Universidad de las Ciencias Informáticas

Facultad 6

Obtención de preferencias de usuario a partir de métodos estadísticos de Minería de Datos

TRABAJO DE DIPLOMA PARA OPTAR POR EL TÍTULO DE INGENIERO EN CIENCIAS INFORMÁTICAS

Autor: Diana Delisle Rivero

Tutor: Ing. Yunier Albrecht Delgado

Co-Tutor: Msc. Yunier Emilio Tejeda Rodríguez

Ciudad de la Habana, Junio de 2011

Año 53 de la Revolución

"La ciencia tiene una característica maravillosa, y es que aprende de sus errores, que utiliza sus equivocaciones para reexaminar los problemas y volver intentar resolverlos, cada vez por nuevos caminos"

Ruy Pérez Tamayo

DEDICATORIA

Dedico esta investigación:

A mis padres, que siempre dieron su vida por mí y a los que debo todo lo que soy.

A mis abuelitos Ondina y Carlos por ser las personas más especiales que han pasado por mi vida.

A mi tía Vivian, Liz y Amandita, ustedes son mi inspiración.

A mi familia y a mi novio Frank por estar presente en cada paso del camino.

A mis hermanitas Mary y Laura que aunque estén lejos, siempre están cerca de mi corazón.

A mis amigas del alma Nixys y Claudia, gracias por estar siempre presentes cuando las necesité.

A Yanet, Yalili y Yuleymis por ser una segunda familia, les aseguro que las voy a extrañar mucho.

A las niñas de mi apto y a todos los amigos que hicieron estos cinco años tan especiales, principalmente Daniel, Alexei y Martha.

Para todos ustedes que ayudaron a realizar este sueño de convertirme en Ingeniera, gracias.

AGRADECIMIENTOS

Agradezco a:

A mis tutores Albrecht y Tejeda, por su guía y paciencia para llevar a cabo esta investigación.

A mi tribunal por la orientación necesaria durante la elaboración del proyecto.

A los profesores Frank y Angel por brindarme su ayuda incondicional.

DECLARACIÓN DE AUTORÍA

Declaro que soy el único autor de este trabajo y autorizo a la Dirección de la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

Para que así conste firmo la presente a los ____ días del mes de _____ del 2011.

"Diana Delisle Rivero"	"Yunier Albrecht Delgado"	"Yunier E. Tejeda Rodríguez"
Autor	Tutor	Co-Tutor

DATOS DE CONTACTO

Tutor: Ing. Yunier Albrecht Delgado.

- Ingeniero en Ciencias Informáticas, Universidad de las Ciencias Informáticas, 2011.
- Jefe del Proyecto Plataforma VideoWeb, Facultad 6, Universidad de las Ciencias Informáticas.
- Profesor del Departamento de Técnicas de Programación, Facultad 6, Universidad de las Ciencias Informáticas.
- Correo electrónico: yalbrecht@uci.cu

Co Tutor: Msc. Yunier Emilio Tejeda.

- Máster en Ciencias Matemáticas.
- Profesor del Departamento de Ciencias Básicas, Facultad 6, Universidad de las Ciencias Informáticas.
- Correo electrónico: yuniere@uci.cu

OPINIONES Y AVALES

OPINIÓN DEL TUTOR

SOBRE EL TRABAJO DE DIPLOMA PRESENTADO PARA OPTAR POR EL TÍTULO DE INGENIERO EN CIENCIAS INFORMÁTICAS

Título: Obtención de preferencias de usuario a partir métodos estadísticos de Minería de datos

Autor(es): Diana Delisle Rivero

El tutor del presente Trabajo de Diploma considera lo siguiente:

El documento presentado se encuentra correctamente estructurado en correspondencia con lo establecido para una tesis de este tipo. Existe correspondencia entre los objetivos planteados y los cumplidos, los contenidos están bien referenciados y la bibliografía utilizada es actual.

El resultado obtenido es de una gran importancia para el proyecto Plataforma VideoWeb porque le brinda al mismo métodos y conocimientos para mejorar los servicios ofrecidos en el producto desarrollado, ofrece un gran valor agregado y esta es una cuestión que es muy valorada en el mercado de software donde pretende insertarse el producto y que es de vital importancia para la economía del país.

El cumplimiento de las metas trazadas no habría sido posible sin la gran responsabilidad mostrada por la estudiante ante las tareas a realizar, su independencia y creatividad le ayudaron en la búsqueda de soluciones ante los problemas encontrados durante el camino del desarrollo de la investigación, demostrando además la capacidad para apropiarse de los conocimientos necesarios para llevar a término completo el objetivo propuesto.

Teniendo en cuenta lo antes expresado, la exposición realizada y el cumplimiento de los objetivos propuestos, se considera que la estudiante se encuentra apta para ejercer como Ingeniera en Ciencias Informáticas y se propone al Tribunal la calificación de 5 puntos.

Ing. Yunier Albrecht Delgado 01/07/11

Firma Fecha

RESUMEN

La Minería de Datos unida a la estadística clásica ofrece diversos métodos de pronóstico para dar apoyo a la toma de decisiones empresariales, lo cual resulta de gran utilidad a la hora de descubrir patrones o elaborar modelos de predicción. En el presente trabajo de diploma se propone un enfoque de minería de uso web basado en el uso de las técnicas estadísticas multivariantes de análisis de componentes principales y análisis por agrupación, con el objetivo de descubrir modelos de navegación comunes a diferentes grupos de usuarios, que permitan a su vez, el establecimiento de las pautas a seguir para la personalización la interfaz de la Plataforma VideoWeb. El documento recoge las principales características de los algoritmos empleados, así como los cambios necesarios en la estructura de almacenamiento del sistema, para obtener la información necesaria que permita llevar a cabo el proceso de minería. Así mismo, se describen las herramientas estadísticas disponibles actualmente para la implementación de las técnicas seleccionadas y se presentan los resultados obtenidos que permiten validar la propuesta de solución escogida.

Palabras clave: análisis de componentes principales, análisis por agrupación, personalización, Minería de Web.

ÍNDICE DE TABLAS

Tabla 1: Descripción de la clase archivo_multimedia.----------20
Tabla 2: Descripción de la clase publicacion_am.----------21
Tabla 3: Descripción de la clase streaming. ----------21
Tabla 4: Descripción de la clase filehtml. ----------22
Tabla 5: Descripción de la clase almacenamiento. ----------22
Tabla 6: Descripción de la clase tipologias_am.----------23
Tabla 7: Descripción de la clase películas.----------24
Tabla 8: Descripción de la clase internos.----------24
Tabla 9: Descripción de la clase docentes. ----------25
Tabla 10: Descripción de la clase videos. ----------25
Tabla 11: Descripción de la clase documentales. ----------26
Tabla 12: Descripción de la clase series.----------26
Tabla 13: Descripción de la clase temporada. ----------27
Tabla 14: Descripción de la clase capítulo. ----------27
Tabla 15: Descripción de la clase users.----------29
Tabla 16: Variables para la base de conocimiento.----------30
Tabla 17: Descripción de la clase historial.----------32
Tabla 18: Descripción de la clase campo.----------32
Tabla 19: Descripción de la clase campo_valor.----------33
Tabla 20: Descripción de la clase usuario_campo. ----------34
Tabla 21: Comparación de programas estadísticos. ----------39
Tabla 22: Resumen de los componentes principales.----------42
Tabla 23: Variables que más contribuyen en cada componente principal.----------45
Tabla 24: Distribución de variables por agrupamiento. ----------52

ÍNDICE DE FIGURAS.

Figura 1: Tipos de *outliers.*----------11
Figura 2: Dendrograma. ----------15
Figura 3: Estructura de almacenamiento de información del sistema. ----------19
Figura 4: Cambios a la estructura de almacenamiento de información del sistema. ----------31
Figura 5: Representación del concepto del método ward. ----------36
Figura 6: Representación del concepto del método de amalgamiento simple.----------36
Figura 7: Representación del concepto del método de amalgamiento completo. ----------37
Figura 8: Diagrama de segmentación del ACP.----------43
Figura 9: Gráfico de los vectores de carga. ----------43
Figura 10: Gráfico de componentes principales.----------44
Figura 11: Gráficos de Diagnóstico. ----------46
Figura 12: Dendrograma del método amalgamiento simple. ----------48
Figura 13: Dendrograma del método distancia promedio ponderada.----------49
Figura 14: Dendrograma del método amalgamiento completo. ----------50
Figura 15: Dendrograma del método ward. ----------51

ÍNDICE DE CONTENIDOS

Capítulo 1. “Fundamentación Teórica” 4
1.1. Introducción 4
1.2. Conceptos asociados al dominio del problema 4
1.3. Análisis de componentes principales y análisis por agrupación. 7
1.3.1. Descripción general. 7
1.3.2. Descripción actual del dominio del problema 15
1.3.3. Situación Problemática 16
1.4. Análisis de otras soluciones existentes 17
1.5. Conclusiones. 17
Capítulo 2. “Solución Propuesta” 18
2.1. Introducción. 18
2.2. Estructura de almacenamiento de información. 18
2.3. Definición de las variables para formar la base de conocimientos. 29
2.4. Cambios necesarios para el diseño de la base de datos. 30
2.5. Algoritmo a utilizar. 34
2.6. Herramientas. 37
2.7. Conclusiones. 40
Capítulo 3. “Resultados Obtenidos” 41
3. Introducción. 41
3.1. Caso de Estudio. 41
3.1.1. Selección y recopilación de datos. 41
3.1.2. Tratamiento previo de los datos 42
3.1.3. Transformación de los datos. 47
3.1.4. Interpretación de los resultados. 52
3.2. Conclusiones. 53
Conclusiones Generales. 54
Recomendaciones. 55
Bibliografía 56
Anexos 58
Glosario 65

Introducción.

El crecimiento masivo de Internet en estos últimos años, ha propiciado un cambio radical en la forma de acceder a la información. La diversidad de contenidos que se pueden encontrar en la red y la facilidad de acceso, ha hecho esta propuesta más atrayente para un gran número de usuarios, que desechando las vías tradicionales de transmisión de información, apuestan por sitios web que presten servicios de distribución de contenidos audiovisuales.

Debido a las oportunidades de mercado que puede ofrecer una aplicación que brinde este tipo de servicio, es necesario analizar el comportamiento de los usuarios, con la finalidad de descubrir tendencias o patrones. Los mismos pueden ser empleados para lograr personalizar los servicios, con el objetivo de anticiparse al abandono de la web, aumentar la fidelidad de los usuarios y convertir visitantes anónimos en clientes repetitivos.

Para lograr personalizar un sistema, es decir, adecuar todas las funcionalidades que brinda una aplicación según las preferencias del usuario, se aprovecha la gran cantidad de datos que los propios visitantes dejan en sus accesos al sitio. Estos pueden ser recogidos a través de comportamientos en sesiones, preferencias reveladas de forma expresa por el propio cliente en foros o encuestas y conocimiento históricos de visitas. Las variables que arroja este análisis deben ser almacenadas y centralizadas en una base de datos.

Sin embargo no es suficiente tener guardada y estructurada esta información, es necesario examinarla eficientemente para extraer de ella el conocimiento oculto en los datos. La Minería de Web se utiliza con este propósito. Esta, unida a la estadística clásica, se integra para ofrecer resultados más completos y precisos que garanticen una correcta toma de decisiones.

En Cuba, la Universidad de las Ciencias Informáticas (UCI) con el proyecto Plataforma VideoWeb, dentro del Centro de Geoinformática y Señales Digitales (Geysed) de la Facultad 6, cuyo objetivo principal es apoyar el proceso educativo de la Universidad a través de la publicación de contenidos audiovisuales, carece de una infraestructura de almacenamiento de datos en la cual guardar todo el historial de navegación que se registra a partir de la interacción de los usuarios con la Plataforma VideoWeb y de un método eficiente que permita el análisis de los datos almacenados. Estas consideraciones permiten reconocer como **problema científico**: ¿cómo obtener patrones o tendencias sobre preferencias de usuario a partir de su interacción con la Plataforma VideoWeb?

Para dar respuesta a esta interrogante la presente investigación define como **objeto de estudio**: algoritmos existentes dentro del análisis de componentes principales como técnica de análisis exploratorio de datos. El **campo de acción** que abarca este trabajo está enmarcado en la obtención de patrones o tendencias sobre preferencias de usuario a partir de su interacción con la Plataforma VideoWeb.

Atendiendo a la situación problemática descrita y al problema planteado, así como al objeto y campo de acción, la presente tesis se estructura y desarrolla en función del siguiente **objetivo general**: utilizar un algoritmo basado en el análisis de componentes principales para el estudio del comportamiento de los usuarios en la Plataforma VideoWeb. Como **idea a defender** se propone: la aplicación de un algoritmo basado en el análisis de componentes principales garantizará la obtención de patrones o tendencias sobre preferencias de usuario en la Plataforma VideoWeb.

El objetivo propuesto indujo a formular, como guías para el desarrollo de la investigación, las siguientes **tareas de investigación**, estas son:

1. Analizar las técnicas de análisis exploratorio de datos.
2. Definir las variables para formar la base de conocimientos
3. Analizar las soluciones existentes.
4. Modificar la base de datos de la Plataforma en base a las variables definidas para la base de conocimiento.
5. Plantear el algoritmo a utilizar.
6. Aplicar el algoritmo en la base de datos de la Plataforma VideoWeb.
7. Valorar los resultados de la aplicación del algoritmo planteado.

Con el correcto cumplimiento de las tareas se esperan obtener como **posibles resultados o aportes prácticos**: el algoritmo a aplicar para la obtención de preferencias de usuario, la definición de patrones o tendencias a obtener con el algoritmo a aplicar y el informe de aplicación del algoritmo en la Plataforma VideoWeb.

Para el desarrollo completo del trabajo y su total entendimiento se hace necesario emplear **métodos de investigación**, dentro de los que se encuentran:

Analítico – Sintético: permitirá inicialmente descomponer el problema en sus diversas partes y ulteriormente descubrir características comunes entre ellas, en el presente trabajo este método se emplea

en el examen de la Minería de Web y de las técnicas estadísticas. En base a los resultados obtenidos a partir de este estudio, se establece mentalmente la unión entre sus componentes y se buscan características generales en ambas.

Inductivo– Deductivo: para ir de lo general a lo particular y viceversa, evidenciándose en el trabajo en el momento de realizar las valoraciones del algoritmo seleccionado y en la aplicación de los conceptos fundamentales al problema particular del trabajo.

Análisis Histórico – Lógico: se usará para conocer la evolución de las tendencias y de las tecnologías de Minería de Web, así como de los métodos estadísticos aplicados a esta, determinando cual es el más indicado para obtener las preferencias de usuarios.

Capítulo 1. “Fundamentación Teórica”

1.1. Introducción

En el presente capítulo se abordan conceptos asociados al dominio del problema que permiten tener una mejor comprensión del mismo. Se brinda información sobre el uso de la Minería de Web y conceptos asociados a las técnicas estadísticas de análisis por agrupación y análisis de componentes principales. Se explica con profundidad aspectos relacionados con la situación problemática existente.

1.2. Conceptos asociados al dominio del problema

Encontrar, extraer, filtrar, y evaluar la información almacenada a partir de las visitas a un portal web es un proceso complejo al cual están asociados varios conceptos que son necesarios exponer, para lograr un mayor entendimiento en el marco del problema.

Minería de Datos (*data mining*).

La Minería de Datos es una etapa dentro de un proceso mayor llamado extracción de conocimiento en bases de datos (KDD). Una definición de Minería de Datos viene de *Gartner Group*, firma de investigación en tecnología de la información: “Minería de Datos es el proceso de descubrir nuevas correlaciones significativas, patrones y tendencias por indagación a través de grandes cantidades de datos almacenados en repositorios, usando tecnologías así como técnicas matemáticas y estadísticas”. Una de las extensiones de la Minería de Datos consiste en aplicar sus técnicas a documentos y servicios de la web, lo que se llama Minería de Web.

Minería de Web (*web mining*).

