mentioned by [2][3][4][5] and the second approach makes reference to the organization of the databases that store multimedia objects as you can see in [1].

A multimedia object is an element composed by alphanumeric characters, graphs, images, animations, videos and audios [6], some examples of them are: webpages, virtual learning objects, and the graphic material online, animations, audio and video, interactive or static. The semantic of a multimedia object must be expressed by all

elements that compose it, text, images, audios, videos as mentioned by [7][8][9].

Although the elements that an object multimedia is composed of, are from different types (text, image, audio, video), there are a great number of correlations between them, as manifested by [8]. For this reason, several studies have been made like [10] with the objective of determining the similarities between objects that form multimedia objects. Based on the previous references, data mining techniques can be used to analyze these elements based on association rules.

Multimedia objects mining must validate the different elements that compose an object, identifying representations that can be interpreted through computational tools.

Text can be represented by a feature vector that holds basic information such as number of words,

number or sentences, number of paragraphs, number of chapters, number of sections, stop words, document text type and size.

Images are represented with some features related to format, image resolution, number of bits by pixel, compression information, histogram of each color matrix that compose RGB.

Audio documents are represented by features such as sample frequency, channels, size sample, coding, tone, intensity. All this generates a feature vector where it is possible to group the features mentioned.

A video has features like duration, frames per second, number of bits by pixel, color, compatibility, coding.

Figure 1 shows the possible representation of a multimedia object, since it can have several elements inside.

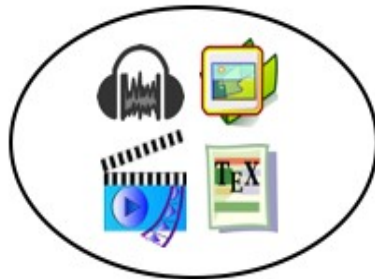


Figure 1. Multimedia Object
Source. Created by authors

Due to their complex structure, multimedia objects require a complex processing to obtain the semantics of their contents. Modelling the structure of the multimedia object is essential because through analysis we can establish criteria on the similarity levels between a group of objects considering the different elements that compose them [6].

In the literature you can clearly see that there have been attempts to design information systems that can handle multimedia objects, which have features like presentation capabilities, storage and communicating the elements of the multimedia object, all this surrounded by the extreme heterogeneity of the analyzed data [11].

This study presents an affinity analysis between two or more elements that compose a multimedia object, implementing the Apriori algorithm using association rules of the data mining methods, the

obtained results will be validated using metrics of support and confidence. The analysis of elements inside a multimedia object allow to obtain a greater understanding of this resource, based on the attributes represented by metadata and the information and semantics that compose it.

Multimedia objects have become a research field in continuous development, as mentioned by [1], to know their structure and identifying the relationship between their elements, would help to show approaches that allow to understand their behavior and improve their organization as proposed in [1]. Furthermore, the comprehension of the structure and behavior allows their use in tasks such as document recovery with images and text, mentioned in [2], and information fusion of heterogeneous data [4][5].

This study has the following structure: section 2, previous works, where you can observe the literacy contributions by different authors for the analysis of multimedia information, in section 3, methods and materials, you can see the concepts and definitions of association rules, the algorithms used to implement them, the validation metrics used for evaluation of the proposed model and finally the description of the dataset used by the authors. In section 4, you can find the experimental process and section 5 presents a discussion, where you can see some considerations about the results obtained and section 6, presents the conclusions of the study.

2. PREVIOUS WORKS

Multimedia objects are a research area of high interest, where you can find several working methods for analysis of the components of an multimedia object. Next, you can find a brief review of the consulted work found in literature where you can evidence the analysis of multimedia objects.

In [2] the paper proposes a framework that applies data mining techniques with the purpose of acquiring new knowledge, using association rules that can perform pattern discovery, the study takes spatial information, stored for future analysis with a set of data mining techniques. The dataset chosen was captured using commercial high definition video cameras with georeferenced data, obtaining a video of one hour in length and was decomposed in 450 frames with a size of 720 x 480 pixels, the first stage of the study was the tagging of the images, based on their texture within these classes: pasture,

forest, agriculture, highway, urban zone. To generate the classes of textures, it was necessary to use the Gabor filter [12] since it is capable of border detection and image orientation. Later, these images were tabulated in a scalable data structure, which produced a spatial event cube, this cube is represented by a matrix of the same size of each image, and it is filled with values with a range between 0 and 4, where each number represents the texture class that position belongs to. The validation metrics used were support and confidence. Finally, the apriori algorithm was used on the dataset.

In [1] they highlight the fast increase in multimedia resources, and this fact has created the need of developing an intelligent method that allows to keep organized the multimedia data contained in a database of big size, the proposed model has the goal of identifying patterns and associations between multimedia resources through the use of link networks, tags and text of multimedia resources are used to measure the semantic relationship between data. The study used 100.000 images taken from Flickr. The model proposed by the authors uses association rules to identify the semantic relationship between data, as the next image shows.

