Universidad de las Ciencias Informáticas

Facultad 6

Título: “Complemento para la plataforma GeneSIG para la identificación de elementos similares.”

Trabajo de Diploma para optar por el Título de

Ingeniero en Ciencias Informáticas

Autor: Ihosbanny Arteaga Aguilar

Tutor: Ing. Carlos Enrique Ramírez Martín

La Habana, Julio de 2015

“Año 56 de la Revolución”

Frase


"No existe una manera fácil. No importa cuán talentoso seas, tu talento te va a fallar si no lo desarrollas. Si no estudias, si no trabajas duro, si no te dedicas a ser mejor cada día."
Will Smith

Declaración de Autoría

Declaro ser autor del trabajo “Complemento para la plataforma GeneSIG para el agrupamiento de elementos similares” y reconozco a la Universidad de las Ciencias Informáticas (UCI) los derechos patrimoniales de la misma, con carácter exclusivo.

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

Ihosbanny Arteaga Aguilar

_____

Firma del Autor

Ing. Carlos Enrique Ramírez Martín

_____

Firma del Tutor

Datos de contacto

Tutor: Ing. Carlos Enrique Ramírez Martín

Correo electrónico: cmartin@uci.cu

Síntesis del Tutor: Graduado de Ingeniero en Ciencias Informáticas en la Universidad de las Ciencias Informáticas en el 2009. Posee la categoría docente de instructor y ha impartido docencia directa. Desarrollador de interfaces de la factoría Interfaz del proyecto Aplicativos SIG del centro GEYSED de la facultad 6 de la UCI desde el año 2011 y profesor del departamento de Programación y Sistemas Digitales a partir del 2014, posee 3 productos registrados en CENDA como parte de un equipo de desarrollo.

Agradecimientos

Las palabras más hermosas que pudiera decir para agradecer a todas las personas que me apoyaron todo este tiempo no son nada en comparación con el amor y respeto tan grande que siento por cada una de ellas. Espero entonces que estas líneas logren llegar a ustedes con la misma fuerza que causó su apoyo en mi corazón:

A mis padres Miriam y Osmani por siempre estar ahí cuando más los necesité, apoyándome en cada momento.

A mi hermana Isis de la C. que aunque no se lo diga a menudo no se imagina lo importante que fue para mí todo este tiempo y lo será para toda la vida.

A mi abuela Caridad (Cari) que me apoyó siempre y me daba ese aliento cuando más lo necesitaba.

A Mayi por ser tan especial y única, me enseñaste muchas cosas de las que nunca me olvidaré.

A Deyli por ser mi prima, hermana, y tía, siempre apoyándome y dándome consejos en todo.

A mis primos Naty, Eliza, Eliani, Yuderkis, Yunior, Leo, Lizbety, Lisandra

A mis tíos y tías Eloy, Yoel, Milo, Mildrey, Maira

A mis amistades del barrio, de Villa Clara.

Al Pollo por ser mi amigo desde que entre en esta Universidad

A Luis D. por joderlo tanto cada vez que me pasaba algo con la PC

Agradecimientos

En fin a todos los que conocí durante estos 5 años de la carrea.

A mis tutores Carlos y Aliana por hacer lo mejor que pudieron por mí. A pesar de que Aliana nos dejó por problemas

Resumen

En Cuba, se encuentra en pleno desarrollo la plataforma soberana para el desarrollo de Sistemas de Información Geográfica, GeneSIG. Dicha plataforma cuenta con numerosas funcionalidades que se dedican al análisis de los datos y al cumplimiento de las necesidades del cliente. Por el impacto económico-social que tiene en nuestro país se le atribuye gran importancia a este producto, pero la misma no cuenta con un complemento capaz de identificar elementos similares en un mapa teniendo en cuenta varios atributos. Por lo cual se convierte en una necesidad suplir esta carencia que posee dicha plataforma. La presente investigación se plantea como objetivo es desarrollar un complemento que permita identificar los elementos geoespaciales similares entre sí con múltiples atributos, dada su información socioeconómica. Para ello se hará uso de la herramienta Weka que dentro de sus funcionalidades tiene implementado el algoritmo K-Means para el agrupamiento de datos. Esta herramienta se ejecutará en la máquina virtual de Java mediante comandos, lo que traería grandes beneficios a la plataforma ya que se ejecutará como un servicio aparte al servicio web y al servicio de Base de Datos y solo en el momento que sea utilizado el complemento. Además se analizarán una serie de algoritmos matemáticos, técnicas de Minería de Datos (MD), así como las herramientas, tecnologías y metodología de desarrollo de software apropiadas para su construcción.

Palabras Clave: agrupamiento de datos, K-Means, Minería de Datos, Sistemas de Información Geográfica, Weka.

Abstract

In Cuba, it is located in the sovereign development platform for the development of GIS, GeneSIG. This platform has many features that are dedicated to data analysis and fulfillment of customer needs. On the socio-economic impact on our country it is attributed great importance to this product, but it does not have a component able to identify similar elements on a map based on several attributes. Thus becomes a need to fill this gap that owns the platform. This research therefore seeks to develop a component to identify similar geospatial elements together with multiple attributes, given their socioeconomic information. For it will make use of the Weka tool within their functionality it has implemented the K-means algorithm for clustering data. This tool will run on the Java virtual machine by commands, which would bring great benefits to the platform as it will run as a separate service to web service and the service of Database and only when it is used the component. In addition a series of mathematical algorithms, data mining techniques (MD) as well as the tools, technologies and software development methodology suitable for construction will be discussed.

Keywords: data clustering, K-Means, Data Mining, Geographic Information Systems, Weka.

Índice

Índice

Índice de Figuras

Introducción

Los Sistemas de Información Geográfica (SIG) constituyen hoy en día una de las herramientas más utilizadas y demandadas por los usuarios en la Red de Redes. Dentro de las extraordinarias potencialidades que brindan es que permiten georeferenciar todo tipo de información que maneja una entidad, es decir, se puede localizar un lugar determinado por referencias geográficas (direcciones postales, coordenadas cartográficas o distribución por municipios, sectores, barrios, secciones censales, etc.). En Cuba esto constituye un auténtico reto, por lo que es objetivo de la Universidad de las Ciencias Informáticas (UCI) que sus profesionales le den solución a los problemas actuales en este sentido, haciendo uso de las Tecnologías de la Información y la Comunicación (TIC). Las Bases de Datos (BD) juegan un papel importante en este reto como principal herramienta para el almacenamiento de los datos. Debido a las características propias de cada tipo de información el hombre ha desarrollado BD específicas para cada uno de estos tipos de datos. Una de las más utilizadas son las BD relacionales, cuyo objetivo principal es modelar problemas reales y administrar un conjunto de datos relacionados entre sí. Aunque estas son las más usadas hoy en día, existen aplicaciones que requieren otros tipos de organización de la información para almacenar datos más complejos como por ejemplo las BD espaciales las cuales son utilizadas esencialmente para almacenar datos geoespaciales, facilitando así el análisis de estos y su visualización en forma de mapas.

