Revista Ingenierías Universidad de Medellín
ISSN: 1692-3324
Universidad de Medellín

Oviedo-Carrascal, Efraín Alberto; Oviedo-Carrascal,
Ana Isabel; Velez-Saldarriaga, Gloria Liliana

Minería multimedia: hacia la construcción de una metodología y una herramienta de analítica de datos no estructurados*

Revista Ingenierías Universidad de Medellín, vol.
16, núm. 31, 2017, Julio-Diciembre, pp. 125-142
Universidad de Medellín

DOI: https://doi.org/10.22395/rium.v16n31a6

Disponible en: https://www.redalyc.org/articulo.oa?id=75055115007

Minería multimedia: hacia la construcción de una metodología y una herramienta de analítica de datos no estructurados*

*Efraín Alberto Oviedo Carrascal***
*Ana Isabel Oviedo Carrascal****
*Gloria Liliana Velez Saldarriaga*****

Recibido: 04/03/2016 • Aceptado: 13/12/2016
DOI: 10.22395/rium.v16n31a6

Resumen

En este trabajo se aborda el desarrollo de proyectos de minería multimedia con la aplicación de técnicas analíticas a textos, imágenes, audios y videos. Para aportar al desarrollo de estos proyectos, se propone una metodología para desarrollar proyectos de minería multimedia llamada Metodología de Analítica Multimedia (MAM). Así mismo, se presenta la construcción de una herramienta software que permite el análisis de minería multimedia llamada Plataforma de Analítica Multimedia (PAM). La metodología y la plataforma son evaluadas con dos casos de estudio sobre predicción de anormalidades en mamografías y análisis de similaridad en imagenología médica. Los resultados obtenidos permitieron validar los pasos propuestos en la metodología MAM y utilizar la plataforma PAM para extraer las características de las imágenes médicas, aplicar técnicas de minería de datos y evaluar satisfactoriamente los resultados obtenidos.

Palabras clave: minería de datos; minería multimedia; metodologías de minería de datos; plataformas para minería de datos.

* Este artículo es resultado de un proyecto de investigación ya finalizado, titulado "Plataforma de minería de datos estructurados y no estructurados - Caso de estudio salud pública", financiado por la Universidad Pontificia Bolivariana, sede Medellín y realizado entre 2015 y 2016.

** Estudiante de Maestría en TIC en la Universidad Pontificia Bolivariana, Medellín. Correo electrónico: eaoc46@gmail.com

*** PhD. Docente Investigadora de la Facultad de Ingeniería en TIC de la Universidad Pontificia Bolivariana, Medellín. Correo electrónico: ana.oviedo@upb.edu.co

**** PhD. Docente Investigadora de la Facultad de Ingeniería en TIC de la Universidad Pontificia Bolivariana, Medellín. Correo electrónico: gloria.velez@upb.edu.co

Multimedia mining: towards the construction of a methodology and a non-structured date analytics tool

Abstract

This research addresses the development of multimedia mining projects by applying analytical techniques to texts, images, audio, and video. In order to develop these projects, a methodology to develop multimedia mining projects (Multimedia Analytical Methodology-MAM) is proposed. Likewise, the construction of a software tool (known as Multimedia Analytical Platform-PAM) which allows the analysis of multimedia mining is introduced. Methodology and platform are evaluated with two study cases on prediction of mammography abnormalities and analysis of medical imaging similarity. Results obtained allowed validating the steps proposed in the MAM methodology and using the PAM platform to extract the characteristics of medical images, to apply data mining techniques, and to satisfactorily evaluate the results obtained.

Keywords: data mining; multimedia mining; data mining methodologies; data mining platforms.

Mineração multimídia: rumo à construção de uma metodologia e de uma ferramenta analítica de dados não estruturados

Resumo

Neste trabalho, aborda-se o desenvolvimento de projetos de mineração de dados multimídia com a aplicação de técnicas analíticas a textos, imagens, áudios e vídeos. Para contribuir para o desenvolvimento desses projetos, propõe-se uma metodologia para desenvolver projetos de mineração multimídia chamada Metodologia de Analítica Multimídia (MAM). Além disso, apresenta-se a construção de uma ferramenta (software) que permite a análise de mineração multimídia chamada Plataforma de Analítica Multimídia (PAM). A metodologia e a plataforma são avaliadas com dois casos de estudo sobre predição de anormalidades em mamografias e análises de similaridade em imagenologia médica. Os resultados obtidos permitiram validar os passos propostos na metodologia MAM e utilizar a PAM para extrair as características das imagens médicas, aplicar técnicas de mineração de dados e avaliar satisfatoriamente os resultados.

Palavras-chave: mineração de dados; mineração multimídia; metodologias de mineração de dados; plataformas para mineração de dados.

INTRODUCCIÓN

La minería de datos se concibe como el proceso mediante el cual es posible descubrir información no trivial a partir de grandes volúmenes de datos, utilizando técnicas de inteligencia artificial y estadística [1]. En una revisión sobre las aplicaciones de la minería de datos, se puede encontrar gran variedad de áreas como el sector salud, finanzas, banca, educación, biología, entre otras. Estas aplicaciones son realizadas en su gran mayoría sobre datos estructurados, es decir, datos organizados en bases de datos. De igual manera, las metodologías y plataformas para minería han centrado sus esfuerzos sobre minería de datos estructurados; de hecho, se puede encontrar diversidad de opciones de acceso libre y licenciado.