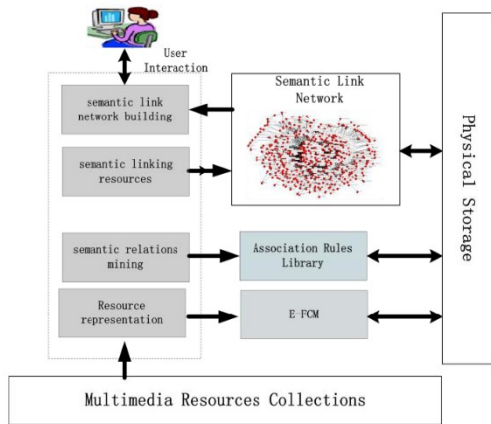


Figure 2. Proposed Model
Source. Taken from [1]

The evaluation metric used to validate the semantic relationship between data was the Jaccard coefficient, implemented considering the relationship degree between data, based on the queries made on the internet between the searched data and the number of results found.

In [3] you can find an analogy between image recovery and text, using a visual approach based on sentences to recover images contained in specific objects. The visual sentence is defined as a couple of patches of adjacent local images and is built using data mining. In this study, the authors propose a method to build visual sentences of images and how to code the visual sentences for indexing and recovery. The research has four stages, first the words creation, second the visual sentences creation, third shows the construction of indexes for each sentence and the fourth and final validates the measure of similarity of the imaged based in sentences. To generate the frequent patches, the used the apriori algorithm, they counted the co-occurrences to analyze the frequent visual words and that amount of words must be higher than a defined threshold.

In [4] the study brings the problem of fusing information located on the web, considering that the processes of analysis, collection and tracking of the information require a bigger effort, usually the studies in this area are focused on generating a multidocumented brief using text features only, meanwhile this study refers to the image theme and text association. This research is developed based on a small training dataset, and the similarity measures between the image and text features. The dataset analyzed contains 300 images of terrorist attacks obtained from the CNN and BBC websites. From those images, the subtitles that refer to the images were also extracted, and the images and texts are categorized manually in 15 classes or domains defined previously. Some of the features extracted from text were obtained after applying stop words and stemming filters, and for images feature extraction was used the Gabor filter, finally obtaining 6 vectors with the extracted features.

In [5] propose information fusion between text and visuals, combining different techniques of data mining such as clustering and association rules. The experimentation was made using as information source 54500 images of ImageCLEF 2011, a Wikipedia collection. Data mining association rules are usually used to discover relationships in big datasets. In this study, the authors applied clustering methods to extract features from text and images. After cluster application, they implement association rules to identify the relationships in the information subject to study. The metrics used to validate the proposed method were support and confidence, apriori was used as algorithm to generate frequent items.

In [13] the investigation is presented about Grouping of images Web using Rules of Association, Measurements of Interest and Partitions Hypergraph, in this case it treats of a new approach for the cluster of images Web, the images are processed to extract the characteristics of the signals like the color in HSV format (tone, saturation, value) and quantized orientation. Web pages that use these images are processed textual features such as stories (key words) and techniques of character reduction stories such as the stem (method to reduce a word to its root), stop the words removed, and the application of The law of Zipf (frequency of appearance of a word), all the visual and the textual characteristics are used to generate the rules of association. The rules for hypergraphs are generated with the characteristics of the texts, in addition, the hypergraph algorithm is used to generate a cluster of characteristics.

[14], mentioned that with the proliferation of multimedia data and ever-increasing requests for multimedia applications, new means emerged for the efficient and effective management of management and access to large audiovisual collections. In the research developed, a new framework for video detection is presented, which plays an essential role in the level of video indexing and retrieval. In this process, the elements that cause the loose video structure and the distribution of data are addressed with effectiveness. In addition, an adaptive mechanism is proposed to determine which are the essential elements that are increasingly used in the traditional approach through association rule mining (ARM). Therefore, to a large extent, the framework has been proposed domain information and the objective of the automatic video content analysis.

3. MATERIALS AND METHODS

3.1 Association Rules

Association rules were mentioned by [15], and they were used for basket case analysis, from this study, they were able to establish criteria for decision making based on the shopping behavior of certain products from customers, and those allowed to set the order or position of the products for sale in different places through the market, obtaining encouraging results. In [16] is mentioned that association rules are part of the data mining techniques and their purpose is based on the association degree of the analyzed information.

In the literature available you can find authors like [15] [17] [18] [19] [20] [21] [22], that define association rules under the same basis, as $I = \{i_1, i_2, i_3, i_4\}$ where I being the items set. A database is a set of transactions where each T is a subset of I . An association rule can be inferred like $X \rightarrow Y$, where X, Y is subset of I and $X \cap Y = \emptyset$. The set of elements X is called predecessor and Y is usually called successor. Based on the literacy analysis, the two metrics most common used to validate the rules generated by the association rules algorithms are support and confidence, as observed in previous works such as [15] [17] [28] [29] [31] [32] [33] [35] [38].