La Minería en la Web o Minería de Web (*web mining*) es definida como: “La aplicación de las técnicas de la Minería de Datos a grandes repositorios de datos de la Web” (Mobasher, y otros, 1996). Una conceptualización más detallada establece que: “la Minería de Web es un proceso complejo que comprende el análisis de información diversa, como el contenido y estructura de los documentos web (HTML, XML), archivos de texto, bases de datos, bitácoras de acceso de usuarios, bitácoras (log) de referencias de otros servidores, perfiles de usuarios y otros, con el fin de encontrar información útil y

relevante de acuerdo a las necesidades de un usuario” (TORRES, 2009). Para esto la Minería de Web pasa por una serie de fases.

Fases de la Minería de Web.

En la bibliografía se encuentra variadas formas de presentar el proceso de minería. En general, son similares, lo que cambia es la forma de agrupación y detalle de las fases o etapas planteadas. Se propone la siguiente (Sánchez Enríquez, 2008).

- Selección y recopilación de datos: En esta fase se decide que se quiere analizar y que datos facilitarán esta información. Posteriormente se localizan los documentos o archivos a adquirir. Estos se capturarán y se almacenarán los datos pertinentes.

- Tratamiento previo de los datos: Esta fase corresponde al filtrado y limpieza de los datos. Se extrae la información y se realizan las tareas de criba sobre estos, eliminando datos erróneos o incompletos. Se obtienen muestras de los mismos (en busca de una mayor velocidad de respuesta del proceso), o se reducen el número de valores posibles y se presentan el resto de manera ordenada y con los mismos criterios formales que los originales.

- Transformación de los datos: En esta fase se utilizan algoritmos para la búsqueda de patrones de comportamiento y para detectar asociaciones. Estos algoritmos se elaboran usando recursos estadísticos y técnicas procedentes de la Minería de Datos. Se procede a transformar los datos para obtener como resultado, información sobre ellos. Los principales algoritmos usados se basan en reunión de grupos homogéneos, seguimiento de rutas o historial de navegación de una persona.

- Interpretación de los resultados: El análisis de la información no tendrá un sentido completo si no se razonan los resultados, es decir estos deben ser utilizados con un objetivo. De esta forma, los modelos obtenidos se incorporan en los sistemas de análisis de información de la organización.

DOMINIOS DE EXTRACCIÓN DE CONOCIMIENTO:

Algunos investigadores han considerado diversos dominios de extracción de conocimiento de Minería de Web de acuerdo al tipo de contenido a minar (Kosala, 2000) (Cooley, et al., November 1997).

Minería de contenido web (*Web content mining*).

La minería de contenido web extrae información relevante sobre el contenido de la web de manera que pueda ayudar a clasificarlo, aumentando así su organización, para posteriormente mejorar el acceso y la recuperación de la información en él contenida.

Minería de estructura web (*Web structure mining*).

Este tipo de minería sirve para conocer la organización de una web, su estructuración y como es la navegación a través de ella. Esto con el propósito de identificar preferencias y clasificaciones de los objetos relacionados, y de esta manera evaluar las relevancias de las páginas.

Minería de utilización (*Web usage mining*).

La minería de utilización emplea las técnicas de Minería de Datos para la extracción de patrones de uso personalizado usando los archivos Log de los servidores Web. El objetivo es sacar patrones de uso de un sitio web de manera que se pueda reestructurar para mejorar los servicios. Otro uso es obtener perfiles de los distintos tipos de usuarios, a través de su comportamiento de navegación, para poder atenderlos de forma más personalizada. Existe una tendencia en dividir este tipo de minería en dos grupos (Kosala, 2000) (Zhu, 2003). Estos son: descubrimiento de patrones de acceso general y descubrimiento de patrones de uso personalizado, siendo este último el desarrollado en el presente trabajo de diploma.

Descubrimiento de patrones de uso personalizado.

Al personalizar un portal web, se identifican los objetivos intrínsecos en las sesiones de navegación. El autor Bamshad Mobasher plantea que “la personalización de la web se puede analizar a partir de tres

tipos de sistemas: los Sistemas de Reglas de Decisión Manual, los Sistemas de Filtraje Basados en el Contenido, y los Sistemas de Filtraje Cooperativo” (Mobasher, y otros, 2004).

Los Sistemas de Filtraje Basados en el Contenido y los Sistemas de Reglas de Decisión Manual utilizan la información obtenida a partir del registro manual de los usuarios, mientras que los Sistemas de Filtraje Cooperativo capturan los intereses de los usuarios para perfeccionar las recomendaciones que se realicen.

El Filtraje Cooperativo (*Collaborative Filtering*) representa las preferencias usadas como una evaluación numérica, las cuales se obtienen a partir de los registros de vistas o los tiempos de acceso a las páginas web (Nakamura, y otros, 2003). Esto permite recomendar objetos preferidos por usuarios similares o predecir la utilidad de ciertos objetos para un grupo de usuarios en particular. Para personalizar un portal web se han desarrollado y aplicado una gran diversidad de algoritmos, entre los que se pueden encontrar el uso de técnicas estadísticas, tales como el análisis de componentes principales y análisis por agrupación.

1.3. Análisis de componentes principales y análisis por agrupación.

1.3.1. Descripción general.

El examen de los datos con la finalidad de obtener información que permita localizar grupos con comportamiento homogéneo y descubrir tendencias, puede realizarse usando diversas técnicas estadísticas. Cada una resuelve problemas de determinadas características y para extraer todo el conocimiento oculto, en general será necesario utilizar una combinación de varias. A continuación se describen las características de dos de estas técnicas, el análisis de componentes principales (ACP) y el análisis por agrupación (CA).

ANÁLISIS DE COMPONENTES PRINCIPALES.

A menudo en la Minería de Web son almacenados un gran número de variables a partir de la interacción de los usuarios con la aplicación. Sin embargo, si se toman demasiadas variables sobre un conjunto de objetos, es difícil visualizar las relaciones entre dichas variables. El análisis de componentes principales pertenece a un grupo de técnicas estadísticas multivariantes, que permite la representación de las medidas numéricas de varias variables en un espacio de pocas dimensiones, donde se puedan percibir

relaciones que de otra manera permanecerían ocultas en dimensiones superiores. Dicha representación debe ser tal que al desechar dimensiones superiores, la pérdida de información sea mínima. Entre las razones que se encuentran para considerar el análisis de componentes principales se hallan (Jhonson, 1998).

- Técnica de análisis exploratorio: El ACP como técnica de análisis exploratorio permite descubrir interrelaciones entre los datos y de acuerdo con los resultados, proponer los análisis estadísticos más apropiados.

- Cribado de datos: el uso del ACP se recomienda como primer paso en el examen de datos. Los análisis de seguimiento sobre las componentes principales son útiles para comprobar las hipótesis que el investigador podría establecer sobre un conjunto de datos multivariados y ayudar a revelar anormalidades en la información recolectada.

- Agrupación: el ACP es también útil si se desea agrupar las unidades experimentales en subgrupos de tipos semejantes o para verificar los resultados de los programas de agrupación.

- Regresión: en la regresión lineal múltiple cuando las variables independientes presentan alta colinealidad es preferible hacer la regresión sobre los componentes principales en lugar de usar las variables originales.

DESCRIPCIÓN

“El ACP comprende un procedimiento matemático que transforma un conjunto de variables correlacionadas de respuesta, en un conjunto menor de variables no correlacionadas llamadas componentes principales” (Zhou, y otros, 2004). Dichas componentes se ordenan en función del porcentaje de varianza explicada. En este sentido, la primera componente será la más importante, por ser la que explica el mayor porcentaje de variabilidad de los datos y los otros componentes tomarán en cuenta la variabilidad restante tanto como sea posible. Es importante resaltar el hecho que a mayor variabilidad de los datos (varianza), se considera que existe mayor información.

En el ACP existe la opción de usar la matriz de covarianzas (∑) o la matriz de correlaciones (P). En la primera opción se le está dando la misma importancia a todas las variables y se aplica en aquellos casos en donde dichas variables surjan de un fundamento igual, es decir que todas deben estar medidas en las mismas unidades y las variables deben tener varianzas que tengan aproximadamente tamaños semejantes. Cuando no parezca que las variables están ocurriendo sobre un fundamento igual, se aplica sobre la matriz de correlaciones. La mayoría de los manuales consultados argumentan que son tres los criterios que emplean los investigadores para determinar el número de componentes (Jolliffe, 1986) (Johnson, 2000).

- El porcentaje total de varianza explicada por los componentes principales debe representar más del 70% ó 75% de la variabilidad de los datos originales.
- El criterio gráfico que parte de la estimación de una gráfica de sedimentación (Scree). Esta se construye al situar el valor de cada valor propio contra el recíproco. Cuando los puntos de la gráfica tienden a nivelarse, estos valores propios suelen estar suficientemente cercanos a cero como para que pueda ignorarse. Por tanto, este método supone que la dimensión del espacio de datos es la que corresponde al valor propio grande más pequeño.
- El último criterio a elegir es cuando los valores propios son mayores o iguales a 1. Esto se debe a que cuando se usan las variables estandarizadas de la matriz de correlaciones para estimar los componentes principales, la varianza explicada de ellos no puede ser inferior a la que resulta de las variables estandarizadas, que en este caso es igual a la unidad. Por lo que frecuentemente se ignoran componentes cuyos valores propios sean menores que 1.

Cuando el porcentaje de variabilidad explicado por dos o tres componentes principales es alto, se puede realizar una representación gráfica de las variables originales y de los individuos de la muestra, que revelan las relaciones de correlación o semejanza entre ellos.

Interpretación de los gráficos.

Se acostumbra a representar gráficamente los puntos-variables y los puntos-individuos tomando como ejes de coordenadas los componentes. A veces, puede facilitar la interpretación de los resultados, el observar la similar ubicación de los puntos en los planos respectivos (Jolliffe, 1986).

Representación de la nube de puntos-variables.

Si se toma en cuenta el plano principal (formado por el primer y segundo componente) entonces un punto-variable viene representado por la coordenada que le corresponda a esa variable en cada uno de esos componentes. La nube de puntos-variables está situada en una superficie circular de radio 1.

Las proximidades entre los puntos-variables indican el grado de correlación que existe entre ellas. Cuando la correlación es igual a uno, los puntos coinciden. La coordenada de un punto sobre un eje es igual a la correlación de esa variable con respecto a éste. Por lo tanto, nos indica la contribución que esta variable tiene en la formación del eje. Para facilitar la interpretación se analizan los siguientes casos extremos:

- En el caso de que todas las p variables estén incorrelacionadas, se obtendrán p componentes igualmente importantes que serán las mismas variables.
- En el caso de que todas las variables tengan una correlación perfecta, se genera un solo componente que es la combinación lineal de las p variables, igualmente ponderadas, y que explica el 100% de la variación total.
- En el caso de que se tenga una variable que no esté correlacionada con las demás, uno de los componentes coincidirá con esta variable.

Representación de la nube de puntos-individuos.

Para interpretar la nube de puntos-individuos en un plano principal, conviene tener en cuenta los siguientes aspectos:

- Los puntos-individuos no quedan encerrados en un círculo de radio uno. Los puntos-variables sí.
- Un punto-individuo situado en el extremo de uno de los ejes, significa que ese individuo está muy relacionado con el respectivo componente.
- Cuando existen puntos-individuos cercanos al origen, significa que estos individuos tienen poca o ninguna relación con los dos componentes.

- Las proximidades entre individuos se interpretan como similitud de comportamiento de los individuos con respecto a las variables. Por ejemplo, dos puntos-individuos que están muy cercanos en el plano, significa que ambos individuos tienen valores próximos en cada una de las respectivas variables.
- Un punto-individuo extremadamente alejado de la nube, puede significar una de dos cosas. Existe un error en la introducción del dato o se trata de un individuo excepcional, el cual conviene sacar del análisis. Para la verificación de la existencia de este tipo de datos (*outlier)*, se puede usar los procedimientos que determina el ACP robusto.

ANÁLISIS DE COMPONENTES PRINCIPALES ROBUSTO.

Para la detección de los *outliers* usando un ACP robusto se calculan las distancias de puntajes y las distancia ortogonales de cada observación, luego se grafican junto con las frontera críticas permitiendo identificar las observaciones regulares de los *outliers*. Básicamente, se distinguen dos tipos de *outliers*: los puntos de palanca (*leverage*) y los *outliers* ortogonales. La Figura 1 muestra datos tridimensionales en un caso en el que las dos primeras componentes se utilizan para aproximarlos.

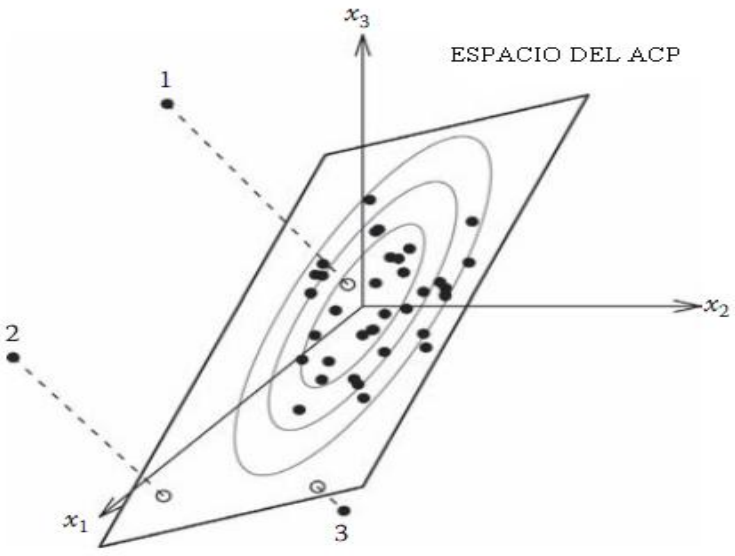

Figura 1: Tipos de *outliers.*

- El punto 1 muestra una distancia ortogonal grande al espacio ACP (*outlier* ortogonal) que no es visible cuando se examinan los datos proyectados en el espacio bidimensional. Estos *outliers* pueden tener un efecto sobre el ACP clásico.
- El punto 2 tiene una distancia ortogonal grande así como de puntaje, lo que significa que su proyección en el espacio ACP está lejos del centro. A *outliers* de este tipo se le llama puntos de leverage malo porque pueden "palanquear" la estimación de las componentes principales.
- El punto 3 tiene una distancia de puntaje alta y una distancia ortogonal pequeña. A *outliers* de este tipo se les llama puntos de *leverage* bueno.

ANÁLISIS POR AGRUPACIÓN.

El análisis por agrupación (*cluster analysis)*, permite catalogar individuos o unidades experimentales en subgrupos (*clusters*) definidos de manera única, cuando no se sabe de antemano de que subgrupos se originan las observaciones. Se puede utilizar en la Minería de Web para evaluar similitudes entre datos, construir un conjunto de prototipos representativos, analizar correlaciones entre atributos, o representar automáticamente un conjunto de datos por pequeños números de regiones, preservando las propiedades topológicas del espacio original de entrada.

Medidas de asociación.

Para poder unir variables o individuos es necesario tener algunas medidas numéricas que caractericen las relaciones entre las variables o los individuos. Cada medida refleja asociación en un sentido particular y es necesario elegir una medida apropiada para el problema concreto que se esté tratando. La medida de asociación puede ser una distancia o una similitud.

- Cuando se elige una distancia como medida de asociación, los grupos se forman como aquellos más parecidos, es decir, que la distancia sea la mínima.
- Cuando se elige una medida de similitud, los grupos se forman maximizando la similitud (Velderrey Sanz, 2010).

Dependiendo del tipo de análisis (por variables o por individuos) que se realiza, existen distintas medidas de asociación aunque, técnicamente, todas las medidas pueden utilizarse en ambos casos (Frakes, 1992).

Entre las medidas de asociación para variables se encuentran:

- Coeficiente de correlación de Spearman: es una medida de asociación lineal que utiliza los rangos, números de orden, de cada grupo de sujetos y compara dichos rangos. El coeficiente de correlación de Spearman es recomendable utilizarlo cuando los datos presentan valores extremos, o ante distribuciones no normales.