Sin embargo, no es en los datos estructurados en donde se encuentra el mayor volumen de información que existe al nivel mundial, sino en multimedia como es el caso de imágenes, texto, audio y vídeos. Así, la minería de datos tiene nuevos requerimientos gracias a lo que hoy se conoce como *big data* [1], que no solo incluye el procesamiento de grandes volúmenes de datos, sino también el procesamiento de variedad de datos multimedia.

De forma general, la minería multimedia requiere la incorporación de distintos tipos de contenidos no estructurados para obtener información no trivial por medio de técnicas de analítica. Para estos análisis se requiere de metodologías y plataformas que brinden soporte para este tipo de datos, los cuales no se encuentran estructurados en bases de datos que permitan aplicarse de forma directa a procesos de analítica. La aplicabilidad de la minería multimedia es tan amplia como las aplicaciones de la minería de datos estructurados. El análisis de imágenes, por ejemplo, puede facilitar las tareas de la medicina ofreciendo apoyo a la toma de decisiones a partir de imágenes diagnósticas. El análisis de texto permite, por ejemplo, identificar la favorabilidad de algún tema o situación en particular en redes sociales.

Actualmente, los proyectos de minería multimedia deben afrontar desafíos no solo con el procesamiento multimedia, sino también con la aplicación de metodologías de minería, las cuales no incluyen procesos de indexación de multimedia, al igual que las plataformas no tienen soporte para procesamiento de imágenes, audios y vídeos. Con el objetivo de apoyar el desarrollo de proyectos de minería multimedia, en este trabajo se propone una metodología y se desarrolla una plataforma (herramienta de *softwar*e) que permiten realizar minería multimedia, haciendo énfasis en la preparación de los datos no estructurados para su tratamiento por técnicas convencionales. Como antecedente, se presentó en [2] un estudio sobre la minería de datos enfocada hacia los servicios de salud en las ciudades inteligentes; dicho estudio permitió identificar los requisitos de la minería multimedia. La organización del artículo es la siguiente. En

la sección 1 se presenta una revisión literaria de minería de datos y multimedia. En la sección 2, de materiales y métodos, se propone una metodología para minería multimedia llamada MAM (Metodología de Analítica Multimedia) y una plataforma para minería multimedia llamada PAM (Plataforma de Analítica Multimedia). En la sección 3 se presentan dos casos de estudios realizados sobre minería de imágenes en el área de la salud. Finalmente, en la sección 4 se presentan las conclusiones y trabajo futuro.

1. REVISIÓN LITERARIA

Para abordar el desarrollo de este artículo, se presenta, a continuación, una revisión literaria que aborda el panorama desde la minería de datos hasta la minería multimedia.

1.1 Minería de datos

En la minería de datos se pueden desarrollar dos tipos de análisis: predictivos y descriptivos [2]. El análisis predictivo abarca las tareas de clasificación, donde se predicen categorías, y la regresión donde se predicen números. Algunas de las técnicas más convencionales de análisis predictivo son las redes neuronales, máquinas de soporte vectorial, árboles de decisión, regresión, bayesianos, y k-vecinos más cercanos.

Por su lado, el análisis descriptivo incluye las tareas de *clustering* donde se describe en forma de grupos, y la asociación donde se describe en forma de reglas. Las técnicas descriptivas más convencionales son k-means, cob-web y a priori.

Para facilitar el desarrollo de proyectos de minería de datos se han formulado diversas metodologías y herramientas *software*. Las metodologías guían el paso a paso del proceso de minería. Algunas de las más destacadas son CRISP-DM (*Cross Industry Standard Process for Data Mining*) creada por un consorcio de empresas europeas dentro de las cuales se destaca IBM siendo una de las metodologías más utilizada [3]; SEMMA (*Sample, Explore, Modify, Model and Assess*) diseñada por *SAS Analytics Solutions* con una etapa de muestreo estadístico no considerada en otras metodologías [4]; KDD (*Knowledge Discovery in Database*) plantea el descubrimiento de conocimiento en bases de datos [5] y Catalyst también conocida como la metodología P3TQ (*Product, Place, Price, Time, Quantity*) que, además, plantea un modelo de negocios [6].

Estas metodologías comparten la mayoría de las etapas: preparación de datos, modelamiento y evaluación. Sin embargo, estas metodologías no ofrecen fases de procesamiento multimedia, aunque algunas de ellas han sido adaptadas para aplicaciones de la minería de texto [7] [8].

Por su parte, las plataformas de minería de datos se han creado para facilitar la aplicación de las técnicas ya que pueden presentar cierta complejidad matemática

y estadística. Algunas de estas plataformas son: WEKA (*Waikato Environment for Knowledge Analysis*) desarrollada en la Universidad de Waikato en Nueva Zelanda distribuida bajo la licencia GNU [9]; RapidMiner creada en la Universidad de Dormund distribuida bajo la licencia GPL [10]; R es una plataforma libre creada en la Universidad de Auckland de Nueva Zelanda con orientación a cálculos estadísticos y gráficas para la presentación de nuevo conocimiento [11]; SPSS consiste en un paquete estadístico que contiene un módulo dedicado a la minería de datos denominado *SPSS Modeler* [12] licenciado por IBM, y *SAS Institute* consiste en una serie de paquetes entre los cuales se destaca *SAS Analytics* que permite realizar análisis predictivo y descriptivo para minería de datos [13].