Las TIC tienen el potencial de brindar nuevas soluciones a los problemas del desarrollo en nuestro país y pueden promover el crecimiento económico, la competitividad, el acceso a la información y los conocimientos, es por ello que surge la Universidad de las Ciencias Informáticas, la cual es una de las principales instituciones encargada de llevar a cabo el proceso de informatización en el país y desarrollar herramientas para la gestión y análisis de los datos. Dentro de la universidad existen varios centros productivos entre los que se encuentra el Centro de Desarrollo Geoinformática y Señales Digitales (GEYSED), cuyo objetivo es desarrollar productos, servicios y soluciones informáticas en el campo del procesamiento de Señales Digitales y la Geoinformática, contribuyendo a la formación integral de profesionales que respondan a las necesidades del progreso científico técnico y socioeconómico. (GEYSED, 2015)

A dicho centro está vinculada la Línea de Producto de Software (LPS) AplicativoSIG, la cual tiene como objetivo la representación de mapas y datos geoespaciales en la web sobre la plataforma soberana para

el desarrollo de Sistemas de Información Geográfica, GeneSIG. Actualmente la plataforma GeneSIG cuenta con una gran cantidad de funcionalidades que se dedican al trabajo con los mapas y al cumplimiento de las necesidades del cliente. Mapa temático es una de estas funcionalidades que se dedican al análisis de los datos, esta funcionalidad puede realizar una tematización en dependencia de lo que desee el cliente ya que cuenta con diferentes clasificaciones como: símbolo proporcional, estilo graduado, categorizado y graficas dinámicas, esta tematización se realiza teniendo en cuenta un atributo por cada capa, a dicho complemento se le realizó una nueva versión para poder realizar la tematización teniendo en cuenta varios atributos de una misma capa, pero aun así no cuenta con una forma de obtener información oculta en los datos. Los complementos desarrollados hasta la fecha solo permiten el análisis de información a través de tematizaciones predefinidas que muestran los elementos estableciendo rangos de valores prestablecido por los usuarios para un determinado parámetro. A través del análisis de estas tematizaciones se dificulta en gran medida agrupar los elementos dado una serie de criterios definidos por los propios usuarios.

Por lo antes planteado surge el siguiente **problema de investigación:** ¿cómo identificar los elementos geoespaciales similares entre sí con múltiples atributos, a partir de su información socioeconómica, en las aplicaciones desarrolladas con GeneSIG?

En consecuencia con lo anterior se determinó como **objeto de estudio:** la gestión de la información socioeconómica adjunta a datos geoespaciales y como **campo de acción:** identificación de elementos geoespaciales similares.

La investigación tiene como **objetivo general:** desarrollar un complemento para la plataforma GeneSIG que permita identificar los elementos geoespaciales similares entre sí con múltiples atributos, dada su información socioeconómica.

Una vez trazado el objetivo general, se determinan los siguientes **objetivos específicos:**

- Analizar las tendencias actuales en el proceso de identificación de elementos geoespaciales similares analizando su información socioeconómica.
- Caracterizar las técnicas y algoritmos de identificación de elementos similares más idóneos para dar solución al problema.

- Desarrollar el análisis, diseño e implementación de la solución planteada.
- Validar la solución desarrollada mediante las realizaciones de pruebas de caja negra.

Con el fin de garantizar el cumplimiento de dichos objetivos se hace necesario el cumplimiento de las siguientes **tareas:**

1. Caracterización de las principales técnicas de identificación de elementos similares en el análisis de información geoespacial para seleccionar el algoritmo que más se adecue para el desarrollo de la solución en la presente investigación.
2. Fundamentación de la metodología de desarrollo de software para guiar el proceso de desarrollo de software durante la investigación.
3. Modelación de los requisitos funcionales y no funcionales de la propuesta de solución para darle cumplimiento a las necesidades del cliente.
4. Realización del análisis y diseño de la solución para lograr el buen funcionamiento del complemento.
5. Implementación de la solución y el desarrollo de la documentación asociada para su uso en la plataforma GeneSIG.
6. Comprobación de la validez de la herramienta desarrollada mediante la realización de pruebas de caja negra.

Como **resultado** de esta investigación se pretende obtener un complemento para identificar los elementos geoespaciales similares entre sí con múltiples atributos, dada su información socioeconómica, en la plataforma GeneSIG.

Para el buen desarrollo de la documentación y elaboración del complemento se hará uso de diferentes métodos científicos. A continuación se dará una breve explicación da la utilización de los mismos en la presente investigación:

Métodos Teóricos:

- **Análisis-Síntesis:** posibilitaron esencialmente la obtención de información teórica acerca de las diferentes herramientas, técnicas y algoritmos para la identificación de elementos similares así como el análisis de los datos geoespaciales**.**

- **Histórico-Lógico:** se empleó con el fin de establecer los antecedentes, tendencias y regularidades del objeto de estudio y el campo de acción para la obtención de un mayor conocimiento sobre la evolución de las técnicas de identificación de elementos similares y evolución de los complementos de la plataforma GeneSIG para poder integrar de forma efectiva el nuevo complemento.

- **Modelación:** contribuyó a definir la estructura, relación de los complementos, funcionalidad de la propuesta de solución. Además permitió la elaboración de los diferentes diagramas que se utilizarán en la investigación y para la construcción de los modelos matemáticos dada su información y elementos seleccionados para cada análisis.

Métodos Empíricos:

- **Observación:** este método se utilizó durante todo el proceso de investigación para percibir los hechos y características manifestadas alrededor de la problemática planteada, los procesos de gestión analizados así como arribar a conclusiones que permitan modelar y aplicar la solución que se propone. Mediante ella se obtienen además datos de las fuentes prácticas.