- Hoddfeding: Es una medida no paramétrica de asociación que detecta salidas mas generales de la independencia. Mediante esta medida de similitud se calcula una matriz del estadístico D de Hoddfeding para todos los pares posibles de columnas de una matriz D. Es una medida de la distancia entre F(x, y) y G(x)H(y), donde F(x,y) es la función de densidad acumulativa conjunta de (x, y) y (G,H) son las funciones marginales.

- Coeficiente de correlación de Pearson: es un índice estadístico que permite medir la fuerza de la relación lineal entre dos variables. Si el coeficiente de correlación de Pearson (r) es cercano a 0, las dos variables no tienen mucho que ver entre sí (no tienen casi ninguna covariación lineal). Si su valor es cercano a +/-1, esto significa que la relación entre las dos variables es lineal y está bien representada por una línea. Un aspecto débil de este análisis es que sólo detecta la parte lineal de las relaciones entre las variables, es decir, una relación que obedece a una ecuación curvilineal pasaría inadvertida.

Las medidas de asociación por individuos son:

- Distancia métrica: Es la distancia clásica, como la longitud de la recta que une dos puntos x_r y x_s en el espacio euclídeo y se define por la siguiente expresión: $d_{rs}= [(x_r - x_s)' (x_r - x_s)]^{1/2}$.

- Distancia métrica estandarizada: Se calcula la distancia euclidiana entre los puntos usando sus valores estandarizados z_r y z_s . Se define como: $d_{rs}= [(z_r - z_s)' (z_r - z_s)]^{1/2}$.

- Distancia de Mahalanobis: Se calcula a partir de los puntos de agrupamiento iniciales, donde Σ se reemplaza por alguna estimación razonable de la misma. La fórmula para calcular esta distancia es: $d_{rs}= [(x_r - x_s)'\Sigma^{-1}(x_r - x_s)]^{1/2}$.

Métodos de análisis por agrupación.

Existen dos tipos básicos de buscar agrupamientos y se distinguen por ser de naturaleza jerárquica o no jerárquica (Jhonson, 1998).

Métodos de análisis por agrupación no jerárquica.

Este tipo de clasificación consiste en seleccionar un conjunto de puntos simientes[1] de los agrupamientos y, a continuación, construir esos agrupamientos en torno a cada una de las simientes. Esto se realiza al asignar cada punto del conjunto de datos a su simiente más cercana, usando las medidas de desemejanza para medir las distancias entre cada uno de los puntos y esas simientes. Este enfoque aunque muy razonable, tiene tres desventajas importantes.

1. El procedimiento exige que se infiera el número de agrupamientos que van a existir.
2. La selección de las simientes iniciales de los agrupamientos influye mucho sobre el procedimiento, por lo que los investigadores podrían realizar un análisis por agrupación sobre el mismo conjunto de datos y producir agrupamientos completamente diferentes.
3. Con frecuencia el procedimiento no es factible desde el punto de vista de cálculo, porque hay demasiadas elecciones posibles, no sólo para el número de agrupamientos, sino también para las ubicaciones de las simientes.

Métodos de análisis por agrupación jerárquica.

Los llamados métodos jerárquicos tienen por objetivo reunir agrupamientos para formar un nuevo *cluster* o bien separar alguno ya existente para dar origen a otros dos, de tal forma que, si sucesivamente se va efectuando este proceso de aglomeración o división se minimice alguna distancia o se maximice alguna

[1] Son los puntos de partida iniciales en la agrupación no jerárquica. Los grupos se construyen alrededor de estas simientes o semillas.

similitud. Los métodos jerárquicos se subdividen en aglomerativos y disociativos. Los métodos aglomerativos comienzan el análisis con tantos grupos como individuos hayan. A partir de estas unidades originales se van formando grupos, de forma ascendente, hasta que al final del proceso todos los casos tratados están englobados en un mismo conglomerado. Mientras que los métodos disociativos, constituyen el proceso inverso al anterior. Comienzan con un conglomerado que engloba a todos los casos tratados y, a partir de este grupo inicial, a través de sucesivas divisiones, se van formando grupos cada vez más pequeños. Al final del proceso se tienen tantas agrupaciones como casos han sido tratados. Los métodos jerárquicos permiten la construcción de un árbol de clasificación, que recibe el nombre de Dendrograma (Figura 1), en el cual se puede seguir de forma gráfica el procedimiento de unión. Este tipo de diagrama contiene ramas que unen puntos datos y muestra el orden en que se asignan los puntos a los agrupamientos (Velderrey Sanz, 2010). Las longitudes de sus ramas son proporcionales a las distancias entre los puntos y agrupamientos, cuando los puntos y los agrupamientos se combinan.

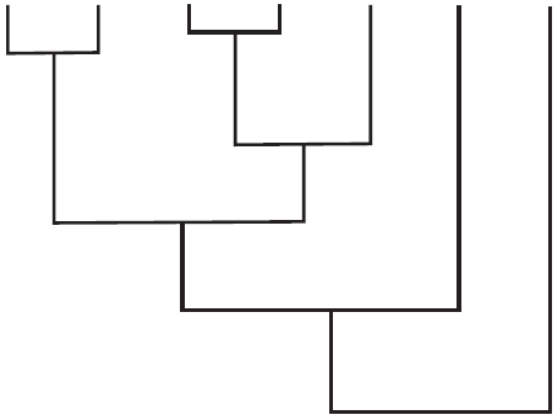

Figura 2: Dendrograma.

1.3.2. Descripción actual del dominio del problema

En la Plataforma VideoWeb se publican archivos multimedia y artículos de contenido. Los archivos multimedia pueden ser de audio o video y su presentación estará definida de acuerdo a la tipología[2] a la

[2] Define el tipo o clasificación de los archivos multimedia (entiéndase serie, película, documental, etc.)

que pertenezca. Mientras que los artículos de contenido son otra forma de publicación que comprenden noticias y avisos.

Los usuarios tienen la posibilidad de autenticarse cuando acceden al sitio, sin embargo no se guarda información que pueda ser utilizada para personalizar los servicios que se le ofrecen. Debido a esto la interfaz se muestra igual para todos los usuarios, ya que el proceso de publicación se realiza siguiendo exclusivamente los criterios del administrador. Los módulos que tienen que ver con la información que se publica generan un determinado número de variables, entre las que se encuentran el tipo de contenido al que se tuvo acceso, la fecha en que se realizó, la operación ejecutada, así como el usuario que efectuó dicha gestión. La aplicación brinda además la posibilidad de votar por una media, registrar la cantidad de veces que se accede a un contenido y guardar las listas de reproducción que desee el usuario.

1.3.3. Situación Problemática

En la versión actual del producto no se personaliza la interfaz con la que interactúa el usuario final. Los contenidos publicados en dicha interfaz están organizados atendiendo a las políticas definidas por los administradores del sistema y para incorporar un valor agregado como el de personalizar la información de acuerdo a las preferencias de los usuarios, es preciso analizar el comportamiento de los clientes al interactuar con la plataforma.

Otra deficiencia que presenta el sitio, es que la interfaz de la plataforma se muestra de igual manera para todos los usuarios, sin tener en cuenta sus gustos, preferencias y necesidades. Los contenidos que se publican en la portada son los mismos para todos los usuarios. Este espacio tan importante podría ser utilizado para hacer énfasis en los contenidos que realmente le interesan al cliente, reduciendo así la navegación requerida para encontrar lo que se desea.

Para el examen del comportamiento de los usuarios, sería necesario que la aplicación registrara una serie de datos asociados a la interacción de los clientes con la aplicación. Sin embargo actualmente no se dispone de una estructura de almacenamiento adecuada, ni un algoritmo que analice esta información, con la finalidad de extraer patrones o tendencias que presente grupos de usuarios en su navegación por el sitio, lo cual a su vez imposibilita el establecimiento de las pautas a seguir para lograr personalizar la interfaz del sistema.

1.4. Análisis de otras soluciones existentes

Para lograr personalizar un portal web se han desarrollado y aplicado una gran diversidad de algoritmos, entre los que podemos encontrar el uso de técnicas estadísticas. El investigador Yanzan Zhou propone un enfoque de minería de uso web para personalizar un sitio basado en el uso modelos de variables latentes (Zhou, et al., 2004). Dicha investigación fundamenta las ventajas que tiene la utilización del método análisis de factores principales (FA). Esta técnica se utiliza frecuentemente para crear nuevas variables que resuman toda la información de los datos originales. Además se emplea para estudiar la estructura de correlaciones entre las variables de un conjunto de datos y determinar si las variables respuestas presentan patrones comunes.

Es importante señalar que aunque el FA puede ser usado para identificar los factores que determinan que un grupo de individuos presenten los mismos patrones de navegación, una revisión bibliográfica más profunda arrojó como resultado, que diversos investigadores y estadísticos creen que el análisis por factores principales no es una técnica estadística válida y útil (Hill, 1997). La mayoría de las críticas que recibe este algoritmo se debe a la no unicidad de sus soluciones y a la subjetividad relacionada con sus numerosos aspectos. En el FA se toman decisiones subjetivas cuando el investigador determina la cantidad de factores subyacentes, determina cómo deben interpretarse y determina cómo deben evaluarse los individuos de la muestra sobre estas nuevas variables. Debido a los numerosos casos en los que la subjetividad puede desempeñar un papel, los críticos del análisis por factores principales sospechan que un investigador puede ser capaz de demostrar cualquier cosa que desee (Jhonson, 1998). Estos argumentos, hacen que la autora de la presente investigación opte por la utilización método estadístico de análisis de componentes principales.

1.5. Conclusiones.

Luego de esta revisión bibliográfica se definió que de acuerdo a las características del problema planteado, sería apropiado emplear en las etapas iniciales del proceso de minería el ACP para visualizar las relaciones existentes entre las variables y los individuos, y a partir de estas formar agrupamientos que permitan crear perfiles de usuarios con gustos semejantes. Para validar los resultados que arroje este examen, se propone emplear métodos de agrupamiento jerárquico de CA, con lo cual se establezcan los modelos necesarios que faciliten la personalización de la plataforma, de acuerdo a las preferencias de los usuarios.

Capítulo2. “Solución Propuesta”

2.1. Introducción.

En este capítulo se describen las herramientas a emplear en la realización de la investigación. Se definen las variables que formarán la base de conocimientos y los cambios que se le harán a la base de datos. Además, se mencionan aspectos esenciales que presentan los algoritmos seleccionados en su fundamentación matemática.

2.2. Estructura de almacenamiento de información.

El modelo de datos de la plataforma VideoWeb está constituido por un conjunto de tablas relacionadas con el almacenamiento de información audiovisual, gestión de usuarios y roles. A continuación se listan dichas tablas.

Listado de tablas:

- streaming
- publicacion_am
- almacenamiento
- archivo_multimedia
- filehtml
- tipologias_am
- peliculas
- internos
- docentes
- videos
- documentales
- series
- temporada
- capitulo
- users

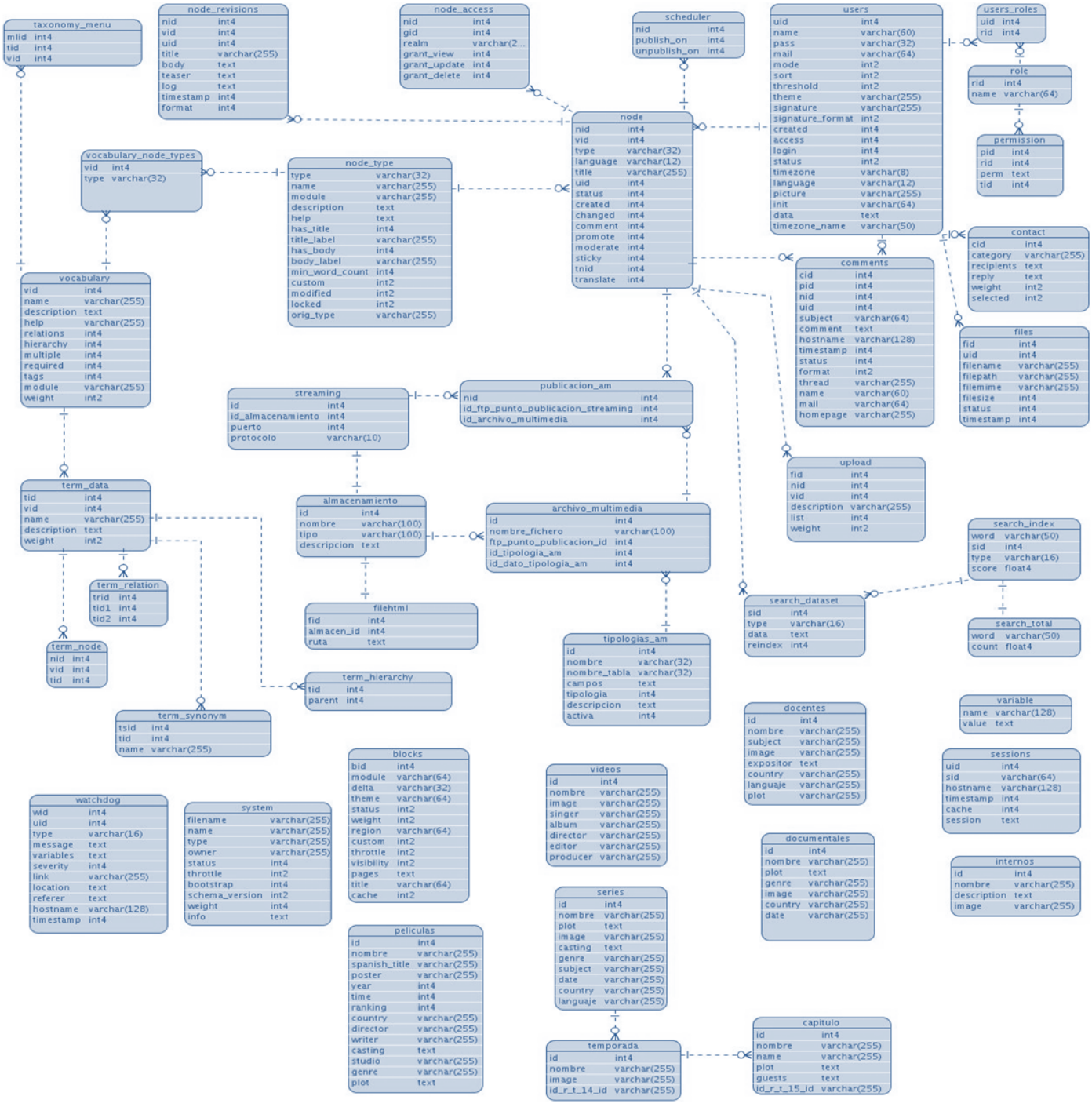

Figura 3: Estructura de almacenamiento de información del sistema.

Descripción de las tablas y atributos.

A continuación se describen de forma general las tablas que almacenan la información del modelo de datos de la Plataforma VideoWeb, al igual que se explica brevemente lo que representan todos sus atributos.

Nombre: archivo_multimedia		
Descripción: En esta tabla se almacenan los datos de las publicaciones de archivos multimedia que posee el sistema.		
Atributo	**Tipo**	**Descripción**
id	int4	Valor numérico único y auto incremental que identifica a cada uno de los archivos multimedia que posee el sistema.
nombre_fichero	varchar(100)	Nombre que posee la publicación para ser mostrada.
ftp_punto_publicacion_id	int4	Identificador del servicio de almacenamiento donde se va a guardar el archivo multimedia.
Id_tipologia_am	int4	Identificador de la tipología del archivo multimedia.
Id_dato_tipologia_am	int4	Identificador del dato del archivo según la tipología.

Tabla 1: Descripción de la clase archivo_multimedia.

Nombre: publicacion_am		
Descripción: En esta tabla se almacenan los datos de donde está publicado cada archivo multimedia.		
Atributo	**Atributo**	**Atributo**
nid	int4	Valor numérico único y auto incremental que identifica a

		cada nodo al que está asociado el archivo multimedia.
id_ftp_punto_publicacion_streaming	int4	Identificador de almacenamiento donde se va a publicar el archivo multimedia.
id_archivo_multimedia	int4	Identificador donde se va a publicar el archivo multimedia.