Estas plataformas han sido comparadas en diversas ocasiones [8]; en general, se destacan WEKA y RapidMiner como las más descargadas de Internet y las más utilizadas. La mayoría de estas plataformas ya han empezado a incluir módulos para minería de textos; sin embargo, es un requerimiento nuevo y exigente incluir soporte para minería de imágenes, audios y vídeos.

1.2 Minería multimedia

En la literatura, se pueden encontrar diversas aplicaciones de la minería a datos multimedia como texto, imágenes, audios y vídeos. La minería de texto estudia la aplicación de técnicas analíticas a documentos de texto. El enfoque más utilizado es llamado bolsa de palabras, donde los documentos son representados por vectores de características mediante la función *TF x IDF*; el primer término *TF* se refiere a la frecuencia de aparición de cada término, mientras que *IDF* es una ponderación como la frecuencia inversa de los términos en el conjunto de documentos [14].

La minería de imágenes, comúnmente, realiza una caracterización por medio de histogramas de color en diversos espacios como RGB, HSV, YUV, entre otros. Adicionalmente, se pueden extraer características de textura y forma [14].

La minería de audio, comúnmente, realiza un enventanamiento para extraer características de bajo nivel en el dominio del tiempo y la frecuencia. Algunas características del dominio del tiempo son la energía, cruces por cero y silencios. Características del dominio de la frecuencia son pitch, centroide de la frecuencia, ancho de banda de la frecuencia, coeficientes cepstrales de la frecuencia de Mel y coeficientes de predicción lineal [14].

Finalmente, la minería de vídeos se presenta como la conjunción de características de textos, imágenes y audio. Como aporte final, se puede identificar que la minería de texto, imágenes y vídeos presenta problemas de dimensionamiento, ya que la multi-

media es representada por vectores de características en alta dimensión, lo que afecta el desempeño de métodos de analítica.

Las aplicaciones más comunes de la minería multimedia se enfocan a procesamiento de texto e imágenes. En [7] se presenta una aplicación de algoritmos de clasificación de minería de textos orientado hacia el sector de la educación para identificar las habilidades de tutores. En [8] se utiliza la minería de texto en el diseño de un modelo de clasificación de opiniones subjetivas, realizando una comparación entre las plataformas de minería de datos WEKA y Rapidminer.

Por su parte, la minería de imágenes se ha convertido en una alternativa muy importante en diversas áreas, entre ellas se destacan los diagnósticos médicos. Se ha encontrado evidencia de este tipo de aplicaciones orientadas a enfermedades como alzheimer, cáncer, apendicitis, problemas digestivos, entre otras. En [15] se presenta un estudio con tres algoritmos de clasificación: árboles de decisión, algoritmos genéticos y k-means, utilizados para el análisis de imágenes biomédicas.

Aunque en la revisión literaria se encuentran diversas aplicaciones de minería multimedia, no se evidencia la existencia de proyectos similares donde se desarrolle una plataforma para minería de texto, imágenes, audios y vídeos. Cabe destacar que algunas plataformas ya incluyen el procesamiento de texto.

2. MATERIALES Y MÉTODOS

La metodología seguida para el desarrollo de este proyecto se basa en los siguientes pasos:

- Definición de requisitos para el desarrollo de proyectos de minería multimedia a través de una revisión literaria sobre el área.
- Formulación de una metodología de minería multimedia que incluya el preprocesamiento y la indexación de material multimedia.
- A través de una metodología ágil de desarrollo de *software*, se crea un prototipo de una plataforma de minería multimedia.
- Finalmente, la metodología de minería multimedia propuesta y la plataforma desarrollada son validadas mediante dos casos de estudio.

2.1 Formulación de una metodología para minería multimedia

A pesar de que las metodologías de minería incluyen etapas de preparación de los datos, solo contemplan análisis estadísticos y transformaciones que se pueden realizar directamente sobre datos estructurados, no sobre datos multimedia.

Teniendo en cuenta que los datos multimedia requieren etapas de pre procesamiento e indexación que permitan representarlos en vectores característicos que se puedan presentar a las técnicas de minería de datos, en este artículo se propone una metodología para minería multimedia llamada Metodología de Analítica Multimedia (MAM).

La figura 1 presenta gráficamente la metodología MAM con un ciclo de 5 etapas: preprocesamiento de multimedia, indexación de multimedia, preparación de datos estructurados, modelamiento analítico y evaluación.

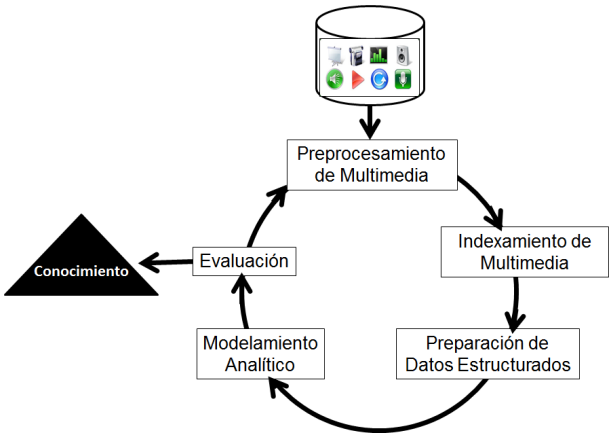

Figura 1: Metodología de minería multimedia MAM

Fuente: elaboración propia

La etapa de p*reprocesamiento de multimedia* realiza procesos de limpieza, dependiendo del tipo de multimedia a analizar. En la tabla 1 se presenta una descripción del preprocesamiento necesario para diferentes tipos de multimedia.