El presente trabajo está conformado por 3 capítulos que se encuentran estructurados de la siguiente forma:

En el **capítulo 1** se analizan los principales conceptos que permitan una mejor comprensión de la investigación a realizar, así como técnicas y algoritmos de agrupamiento. Se abordan además las tecnologías, metodologías y herramientas existentes que se utilizarán en la propuesta para el desarrollo de la solución.

En el **capítulo 2** se realiza el modelado y el levantamiento de los requisitos funcionales y no funcionales del complemento. A partir de los mismos se procede al diseño detallado de las clases generando el modelo de dominio, el diagrama de paquetes, el diagrama de despliegue y los diagramas de clases del diseño que sientan las bases para las futuras etapas de implementación y prueba.

En el **capítulo 3** se describen algunos de los artefactos generados según la metodología previamente seleccionada y se realizan las pruebas de software que permitan garantizar el cumplimiento de los requisitos de la investigación.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA

Introducción

En el capítulo se analizan los distintos conceptos y definiciones asociados al dominio de la investigación para lograr una mayor comprensión. Se analizan además algunas herramientas existentes que dan soporte a la Minería de Datos (MD), fundamentales para garantizar el sustento teórico en la construcción del complemento que se desea desarrollar.

1.1 Conceptos asociados al dominio de la investigación.

En el presente epígrafe se describen términos y conceptos afines al campo de acción en el que se enmarcó la investigación.

1.1.1 Sistema de Información Geográfica

La propuesta por el Centro Nacional de Información y Análisis Geográfico (NCGIA por sus siglas en inglés) de los Estados Unidos plantea que: *"Un SIG es un sistema de información compuesto por hardware, software y procedimientos para capturar, manejar, analizar, modelar y representar datos Georreferenciados, con el objetivo de resolver problemas de gestión y planificación".* (NCGIA, 2015)

1.1.2 Datos

Debido a la importancia que tiene la investigación y que va a estar enfocada principalmente en el trabajo con los datos asociados a los mapas y elementos geoespaciales con los que trabaja GeneSIG, se hace necesario definir qué son estos desde el punto de vista de varios autores.

Por su parte Gil Flores en su libro "Análisis de Datos Cualitativos. Aplicaciones a la Investigación Educativa", Barcelona, Edit. PPU, 1994, Cap. 1, plantea que: *"Se puede definir los datos como aquella información extraída de la realidad que tiene que ser registrada en algún soporte físico o simbólico, que implica una elaboración conceptual y además que se pueda expresar a través de alguna forma de lenguaje".* (Flores, 1994)

Por otro lado, Kruse en su libro "Estructura de Datos y Diseños de Programas" plantea que: *"Los datos son los hechos que describen condiciones, situaciones, valores, sucesos o entidades, son comunicados*

por varios tipos de símbolos tales como las letras del alfabeto, números, movimientos de labios, puntos y rayas, señales con la mano, dibujos, entre otros". (Kruse, 2014)

La presente investigación se regirá por el concepto de Flores ya que se considera que este es el que cumple de una manera más abarcadora los elementos referentes con la investigación. Un mismo dato puede tener varios significados en dependencia del contexto donde se esté usando y de quien lo está analizando. La plataforma GeneSIG opera con los datos geoespaciales por lo que es necesario conocer qué son los datos geoespaciales.

1.1.3 Datos Geoespaciales

Según Gastelú en su libro "Servicio y localización espacial" plantea que: *"Los datos geoespaciales se utiliza para representar puntos, líneas y áreas en una superficie. Por lo general, estos elementos se refieren a la actual ubicación física en la Tierra, por lo que se puede describir un conjunto de datos geoespaciales"*. (Gastelú, 2011)

De acuerdo con la empresa Española "Ager Ingenieros": *"Los datos espaciales o geodatos presentan dos tipos de propiedades: las geométricas y las descriptivas. Estas propiedades son las que les proporcionan su utilidad, constituyendo así el núcleo de los Sistemas de Información Geográfica".* (Ager, 2003)

Una vez analizados cada uno de los conceptos referentes a los datos geoespaciales, la presente investigación se guiará según lo planteado por Gastelú. Con el desarrollo y aumento que han alcanzado los datos hoy en día, es necesario almacenarlos para luego poder convertirlos en información. Una de las herramientas que realizan esta función son las BD, cuyo concepto se define a continuación.

1.1.4 Base de Datos

Según la entidad Microsoft una base de datos es: *"Una herramienta que es utilizada para recopilar y organizar información sobre personas, productos, pedidos, o cualquier otra cosa".* (Microsoft, 2014)

Por otro lado C. J. Date, en su libro "Introducción a los Sistemas de Bases de Datos", define una base de datos como *"Un sistema computarizado para llevar registros. Es posible considerar a la propia base de datos como una especie de armario electrónico para archivar; es decir, es un depósito o contenedor de una colección de archivos de datos computarizados".* (Date, 2001)

De los diferentes conceptos de base de datos y datos geoespaciales previamente mencionados según distintos autores o instituciones, este trabajo considerará lo planteado por Date. Una vez visto estos conceptos, es necesario conocer donde pudieran ser almacenados los datos geoespaciales para trabajar con ellos.

1.1.5 Base de Datos Espaciales

Ralf Hartmut Güting plantea en su libro "An Introduction to Spatial Database Systems": *"Una base de datos espacial es un sistema de bases de datos que ofrece tipos de datos espaciales en su modelo de datos y lenguaje de consulta, soporta diversos tipos de datos espaciales en su implementación, proporcionando al menos indexación espacial y métodos de unión espacial"*. (Güting, 1994)

John Wiley plantea en su libro "Wiley Encyclopedia of Computer Science and Engineering" que: "Base de Datos Espacial es un módulo de software que puede trabajar con un sistema de gestión de base de datos subyacente y que logra soportar múltiples modelos de datos espaciales". (Wiley, 2008)

Lo planteado por John Wiley se considera que aborda de una forma más amplia lo relacionado con la investigación. Una vez conocidos donde van a estar almacenados estos tipos de datos se hace necesario hacerle un análisis para poder realizar una buena toma de decisiones, para ello se debe conocer qué es el análisis de datos.