The problem of discovering all the association rules can be broken down into two sub-problems according to what is mentioned in [15]. First, the number of transactions that contain a set of elements, since many rules can be generated that are of little interest. The second problem refers to the use of large data sets to generate the desired rules. On the other hand, in [23] it is stated that the mining of association rules has contributed to many advances in the area of knowledge discovery. However, the quality of the association rules discovered is a major concern and has attracted more and more attention recently. One of the problems with the quality of the discovered association rules is the enormous size of the extracted rule set.

In [24] it is mentioned that the association rules are intended to find trends that can be used to understand and explore patterns of behavior in the data that are analyzed, it should be noted that not all association rules represent a pattern in the data studied. A rule will represent a pattern if it meets certain criteria defined in the induction algorithms, which also express the reliability of the rules.

3.2 Association Rules Algorithms

In the literature available and consulted, you can appreciate several algorithms designed to generate association rules from various massive data source, within these algorithms is possible to highlight the relevance of the Apriori Algorithm [15], since has become the basis for highly efficient methods like ECLAT [25] and FPGrowth [26] and others. In Table 1, you can find the association rules algorithms found on the literature consulted.

Table 1. Association rules algorithms

Algorithm Name	Author
Apriori	Agrawal, 1993 [15]
Algorithm Partition	Savasere, 1995 [27]
Algorithm ECLAT	Zaki, 1997 [25]
Algorithm CBA	Liu, 1998 [28]
Frequent Parent Growth	Han, 2000 [26]
Algorithm RARM	Das, 2001 [29]
Algorithm CMAR	Li, 2001 [30]
Algorithm CPAR	Yin, 2003 [31]
Algorithm MMAC	Thabtah, 2004 [32]
Algorithm QFP	Juan, 2010 [33]

In the proposed study, we use the Apriori algorithm to analyze multimedia objects and it is described next.

Apriori is an algorithm design to discover big sets of items through several passes over data [15]. First step counts the number of occurrences of the item with the purpose of setting the size of the 1-itemset. Then, the k interaction is invoked, and it is composed by two phases. First, the size of itemset Lk-1 is found in interaction k-1, this step is used to generate the candidate of the itemset Ck. The previous step is made using the function of candidate generation apriori (apriori-gen) which takes the arguments Lk-1. Then, the database is scanned and the count of the support candidates in Ck is made. To obtain an efficient count, it is necessary to determine the candidates Ck that are contained in the given transaction t. Next, you can find the function to generate the candidates Apriori-gen and the Apriori Algorithm.

In [34] is mentioned that Apriori is part of the Association Rules methods, which generates a number of rules considered as good when analyzing small databases, on the contrary it is considered as a disadvantage of the algorithm that As the size of the database increases, the performance of said algorithm decreases, this is because it must scan the entire database each time a transaction is scanned, which is why the algorithm's main problem is considered to be Apriori is the treatment of data in large databases, a problem also mentioned [35] who states that the performance of the Apriori algorithm will be very low and inefficient when the memory capacity is limited with a large number of transactions, another of the problems of the Apriori algorithm is that it suffers from some deficiencies despite being clear and simple, the main limitation

is costly waste of time to maintain a number of candidate sets with lots of frequent element sets.

3.3 Evaluation Metrics

Association rules allow to identify trends or relationships in the data found on a database, from these coincidences rules are generated, which makes important to determine the quality of the rules, so they can guarantee an optimal solution for the analysis of the data behavior. With this in mind, we present the evaluation metrics for association rules and some measures that come up as alternatives to validate association rules. In Table 2, you can find the most common metrics for validating association rules.

Table 2. Evaluation Metrics

Metric	Equation
Support	$\frac{X \cup Y}{D}$
Confidence	$\frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$

Support is a metric of interest to measure the quality of association rules. Given that the frequency of X is the number of rows that contain that item in the database. Support in association rules is represented by union of predecessor X and successor Y. Authors as [36] [37], define support as a metric that counts the frequency the items of an association rule appear in the data, in other words the number of transactions that contains the items in a rule all together in the data in relationship with the total number of transactions.

In [15] is mentioned that support is a norm that is defined as the fraction of the transactions in T that satisfy the union of the successor and predecessor elements in the rule, not to be confused with confidence because confidence is a metric that allows to evaluate the strength of a rule, meanwhile support corresponds to its statistical significance.

Authors as [37] manifest that confidence of a single association rule is define as the percentage / fraction of the number of transactions that contains X U Y compared to the total number of records that contains X. The range of the values that can take the confidence metric is given between 0 and 1 according to [38].

3.4 Dataset

With the objective of analyzing at least two elements that compose a multimedia object, we present the methodology to create the dataset.

A first step, before applying association rules, is to identify the dataset to be used, for this it was necessary to search through information sources and specialized databases for research articles. Based on this, we used as dataset the information describe in paper [13], which corresponds to the data stored in a website and contains elements of image and text types, and together they can be analyzed as a multimedia object. In the web repository analyzed, you can find a series of images of animals, with the description in text of the habitat of each animal, as their growing process, feeding and other information.