Tabla 1: Metodología MAM. Etapa de preprocesamiento

Multimedia	*Pre procesamiento*
Textos	En el enfoque conocido como bolsa de palabras: • Se eliminan stopwords. • Reducción de raíces (stemming).
Imágenes	Se pueden realizar diferentes actividades dependiendo de las imágenes a analizar: • Cambiar el formato de la imagen a png, jpg, etc. • Segmentación para extraer el segmento de la imagen de nuestro interés y así eliminar ruido. • Cambiar el espacio de color a binario, escala de grises, HSV, YUV, etc. • Remuestreo de las imágenes al mismo tamaño.

Multimedia	*Pre procesamiento*
Audios	A todos los archivos de audio se debe realizar: • Remuestreo de los audios a la misma resolución.
Videos	Para preprocesar videos, las acciones más comunes son: • Detección de escenas. • Selección de una imagen por escena. • Preprocesamiento de las imágenes.

Fuente: elaboración propia

En la etapa de *indexación de multimedia*, se representan los datos no estructurados en vectores numéricos por medio de procesos de extracción de características. Esta etapa también depende del tipo de recurso multimedia analizado como se presenta en la tabla 2.

Tabla 2: Metodología MAM. Etapa de Indexamiento

Multimedia	*Indexamiento*
Textos	En el enfoque conocido como bolsa de palabras, se calcula: • Frecuencia de las palabras en cada documento. • Frecuencia global de las palabras. También se aplica para términos o frases en lugar de palabras.
Imágenes	Se pueden calcular diferentes características como: • Características de color. • Características de textura. • Características de forma.
Audios	Se divide la señal de audio en ventanas y se extraen: • Características en el dominio del tiempo. • Características en el dominio de la frecuencia.
Videos	Se extraen características de las imágenes estáticas de las escenas. De igual forma, se pueden incluir el texto asociado al video por medio del enfoque de bolsa de palabras. Para un mayor cubrimiento de las características, también se pueden incluir características del audio.

Fuente: elaboración propia

En la etapa de p*reparación de datos estructurados*, se desarrollan las actividades convencionales de minería de datos estructurados, ya que en esta instancia la multimedia fue convertida y representada como tablas de datos. Las actividades recomendadas

en esta etapa son: detección de nulos y atípicos, análisis de correlaciones, selección de variables, transformaciones y balanceo de datos.

En la etapa de *modelamiento analítico* se aplican las técnicas de minería de datos dependiendo del tipo de análisis a realizar. En el análisis predictivo se resaltan técnicas como redes neuronales, árboles de decisión, máquinas de soporte vectorial, regresión y métodos perezosos como K-vecinos más cercanos. En el análisis descriptivo se resaltan técnicas como k-means para clustering y a-priori para reglas de asociación.

Finalmente, en la etapa de *evaluación*, se verifica la confianza de los resultados obtenidos y se realiza un proceso de interpretación para encontrar nuevo conocimiento. La evaluación depende del tipo de análisis realizado; para análisis predictivos se utilizan diferentes mediciones del error, la matriz de confusión, *precisión, recall* y área ROC. La matriz de confusión presenta el desempeño de un clasificador en términos de las cantidades: verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. La medida *precisión* indica el porcentaje de aciertos al predecir los datos pertenecientes a la clase de interés. La medida *recall* determina el grado de cobertura de los datos de la clase de interés. El área ROC (*Receiver Operating Characteristic*) grafica la razón de verdaderos positivos frente a la razón de falsos positivos.

En análisis descriptivos se utilizan índices como la silueta de los datos cuando se crean *clústeres* y la confianza cuando se crean reglas de asociación. La silueta de los datos se define como:

$$Silueta\left(clustering\right)=\frac{\sum_{i=1}^{n} sil_i}{n}, y\, sil_i=\frac{\left(b_i-a_i\right)}{\max\left(b_i,a_i\right)}, \tag{1}$$

donde n es la cantidad de datos, a_i es el promedio de las distancias del dato x_i a los registros en el mismo *cluster*, y b_i es el promedio de las distancias del dato x_i a los objetos en otros *clusters*. El valor óptimo se encuentra en una cantidad cercana a 1. Por su parte, la confianza de las reglas de asociación se define como la proporción de registros en donde el consecuente aparece acompañando el precedente.

2.2 Desarrollo de una plataforma para minería multimedia

En esta sección se describe el diseño e implementación de una plataforma de minería multimedia llamada Plataforma de Analítica Multimedia (PAM). La plataforma se desarrolló en el lenguaje de programación JAVA y se utilizaron librerías bajo licencia GNU-GPL como Lucene para la extracción de características de texto, LIRE para la extracción de características de imágenes y WEKA para las técnicas de minería de datos.