1.1.6 Análisis de Datos

Margaret Rouse plantea en su libro "Análisis de Datos" que: *"La ciencia que examina datos en bruto con el propósito de sacar conclusiones sobre la información, se distingue de la extracción de datos por su alcance, su propósito y su enfoque sobre el análisis, además se centra en la inferencia, o sea, el proceso de derivar una conclusión basándose solamente en lo que conoce el investigador."* (Rouse, 2012)

Por su parte Rodríguez Gómez plantea en su libro "Metodología de la Investigación cualitativa" que el análisis de datos es*: "Un conjunto de manipulaciones, transformaciones, operaciones, reflexiones y comprobaciones que realizamos sobre los datos con el fin de extraer significado relevante en relación a un problema de investigación."* (Rodríguez, 1999)

La presente investigación se guiará por el concepto de Rodríguez Gómez ya que se considera que este es el que cumple de una manera más abarcadora los elementos referentes con la investigación. Para realizar un buen análisis de datos que lleve a una efectiva toma de decisiones es necesario definir qué técnicas de análisis de datos utilizar.

1.1.7 Técnicas de análisis de datos

Las técnicas de análisis de datos son herramientas útiles para organizar, describir y analizar los datos recogidos con los instrumentos de investigación. Sirven además para obtener y analizar los datos que luego se convertirán en conocimiento útil, pues dependen de la naturaleza del objeto de estudio. El análisis de datos encierra dos procedimientos básicos:

- La organización de los datos.
- La descripción y análisis de los datos.

Según Hernández, Fernández y Baptista en su libro "Metodología de la Investigación" plantean que: *"Las técnicas de análisis de los datos pueden ser agrupadas bajo tres tipos: Observación, Entrevista y Subjetiva."* (Hernández y otros, 2003)

Por su parte Rodríguez Peñuelas, plantea en su libro "Metodología de Investigación" que las técnicas son: *"los medios empleados para recolectar información, entre las que destacan la observación, cuestionario, entrevistas, encuestas".* (Peñuelas, 2005)

Una vez analizados cado uno de los conceptos referentes a las técnicas de análisis de datos, la presente investigación se guiará según lo planteado por Hernández, Fernández y Baptista ya que se considera que abarca de manera más amplia lo relacionado con la misma.

1.1.8 Técnicas de análisis de datos espaciales

Uno de los campos que cobra más terreno en estos días es la extracción de conocimiento a partir de fuentes masivas de datos. Para ello se emplean las denominadas técnicas de MD de datos, que son algoritmos capaces de obtener relaciones entre distintos atributos o conceptos para ayudar, por ejemplo, a la toma de decisiones. La identificación de elementos similares pudiera considerarse como una de las

técnicas de análisis de datos debido a la importancia que tiene para la toma de decisiones. Se puede decir que 2 o más elementos son similares según las características que tengan y la información socioeconómica asociada a estos. El objetivo de la identificación de elementos similares es agruparlos en grupos en los cuales los elementos de cada grupo son muy similares entre sí y a la vez muy diferentes de los elementos de los otros grupos. Una de las ramas de la MD son los algoritmos de agrupamiento, los cuales precisamente se emplean para identificar elementos similares y es por ello que son los algoritmos idóneos a emplear. Existen múltiples algoritmos de agrupamientos los cuales se analizarán más adelante.

Otra de las técnicas utilizadas para el análisis de los datos espaciales es: Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Databases, KDD por sus siglas en inglés), que se refiere al proceso no-trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información. No es un proceso automático, es un proceso iterativo que exhaustivamente explora volúmenes muy grandes de datos para determinar relaciones. Es un proceso que extrae información que puede usarse para dibujar conclusiones basadas en relaciones o modelos dentro de los datos.

La siguiente figura ilustra las etapas del proceso KDD:

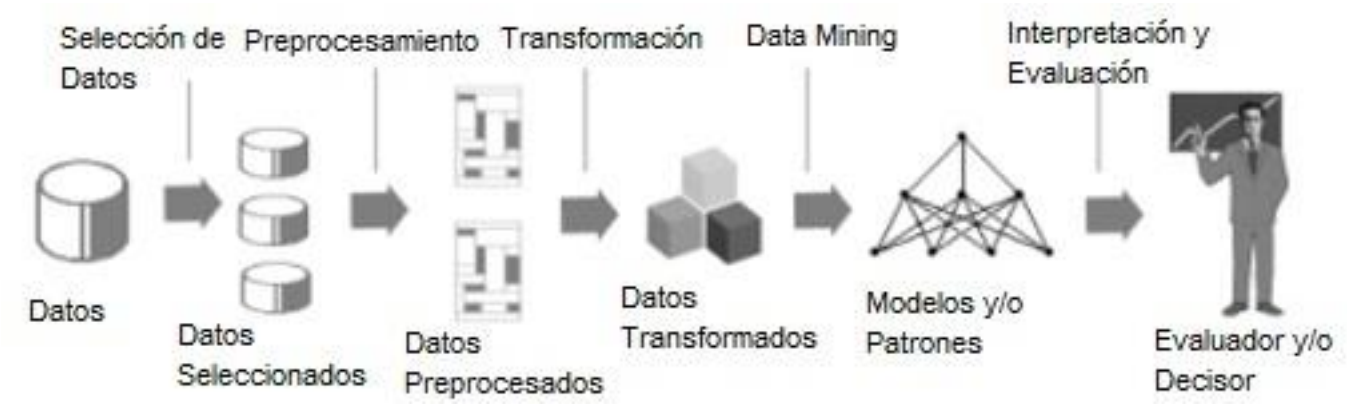

Fig. 1: Proceso de KDD (Han, 2001).

- **Selección de datos:** en esta etapa se determinan las fuentes de datos y el tipo de información a utilizar. Es la etapa donde los datos relevantes para el análisis son extraídos desde la BD.

- **Preprocesamiento:** esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos necesaria para las fases posteriores. En esta etapa se utilizan diversas estrategias para manejar datos faltantes o en blanco, datos inconsistentes o que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.

- **Transformación:** consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente.

- **Data Mining:** es la fase de modelamiento propiamente, en donde métodos inteligentes son aplicados con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles y que están contenidos u "ocultos" en los datos.

- **Interpretación y Evaluación:** se identifican los patrones obtenidos y que son realmente interesantes, basándose en algunas medidas y el analista realiza una evaluación de los resultados obtenidos.

El complemento a desarrollar, según sus características, se ve reflejado durante todas las etapas del proceso de KDD previamente mencionadas.