Tabla 2: Descripción de la clase publicacion_am.

Nombre: streaming		
Descripción: En esta tabla se almacenan los datos de los puntos de publicación streaming que posee el sistema.		
Atributo	**Atributo**	**Atributo**
id	int4	Valor numérico único y auto incremental que identifica a cada uno de los puntos de publicación de streaming.
id_almacenamiento	int4	Id que identifica el almacenamiento que se va a utilizar como streaming.
puerto	int4	Puerto que se utiliza.
protocolo	varchar(10)	Protocolo que se utiliza.

Tabla 3: Descripción de la clase streaming.

Nombre: filehtml		
Descripción: En esta tabla se almacenan los datos del servidor de almacenamiento filehtml.		
Atributo	**Tipo**	**Descripción**
fid	int4	Identificador del servidor de almacenamiento.
almacenamiento_id	int4	Identificador del almacenamiento al cual pertenece.
ruta	int4	Ruta donde se van a almacenar los archivos multimedia.

Tabla 4: Descripción de la clase filehtml.

Nombre: almacenamiento		
Descripción: En esta tabla se almacenan los datos de los servidores de almacenamiento que utiliza la plataforma VideoWeb.		
Atributo	**Atributo**	**Atributo**
id	int4	Valor numérico único y auto incremental que identifica a cada uno de los almacenamientos de acuerdo al nombre y tipo.
nombre	varchar(100)	Nombre que identifica cada uno de los almacenamientos
tipo	varchar(100)	Identifica que tipo de servidor de almacenamiento se va a utilizar (ftphtml, filehtml).
descripcion	text	Descripción del almacenamiento.

Tabla 5: Descripción de la clase almacenamiento.

Nombre: tipologias_am		
Descripción: En esta tabla se almacenan los datos de las tipologías de archivo multimedia que posee el sistema.		
Atributo	**Atributo**	**Atributo**
id	int4	Valor numérico único y auto incremental que identifica a cada una de las tipologías de archivo multimedia que posee el sistema.
nombre	varchar(32)	Nombre de la tipología.
nombre_tabla	varchar(32)	Nombre de la tabla en la base de datos.
campos	text	Arreglo con la estructura de cada uno de los campos que conforman la tipología.
tipologia	int4	Identifica si es una tipología que tiene asociado video o no.
descripcion	text	Breve descripción de la tipología.
activa	int4	Identifica si fue creada o no la tipología.

Tabla 6: Descripción de la clase tipologias_am.

Nombre: peliculas		
Descripción: En esta tabla se almacenan los datos de un archivo multimedia definido como película.		
Atributo	**Atributo**	**Atributo**
id	int4	Valor numérico único y auto incremental que identifica a cada una de las películas.
nombre	varchar(255)	Título original de la película.
spanish_title	varchar(255)	Título en español de la película.
poster	varchar(255)	Imagen de la película.
year	int4	Año en que fue producida la película.
time	int4	Duración de la película.

ranking	int4	Valor de la votación que le han dado los usuarios al material.
country	varchar(255)	Nombre del país al que pertenece la película.
director	varchar(255)	Nombre del director de la película.
writer	varchar(255)	Nombre del escritor de la película.
casting	text	Nombre de los actores de la película.
studio	varchar(255)	Estudio donde se produjo la película
genre	varchar(255)	Género de la película.
plot	text	Resumen de la película.

Tabla 7: Descripción de la clase películas.

Nombre: internos		
Descripción: En esta tabla se almacenan los datos de un archivo multimedia definido como interno.		
Atributo	**Atributo**	**Atributo**
id	int4	Valor numérico único y auto incremental que identifica a cada uno de los materiales internos.
nombre	varchar(255)	Nombre del archivo multimedia interno.
description	text	Descripción del archivo multimedia interno.
image	varchar(255)	Imagen del archivo multimedia interno.

Tabla 8: Descripción de la clase internos.

Nombre: docentes		
Descripción: En esta tabla se almacenan los datos de un archivo multimedia definido como docente.		
Atributo	**Atributo**	**Atributo**
id	int4	Valor numérico único y auto incremental que identifica a cada uno de los materiales docentes.
nombre	varchar(255)	Nombre del archivo multimedia docente.
subject	varchar(255)	Tema del archivo multimedia docente.
image	varchar(255)	Imagen del archivo multimedia docente.
expositor	text	Ponente del archivo multimedia docentes.
country	varchar(255)	País del archivo multimedia docentes.
languaje	varchar(255)	Idioma del archivo multimedia docentes.

Tabla 9: Descripción de la clase docentes.

Nombre: videos		
Descripción: En esta tabla se almacenan los datos de un archivo multimedia definido como video.		
Atributo	**Tipo**	**Descripción**
id	int4	Valor numérico único y auto incremental que identifica a cada uno de los videos.
nombre	varchar(255)	Nombre que identifica el video.
imagen	varchar(255)	Imagen del video.
singer	varchar(255)	Nombre del cantante o agrupación que interpreta la canción del video.
album	varchar(255)	Nombre del álbum del video.
director	varchar(255)	Nombre del director del video.
editor	varchar(255)	Nombre del editor del video.
producer	varchar(255)	Nombre del productor del video.

Tabla 10: Descripción de la clase videos.

Nombre: documentales		
Descripción: En esta tabla se almacenan los datos de un archivo multimedia definido como documental.		
Atributo	**Tipo**	**Descripción**
id	int4	Valor numérico único y auto incremental que identifica a cada uno de los materiales documentales.
nombre	varchar(255)	Nombre del documental.
plot	text	Resumen del documental.
genre	varchar(255)	Género del documental.
image	varchar(255)	Imagen del documental.
country	varchar(255)	País del documental.
date	varchar(255)	Fecha del documental.

Tabla 11: Descripción de la clase documentales.

Nombre: series		
Descripción: En esta tabla se almacenan los datos del archivo multimedia definido como serie.		
Atributo	**Tipo**	**Descripción**
id	int4	Valor numérico único y auto incremental que identifica a cada una las series.
nombre	varchar(255)	Nombre que identifica a la serie.
plot	text	Resumen de la serie.
imagen	varchar(255)	Imagen de la serie.
casting	text	Nombre de los actores que interpretan la serie.
genre	varchar(255)	Género de la serie.
subject	varchar(255)	Tema de la serie.
date	varchar(255)	Fecha de la serie.
country	varchar(255)	Nombre del país de la serie.
languaje	varchar(255)	Idioma de la serie.

Tabla 12: Descripción de la clase series.

Nombre: temporada		
Descripción: En esta tabla se almacenan los datos de la temporada de una serie.		
Atributo	**Tipo**	**Descripción**
id	int4	Valor numérico único y auto incremental que identifica a cada temporada de la serie.
nombre	varchar(255)	Nombre que identifica a la temporada de la serie.
imagen	varchar(255)	Imagen de la temporada de la serie.
id_r_t_14_id	varchar(255)	Id de la serie a la que pertenece.

Tabla 13: Descripción de la clase temporada.

Nombre: capitulo		
Descripción: En esta tabla se almacenan los datos de un capítulo de una serie.		
Atributo	**Tipo**	**Descripción**
id	int4	Valor numérico único y auto incremental que identifica a cada capítulo de la temporada de una serie.
nombre	varchar(255)	Nombre que identifica al capítulo.
imagen	varchar(255)	Imagen identifica al capítulo.
id_r_t_14_id	varchar(255)	Id de la temporada a la que pertenece.

Tabla 14: Descripción de la clase capítulo.

Nombre: users		
Descripción: En esta tabla se los datos de los usuarios registrados.		
Atributo	**Tipo**	**Descripción**
uid	int4	Valor numérico único y auto incremental que identifica a cada uno de los usuarios.
name	varchar(60)	Nombre único del usuario.
pass	varchar(32)	Contraseña del usuario.
mail	varchar(64)	Dirección de correo del usuario.
mode	int2	Modo de visualización de los comentarios para cada usuario.
sort	int2	Forma de ordenarlos comentarios por fecha.
threshold	int2	Anteriormente usado para las preferencias de usuario, ya no se usa.
theme	varchar(255)	Tema por defecto.
signature	varchar(255)	Firma de usuario.
signature_format	int2	Formato de la firma del usuario.
created	int4	Marca de tiempo de la creación del usuario.
acess	int4	Marca de tiempo para acceso anterior del usuario al sitio.
login	int4	Marca de tiempo para el último acceso del usuario.
status	int2	Valor que muestra si el usuario está activo (1) o bloqueado (0).

timezone	varchar(8)	Zona horaria del usuario.
language	varchar(12)	Lenguaje por defecto del usuario.
picture	varchar(255)	Dirección de la imagen cargada por el usuario.
init	varchar(64)	Dirección de correo usada para la creación inicial de una cuenta.
data	text	Arreglo de nombre y valores relacionados con el usuario. No se recomienda el uso de este campo.
timezone_name	varchar(50)	Nombre de la zona horaria.

Tabla 15: Descripción de la clase users.

El modelo de datos disponible en la actualidad, no cuenta con la estructura adecuada para establecer patrones de comportamiento según la interacción de los clientes con la aplicación. A partir de cada tipología no se registran los campos definidos como preferencias. Tampoco se almacena la interacción de los usuarios con estas variables, lo que imposibilita a su vez el establecimiento de la frecuencia con que grupos de usuarios utilizan los mismos recursos. A partir de estos resultados, se hace necesario establecer una serie de cambios en el modelo de datos, siendo el primer paso la definición de las variables para formar la base de conocimientos.

2.3. Definición de las variables para formar la base de conocimientos.

Un modelo de datos de usuarios que permita personalizar un portal web, está diseñado en base a la información que sobre un cliente se tiene y define la adaptación de contenidos, servicios y funcionalidades a cada cliente de forma individual. La información almacenada puede ser clasificada en tres tipos de variables: variables demográficas, de sesión y de operaciones. Las variables demográficas representan información relacionada con características propias del cliente como persona física, mientras que la información recogida sobre cómo se relaciona el usuario con el web son representadas por las variables de sesión, con ellas será posible conocer hábitos y preferencias de los clientes en base a la utilización de los contenidos, los servicios y la interfaz de usuario. Por último las variables de operaciones representan

información sobre el conjunto de operaciones a nivel de transacción que el cliente ha efectuado durante su histórico de sesiones (IWorld.com, Sciences de l' Evolution, 2003). A este último grupo pertenecen las variables seleccionadas para obtener las preferencias. Dichas variables están divididas de acuerdo a la tipología a la que pertenecen y las cuales se muestran a continuación (Tabla 16):

películas	series	documentales	docentes	videos
género	reparto	género	exponente	cantante
director	género	país	país	álbum
reparto	país		idioma	editor
escritor	idioma			productor
país				

Tabla 16: Variables para la base de conocimiento.

2.4. Cambios necesarios para el diseño de la base de datos.

A partir del análisis del modelo de datos de la plataforma VideoWeb se definieron las variables relevantes para la obtención de preferencias de usuarios. Sin embargo es necesario adecuar la estructura de almacenamiento del sistema para guardar estas variables y así modelar las relaciones que se establezcan entre los usuarios y las preferencias seleccionadas. Es debido a esto que se propone la inclusión de cuatro tablas al modelo de datos, las cuales se numeran y describen a continuación.

Listado de tablas:

- historial
- campo
- campo_valor
- usuario_campo

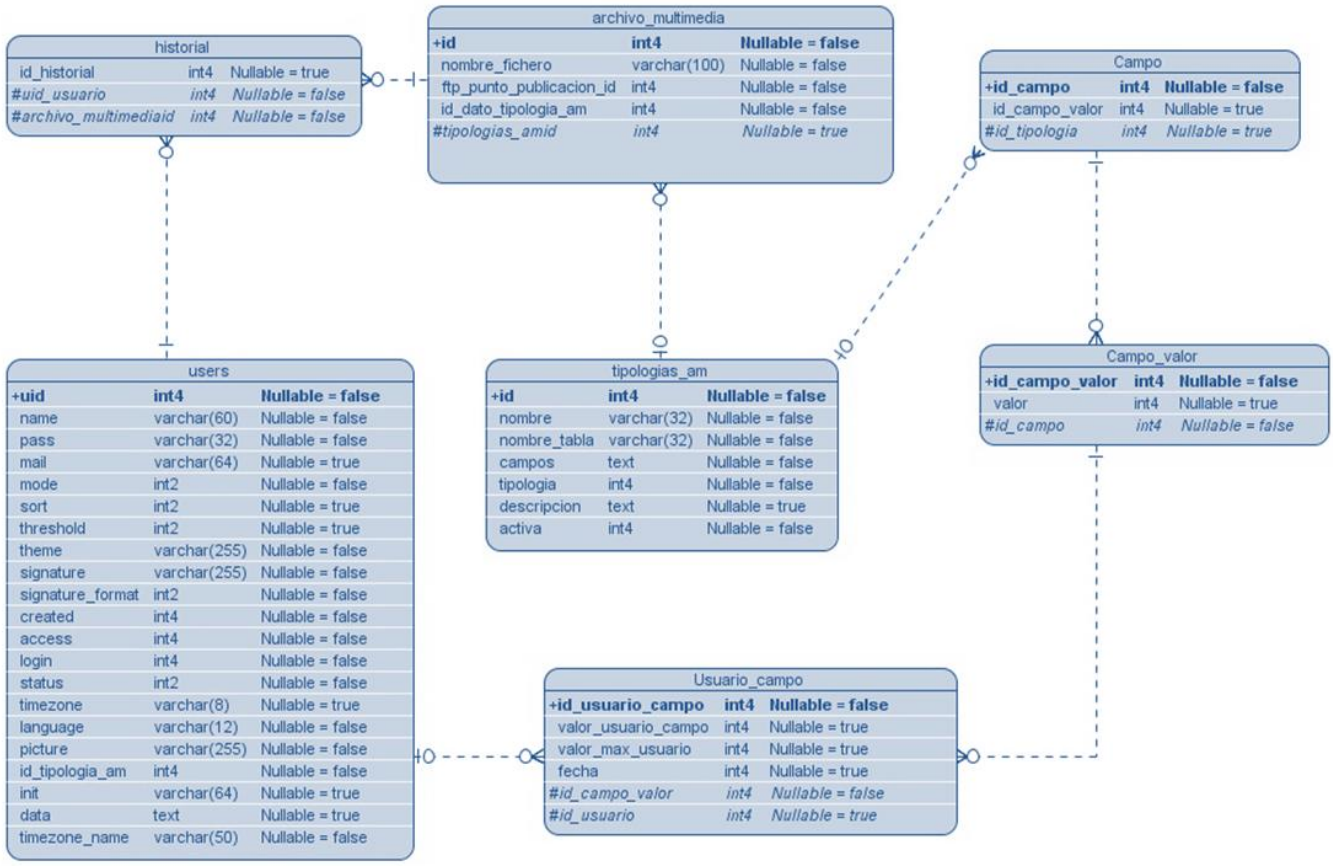

Figura 4: Cambios a la estructura de almacenamiento de información del sistema.

Descripción de las tablas y atributos.

A continuación se describen de forma general los datos que se almacenan en las tablas que se deben incluir en el modelo de datos de la Plataforma VideoWeb, al igual que se explica brevemente lo que representan los atributos de cada una. Las tablas users, archivo_multimedia y tipologias_am presentes en el diagrama precedente, fueron descritas con anterioridad.

Nombre: historial		
Descripción: En esta tabla se almacenan los historiales de acceso a los archivos multimedia.		
Atributo	**Tipo**	**Descripción**
id_historial	int4	Valor numérico único y auto incremental que identifica a cada uno de los campos de esta tabla.
uid_usuario	int4	Identificador del usuario.
archivo_multimediaid	int4	Identificador del archivo multimedia.

Tabla 17: Descripción de la clase historial.