To build the dataset, it was mandatory to store the images and the text documents found in the website of the repository, we downloaded 100 images with their correspondent text documents, next you can see the data.

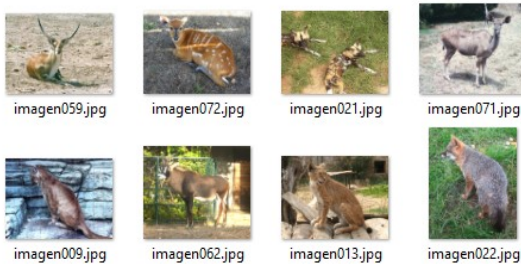


Figure 3. Image type items for dataset construction
Source: Created by authors

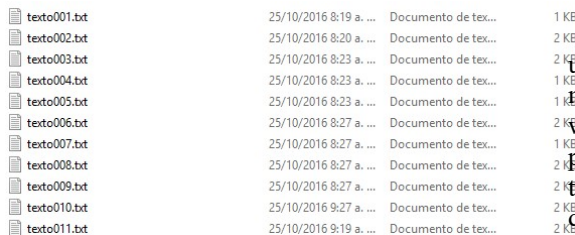


Figure 4. Text type items for dataset construction
Source: Created by authors

After storing all the documents, we started with the dataset creation, for this is necessary to determine the different attributes to extract of each

item that compose a multimedia object. For the images, the following features were chosen:

- Main color: Refers to the color that dominates the image, considering its three matrices that its composed of, in this case they are mostly RGB.
- Secondary color: Refers to the color that is secondary in the image, considering its three matrices that its composed of, in this case they are mostly RGB.
- Tertiary color: Refers to the color that is less frequent in the image, considering its three matrices that its composed of, in this case they are mostly RGB.
- Histogram: The histogram generates a vector with the number of color pixels that has the image, this number ranges from 0 to 255 where 0 represents absence of color and 255 the white color.

For the extraction of the histogram vectors in each image, we used the following commands in Matlab.

```
clear all
clc
I=imread('imagen100.jpg');
imshow(I)
a=rgb2gray(I);
b=imhist(a);
```

Figure 5. Commands for the histogram extraction
Source: Created by authors

When you need to represent the features of a multimedia object, the size of the images becomes a relevant factor, because the proportion or size of the image affects directly the number of features that can be generated when the pixel count or the histogram is made.

To perform the text document processing we used a stops word, additionally we applied several methods like stemmer, discretization and the bag of words methodology. The generated attributes were passed through a verification process to validate that the analysis would be made with the correct data and not data with noise. For this previous process we used the Weka tool, obtaining the final dataset to be used with the algorithm of association rules.

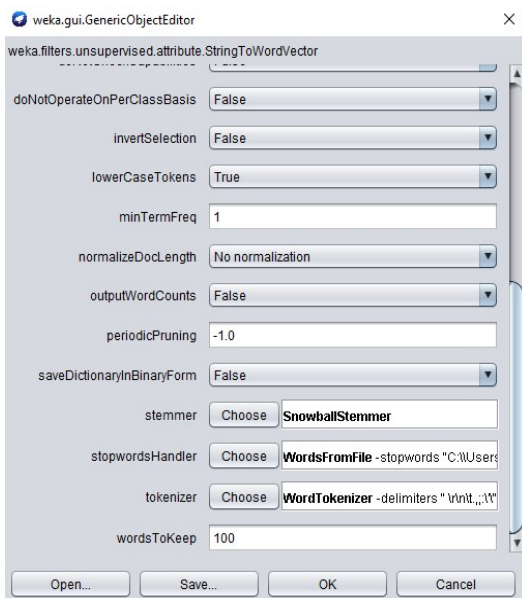
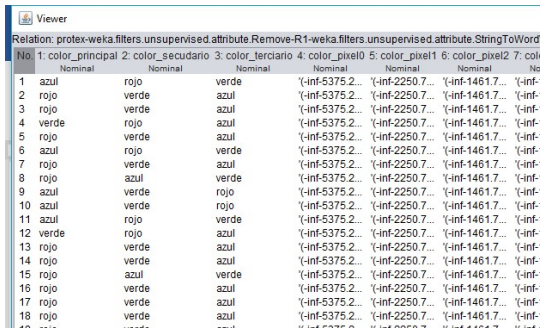


Figure 6. StringToWordVector Settings
Source. Created by authors



The screenshot shows the 'Viewer' window in Weka Explorer. The relation is 'protex-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.StringToWordV'. The data is as follows:

No	1. color_principal	2. color_secudario	3. color_tercario	4. color_pixel0	5. color_pixel1	6. color_pixel2	7. color...
1	azul	rojo	verde	'-inf-5375.2...	'-inf-2250.7...	'-inf-1461.7...	'-inf-1...
2	rojo	verde	azul	'-inf-5375.2...	'-inf-2250.7...	'-inf-1461.7...	'-inf-1...
3	rojo	verde	azul	'-inf-5375.2...	'-inf-2250.7...	'-inf-1461.7...	'-inf-1...
4	verde	rojo	azul	'-inf-5375.2...	'-inf-2250.7...	'-inf-1461.7...	'-inf-1...
5	rojo	verde	azul	'-inf-5375.2...	'-inf-2250.7...	'-inf-1461.7...	'-inf-1...
6	azul	rojo	verde	'-inf-5375.2...	'-inf-2250.7...	'-inf-1461.7...	'-inf-1...
7	rojo	verde	azul	'-inf-5375.2...	'-inf-2250.7...	'-inf-1461.7...	'-inf-1...
8	rojo	azul	verde	'-inf-5375.2...	'-inf-2250.7...	'-inf-1461.7...	'-inf-1...
9	azul	verde	rojo	'-inf-5375.2...	'-inf-2250.7...	'-inf-1461.7...	'-inf-1...
10	azul	verde	rojo	'-inf-5375.2...	'-inf-2250.7...	'-inf-1461.7...	'-inf-1...
11	azul	rojo	verde	'-inf-5375.2...	'-inf-2250.7...	'-inf-1461.7...	'-inf-1...
12	verde	rojo	azul	'-inf-5375.2...	'-inf-2250.7...	'-inf-1461.7...	'-inf-1...
13	rojo	verde	azul	'-inf-5375.2...	'-inf-2250.7...	'-inf-1461.7...	'-inf-1...
14	rojo	verde	azul	'-inf-5375.2...	'-inf-2250.7...	'-inf-1461.7...	'-inf-1...
15	rojo	azul	verde	'-inf-5375.2...	'-inf-2250.7...	'-inf-1461.7...	'-inf-1...
16	rojo	verde	azul	'-inf-5375.2...	'-inf-2250.7...	'-inf-1461.7...	'-inf-1...
17	rojo	verde	azul	'-inf-5375.2...	'-inf-2250.7...	'-inf-1461.7...	'-inf-1...
18	rojo	verde	azul	'-inf-5375.2...	'-inf-2250.7...	'-inf-1461.7...	'-inf-1...

Figure 7. Bag of Words Implementation
Source. Created by authors

Finally, after applying the bag of words to the text documents and joining this information with the attributes extracted from the images, we created a dataset with 100 records and 2556 attributes, which can be reduced since the bag of words application introduces some errors and this decreases the number of attributes to select.

```

255 @attribute color_pixel248 numeric
256 @attribute color_pixel249 numeric
257 @attribute color_pixel250 numeric
258 @attribute color_pixel251 numeric
259 @attribute color_pixel252 numeric
260 @attribute color_pixel253 numeric
261 @attribute color_pixel254 numeric
262 @attribute color_pixel255 numeric
263 @attribute color_pixel256 numeric
264 @attribute tipoanimal (carnivoro,herbivoro,omnivoro)
265
266 @data
267
268 1, "Lions are the largest cats in Africa. This male lion was actually not roaring, b
269 2, "Siberian tigers are the largest cats in the world, and among the most beautiful i
270 3, "Jaguars are the largest cats of the American continent, living 15-20 years. They
271 4, "There are five subspecies of tigers and three more have become extinct in recent
272 5, "Leopards hunt most successfully at night and from ambush. During the daytime they
273 6, "Servals are found throughout most of Sub-Saharan Africa, but small isolated popu
274 7, "Cheetahs are well known for being the fastest animals on land, reaching speeds of
275 8, "Kudu are found throughout Africa, and in Asia from Turkey through India to

```

Figure 8. ARFF File generated
Source. Created by authors

4. EXPERIMENTATION AND RESULTS

To make the implementation of the association rule algorithm Apriori, we used the Weka tool. First, the dataset creation was made, process that is described in section 3.4 (dataset).

After the dataset was generated, we validated the data obtained in the ARFF file, which holds the attributes of the items that compose the multimedia objects.

Then, a preprocessing of the data was performed with the goal of deleting variables or attributes that were irrelevant. The original dataset had 100 records and 2556 attributes, after the preprocessing final the new ARFF file had 100 records and 524 attributes, as you can see in Figure 8.

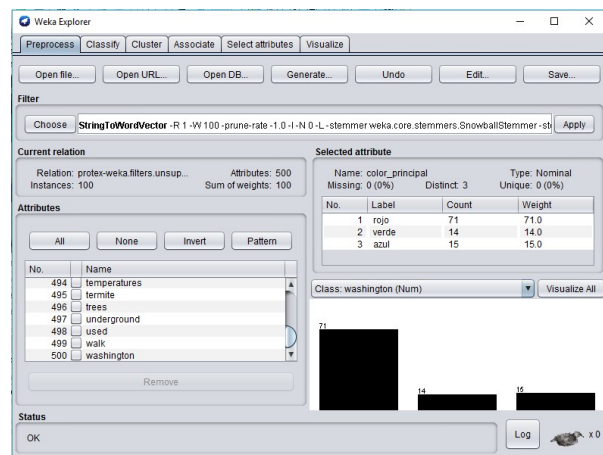


Figure 8. Number of attributes after preprocessing
Source. Created by authors

Then, with the dataset finally processed and revised, we proceeded to the configuration of the apriori algorithm, as you can see in Figure 9.