En una primera versión de PAM, se tienen disponibles procesos de minería de texto e imágenes, con las condiciones presentadas en la tabla 3.

Tabla 3: Plataforma PAM. Condiciones implementadas

Condición	Descripción
Perfiles de usuario	• Los usuarios con perfil de administración tienen acceso a la gestión de usuarios, a la configuración del proceso de indexamiento multimedia y las técnicas de minería. • Los usuarios con perfil de operador pueden cargar un conjunto de multimedia y realizar dos tipos de análisis: predictivo y descriptivo.
Indexamiento de texto	Por medio de la librería Lucene se pueden extraer las siguientes características textuales: • Representación binaria de la aparición de los términos. • Frecuencia de los términos. • Función TF x IDF con la frecuencia de los términos y la ponderación de la frecuencia inversa de los términos en los documentos.
Indexamiento de imágenes	Por medio de la librería LIRE se pueden extraer características de color, textura y forma como: • Auto Color Correlogram: correlación del color en una imagen en el espacio de color HSV. • Color Layout: distribución espacial del color en una imagen. • Scalable Color: histograma de color en el espacio HSV utilizando la transformada Haar. • Simple Color Histogram: histograma del color de una imagen en el espacio de color RGB. • Fuzzy Color Histogram: histograma de color que se caracteriza por ser poco sensible a la interferencia de ruidos generados por factores como la intensidad de la iluminación. • Edge Histogram: distribución espacial de cinco tipos de bordes. Para ello divide la imagen en 16 sub imágenes y para cada una de ellas genera el histograma. • Tamura: utiliza seis diferentes tipos de textura que puede percibir el ser humano mediante las cuales busca identificar el tipo de superficies presentes en una imagen. • PHOG: apariencia global de una imagen representando la forma y distribución espacial. • Local Binary Patterns: textura de una imagen a partir de la conversión de la misma a escala de grises.
Técnicas implementadas	• Análisis predictivo: máquinas de soporte vectorial y árboles de decisión. • Análisis descriptivo: método k-means para realizar clustering.

Fuente: elaboración propia

Las técnicas a implementar fueron seleccionadas en un estudio previo [4], donde se examinaron los algoritmos de minería de datos más utilizados para procesamiento multimedia.

En la figura 2 se presenta un diagrama de actividades donde se describe el proceso de minería multimedia en la plataforma PAM. Al iniciar la sesión, se debe indicar qué tipo de multimedia se desea procesar (imágenes o texto), se procede con la carga de la multimedia a la plataforma y posteriormente se realiza el respectivo preprocesamiento e indexación. En este punto se debe tomar una decisión sobre el tipo de análisis que se desea realizar. En caso de realizar análisis predictivo se debe seleccionar la técnica a utilizar, ingresar el número de clases indicado una etiqueta para cada una y realizar las etapas de aprendizaje y evaluación de resultados. Como resultado del proceso, la plataforma presenta la matriz de confusión obtenida de la evaluación del modelo de clasificación creado. En caso de realizar análisis de tipo descriptivo, solo se debe indicar el número de *clusters* esperados. Esta información es suficiente para que la plataforma realice el proceso de agrupamiento e indique para cada una de los datos el *clúster* asignado y el índice de la silueta respectivo.

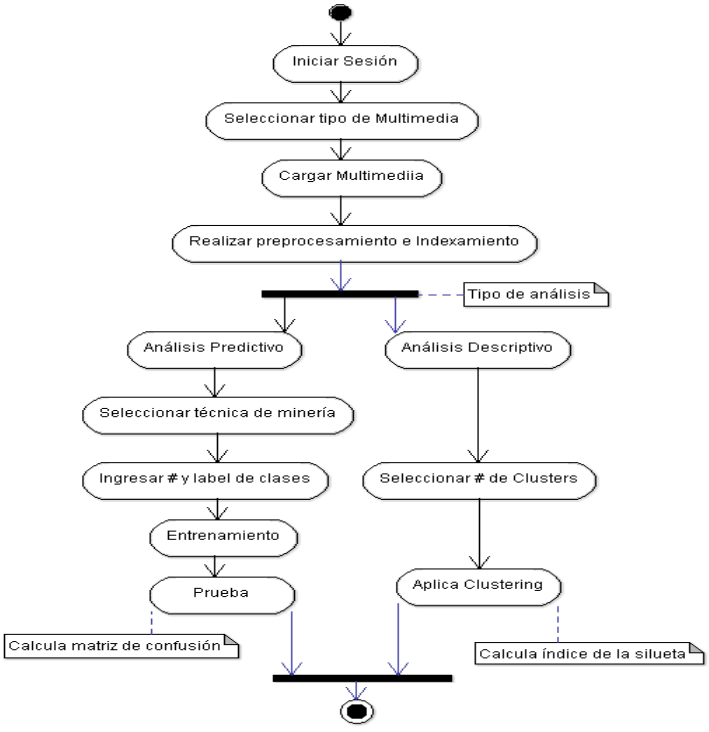

Figura 2: Plataforma PAM. Diagrama de actividades

Fuente: elaboración propia

En la figura 3 se presenta la interfaz principal de la plataforma PAM, cuando se realiza minería a imágenes. A la izquierda se puede observar un menú de opciones que permiten desarrollar los pasos de la metodología MAM propuesta. Para realizar minería de texto se utiliza una interfaz similar.