Nombre: campo		
Descripción: En esta tabla se almacenan los campos relevantes en la obtención de preferencias de usuario.		
Atributo	**Tipo**	**Descripción**
id_campo	int4	Valor numérico único y auto incremental que identifica a cada uno de los campos de esta tabla.
id_tipologia	int4	Identificador de la tipología de archivo multimedia.
id_campo_tipologia	int4	Identificador del campo perteneciente a la tipología de archivo multimedia.

Tabla 18: Descripción de la clase campo.

Nombre: campo_valor		
Descripción: En esta tabla se almacenan los valores que toman los campos relevantes en la obtención de preferencias de usuario.		
Atributo	**Tipo**	**Descripción**
id_campo_valor	int4	Valor numérico único y auto incremental que identifica a cada uno de los campos de esta tabla.
id_campo	int4	Identificador de los campos relevantes para la obtención de preferencias.
valor	int4	Valor a tomar por el campo para obtener las preferencias.

Tabla 19: Descripción de la clase campo_valor.

Nombre: usuario_campo		
Descripción: En esta tabla se almacenan los valores que toman los campos relevantes en la obtención de preferencias de usuario.		
Atributo	**Tipo**	**Descripción**
id_usuario_campo	int4	Valor numérico único y auto incremental que identifica a cada uno de los campos de esta tabla.
id_usuario	int4	Identificador de los usuarios.
id_campo_valor	int4	Identificador del valor que toman los campos relevantes para la obtención de preferencias de usuarios.
valor_usuario_campo	int4	Valor de preferencia del usuario por el campo.

valor_max_usuario	int4	Valor máximo histórico del usuario por el campo.
fecha	int4	Fecha en que se registra.

Tabla 20: Descripción de la clase usuario_campo.

A raíz de la solución propuesta, el sistema deberá contar con la estructura adecuada para el almacenamiento de la información y su correspondiente análisis. A partir de este punto se está en condiciones de seleccionar el algoritmo a utilizar para lograr el objetivo general planteado.

2.5. Algoritmo a utilizar.

El análisis de los datos con la finalidad de obtener información que permita localizar grupos con comportamiento homogéneo y establecer relaciones, puede realizarse usando diversas técnicas. Cada una resuelve problemas de determinadas características, y para obtener el conocimiento oculto que se encuentra en una base de datos, la elección dependerá de las condiciones y necesidades de cada sistema en particular. Un extenso análisis bibliográfico permitió reconocer como la más factible para lograr personalizar los servicios de la Plataforma VideoWeb, las técnicas estadísticas de análisis por agrupación y análisis de componentes principales. A continuación se mencionan los diferentes algoritmos para el ACP y CA respectivamente, dando sus ideas básicas.

ALGORITMOS DE ACP.

Rotación de Jacobi.

Este algoritmo también conocido como descomposición espectral, es muy usado debido a su simplicidad. Consiste en calcular los vectores carga *(loading)* de la matriz *P* como los vectores propios de la matriz de covarianza o correlación de muestra, respectivamente de *X*. Luego de calcular la matriz de vectores propios *P*, la matriz de los puntajes *(scores)* de las componentes principales se obtienen multiplicando con la matriz de datos centrada *X, T=X*P.*

El software estadístico R al calcular las componentes principales según la Rotación de Jacobi, no funciona cuando la matriz de datos tiene menos filas que columnas (Varmuza, 2008).

Descomposición de Valores Singulares (VSD).

Este algoritmo es ampliamente usado debido a que descompone cualquier matriz *X* de orden *nxm* en un producto de tres matrices *X=T*S*P´*. Para la matriz de datos centrada *X* la matriz *T* de orden *nxm* contiene los *scores* normalizados del ACP de longitud 1. *S* es una matriz diagonal de orden *mxm* conteniendo los valores singulares en su diagonal que son igual a las desviaciones estándar sqrt (lamda sub j) de los *scores*. *P´* es la traspuesta de la matriz de carga del análisis de componentes principales de orden *mxm.* La matriz de *scores T* se calcula por *T=To*S*.

Mínimos Cuadrados Parciales Iterativos no Lineal (NIPALS).

Este algoritmo es menos conocido que el VSD y la Rotación de Jacobi, sin embargo, es muy popular en aplicaciones del ACP en quimiometría. En este método se calculan las componentes paso a paso y a diferencia de los otros algoritmos mencionados, se pueden elegir el número de componentes a calcular. Debido a que el proceso es iterativo, se vuelve ineficiente si se hace el proceso para todos los componentes (Varmuza, 2008).

Procedimiento.

Este procedimiento viene dado a partir del uso del software estadístico R (Castillo, 2009).

- Si la matriz de datos centrada ***X*** tiene más filas que columnas se usará el algoritmo de Rotación de Jacobi.
- Si la matriz de datos centrada ***X*** tiene menos filas que columnas se usará el algoritmo VSD.
- Si lo que quiere es calcular unas cuantas componentes principales se usará el algoritmo NIPALS.

A partir del análisis precedente de las características de los algoritmos NIPALS, Rotación de Jacobi y VSD se escogió este último como herramienta matemática a utilizar para obtener los componentes principales. VSD es una técnica de compresión potente y ampliamente utilizada en la comunidad científica. Es además el algoritmo más abarcador de los tres anteriormente propuestos, siendo una generalización de la descomposición espectral. Estas características permitirán brindar una solución más eficiente y completa en la obtención de preferencia de usuarios.

ALGORITMOS DE ANÁLISIS POR AGRUPACIÓN.

Métodos de análisis por agrupación jerárquica.

Ward.

El método de ward es un procedimiento jerárquico en el cual, en cada etapa, se unen los dos agrupamientos para los cuales se tenga el menor incremento en el valor total de la suma de los cuadrados de las diferencias, dentro de cada agrupamiento, de cada individuo al centroide[3] del grupo. A diferencia de los demás métodos, este utiliza el cálculo de la varianza para determinar la distancia entre agrupamientos. El método ward es bastante eficaz, aunque tiende a crear grupos muy pequeños (Figura 5).

Figura 5: Representación del concepto del método ward.

Amalgamiento simple (*single linkage*).

Este método, se inicia con N agrupamientos, en donde cada uno de éstos contiene una observación y continúa combinando los puntos y agrupamiento hasta que todas las observaciones están dentro de un agrupamiento. En este método se considera que la distancia entre dos agrupamientos viene dada, respectivamente, por la mínima distancia entre sus componentes (Figura 6).

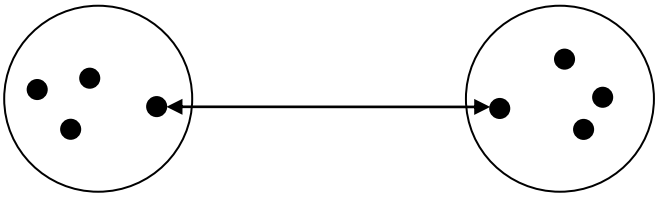

Figura 6: Representación del concepto del método de amalgamiento simple.

[3] Valores medios de los objetos que contiene el grupo, en cada una de las variables.

Distancia promedio ponderada. (*Unweighted aritmethic average*).

En esta estrategia la distancia entre los agrupamientos se obtiene como el promedio ponderado de las distancias, o similitudes, de los componentes de un agrupamiento respecto a los del otro. Este método tiende a tener buenos resultados cuando los componentes se encuentran en diferentes grupos o cuando forman una cadena alongada.

Amalgamamiento completo (*complete linkage*).

En este método la distancia entre dos agrupamientos hay que medirla atendiendo a sus elementos más dispares, o sea, la distancia entre las observaciones viene dada, respectivamente, por la máxima distancia entre sus componentes (ver Figura 7). Esta regla tiene buenos resultados cuando los agrupamientos se encuentran aislados y bien definidos. Si los *clusters* se encuentras alongados, el resultado será pobre.

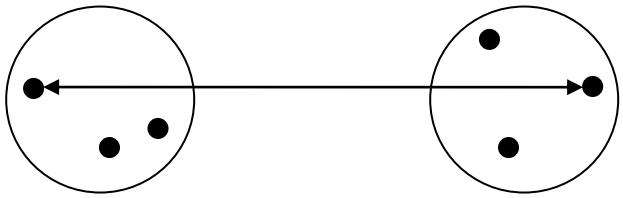

Figura 7: Representación del concepto del método de amalgamiento completo.

El análisis precedente, arrojó como conclusión que no existe un algoritmo de agrupamiento que funcione eficientemente para cualquier conjunto de datos. Además, utilizando distintos algoritmos sobre el mismo conjunto de datos, se obtienen resultados diferentes. Una alternativa válida, es implementar más de una solución y elegir como resultado un promedio de las mismas. Si varios métodos dan resultados semejantes, entonces se puede suponer que en realidad existen agrupaciones naturales.

2.6. Herramientas.

La evolución de la tecnología ha facilitado y automatizado en gran medida las tareas de análisis de información. Cada progreso ha abierto nuevas posibilidades de exploración y ha aumentado la calidad de los resultados obtenidos. Para decidir cuál es la herramienta más adecuada para un análisis estadístico eficiente de la información almacenada, se examinan las características y funcionalidades de varios

programas, cuya selección se realiza de acuerdo a las necesidades y problemas a resolver. A continuación se muestran algunas características de los mismos.

SPSS (*Statistical Package for the Social Sciences)*

Programa estadístico muy usado en las ciencias sociales y las empresas de investigación. Su uso es muy popular, debido a la capacidad de trabajar con bases de datos de gran tamaño. Cuenta con una interfaz gráfica de usuario amigable, que brinda la posibilidad de acceder a las opciones de la aplicación a través de botones de la interfaz gráfica. Permite además la decodificación de las variables y registros según las necesidades del usuario. El programa consiste en un módulo base y módulos anexos que se han ido actualizando constantemente con nuevos procedimientos estadísticos. Cada uno de estos módulos se compra por separado. El módulo SPSS *Programmability Extension* permite utilizar el lenguaje de programación Python para un mejor control de diversos procesos dentro del programa que hasta ahora eran realizados principalmente mediante scripts (con el lenguaje SAX Basic). Existe también la posibilidad de usar las tecnologías NET de Microsoft para hacer uso de las librerías del SPSS (SALAS, Agosto 2008).

SAS *(Statistical Analysis System)*

Software estadístico que permite crear gráficos, trabajar como una hoja de cálculo o compilar programas en lenguaje C. Requiere el ingreso de comandos para ejecutar gran parte de sus rutinas y opciones. SAS comprende amplias posibilidades de procedimientos estadísticos (métodos multivariados, regresión múltiple con posibilidades diagnósticas, análisis de supervivencia con riesgos proporcionales y regresión logística) y contiene potentes posibilidades gráficas. Los resultados pueden guardarse como archivos y usarse como entradas para futuras ejecuciones (SALAS, Agosto 2008).

R

Programa estadístico y un lenguaje de programación de uso libre, de código abierto y distribución gratuita. El código de R está disponible bajo las condiciones de la licencia GNU-GPL. Abarca una amplia gama de técnicas estadísticas que van desde los modelos lineales a técnicas de clasificación, así como análisis de series temporales. Proporciona una amplia gama de gráficos que además son fácilmente adaptables y extensibles. La calidad de los gráficos producidos y la posibilidad de incluir en ellos símbolos y fórmulas matemáticas, posibilitan su inclusión en publicaciones que suelen requerir gráficos de alta calidad. El

código fuente de R está escrito en C y se encuentra disponible en varias formas fundamentalmente para máquinas Unix y Linux, o como archivos binarios pre-compilados para Windows, Linux (Debian, Mandrake, RedHat, SuSe), Macintosh y Alpha Unix. R es un lenguaje Orientado a Objetos. Una diferencia importante entre R con el resto del software estadístico es el uso del objeto como entidad básica. Cualquier expresión evaluada por R tiene como resultado un objeto. Cada objeto pertenece a una clase, de forma que las funciones pueden tener comportamientos diferentes en función de la clase a la que pertenece su objeto argumento. Debido a la variedad de aplicaciones estadísticas que presenta el mercado, se realiza una comparación de aspectos generales entre los programas SPSS, SAS y R (Tabla 21) (SALAS, Agosto 2008).

Aspecto	**Programas Estadísticos**		
	SPSS	SAS	R
Amigabilidad con el usuario	Excelente	Baja-Regular	Baja-Regular
Calidad de gráficos	Regular	Buena-Excelente	Excelente
Control de procesos	Baja	Excelente	Excelente
Costo	U$S 1500	U$S 7200	Gratis
Código fuente disponible	No	No	Sí
Variedad análisis estadísticos	Buena	Buena-Excelente	Excelente
Documentación	Excelente	Buena	Buena-Excelente
Soporte técnico	Bueno	Bueno	Bajo
Sistema operativo	Windows®	Macintosh® Linux Macintosh®	Windows® Macintosh® Linux

Tabla 21: Comparación de programas estadísticos.

El estudio de las características de las herramientas estadísticas SPSS, SAS y R permite adoptar a esta última para el análisis de la información almacenada en la estructura de datos de la Plataforma VideoWeb. R reúne mayor cantidad de recursos y proporciona una manejabilidad superior frente a las otras opciones mencionadas, además de ser el que tiene una mayor implantación en la comunidad científica. Para la elección de R se han evaluado distintos aspectos, siendo especialmente destacables en lo que se refiere a calidad, a la cantidad de técnicas y funciones implementadas. Por otra parte, la transparencia en la construcción de R permite un mayor control del proceso de generación de conocimiento por parte de los usuarios

2.7. Conclusiones.

En el presente capítulo se analizó la estructura de almacenamiento de la información de la Plataforma VideoWeb, se concluyó que esta base de datos era insuficiente para guardar la información necesaria para personalizar la interfaz de dicha aplicación. A raíz de esto se propuso la inclusión de cuatro tablas en el modelo de datos, que permitan el almacenamiento de las variables seleccionadas como preferencias de usuario.

En base a la necesidad de analizar eficientemente los datos almacenados y de acuerdo con los procedimientos para la aplicación de un algoritmo dentro de la técnica de análisis de componentes principales, se escogió el algoritmo VSD, ya que proporciona un modelo de distribución normalizado de los datos, el que redunda en una representación matemática más compacta y, por tanto, más simple de computar. A partir de estas interpretaciones y para validar los resultados del ACP, se utiliza los algoritmos de análisis por agrupación de amalgamiento simple, amalgamamiento completo, distancia promedio ponderada y ward. Por último y en base a la necesidad de analizar eficientemente los datos mediante los métodos estadísticos escogidos, se estudiaron tres herramientas de análisis estadístico, siendo la elegida R debido a su facilidad de uso, implementación y su gratuidad.

Capítulo 3. “Resultados Obtenidos”

3. Introducción.

En el presente capítulo se muestran los resultados obtenidos al examinar un conjunto de datos de prueba, con el fin de detectar patrones de comportamiento en el historial de navegación de un grupo de usuarios. Para esto se analiza un caso de estudio, se describe el proceso de selección de los datos y la aplicación de los métodos estadísticos ACP y CA, a dichos datos.

3.1. Caso de Estudio.

3.1.1. Selección y recopilación de datos.

A partir de la Base de Datos de la Plataforma VideoWeb, se creó una réplica de la misma, incluyendo en su organización los cambios propuestos en el capítulo anterior. La información contenida en esta estructura de almacenamiento, se generó con la herramienta *Data Generator for Postgre SQL.* Los datos utilizados para este caso de estudio fueron obtenidos de la tabla campo_valor, que almacena el número de accesos de un archivo multimedia para cada usuario. La tabla resultante contiene 3631 registros correspondientes a los accesos de los usuarios a la tipología película, de acuerdo al país al que pertenece (ver Anexo1). Para posibilitar el procesamiento de los datos por el software estadístico R, se exportó dicha tabla en formato txt. Se midieron catorce (14) variables en una muestra de 48 usuarios. A continuación se enumeran las variables recogidas para este análisis.

- ID: id del usuario
- EU: Estados Unidos.
- FRA: Francia.
- IND: India.
- JAP: Japón.
- ARG: Argentina.
- COR: Corea.
- COL: Colombia.
- ING: Inglaterra.
- CUB: Cuba
- MEX: México.
- VEN: Venezuela.
- CHL: Chile.
- CAN: Canadá.
- ESP: España.