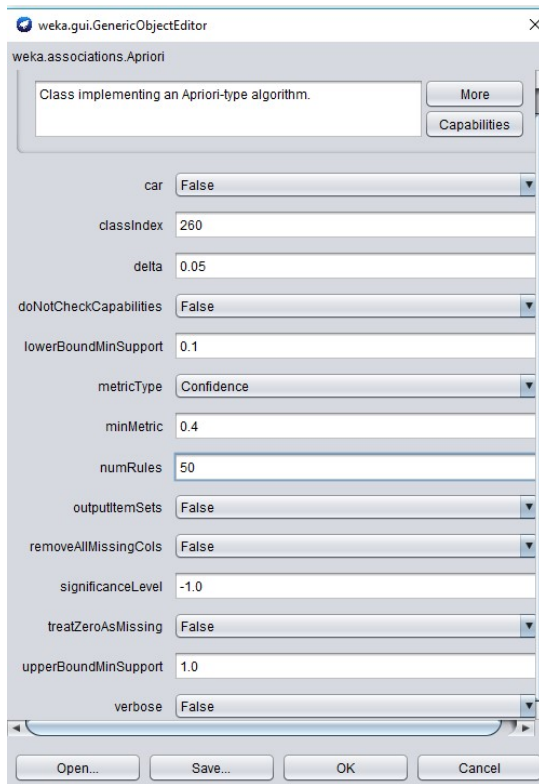


Figure 9. Apriori Algorithm Configuration
Source. Created by authors

To configure the Apriori algorithm the input data must be nominal or binary, and the attributes generated are continuous, and for this the variables had to be discretized to be able to implement the Apriori. In the discretization process you need to choose the number of uncertainty groups, considering that is not possible to be completely accurate in the number of right divisions to generate the value intervals for the analyzed attributes.

To implement the apriori algorithm, we encountered some difficulties looking for a relationship between the items that compose the multimedia object based on the documents (text and image).

When we combined the image and text attributes in a single ARFF file and implement the apriori method, this does not generate rules, and it is an indication that the analysis of the data is ineffective, to solve this, it became necessary to process separately the images and text documents. Next you can see the rules obtained.

```
color_pixel11=(-inf-3001]' 98 ==> color_pixel0=(-inf-7167]' 98 conf:(1)
color_pixel18=(-inf-10428.333333]' 98 ==> color_pixel17=(-inf-12213]' 98 conf:(1)
color_pixel246=(-inf-2665.333333]' 98 ==> color_pixel247=(-inf-31244]' 98 conf:(1)
color_pixel3=(-inf-1952]' 97 ==> color_pixel0=(-inf-7167]' 97 conf:(1)
color_pixel4=(-inf-2329.666667]' 97 ==> color_pixel0=(-inf-7167]' 97 conf:(1)
color_pixel3=(-inf-1952]' 97 ==> color_pixel11=(-inf-3001]' 97 conf:(1)
color_pixel4=(-inf-2329.666667]' 97 ==> color_pixel11=(-inf-3001]' 97 conf:(1)
color_pixel4=(-inf-2329.666667]' 97 ==> color_pixel3=(-inf-1952]' 97 conf:(1)
color_pixel3=(-inf-1952]' 97 ==> color_pixel4=(-inf-2329.666667]' 97 conf:(1)
color_pixel19=(-inf-8765.666667]' 97 ==> color_pixel17=(-inf-12213]' 97 conf:(1)
```

Figure 10. Association rules generated with attributes of image type
Source. Created by authors

```
bamboo=(-inf-0.677902]' 98 ==> pandas=(-inf-0.798015]' 98 conf:(1)
Brown=(-inf-0.798015]' bamboo=(-inf-0.677902]' 97 ==> pandas=(-inf-0.798015]' 97 conf:(1)
fox=(-inf-0.557789]' 96 ==> fox=(-inf-0.557789]' 96 conf:(1)
fox=(-inf-0.557789]' 96 ==> fox=(-inf-0.557789]' 96 conf:(1)
dogs=(-inf-0.557789]' 96 ==> breeds=(-inf-0.677902]' 96 conf:(1)
pack=(-inf-0.557789]' 96 ==> breeds=(-inf-0.677902]' 96 conf:(1)
skin=(-inf-0.557789]' 96 ==> pandas=(-inf-0.798015]' 96 conf:(1)
bamboo=(-inf-0.677902]' breeds=(-inf-0.677902]' 96 ==> pandas=(-inf-0.798015]' 96 conf:(1)
bamboo=(-inf-0.677902]' chimps=(-inf-0.677902]' 96 ==> pandas=(-inf-0.798015]' 96 conf:(1)
bamboo=(-inf-0.677902]' sheep=(-inf-0.677902]' 96 ==> pandas=(-inf-0.798015]' 96 conf:(1)
```

Figure 11. Association rules generated with attributes of text type Source. Created by authors

The rules generated can be used to establish a relationship between the elements that compose a multimedia object. Next, we describe some of the rules created by the apriori method.