Figura 3: Interfaz de la plataforma PAM

Fuente: elaboración propia

3. CASOS DE ESTUDIO

Con el objetivo de validar la metodología MAM y la plataforma PAM, se desarrollan dos casos de estudio relacionados con la minería de imágenes.

4. PREDICCIÓN DE ANORMALIDADES EN MAMOGRAFÍAS

La mamografía es un examen médico que consiste en tomar una radiografía de los senos con el fin de detectar signos de cáncer. Mediante este examen es posible detectar micro calcificaciones que son pequeños depósitos de calcio que pueden indicar la presencia de cáncer de seno. Para este caso de estudio se utilizó la base de datos *Mamography Image Analyze Society* (MIAS), que es una sociedad inglesa que se dedica a la investigación de las mamografías. Esta base de datos [16] incluye exámenes de ambos senos de 161 pacientes para un total de 322 imágenes. Las anormalidades consisten en una pequeña masa que muestra presencia de cáncer maligno. Los diagnósticos de la base de datos MIAS han sido realizados por radiólogos expertos.

En la figura 4 se presentan dos mamografías seleccionadas de la base de datos MIAS. El examen de la izquierda corresponde a un caso diagnosticado como normal,

mientras que el examen de la derecha presenta una anormalidad que ha sido encerrada con fines ilustrativos.

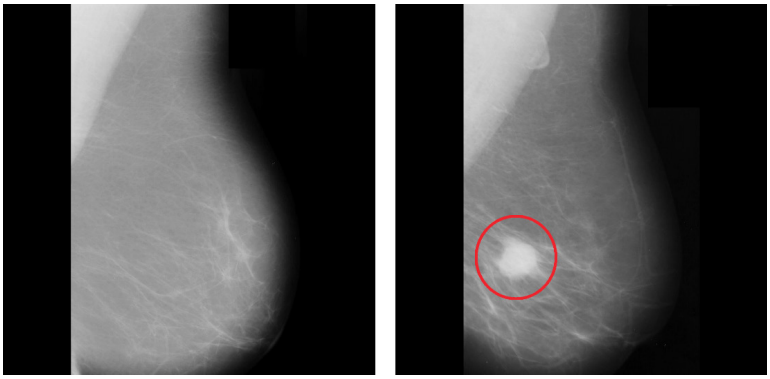

Figura 4: Mamografías de la base de datos MIAS del caso de estudio de predicción de anormalidades

Fuente: imágenes obtenidas de la base de datos MIAS [16]

A continuación, se describe la aplicación de la metodología MAM para este caso de estudio. En la etapa de *preprocesamiento de multimedia*, las imágenes de esta base de datos se encuentran en formato PGM. Este formato no es compatible con la plataforma; por este motivo fue necesario realizar una etapa de conversión de las imágenes del formato PGM al formato JPG compatible con la plataforma.

En la etapa de *indexación de multimedia,* se realiza la extracción de características de cada una de las imágenes. Las características extraídas deben permitir la identificación de las imágenes en las dos clases que se pretende clasificar: diagnóstico normal y diagnóstico anormal. En las imágenes seleccionadas se observa que las mamografías con diagnóstico anormal presentan algunas masas con una textura, forma y color representativos. Por esta razón se optó por extraer simultáneamente características de color, forma y textura.

En la etapa de *preparación de datos estructurados,* se selecciona la variable objetivo, la cual indica la existencia de anormalidades.

En la etapa de *modelamiento analítico,* para tareas de clasificación la plataforma ofrece dos técnicas: máquinas de soporte vectorial y árboles de decisión. Para cada una de estas técnicas se crearon modelos analíticos, utilizando el 70% de los datos para la etapa de entrenamiento y un 30% para la etapa de evaluación.

En la etapa de *evaluación,* el modelo que presentó un mejor resultado utiliza la técnica de árboles de decisión con un 86% de instancias clasificadas correctamente. En la tabla 4 se presenta la matriz de confusión generada, de donde la precisión indica que el 90% de anormalidades reconocidas estaban correctas, es decir, que se tuvieron el 10%

de falsos positivos. La cobertura indica que se reconocieron 75% de anormalidades, es decir, que 25% de las anormalidades no fueron reconocidas. Finalmente, el área ROC indica un desempeño general del 85.6%, lo cual lo posiciona en un buen clasificador.

Tabla 4: Matriz de confusión obtenida en el caso de estudio de predicción de anormalidades en mamografías.