Tratamiento previo de los datos

En muchas situaciones se requiere tratamientos previos del conjunto de datos que representen y expliquen el comportamiento de un número de unidades de observación. El uso del ACP se recomienda como primer paso en este examen, ya que permite reducir la dimensión de la matriz de datos con el fin de evitar redundancias y destacar relaciones.Esto aumenta la calidad de la información recolectada, de modo que las técnicas de extracción de conocimiento puedan obtener mayor y mejor información.

Análisis de los datos utilizando ACP.

Se utilizó la descomposición del valor singular para realizar el ACP usando la función *princomp* que utiliza el software R para calcular los componentes principales a partir de la matriz de correlaciones. A pesar de que el problema original se ubica en el espacio vectorial $\mathbb{R}^{14}$; en la Tabla 22 se observa que los tres primeros componentes conservan un 74% de la información original de la nube de puntos. En la gráfica de sedimentación (ver Figura 8) se observa una ruptura entre el tercero y el cuarto valor propio, de modo que parecería que los datos tienden a caer dentro de un subespacio tridimensional del espacio muestral 14-dimensional. Estos criterios permiten definir que tres componentes principales son suficientes para resumir la información de las variables originales.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Desviación Estándar	2.654	1.430	1.126	1.019	0.850	0.701	0.576	0.555
Proporción de Varianza	0.503	0.146	0.090	0.074	0.051	0.035	0.023	0.022
Proporción Acumulada	0.503	0.649	0.740	0.814	0.866	0.901	0.925	0.947

Tabla 22: Resumen de los componentes principales.

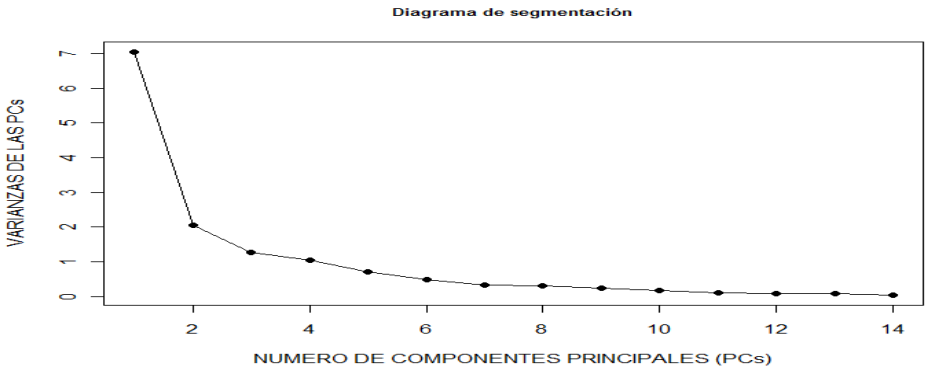

Figura 8: Diagrama de segmentación del ACP.

La fortaleza de los métodos multivariados son las gráficas. Algunas de las salidas numéricas se usan para establecer las coordenadas, mientras que otras ayudan en su interpretación. A continuación se muestran dos gráficas que representan los puntos variables (ver Figura 9) y los puntos individuos (ver Figura 10).

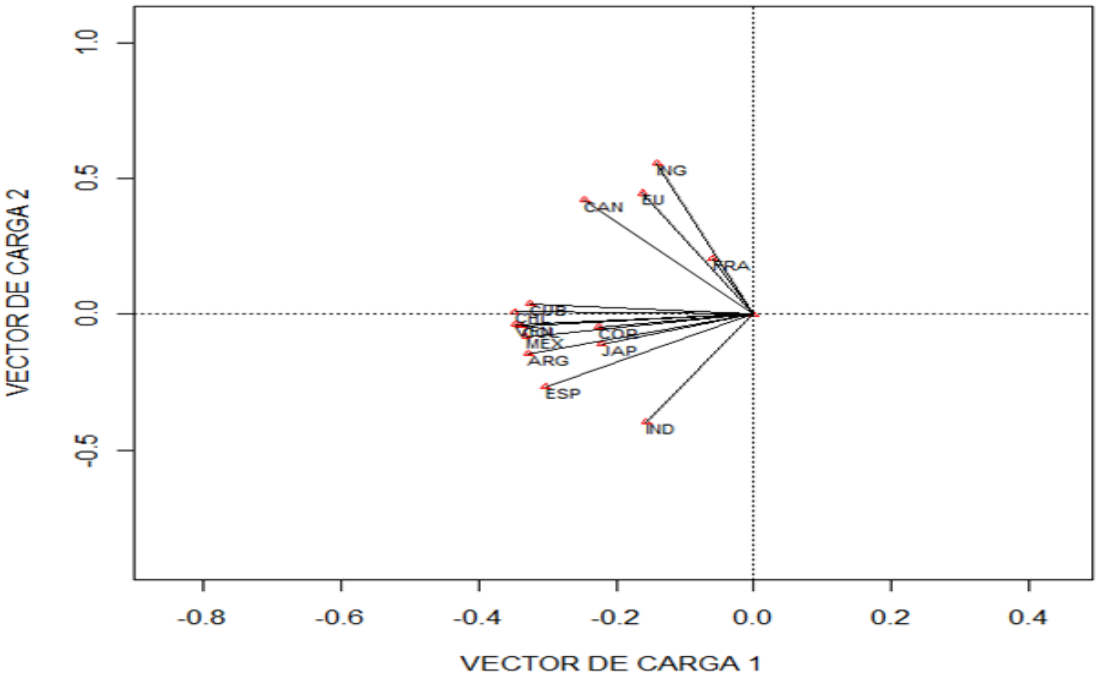

Figura 9: Gráfico de los vectores de carga.

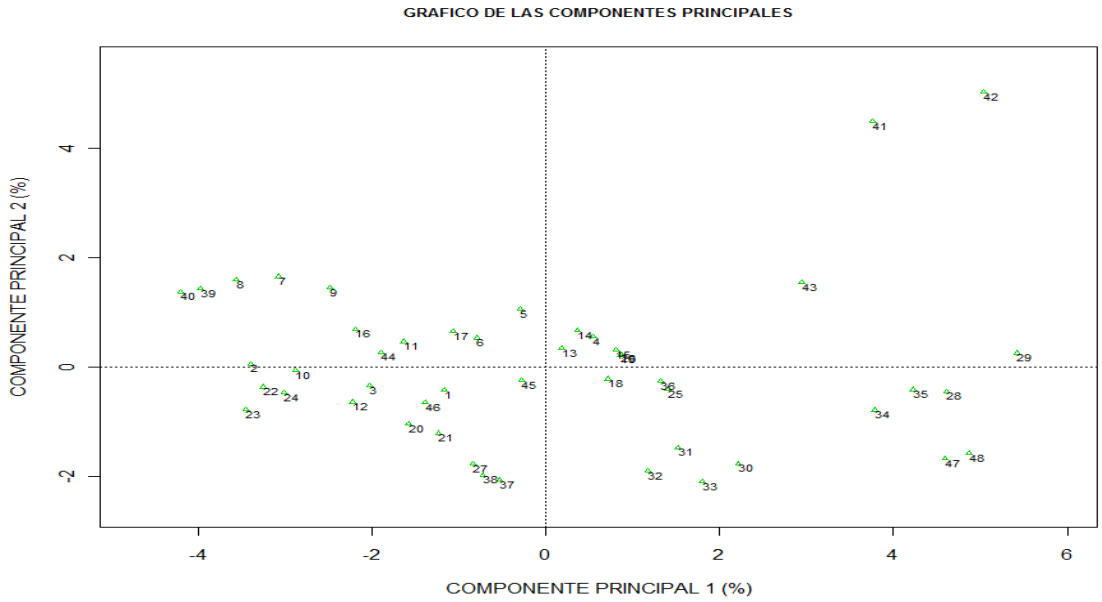

Figura 10: Gráfico de componentes principales.

Al graficar los componentes (ver Figura 9) (ver Figura 10), se puede visualizar las relaciones entre las variables estudiadas y las componentes extraídas. Los valores de cada variable en las coordenadas corresponden a los pesos de las mismas en los ejes de cada componente. En el gráfico se puede observar que las variables están más cerca de aquella componente sobre la que cargan más alto.

En la siguiente tabla se muestra el grado de correlación entre cada variable y la componente principal, indicando la contribución que dicha variable tuvo en la formación del eje correspondiente (Tabla 23).

Países	Comp.1	Comp.2	Comp.3
EU	0	1	0
Corea	1	0	1
Francia	0	0	0
Japón	1	0	1
España	1	0	0
Argentina	1	0	0
India	0	1	

Colombia	1	0	0
Inglaterra	0	1	0
Cuba	1	0	0
México	1	0	0
Venezuela	1	0	0
Chile	1	0	0
Canadá	1	1	0

Tabla 23: Variables que más contribuyen en cada componente principal.

A partir del gráfico de los vectores de carga y el gráfico de los componentes principales, se establecen las siguientes conclusiones en la aplicación del ACP sobre la muestra:

a. Se observó que la primera componente se caracteriza por los usuarios que presentan un alto número de accesos a las películas de lengua hispana (CHL, VEN, MEX, CUB, COL, ARG, ESP), ya que las variables que representan dichos valores están altamente correlacionados con la primera componente principal.
b. La segunda componente está dominada por las variables que representan las películas de lengua inglesa. La variable IND también presenta una alta correlación con esta componente, pero en sentido contrario (ver Figura 9), lo que significa que los usuarios que prefieren las películas pertenecientes a EU, CAN e ING, no ven películas de IND.
c. El análisis de los valores que contribuyen en la formación de cada componente (Tabla 23), permitió observar que un grupo pequeño de usuarios prefieren las películas asiáticas, que vienen representadas por las variables COR y JAP.
d. En la gráfica de los vectores de carga (ver Figura 9) y en las cifras de la Tabla 23, se observa una alta correlación entre las variables CUB, ESP, COL, MEX, CHL, ARG y VEN y estas a su vez presentan baja correlación con las formadas por las variables EU, ING y CAN. También se puede identificar 3 subgrupos más, conformados por COR y JAP, FRA e IND respectivamente. Por lo cual se puede presumir que existen 5 subgrupos de variables.
e. Por último, en la gráfica de los componentes (ver Figura 10), los individuos 41, 42 y 29 se encuentran alejados de la nube de puntos, lo que puede significar la presencia de datos *outlier*.

Para comprobar los posibles datos *outlier*, se utilizó un ACP robusto, empleando los gráficos de diagnóstico (ver Figura 11), mostrando que los usuarios 29, 41 y 42 (ver Anexo 2) son puntos *leverage* bueno, debido a que las distancias de puntajes son mayores que las frontera crítica. Sin embargo no es necesario sacar del análisis a estos individuos, ya que no determinan un incremento significativo de la varianza.

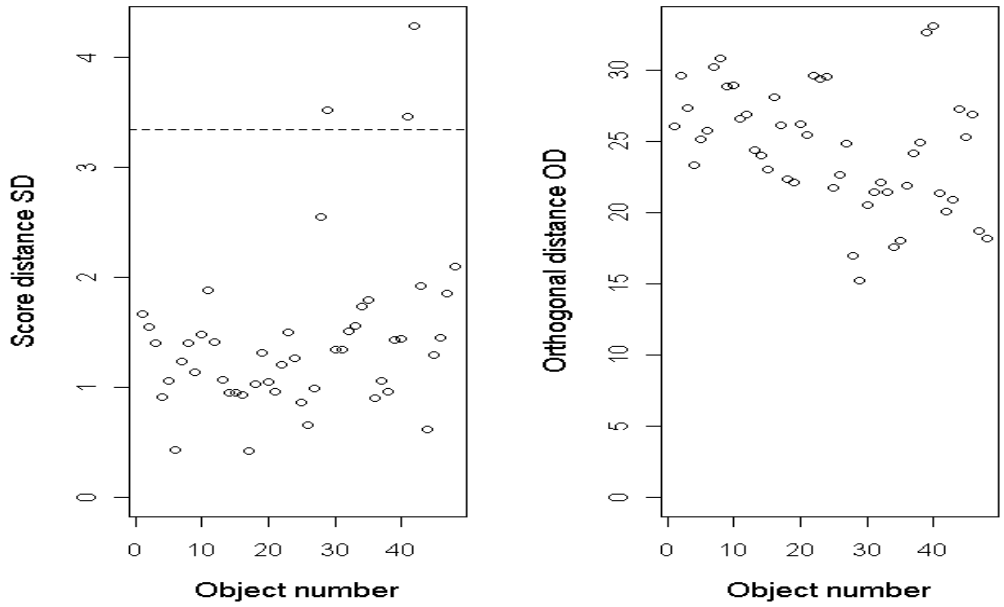

Figura 11: Gráficos de Diagnóstico.

A modo de resumen, el empleo de las técnicas de ACP y ACP robusto brindó resultados que permiten hacer interpretaciones a nivel de grupos de variables. Esto se completó con el señalamiento de individuos que presentaban un comportamiento atípico.

Transformación de los datos.

A partir de las interpretaciones obtenidas del estudio anterior, se deben emplear algoritmos que permitan transformar los datos para obtener información relevante de ellos. Debido a que el ACP fue usado con ánimo puramente exploratorio, con el fin de saber que grupos podían generarse a partir de las variables medidas, se hace necesario confirmar los resultados arrojados en este análisis y validar los subgrupos de variables identificados por el ACP. Para esto se emplean los algoritmos de agrupamiento.

Análisis de los datos utilizando el análisis por agrupación.

Se utilizó la medida de similitud de Hoeffding para realizar agrupamientos jerárquicos utilizando los algoritmos de amalgamiento simple, amalgamiento completo, distancia promedio ponderada y ward usando la función *varclus* que utiliza el software R para realizar análisis de *cluster* sobre variables. El proceso de unión de los *clusters* se puede observar a partir de los Dendogramas que forman los distintos métodos utilizados, los cuales a muestran a continuación.

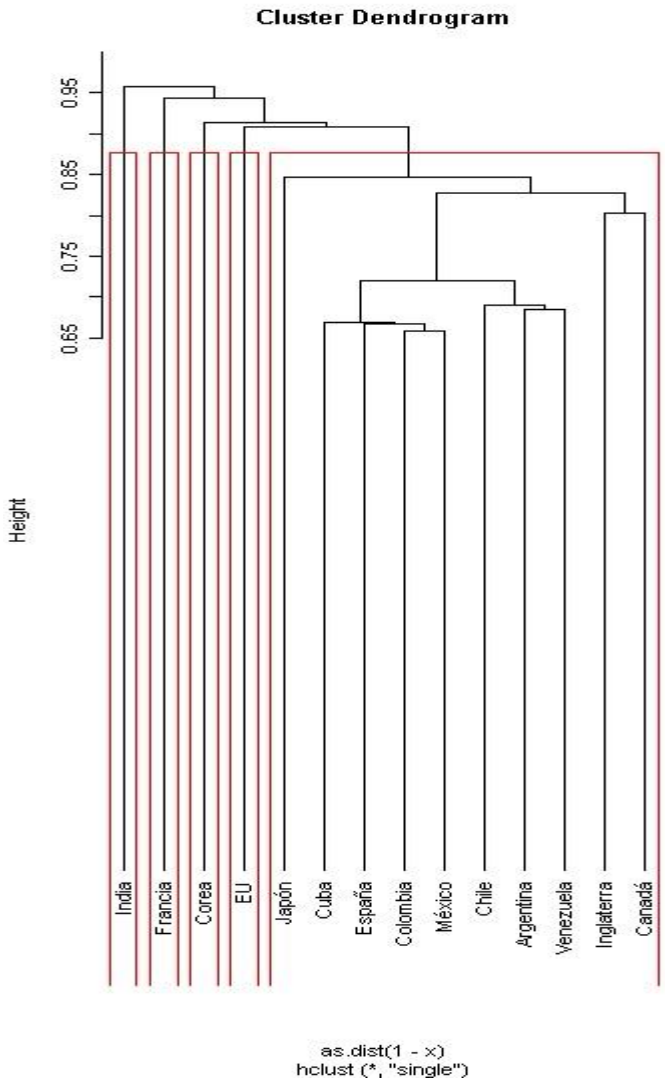

Figura 12: Dendrograma del método amalgamiento simple.