La estadística es otra de las técnicas para el análisis de los datos, esta es la rama de las matemáticas que se dedica al análisis e interpretación de series de datos, generando unos resultados que se utilizan básicamente en dos contextos: la toma decisiones y la proyección de situaciones futuras (García, 1993). Tradicionalmente la estadística se ha dividido en dos ramas diferentes: la estadística descriptiva y la inferencia estadística. La estadística descriptiva es aquella que recopila, analiza, estudia y describe los datos para que puedan ser interpretados cómoda y por lo tanto, pueden utilizarse eficazmente para el fin que se desee (Minotta, 2006). La inferencia estadística es una técnica mediante la cual se obtienen generalizaciones o se toman decisiones en base a una información parcial o completa obtenida mediante técnicas descriptivas (Parra, 2013).

Estas técnicas de análisis de datos son de gran importancia en la MD debido a que ayudan a descubrir información oculta en los datos. Haciendo uso de la inteligencia de negocio se puede obtener un gran volumen de información que pudiera ser de gran ayuda a las entidades en la toma de decisiones.

1.1.9 Inteligencia de Negocio Geoespacial

La inteligencia de negocio geoespacial está muy relacionada con los Sistemas de Información Geográfica (SIG), a tal punto que le agrega la variable dimensión del espacio y permite la recuperación de la información de una manera más precisa.

Parr en su libro "*Data Mining Cookbook Modeling Data for Marketing, Risk, and Customer Relationship Management*" plantea que *"La inteligencia de negocios se define como la habilidad corporativa para tomar decisiones que se logra mediante el uso de metodologías, aplicaciones y tecnologías que permiten reunir, depurar, transformar datos, y aplicar en ellos técnicas analíticas de extracción de conocimiento".* (Parr, 2000)

En tanto Juan Vicenteño, Francisco González y Carlos Álvaro plantean en su libro "Inteligencia Geoespacial" que la inteligencia de negocio geoespacial es: *"Un proceso mediante el cual los datos geoespaciales se convierten en información valiosa para la toma de decisiones dentro de una organización".* (Vicenteño y otros, 2013)

La presente investigación se guiará por el concepto de Parr ya que se considera que este es el que cumple de una manera más amplia los elementos referentes al tema de la investigación.

1.2 Técnicas de Agrupamiento

En la presente investigación se analizan diferentes técnicas de agrupamiento para ver cuál de ellas se adecua más al tema de la investigación, para así poder realizar de una manera eficiente la identificación de elementos similares.

El problema de formar grupos en un conjunto de datos es muy importante para el conocimiento del comportamiento de una población de la cual solo se tiene una cantidad "n" de sus elementos. La solución

de estos problemas se realiza mediante la creación de algoritmos de agrupamiento. Estos permiten agrupar los objetos de una base de datos en grupos (clústeres), conformados por elementos tan similares como sea posible según la cantidad de clases en las que se desee dividir los elementos. Los métodos de agrupamiento pueden dividirse en tres grupos fundamentales:

Jerárquicos: son aquellos en los que se va particionando el conjunto de datos por niveles, de modo tal que en cada nivel generalmente se unen o se dividen dos grupos del anterior, según si es un algoritmo aglomerativo o divisivo, los cuales determinan principalmente el orden de ascendencia de los elementos al cambiar de jerarquía. (Duda y otros, 2001)

Particionales: estos realizan una división inicial de los datos en grupos y luego mueven los objetos de un grupo a otro según se optimice alguna función objetivo. Para su implementación pueden utilizarse los algoritmos K-Medoid y/o K-Means principalmente haciendo uso del cálculo de la distancia euclidiana para calcular la distancia mínima entre los elementos y poder asignarlos a cada grupo. (Duda y otros, 2001)

Basados en Densidad: enfocan el problema de la división de una BD en grupos, teniendo en cuenta la distribución de densidad de los puntos, de modo tal que los grupos que se formen tienen una alta densidad de puntos en su interior mientras que entre ellos aparecen zonas de baja densidad. Tiene la ventaja de que la BD puede ser examinada en un solo paso. (Hinneburg, 1998)

Según fueron analizadas las características anteriormente descritas y el objetivo que sigue la presente investigación, esta empleará la técnica de agrupamiento Particionales, enfocándose en el uso del algoritmo K-Means. Más adelante se detallarán algunos de los algoritmos, que fueron analizados, pertenecientes a esta técnica.

1.2.1 Principales Algoritmos de Agrupamiento

En el área de la MD son muchos los algoritmos de agrupamiento que son utilizados para al trabajo con los datos espaciales y análisis de los datos. A continuación se muestra una figura a modo de mapa conceptual para evidenciar la relación que tienen estos algoritmos entre sí pudiendo observarse el porqué de la selección del algoritmo K-Means perteneciente a la técnica Particional.