Clasificado como		*Diagnóstico real*
Normal	*Anormal*	
34	2	Normal
6	18	Anormal

Fuente: elaboración propia

5. ANÁLISIS DE SIMILARIDADES EN IMAGENOLOGÍA MÉDICA

Una gran cantidad de exámenes médicos utilizan imágenes para el diagnóstico y toma de decisiones relacionados con la salud de un paciente. Cada tipo de examen presenta unas características propias que lo identifican de los demás tipos de exámenes, relacionadas principalmente con la estructura de la parte del cuerpo a la que se realiza el examen. En este caso de estudio se pretende realizar *clustering* para agrupar las imágenes de acuerdo con el tipo de examen médico al cual corresponden. Se utilizan imágenes de tres bases de datos distintas: 1) *Repository for Molecular BRAin Neoplasia DaTa* (Rembrandt) [17]; se trata de una serie de imágenes correspondientes a 110 casos de exámenes de resonancia magnética donde se diagnostica la presencia de tumores cerebrales; 2) JSRT perteneciente a la Sociedad Japonesa de Radiología Tecnológica [18] que tiene como objetivo estudiar la presencia de nódulos pulmonares a través del estudio de radiografías del tórax, y 3) MIAS [16] tiene como objetivo el análisis de mamografías para detectar la presencia de cáncer de seno. En la figura 5 se presenta una imagen de cada una de las bases de datos utilizadas. La imagen de la izquierda pertenece a la base de datos JSRT, la del medio pertenece a Rembrandt y la de la derecha pertenece a MIAS.

Las tres bases de datos mencionadas han sido utilizadas de forma independiente en diversos estudios de procesamiento de imágenes. En este caso de estudio se toman 27 muestras de cada una de ellas para conformar una base de datos de 81 imágenes para realizar *clustering* de tipos de exámenes médicos.

A continuación, se describe la aplicación de la metodología MAM para este caso de estudio. En la etapa de *preprocesamiento de multimedia*, las imágenes de la base de datos JSRT se encuentran originalmente en el formato IMG que no es compatible con la plataforma de *software*; lo mismo sucede con las imágenes de la base de datos

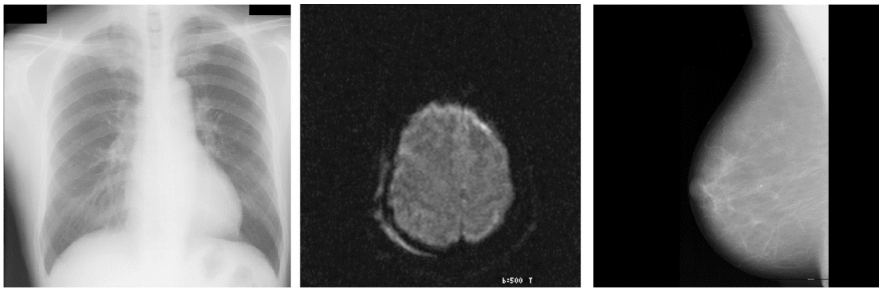

Figura 5: Exámenes médicos analizadas en el caso de estudio de clustering en imagenología médica
Fuente: imágenes obtenidas de las bases de datos JSRT [18], Rembrandt [17] y MIAS [16].

Rembrandt que se encuentran en formato DICOM y las imágenes de la base de datos MIAS que están en formato PGM. Por lo anterior, las imágenes seleccionadas fueron convertidas de su formato original a JPG.

En la etapa de *indexación de multimedia,* se realizó la extracción de características de color, forma y textura.

En la etapa de *preparación de datos estructurados,* se representan los datos en un archivo con formato *arff* compatible con WEKA.

En la etapa de *modelamiento analítico,* utilizando la técnica K-Means, se crean varios modelos para su posterior evaluación.

Finalmente, en la etapa de e*valuación,* se presenta la medida de silueta para analizar la calidad de los resultados. En la figura 6 se presenta un gráfico con el índice de la silueta calculado para cada una de las imágenes. En el caso del *cluste*r 1 se observa que todas la imágenes presentan un índice superior a 0.75; el *cluster* 2 presenta índices

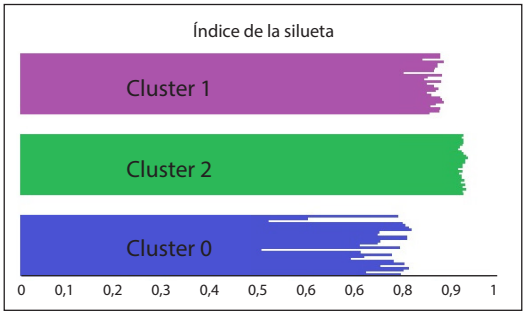

Figura 6: Índice de silueta obtenido en el caso de estudio de clustering en imagenología médica
Fuente: elaboración propia

superiores a 0.9, mientras que el *cluster* 0 presenta índices más bajos, pero siempre superiores a 0.5. En promedio, el índice de la silueta para este proceso de *clustering* es de 0.845, un valor que indica un desempeño alto al agrupar las imágenes de esta base de datos.

6. CONCLUSIONES

Con el objetivo de aportar en el desarrollo de proyectos de minería multimedia, en este trabajo se proponen una metodología y una plataforma de analítica para multimedia. La metodología propuesta, llamada Metodología de Analítica Multimedia (MAM), define etapas de preprocesamiento e indexación de multimedia. Con estas etapas se obtiene una representación de los datos en vectores característicos permitiendo aplicar las técnicas convencionales de minería de datos. Asimismo, se presenta la construcción de una herramienta *software* que permite el análisis de minería multimedia llamada Plataforma de Analítica Multimedia (PAM).

La metodología y la plataforma fueron validadas por medio de dos casos de estudio sobre imagenología médica que permitieron: 1) Validar los pasos propuestos en la metodología MAM, definiendo fases reproducibles en nuevos proyectos de minería multimedia, 2) Utilizar la plataforma PAM para extraer las características de las imágenes médicas, aplicar técnicas de minería de datos y evaluar satisfactoriamente los resultados obtenidos.