Conglomerados que se forman usando el método amalgamiento simple:

Conglomerado 1: IND.

Conglomerado 2: FRAN.

Conglomerado 3: COR.

Conglomerado 4: EU.

Conglomerado 5: JAP, CUB, ESP; COL, MEX, CHL, ARG, VEN, ING y CAN.

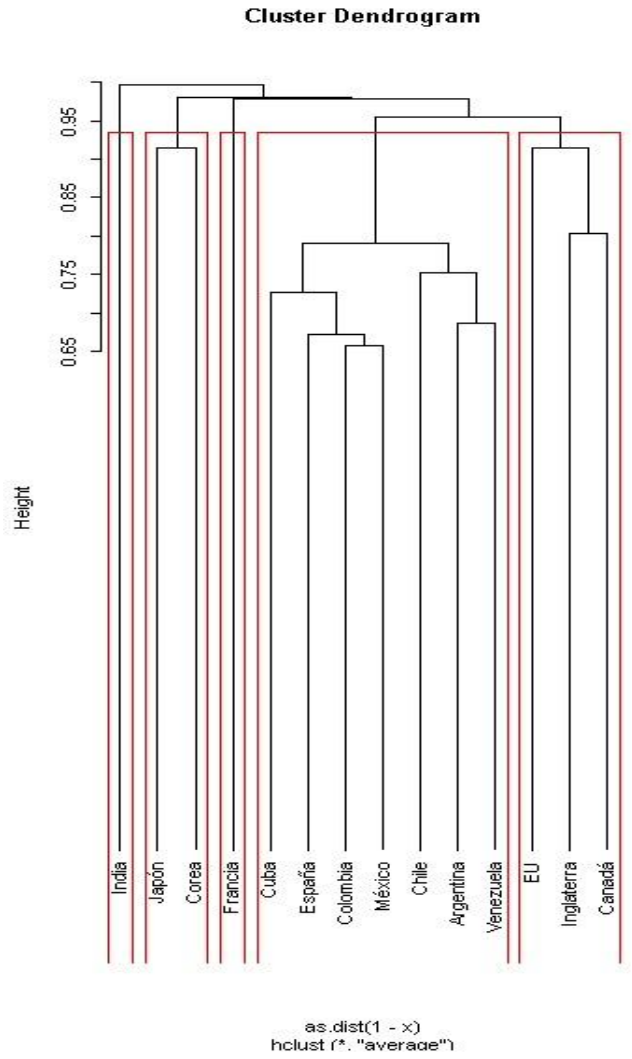

Figura 13: Dendrograma del método distancia promedio ponderada.

Conglomerados que se forman usando el método distancia promedio ponderada.

Conglomerado 1: IND.

Conglomerado 2: JAP y COR.

Conglomerado 3: FRAN.

Conglomerado 4: CUB, ESP, COL, MEX, CHL, ARG, y VEN.

Conglomerado 5: EU, ING y CAN.

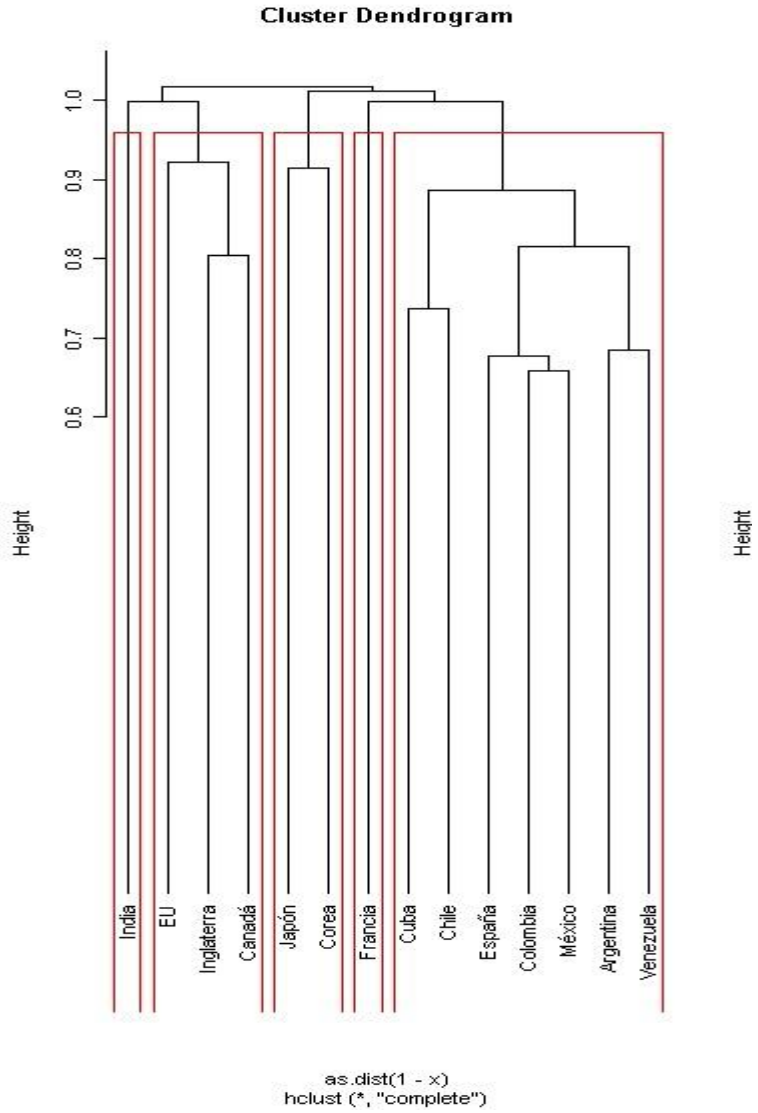

Figura 14: Dendrograma del método amalgamiento completo.

Conglomerados que se forman usando el método amalgamiento completo:

Conglomerado 1: IND.

Conglomerado 2: EU, ING y CAN.

Conglomerado 3: JAP y COR.

Conglomerado 4: FRAN.

Conglomerado 5: CUB, ESP, COL, MEX, CHL, ARG y VEN.

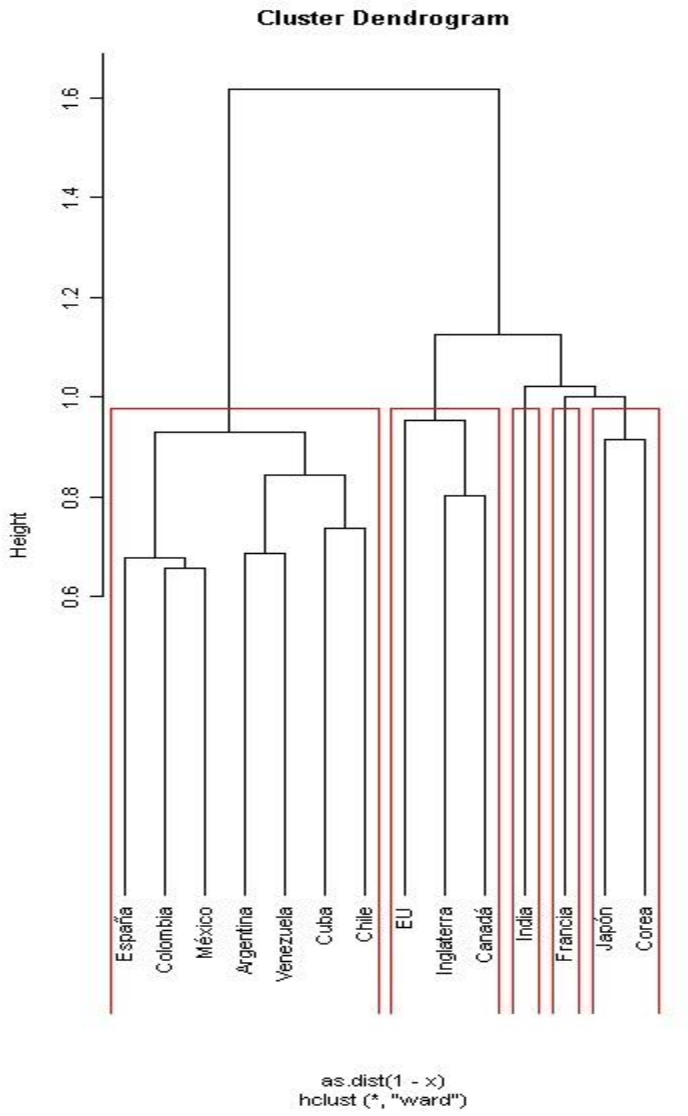

Figura 15: Dendrograma del método ward.

Conglomerados que se forman usando el método ward:

Conglomerado 1: CUB, ESP, COL, MEX, CHL, ARG y VEN.

Conglomerado 2: EU, ING y CAN.

Conglomerado 3: IND.

Conglomerado 4: FRAN.

Conglomerado 5: JAP y COR.

En los Dendogramas pertenecientes a los métodos amalgamiento simple (Figura 12), distancia promedio ponderada (Figura 13), amalgamiento completo (Figura 14) y ward (Figura 15), se observó que coinciden las variables que conforman los conglomerados formados por los tres últimos métodos. Los grupos finales de variable son:

Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
IND	FRAN	COR	EU	CUB
		JAP	ING	ESP
			CAN	COL
				MEX
				CHL
				COL
				VEN
				ESP

Tabla 24: Distribución de variables por agrupamiento.

3.1.2. Interpretación de los resultados.

La formación de estos grupos de variables pueden tener influencia en la forma que se diseñen los perfiles de usuarios, ya que no se están midiendo 14 variables para cada cliente como se pensaba inicialmente, si no que los agrupamientos precedentes determinan que a partir de 5 características diferentes que identifican a los países que conforman cada conglomerado, pueden diseñarse las interfaces que se muestren a los clientes. Al reducir la información de la muestra sobre pequeños grupos específicos, se logra una mayor concisión y una descripción más comprensible de las observaciones, con una mínima pérdida de información. De esta forma disminuye la recolección de datos y los costos de procesamiento en la implementación de un modelo en una gran base de datos. En este caso, como se señaló en el análisis de componentes principales, los grupos encontrados pueden representarse en función del idioma que presentan los países que conforman cada agrupamiento. También al dividir los países de origen de las películas en grupos homogéneos, es posible seleccionar los conglomerados encontrados a fin de probar diversas estrategias de mercadotecnia. Una estrategia válida podría ser aumentar la publicación de

las películas pertenecientes a los grupos 4 y 5 (ver Tabla 24), ya que en el ACP se observó que existía una gran densidad en la nube de puntos-individuos en función de la componente 1 (ver Figura10), lo que significa que las películas con más audiencia son las de lengua hispana y un aumento en la publicación de los filmes que presentan estas características, podría derivar en un consumo mayor de los servicios brindados por la Plataforma VideoWeb.

3.2. Conclusiones.

En este capítulo se desarrollaron y validaron los algoritmos de análisis de componentes principales, análisis de componentes principales robusto y análisis por agrupación. Con los resultados arrojados por estas técnicas, se pudo identificar patrones y tendencias en el comportamiento de la navegación de los usuarios con la Plataforma VideoWeb, y de esta forma cumplir con el objetivo de la investigación.

Conclusiones Generales.

Con la culminación del presente trabajo de diploma, se obtuvieron los resultados que se detallan a continuación.

- Se realizó un estudio de las características y estructura de la Minería de Datos, tomando un enfoque de minería de uso web para personalizar los servicios prestados por la Plataforma VideoWeb.

- Se llevó a cabo un estudio del arte de las distintas herramientas estadísticas que están disponibles actualmente para el análisis de datos, donde se obtuvo que el software y lenguaje de programación R es el más adecuado para la implementación de los algoritmos de análisis multivariante escogidos.

- Un análisis de los datos permitió concluir que son aplicables los métodos estadísticos de análisis de componentes principales para realizar agrupamientos de variables que permitan crear perfiles iguales para usuarios que consuman los mismos recursos, lo que permite personalizar la información que se le muestra a estos y tomar decisiones desde el punto de vista administrativas para mejorar el servicio a la comunidad cliente.

- Se utilizaron distintos métodos de análisis por agrupación, para probar la fiabilidad de los resultados obtenidos por el ACP y validar el trabajo realizado.

Recomendaciones.

Se recomienda la aplicación de las técnicas estadísticas multivariantes planteadas en función de la informática, específicamente en la Plataforma VideoWeb, ya que le permiten a este sistema contar con un valor agregado que hace de él un producto con una gran oportunidad de mercado. Integrando los resultados de esta investigación a la aplicación, le proveen de mecanismos autónomos para gestionar la atención a los usuarios y brindar satisfacción en el servicio a los mismos, lo cual a su vez, garantiza una oportunidad mucho más clara de ventas; cuestión muy importante para el país en materia de exportaciones o sustitución de importaciones.

Es además necesario recomendar que esta investigación se pueda ampliar teniendo en cuenta otros criterios para la selección de preferencias de usuario, ya que solo está enfocada al análisis de datos relacionados a los contenidos, pero se pudiera conceptualizar teniendo en cuenta criterios diversos como características personales obtenidas a partir de perfiles de usuarios y variables temporales de consumo de información. Mientras más se conozca acerca de los clientes sin llegar a ser invasivo se podrá llegar a brindar un servicio mucho más completo y orientado al usuario.

Bibliografía

Berry, M. and Linoff, G. 1997. *Data Mining Techniques for Marketing, Sales and Customer Support.* New York : John Wiley & Sons, 1997.

Castillo, José Antonio Sáez. 2009. *Métodos Estadísticos con R y R Commander.* Dpto. Estadítica e Investigación Operativa. Universidad de Jaén : s.n., 2009.

Cooley, R. and Mobasher, B and Srivastava, J. November 1997. *Web Mining: Information and Pattern Discovery on the World Wide Web.* November 1997.

Emmanuel, Paradis. 2003. *R para Principiantes.* Universit Montpellier II, France : s.n., 2003.

Fayyad, U., G. Piatetsky-Shapiro y P. Smyth. 1996. *Data Mining and Knowledge Discovery in Databases: An overview, Communications of ACM.* 1996.

Frakes, William B. Baeza-Yates, R. 1992. *Information retrieval : data structures & algorithms .* 1992. 0134638379 .

IWorld.com. **web, Técnicas y modelos de personalización de sitios. Sciences de l' Evolution, 2003.** Sciences de l' Evolution, 2003. F-34095 Montpellier cdex 05..

Jhonson, Dallas E. 1998. *Métodos Multivariados aplicados al Análisis de Datos.* Kansas State University : International Thomson Editores, 1998.

—. 1998 . *Métodos Multivariados Aplicados al Análisis de Datos.* Kansas State University : International Thomson Editores, Brooks Cole Publishing Company, 1998 . ISBN 0-534-23796-7.

Johnson, Dallas E. 2000. *Applied Multivariate Methods for Data Analysts.* s.l. : International Thomson Editores , 2000. 968-7529-90-3.

Jolliffe, I. T. 1986. *Principal component analysis.* New York: Springer : s.n., 1986. 0-387-96269-7.

Kosala, R. 2000. *Web Mining Research: A Survey. ACM SIGKDD Explorations Newsletter.* 2000.

LUCENA, MARÍA ALDEHUELA. Junio 2005. *ANÁLISIS COMPARATIVO ENTRE MÉTODOS ESTADÍSTICOS Y DE MINERÍA DE DATOS.* MADRID : PROYECTO FIN DE CARRERA, Junio 2005.

Mobasher, B, Jin, X. and Zhou, Y. 2004. *Semantically Enhanced Collaborative Filtering on the Web.* Proceedings of the European Web Mining Forum, LNAI, Springer : s.n., 2004.

Mobasher, B., Jain, N. and Han, E. and Srivastava, J. 1996. *Web mining: Pattern discovery from world wide web.* Department of Computer Science, University of Minnesota : s.n., 1996.