- In the text items that compose the multimedia object, each time that appears the word bamboo also appears the word chimps, this rule is evaluated through the metric confidence with a value of 100%.
- In the text items that compose the multimedia object, each time that appears the word bamboo also appears the word breeds, this rule is evaluated through the metric confidence with a value of 100%.
- In the text items that compose the multimedia object, each time that appears the word dogs also appears the word breeds, this rule is evaluated through the metric confidence with a value of 100%.
- In the image items that compose the multimedia object, each time that appears the pixel 11 also appears the pixel 10, this rule is evaluated through the metric confidence with a value of 100%.
- In the image items that compose the multimedia object, each time that appears the pixel 119 also appears the pixel 117, this rule is evaluated through the metric confidence with a value of 100%.

5. DISCUSSION

The results obtained by the study show a clear relation between the elements that compose a multimedia object, which means documents with text and images for this instance. This translates into a contribution to approach heterogeneous information fusion in a dataset, and then to detect possible affinities between the attributes subject to study. For example, in figure 10, you can see a clear co-occurrence between attributes called pixels, where pixel 10 and pixel 11 show a strong relationship, which is supported by the validation metrics with a value of 100%, also in figure 11 you can see the relationship or affinity levels between information of text type, like the co-occurrence between the words *bamboo* and *pandas*.

Several approaches have been made by different authors, where you can highlight the contributions made [4] and [5], studies with the goal of solving the problem of fusing heterogeneous data, although the authors have some concerns, their proposal can be considered a valuable approach in this scientific field.

In [1] stands out the increase of multimedia resources, which have brought a greater need for extracting new knowledge from these resources, and their organization, showing future works such as the design of an intelligent system to generate association rules, as this study presents, to search rapidly for similar multimedia objects or to organize them based on themes related to keywords as the results obtained in figure 11.

The affinity analysis proposed, using association rules, has obtained interesting results and encourages to build heterogeneous datasets with a greater amount of data types such as images, text, audio and video, with the goal of discovering relationships between their elements or other multimedia objects.

6. CONCLUSIONS

Multimedia objects are resource available with permanent growth, currently they are heavily used to supply different needs, data mining allows to analyze massive information sources, and represents an alternative to explore multimedia objects searching to acquire relevant knowledge from them. The apriori algorithm is a method of association rules, and it is used to solve multiple problems according to the literature consulted, and

it becomes an alternative to the affinity analysis between multimedia objects with excellent results. The rules generated from Apriori have high values with the metric confidence, when identifies relationships between the documents that have image and text items and compose a multimedia object. Our experimentation permits to propose the association rules as a method to analyze all the components of a multimedia object, achieving new knowledge based in their structure or attributes. In future works would be possible to implement association rules to analyze elements as text, images, audio and video that can form a multimedia object.

REFERENCES:

- [1] Hu, C., Xu, Z., Liu, Y., Mei, L., Chen, L., & Luo, X. (2014). Semantic link network-based model for organizing multimedia big data. *IEEE Transactions on Emerging Topics in Computing*, 2(3), 376-387.
- [2] Tešić, J., Newsam, S., & Manjunath, B. S. (2003). Mining image datasets using perceptual association rules. In *Proc. SIAM Sixth Workshop on Mining Scientific and Engineering Datasets in conjunction with SDM*.
- [3] Zheng, Q. F., Wang, W. Q., & Gao, W. (2006, October). Effective and efficient object-based image retrieval using visual phrases. In *Proceedings of the 14th ACM international conference on Multimedia* (pp. 77-80). ACM.
- [4] Jiang, T., & Tan, A. H. (2009). Learning image-text associations. *IEEE Transactions on Knowledge and Data Engineering*, 21(2), 161-177.
- [5] Alghamdi, R. A., Taileb, M., & Ameen, M. (2014, April). A new multimodal fusion method based on association rules mining for image retrieval. In *Mediterranean Electrotechnical Conference (MELECON), 2014 17th IEEE* (pp. 493-499). IEEE.
- [6] Grosky, W. I. (1997). Managing multimedia information in database systems. *Communications of the ACM*, 40(12), 72-80.
- [7] Yang, Y., Zhuang, Y. T., Wu, F., & Pan, Y. H. (2008). Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. *Multimedia, IEEE Transactions on*, 10(3), 437-446.
- [8] Zhuang, Y. T., Yang, Y., & Wu, F. (2008). Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval.