Como trabajo futuro se propone: 1) En la plataforma incluir una etapa de transformación donde se permita realizar actividades como la normalización y la discretización de datos, 2) Adicionar nuevas técnicas de minería y nuevos tipo de análisis como la regresión y la asociación, 3) En el procesamiento de imágenes, automatizar nuevas actividades como la segmentación y remuestreo de imágenes para que tengan el mismo tamaño, y 4) En el procesamiento multimedia, incluir en la plataforma el análisis de audios y vídeos.

7. AGRADECIMIENTOS

Este documento es resultado del proyecto de investigación *"Plataforma de minería de datos estructurados y no estructurados: caso de estudio salud pública",* registrado en el Centro de Investigación para el Desarrollo y la Innovación –CIDI– de la Universidad Pontificia Bolivariana (UPB). Los autores expresan su agradecimiento al grupo de investigación Gidati y al CIDI de la UPB.

REFERENCIAS

[1] X. Wu, X. Zhu, G. Wu y W. Ding, «Data mining with big data», *IEEE transactions on knowledge and data engineering,* vol. 26, n.º 1, pp. 97-107, 2014.

[2] E. A. Oviedo, A. I. Oviedo y G. L. Vélez, «Minería de datos: aportes y tendencias en el servicio de salud de ciudades inteligentes», *Revista Politécnica,* vol. 11, n.º 20, pp. 111-120, 2015.

[3] J. Moine, «Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo. Tesis de Maestría» Universidad Nacional de la Plata, Argentina, 2013.

[4] A. Azevedo y L. Rojão, «KDD, SEMMA and CRISP-DM: a parallel overview», *IADS-DM,* pp. 182-185, 2008.

[5] O. Maimon y L. Rokach, Data mining and knowledge discovery handbook, New Rork: Springer, 2005.

[6] D. Pyle, Business modeling and data mining, Morgan Kaufmann, 2003.

[7] P. Santana, R. Costaguta y D. Missio, «Aplicación de algoritmos de clasificación de minería de textos para el reconocimiento de habilidades de e-tutores colaborativos», *Revista Iberoamericana de Inteligencia Artificial,* pp. 57-67, 2014.

[8] M. Tapia, O. Ruiz y C. Chirinos, «Modelo de clasificación de opiniones subjetivas en redes sociales», *Ingeniería: Ciencia, Tecnología e Innovación,* 2014.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann y I. H. Witten, «The WEKA Data Mining Software: An Update», *SIGKDD Explorations,* pp. 10-18, 2009.

[10] M. Hofmann y K. Ralf, RapidMiner: Data Mining Use Cases and Business Analytics Applications, CRC Press, 2013.

[11] L. Torgo, Data mining with R: learning with case studies, Chapman & Hall / CRC., 2010.

[12] B. Devi, K. Rao, S. Setty y M. Rao, «Disaster Prediction System Using IBM SPSS Data Mining Tool», *International Journal of Engineering Trends and Technology (IJETT),* pp. 3352-3357, 2013.

[13] G. Fernandez, Data mining using SAS applications, CRC Press, 2010.

[14] A. I. Oviedo, J. Perea-Ortega, O. Ortega y E. Sanchis, «Video clustering based on the collaboration of multimedia clusterers,» de *CLEI 2012 XXXVIII Conferencia Latinoamericana en Informática*, Medellín, 2012.

[15] S. Suganthira, P. Thamilselvan, J. G. R. Sathiaseelan y M. Lakshmiprabha, «A Technical Study on Biomedical image Classification using Mining Algorithms,» de *National Conference on Recent Advancements in Software Development (NCRASD-2015)*, Karaikudi, 2015.

[16] J. Suckling, J. Parker, D. R. Dance, S. Astley, I. Hutt, C. Boggis y J. Savage, «The mammographic image analysis society digital mammogram database,» *In Exerpta Medica. International Congress Series,* pp. 375-378, 1994.

[17] D. A. Wainwright, I. V. Balyasnikova, A. L. Chang, A. U. Ahmed, K. S. Moon, B. Auffinger y M. S. Lesniak, «IDO Expression in Brain Tumors Increases the Recruitment of Regulatory T Cells and Negatively Impacts Survival,» *Clinical cancer research,* vol. 18, n.º 22, pp. 6110-6121, 2012.

[18] J. Shiraishi, H. Abe, R. Engelmann y K. Doi, «Effect of high sensitivity in a computerized scheme for detecting extremely subtle solitary pulmonary nodules in chest radiographs: observer performance study», *Academic radiology,* vol. 10, n.º 11, pp. 1302-1311, 2003.

[19] J. Mena, Data mining your website, Digital Press, 1999.

[20] D. Corrales, A. Ledesma, A. Peña, J. Hoyos, A. Figueroa y J. Corrales, "A new dataset for coffee rust detection in Colombian crops base on classifiers," Revista S&T, pp. 9-23, 2014.

[21] J. Riquelme, R. Ruiz y K. Gilbert, "Minería de datos: Conceptos y tendencias," vol. 10, nº 29, pp. 11-18, 2006.

[22] D. Torres, "Diseño y aplicación de una metodología para análisis de noticias policiales utilizando minería de textos," Universidad de Chile, 2013.