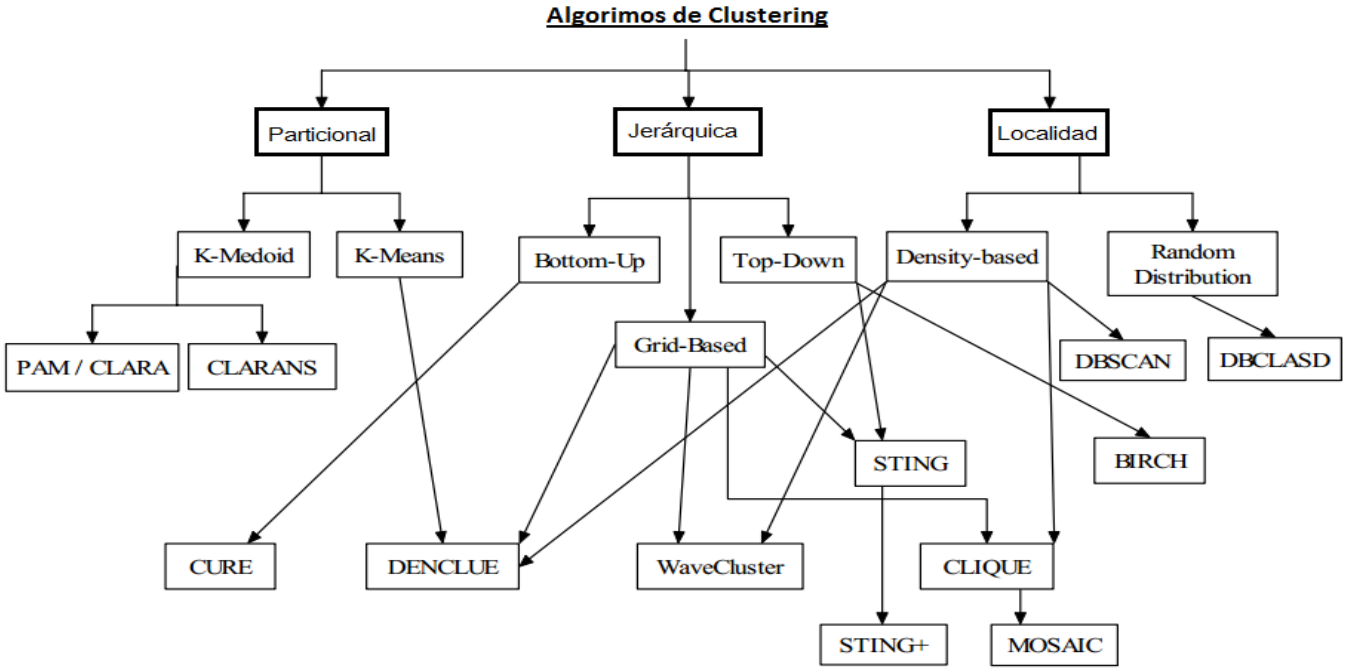

Fig. 2: Principales algoritmos de agrupamiento empleados en la Minería de Datos Espacial (SDM) (Kolatch, 2001).

PAM (Partitioning Around Medoides)

Es un algoritmo que fue desarrollado por Kaufman y Rousseeuw (1990). Este trata de determinar k particiones de n objetos. Después de una selección aleatoria inicial de k objetos representativos, el algoritmo intenta varias veces hacer una mejor elección de los representantes del clúster. Todos los posibles pares de objetos son analizados, donde un objeto en cada par se considera un objeto representativo. El conjunto de mejores objetos para cada grupo en una iteración forma los objetos representativos para la siguiente iteración. El último conjunto de objetos representativos son los respectivos Medoides de los clústeres. La complejidad de cada iteración es $O(k (n-k)^2)$ (Kaufman, 1990).

CLARA (Clustering Large Applications)

Este algoritmo es muy similar al PAM, solo con la diferencia de que la agrupación de objetos con dicho algoritmo se lleva a cabo en dos pasos. El algoritmo CLARA divide la base de datos original en muestras de tamaño s, aplicando el algoritmo PAM sobre cada una de ellas, seleccionando la mejor clasificación de las resultantes. Este algoritmo está indicado para bases de datos con gran cantidad de objetos, y su principal motivación es la de minimizar la carga computacional, en detrimento de una agrupación óptima y precisa. (Andritsos, 2002)

K-Medoid

Técnica de partición que agrupa el conjunto de datos de "n" objetos en "k" grupos conocidos a priori. El algoritmo k-modos está diseñado para datos cualitativos. A diferencia del K-Means este selecciona los puntos como centros y trabaja con una matriz arbitraria las distancias entre los puntos, además de que en el K-Means se conoce a priori cuál es el centroide (punto medio de cada clúster) y en el K-Medoid se toma aleatorio para el número de clases. Otras de las diferencias entre estos algoritmos es que (Zhexue, 1990):

- Usa una medida de la distancia distinta.
- Las medias son reemplazadas por modos.
- Los modos se actualizan con un método basado en frecuencia.

A continuación la figura muestra el proceso de dicha técnica.

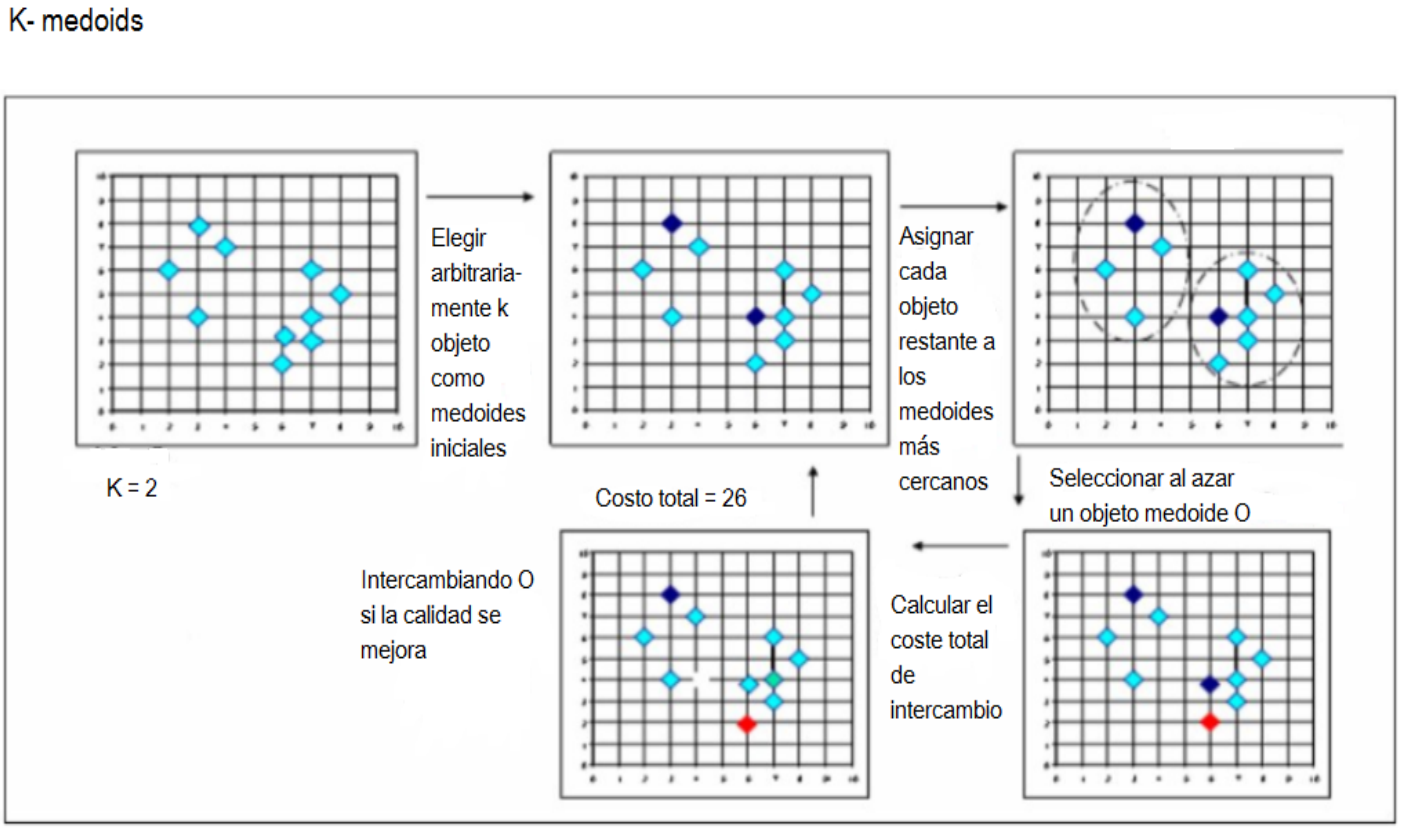

Fig. 3: Proceso de la técnica K-Medoid (Shalini S. Singh, 2011).