Molina, L. Nov. 2002. *confiesen, Data Mining: Torturando a los datos hasta que.* Nov. 2002.

Nakamura, A., Kudo, M. and Tanaka, A. 2003. Collaborative Filtering Using Restoration Operators. Springer-Verlag Berlin Heidelberg : s.n., 2003.

O'Conner, Mark and Herlocker Jon. 1999. *Clustering Items for Collaborative Filtering. Proceedings of the ACM.* 1999.

SALAS, CHRISTIAN. Agosto 2008. *¿Por qué comprar un programa estadístico si existe R?* Argentina : s.n., Agosto 2008.

Sánchez Enriquez, Hyder Ysaias. 2008. *Aplicación en Minería de Datos. Web Mining.* 2008.

TORRES, ANTONIO GONZÁLEZ. 2009. zarza.usal.es. [Online] 2009. http://zarza.usal.es/fgarcia/doctorado/iweb/05-07/Trabajos/MineriaWeb %20y %20/Personalizacion.pdf.

Varmuza, K. & Filzmoser. 2008. *Introduction to multivariate statistical analysis in chemometrics.* Boca Raton : CRC Press, 2008.

Velasquez, J., Bassi, A. and Yasuda, H. and Aoki, T. 2004. *Mining web data to create online navigation.* 2004.

Velderrey Sanz, Pablo. 2010. *Técnicas de segmentación de mercados.* s.l. : STARBOOK, 2010. 978-84-92650-28-6.

Web Mining: Information and Pattern Discovery on the World Wide Web. **Cooley, R. and Mobasher, B and Srivastava, J. November 1997.** s.l. : Proceedings of The 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), November 1997.

Zhou, Y., Jin, X. and Mobasher, B. 2004. *A Recommendation Model Based on Latent Principal Factors in Web Navigation Data.* New York : s.n., 2004.

Zhu, T. 2003. *Learning Browsing Behavior Model for Web Recommendation. Doctor of Philosophy Thesis.* University of Alberta, Edmonton, Canada, : s.n., 2003.

Anexos.

Anexo 1: Conjunto de datos.

ID	EU	FRA	IND	JAP	ESP	ARG	COR	COL	ING	CUB	MEX	VEN	CHL	CAN
1	6	7	2	5	8	7	8	8	3	8	9	7	5	10
2	9	10	5	8	10	9	9	10	5	9	9	8	8	10
3	7	8	3	6	9	8	9	7	4	9	9	8	6	10
4	5	6	8	5	6	5	9	2	8	4	5	8	7	5
5	6	8	8	8	4	4	9	5	8	5	5	8	8	7
6	7	7	7	6	8	7	10	5	9	6	5	8	6	6
7	9	9	8	8	8	8	8	8	10	8	10	8	9	10
8	9	9	9	8	9	9	8	8	10	9	10	9	9	10
9	9	9	7	8	8	8	8	5	9	8	9	8	8	10
10	4	7	10	2	10	10	7	10	3	10	10	10	9	10
11	4	7	10	0	10	8	3	9	5	9	10	8	10	5
12	4	7	10	4	10	10	7	8	2	8	8	10	10	7
13	6	9	8	10	5	4	9	4	4	4	5	4	7	8
14	8	9	8	9	6	3	8	2	5	2	6	6	7	6
15	4	8	8	7	5	4	10	2	7	5	3	6	6	6
16	6	9	6	7	8	9	8	9	8	8	7	6	8	10
17	8	7	7	7	9	5	8	6	6	7	8	6	6	8
18	6	8	8	4	8	8	6	4	3	3	6	7	2	4
19	6	7	8	4	7	8	5	4	4	2	6	8	3	4
20	4	8	7	8	8	9	10	5	2	6	7	9	8	9
21	3	8	6	8	8	8	10	5	3	6	7	8	8	8
22	9	8	7	8	9	10	10	10	3	10	8	10	8	8
23	7	10	7	9	9	9	10	10	3	9	9	10	9	8

24	9	8	7	10	8	10	10	10	2	9	7	9	9	8
25	6	9	7	7	4	5	9	3	2	4	4	4	4	4
26	7	8	7	8	5	4	8	2	3	4	5	6	5	6
27	2	10	7	9	8	9	10	5	3	5	6	7	6	5
28	6	3	5	3	5	3	5	0	0	3	3	0	0	0
29	4	3	4	3	3	0	0	0	0	4	4	0	0	0
30	4	6	5	6	9	4	10	3	1	3	3	2	2	3
31	5	5	4	7	8	4	10	3	2	5	5	3	4	3
32	3	3	5	7	7	9	10	3	2	5	3	7	5	2
33	2	3	5	7	7	9	10	3	2	2	3	6	4	2
34	3	4	6	4	3	3	8	1	1	3	3	3	2	2
35	6	7	4	3	3	0	9	0	1	0	2	3	1	3
36	9	8	5	5	6	6	8	2	2	2	4	5	6	3
37	4	9	6	4	10	8	8	9	1	3	9	7	5	2
38	4	9	6	6	9	9	7	9	1	2	10	8	5	2
39	10	6	9	10	9	10	10	10	10	10	8	10	10	10
40	10	6	9	10	9	10	10	10	10	10	10	10	10	10
41	10	7	8	0	2	1	2	0	10	2	0	3	0	10
42	10	3	8	0	1	1	0	0	10	0	0	0	0	10
43	3	4	9	8	2	4	5	3	6	2	1	3	3	8
44	7	7	7	6	9	8	8	6	8	8	10	8	8	5
45	9	6	10	9	7	7	10	2	1	5	5	7	8	5
46	9	8	10	10	7	9	10	3	1	5	7	9	9	4
47	0	7	10	3	5	0	10	0	0	2	2	0	0	0
48	0	6	10	1	5	0	10	0	0	2	2	0	0	0

Anexo 2: Tabla de datos *outlier*.

ID	
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	0
17	0
19	0
20	0
21	0
22	0
23	0
24	0
25	0

ID	
26	0
27	0
28	0
29	P.L.B
30	0
31	0
32	0
33	0
34	0
35	0
36	0
37	0
38	0
39	0
40	0
41	P.L.B
42	P.L.B
43	0
44	0
45	0
46	0
47	0
48	0

Anexo 3: Código en R para ACP.

```
#Entrada de datos
misdatos<-Datos[,-c(1)]
misdatos
#Método de análisis de componentes principales
pca1<-princomp(misdatos,cor=T)
summary(pca1)

#Screeplots
plot(pca1$sd^2,type='o',pch=19,xlab="NUMERO DE COMPONENTES PRINCIPALES (PCs)",ylab="VARIANZAS DE LAS PCs",cex.lab=1.0,main="Diagrama de segmentación",cex.main=0.8);

#ScatterPlot de los vectores de cargas
variables<-
c("EU","Corea","Francia","Japón","España","Argentina","India","Colombia","Inglaterra","Cuba","México","Venezuela","Chile","Canadá")
x<-c(0,pca1$loading[,1]);y<-c(0,pca1$loading[,2])
plot(x,y,pch=2,col=2,cex=0.5,ylim=c(min(pca1$loading[,2])-
0.5,max(pca1$loading[,2])+0.5),xlim=c(min(pca1$loading[,1])-0.5,max(pca1$loading[,1])+0.5),
xlab="VECTOR DE CARGA 1",ylab="VECTOR DE CARGA 2",cex.lab=1.0,main="GRAFICO DE LOS VECTORES DE CARGAS",cex.main=0.8);
text(x[-1],y[-1],labels=variables,adj=c(0,1),cex=0.7);abline(h=0,lty=3);abline(v=0,lty=3);
plot_var<-function(a)(segments(0,0,pca1$loading[,1][a],pca1$loading[,2][a]))
for (i in 1:length(variables)) (plot_var(i))

#Selección de las variables que más contribuyen a cada componente principal
funcion1<-function(p){pca1$loading[,1:p]}
vectores<-funcion1(4)#Vectores propios de las componentes principales
Selec.var<-matrix(0,dim(vectores)[1],dim(vectores)[2])
for (j in 1:dim(vectores)[2]) {for (i in 1:dim(vectores)[1])
```

```
if ((abs(vectores[i,j]))>=max(abs(vectores[,j]))/2) Selec.var[i,j]<-1 else Selec.var[i,j]<-0
}
rownames(Selec.var)<-variables;Selec.var

#Diagrama de dispersión de los scores de las componentes principales
usuarios<-1:48
plot(pca1$scores[,1],pca1$scores[,2],pch=2,col=2,cex=0.5,ylim=c(min(pca1$scores[,2])-
0.5,max(pca1$scores[,2])+0.5),xlim=c(min(pca1$scores[,1])-0.5,max(pca1$scores[,1])+0.5),
xlab="COMPONENTE PRINCIPAL 1 (50%)",ylab="COMPONENTE PRINCIPAL 2
(15%)",cex.lab=1.0,main="GRAFICO DE LAS COMPONENTES PRINCIPALES",cex.main=0.8);
text(pca1$scores[,1],pca1$scores[,2],labels=usuarios,adj=c(0,1),cex=0.7);abline(h=0,lty=3);abline(v=0,lty=
3);
```

Anexo 4: Código en R para ACP Robusto.

```
#Análisis de componentes principales robusto
library(chemometrics)
library(pcaPP)
X.rpc<-function(X,m)(PCAgrid(X,k=m,scale=mad,center = l1median(X,trace = -1)))#X está escalado a varianza 1 usando estimador robusto
DCP<-X.rpc(misdatos,14)$sdev# las desviaciones estándar (robustas) de las componentes principales
prop_var<-DCP^2/sum(DCP^2)# proporción de varianzas (robustas) de las componentes principales
prop_acum<-numeric(length(DCP));prop_acum[1]<-prop_var[1]
for (i in 2:length(DCP))(prop_acum[i]<-prop_var[i]+prop_acum[i-1])
summary<-rbind(DCP,prop_var,prop_acum)#matriz resumen conteniendo las desviaciones estándar (robustas),

#Diagrama de segmentación
plot(DCP^2,type='o',pch=19,xlab="NUMERO DE COMPONENTES PRINCIPALES (PCs)",ylab="VARIANZAS DE LAS PCs",cex.lab=1.0,main="Diagrama de segmentación",cex.main=0.8);

#gráficos de diagnósticos para PCA ROBUSTO
res<-pcaDiagplot(misdatos,X.rpc(misdatos,14),a=4)#genera gráficos de diagnósticos
#usando cuatro componentes principales
SDist<-res$SDist#distancia de scores
ODist<-res$ODist#distancia ortogonal
outliers<-numeric(length(SDist))
for (i in 1:length(SDist)){
if (SDist[i]>res$critSD&&ODist[i]>res$critOD) outliers[i]<-"P.L.M"
if (SDist[i]>res$critSD&&ODist[i]<=res$critOD) outliers[i]<-"P.L.B"
if (SDist[i]<=res$critSD&&ODist[i]>res$critOD) outliers[i]<-"O.O"
}
OUTLIERS<-as.matrix(as.factor(outliers));OUTLIERS
```

Anexo 5: Código en R para análisis por agrupación.

```
#métodos de agrupamientos gerárquico
library(Hmisc)

cluster<-function(M,n){
csine<-varclus(as.matrix(M), similarity=c("hoeffding"),type=c("data.matrix"), method =
"single")#acoplamiento único
cwarde<-varclus(as.matrix(M), similarity=c("hoeffding"),type=c("data.matrix"),method = "ward")#ward
ccome<-varclus(as.matrix(M),
similarity=c("hoeffding"),type=c("data.matrix"),method="complete")#acoplamiento completo
caverage<-varclus(as.matrix(M), similarity=c("hoeffding"),type=c("data.matrix"),method=
"average")#distancia promedio ponderada

#dendogramas de
par(mfrow = c(1, 4))
plot(csine$hclust, hang = -1)#amalgamiento simple
rect.hclust(csine$hclust, n)
plot(cwarde$hclust , hang = -1)#ward
rect.hclust(cwarde$hclust, n)
plot(ccome$hclust , hang = -1)#amalgamiento completo
rect.hclust(ccome$hclust, n)
plot(caverage$hclust, hang = -1)# average
rect.hclust(caverage$hclust, n)

clases<-as.matrix(cbind(cutree(csine$hclust, n),cutree(ccome$hclust, n),cutree(cwarde$hclust,
n),cutree(caverage$hclust, n)))
print(clases)
}
cluster(misdatos,5)
```

Glosario

Centroides de agrupamiento: son los valores medios (medias) de las variables para todos los casos u objetos de un grupo particular.

Coeficiente de correlación: entre parejas de variables, permite agrupar variables de tal manera que variables en el mismo grupo tengan correlaciones altas y variables en grupos diferentes tengan correlaciones bajas.

Conocimiento: es el resultado del proceso de minado.

Colineales: dos cuerpos o componentes que están ubicados en la misma dirección y sobre una misma línea recta.

Correlación: es una medida de la relación entre dos o más variables. La correlación puede tomar valores entre –1 y +1. El valor de –1 representa una correlación negativa perfecta mientras un valor de +1 representa una correlación perfecta positiva. Un valor de 0 representa una falta de correlación.

Covarianza: relación sistemática entre dos variables, en la cual el cambio en una implica un cambio correspondiente en la otra.

Datos continuos: datos que pueden pasar de una clase a la siguiente sin interrumpirse y que pueden expresarse mediante números enteros o fraccionarios.

Dendrograma: llamado también gráfica de árbol jerárquico, es un dispositivo gráfico para presentar los resultados del conglomerado. Las líneas verticales representan los grupos que están unidos. La posición de la línea en la escala indica las distancias en las que se unieron los grupos. Se lee de izquierda a derecha.

Desviación estándar: raíz cuadrada positiva de la varianza; medida de dispersión con las mismas unidades que los datos originales, más bien en las unidades al cuadrado en que está la varianza.

Dispersión: es la extensión o variabilidad de un conjunto de datos.

Espacio euclídeo: es un espacio vectorial dotado de un producto escalar.

Incorrelacionadas: inexistencia de correlación entre las variables medidas.

Inferencia estadística: proceso de generalizar los resultados de la muestra a los resultados de la población.

KDD: acrónimo de descubrimiento de conocimiento en bases de datos, que es el proceso de identificar patrones útiles a partir de cantidades de datos para encontrar conocimiento útil en ellos (Kosala, 2000).

Media: el promedio, valor que se obtiene al sumar todos los elementos en un conjunto y dividirlos entre el número de elementos.

Normalización: estandarización de los datos, con el objetivo de lograr un mejor acoplamiento de estos.

Multicolinealidad: problema estadístico que se presenta en el análisis de regresión múltiple, en el que la confiabilidad de los coeficientes de regresión se ve reducida debido a un alto nivel de correlación entre las variables independientes.

R: software estadístico y lenguaje de programación de uso libre, de distribución gratuita y de código abierto (SALAS, Agosto 2008).

Repositorio: es un sitio centralizado donde se almacena y mantiene información digital, habitualmente bases de datos o archivos informáticos.

Regresión: proceso general que consiste en predecir una variable a partir de otra mediante medios estadísticos, utilizando datos anteriores.

Simientes: son los puntos de partida iniciales en el análisis por agrupación no jerárquica. Los grupos se construyen alrededor de estas simientes o semillas.

Streaming: consiste en la distribución de audio o video por Internet, sin necesidad de descargar lo que se está viendo.

Variable: propiedad o rasgo de un hecho u objeto (no constante) por la que puede ser caracterizado o clasificado. Representación de una característica, de un atributo, que posee alguna realidad.

Varianza: promedio de la desviación al cuadrado de una variable respecto de su media.

Valores propios: de una matriz Σ son las raíces de la ecuación polinomial definida por $\Sigma - \lambda I = 0$

Vectores propios: valor no cero correspondiente a (una columna de números), al valor propio de la matriz Σ, que satisface la expresión $\Sigma a = \lambda a$.

Vectores ortogonales: dos vectores son ortogonales si su producto escalar es cero.