- Multimedia, IEEE Transactions on, 10(2), 221-229.
- [9] Hunter, J., & Choudhury, S. (2003). Implementing preservation strategies for complex multimedia objects. In *Research and Advanced Technology for Digital Libraries* (pp. 473-486). Springer Berlin Heidelberg.
- [10] Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International journal of computer vision*, 7(1), 11-32.
- [11] Little, T. D., & Ghafoor, A. (1990). Synchronization and storage models for multimedia objects. *Selected Areas in Communications, IEEE Journal on*, 8(3), 413-427.
- [12] W. Ma and B. S. Manjunath, "A texture thesaurus for browsing large aerial photographs," *Journal of the American Society of Information Science*, 1998.
- [13] Malik, H. H., & Kender, J. R. (2006, July). Clustering web images using association rules, interestingness measures, and hypergraph partitions. In *Proceedings of the 6th international conference on Web engineering* (pp. 48-55). ACM.
- [14] Chen, C. L., Tseng, F. S., & Liang, T. (2010). An integration of WordNet and fuzzy association rule mining for multi-label document clustering. *Data & Knowledge Engineering*, 69(11), 1208-1226.
- [15] Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *ACM SIGMOD Record* (Vol. 22, No. 2, pp. 207-216). ACM.
- [16] Mustafa, M. D., Nabila, N. F., Evans, D. J., Saman, M. Y., & Mamat, A. (2006). Association rules on significant rare data using second support. *International Journal of Computer Mathematics*, 83(1), 69-80.
- [17] Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).
- [18] Agrawal, R., & Shafer, J. C. (1996). Parallel mining of association rules. *IEEE Transactions on Knowledge & Data Engineering*, 6(6), 962-969.
- [19] Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1), 53-87.
- [20] Hanguang, L., & Yu, N. (2012). Intrusion detection technology research based on apriori algorithm. *Physics Procedia*, 24, 1615-1620.
- [21] Xiang, L. I. (2012). Simulation System of Car Crash Test in C-NCAP Analysis Based on an Improved Apriori Algorithm*. *Physics Procedia*, 25, 2066-2071.
- [22] Tsuji, K., Takizawa, N., Sato, S., Ikeuchi, U., Ikeuchi, A., Yoshikane, F., & Itsumura, H. (2014). *Book Recommendation Based on Library Loan Records and Bibliographic Information*. *Procedia-Social and Behavioral Sciences*, 147, 478-486.
- [23] Xu, Y., Li, Y., & Shaw, G. (2011). Reliable representations for association rules. *Data & Knowledge Engineering*, 70(6), 555-575.
- [24] Domingues, M. (2004). *Generalization of association rules* (Tesis de Maestria). Escola de Engenharia de São Carlos, Brasil.
- [25] Zaki, M. J., Parthasarathy, S., Ogihara, M., & Li, W. (1997, August). New Algorithms for Fast Discovery of Association Rules. In *KDD* (Vol. 97, pp. 283-286).
- [26] Han, J., Pei, J., & Yin, Y. (2000, May). Mining frequent patterns without candidate generation. In *ACM SIGMOD Record* (Vol. 29, No. 2, pp. 1-12). ACM.
- [27] Savasere, A., Omiecinski, E. R., & Navathe, S. B. (1995). An efficient algorithm for mining association rules in large databases.
- [28] Liu, B., Hsu, W., & Ma, Y. (1998, August). Integrating classification and association rule mining. In *Proceedings of the fourth international conference on knowledge discovery and data mining*.
- [29] Das, A., Ng, W. K., & Woon, Y. K. (2001, October). Rapid association rule mining. In *Proceedings of the tenth international conference on Information and knowledge management* (pp. 474-481). ACM.
- [30] Li, W., Han, J., & Pei, J. (2001). CMAR: Accurate and efficient classification based on multiple class-association rules. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on* (pp. 369-376). IEEE.
- [31] Yin, X., & Han, J. (2003, May). CPAR: Classification based on Predictive Association Rules. In *SDM* (Vol. 3, pp. 369-376).
- [32] Thabtah, F., Cowling, P., & Peng, Y. (2004, November). MMAC: A new multi-class, multi-label associative classification approach. In *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on* (pp. 217-224). IEEE.

- [33] Juan, L., & De-ting, M. (2010, October). Research of an association rule mining algorithm based on FP tree. In Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on (Vol. 1, pp. 559-563). IEEE.
- [34] Narvekar, M., & Syed, S. F. (2015). An Optimized Algorithm for Association Rule Mining Using FP Tree. *Procedia Computer Science*, 45, 101-110.
- [35] Bhandari, A., Gupta, A., & Das, D. (2015). Improved Apriori Algorithm using frequent pattern tree for real time applications in data mining. *Procedia Computer Science*, 46, 644-651.
- [36] Pinho., J. (2010). Métodos de Clasificación basados en asociación aplicados a sistemas de Recomendación (Tesis de Doctorado). Universidad de Salamanca, España.
- [37] Kotsiantis, S., & Kanellopoulos, D. (2006). Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, 32(1), 71-82.
- [38] Azevedo, P. J., & Jorge, A. M. (2007). Comparing rule measures for predictive association rules. In *Machine Learning: ECML 2007* (pp. 510-517). Springer Berlin Heidelberg.