K-Means

Es un método de agrupamiento que tiene como objetivo la partición de un conjunto "n" de elementos en "k" grupos (clústeres) en el que cada iteración de este algoritmo los elementos cambian de grupo en dependencia de la distancia que tenga cada uno de estos elementos respecto al centroide (punto medio de cada clúster). K-Means sigue un procedimiento simple de clasificación de un conjunto de objetos en un determinado número "k" de clústeres, donde "k" es determinado a priori. Se realiza en 4 etapas: (Díez., 2003)

1. Elegir aleatoriamente "n" elementos que estarán agrupados en "k" clústeres iniciales. Para cada clúster "k", el valor inicial del centro es "x_i".

2. Reasignar los objetos del clúster según el cálculo de la distancia seleccionada respecto al centroide de cada clúster.

3. Una vez que todos los objetos son colocados, recalcular los centros de “k” clúster.

4. Repetir las etapas 2 y 3 hasta que no se hagan más reasignaciones. Aunque el algoritmo termina siempre, no se garantiza obtener la solución óptima. En efecto, el algoritmo es muy sensible a la selección aleatoria de los “k” centros iniciales. Esta es la razón por la que se utiliza el algoritmo de K-Means numerosas veces sobre un mismo conjunto de datos para intentar minimizar este efecto, sabiendo que a centros iniciales lo más espaciados posibles dan mejores resultados.

A continuación se describen algunas de las características de este método:

- Agrupar los datos en “k” grupos (clústeres) siendo “k” parámetro a priori del método.
- Cada grupo está asociado a su centroide (media de los puntos del grupo).
- Cada punto (objeto) se asigna al grupo más cercano.
- El criterio de agrupamiento es minimizar la suma de las distancias al cuadrado de todos los puntos al centro de su clúster.

La figura que se muestra a continuación refleja el proceso de dicha técnica.

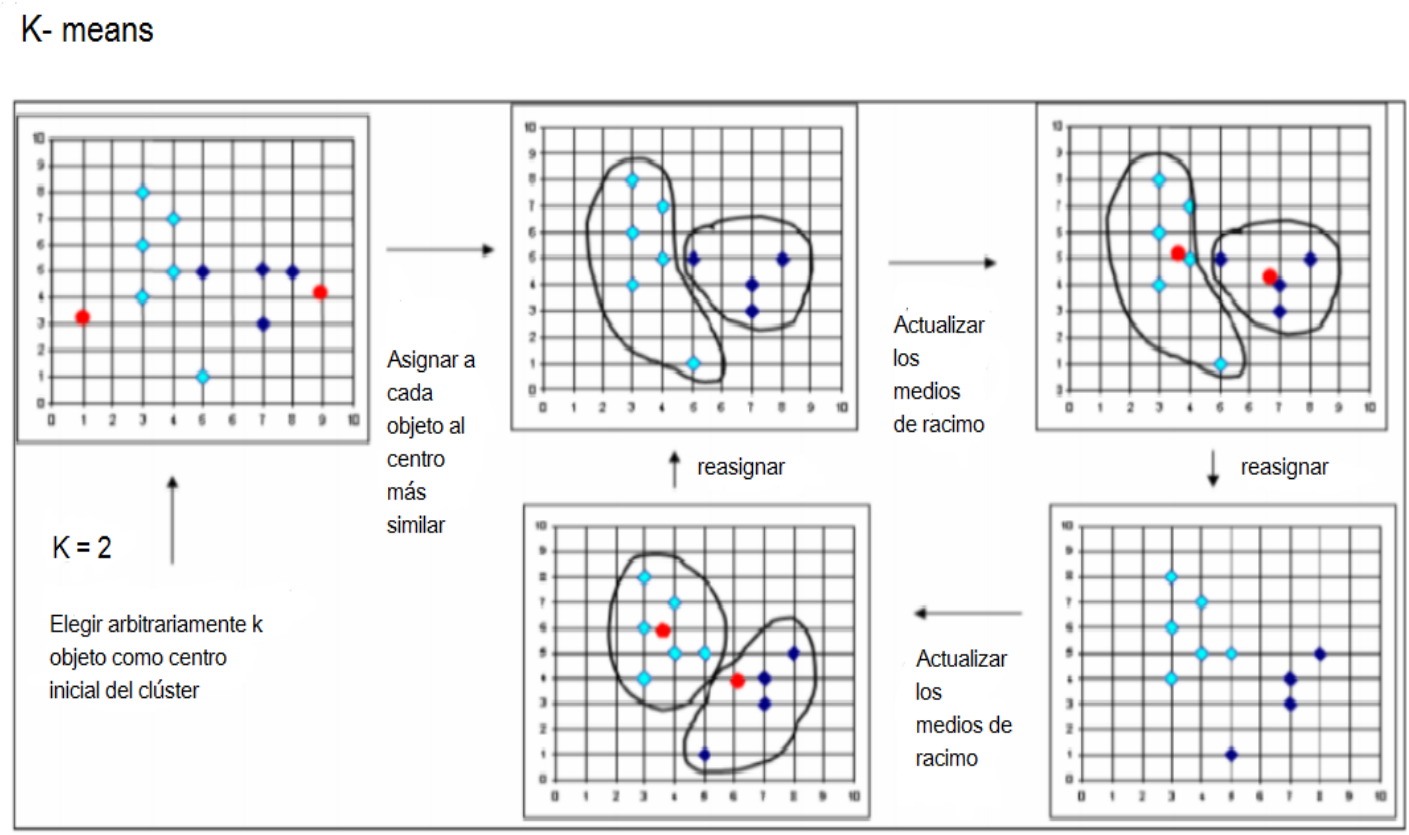

Fig. 4: Proceso de la técnica K-Means (Shalini S. Singh, 2011).

A continuación se muestra una tabla comparativa de las características de los algoritmos K-Means o K-Medoid.

Tabla 1: Comparación entre K-Means y K-Medoid. (Shalini S. Singh, 2011)

Diferentes ajustes	*K-Means*	*K-Medoid*
Complejidad	*O (i k n)*	*O (i k(n-k)^2)*
Eficiencia	*Comparativamente más*	*Comparativamente menos*
Implementación	*Fácil*	*Complicada*
Sensible a los valores atípicos	*Si*	*No*

Luego de un análisis de las principales características de estos algoritmos de agrupamiento se llega a la conclusión que el más recomendado a utilizar en la presente investigación es el K-Means ya que según sus características es el que más se ajusta a emplear en el desarrollo del complemento, su complejidad temporal es menor y además brinda como resultado final una información más exacta. Otra de las ventajas por las que se decidió utilizar este algoritmo es que para la realización de la identificación de los elementos similares en el sistema se hará uso de una biblioteca externa: Weka, la cual tiene implementado, dentro de todos los algoritmos propios del análisis geoespacial el K-Means.

Para determinar la similitud de los elementos en cada clúster, según los algoritmos vistos con antelación, se emplean diferentes algoritmos matemáticos. Debido a que se escogió el método K-Means, para conocer la distancia entre los elementos se utiliza el cálculo de la Distancia Euclidiana. A continuación se muestra el funcionamiento de dicho cálculo.

Distancia Euclidiana (Ritzman, 2000):

Es la distancia ordinaria entre dos puntos, se calcula a partir del teorema de Pitágoras. Dado dos puntos A y B medidos según las variables X y Y la distancia euclidiana seria:

$$d_{A-B} = \sqrt{(A_x - B_x)^2 + (A_y - B_y)^2}$$

Si se tienen 2 puntos A y B, están medidas con un número n de dimensiones y no solo X;Y, el cálculo de la Distancia Euclidiana sería:

$$d_{A-B} = \sqrt{\sum_{1}^{n} (A_n - B_n)^2}$$

1.3 Análisis de herramientas con soporte para Minería de Datos geoespaciales y posibles soluciones existentes

Con el fin de dar solución a la problemática planteada se hace necesario realizar un estudio de las principales herramienta existentes que dan soporte a la MD o que tengan como base dichos procesos para obtener resultados específicos. A través de este estudio se espera encontrar características y experiencias que apoyen el desarrollo de la solución deseada.

1.3.1 SD-Miner

Algunas de las características de esta herramienta se describen a continuación: (Agudelo, 2011)

- Soporte para técnicas de MD espaciales tales como agrupamiento, clasificación espacial, caracterización espacial y espacio-temporal y reglas de asociación espacial.
- Implementa sus algoritmos en una librería que permite que sean utilizados por otros sistemas.
- Se divide en tres módulos: Interfaz Gráfica de Usuario, el módulo SD-Miner y el módulo de administración de bases de datos.
- Capacidad de detectar la naturaleza de los datos.

Esta herramienta posibilita la realización de MD, no permitiendo la identificación de elementos similares.

1.3.2 RapidMiner

Algunas de las características de esta herramienta se describen a continuación: (Mierswa y otros, 2006)

- Desarrollado en Java.

- Multiplataforma.
- Puede usarse de diversas maneras:
    - A través de un GUI (Graphical User Interface por sus siglas en inglés).
    - En línea de comandos.
    - Desde otros programas a través de llamadas a sus bibliotecas.
- Extensible.

También posibilita la realización de MD, pero no permite la identificación de elementos similares.

1.3.3 MatLab

MATLAB es un lenguaje de alto desempeño diseñado para realizar cálculos técnicos. Integra el cálculo, la visualización y la programación en un ambiente fácil de utilizar donde los problemas y las soluciones se expresan en una notación matemática. MATLAB es un sistema interactivo cuyo elemento básico de datos es el arreglo que no requiere de dimensionamiento previo. Esto permite resolver muchos problemas computacionales, específicamente aquellos que involucren vectores y matrices. (Elizondo, 2002)

MatLab es usado principalmente para:

- Cálculos numéricos.
- Análisis de datos, exploración y visualización.
- Graficar datos con fines científicos o de ingeniería.
- Desarrollo de aplicaciones que requieran de una GUI

Esta es otra de las herramientas analizadas que, aun cuando no posibilita la identificación de elementos similares, es de gran utilidad para la MD.

1.3.4 Geographical Data Mining Analyst (GeoDMA)

- Soporta el uso de datos espaciales para la comparación de imágenes y regiones obtenidas en los procesos de segmentación y análisis de imágenes.
- Utiliza árboles de decisión y algoritmos para mapas auto-organizados.
- Su implementación es en C++ e interfaz en QT.

1.4 Herramientas y metodologías a utilizar en el desarrollo del complemento

Antes de conocer cuáles serán las herramientas, tecnologías y la metodología a utilizar en el desarrollo del complemento es necesario conocer algunas características de la plataforma GeneSIG, para comprender el porqué de la selección de estas herramientas. “La Plataforma GeneSIG es un producto encaminado a realizar la representación y análisis geoespacial de información geográfica y su estructura arquitectónica permite personalizar sus funcionalidades a cualquier negocio que lo requiera a través de la reutilización de sus complementos. Puede ser considerado como un Sistema de Información Geográfica Web único y extensible, basado en estándares OpenGIS que incluye funcionalidades operativas de las aplicaciones de esta tecnología”. (Pantoja, 2010)

1.4.1 Herramientas y Tecnologías predefinidas según GeneSIG

A continuación se describirán cada una de las herramientas y tecnologías que se emplearán en el desarrollo del complemento, seleccionadas debido a que el complemento que se desarrollará será integrado a la plataforma GeneSIG. Pudiendo tener acceso a estas en el Manual de GeneSIG.

PostgreSQL 8.4

A través de los Sistemas Gestores de Base de Datos (SGBD) se puede almacenar y posteriormente acceder a los datos. PostgreSQL es el SGBD utilizado en la presente investigación, el cual tiene integrado el módulo PostGis para el trabajo con los datos espaciales. Esta herramienta será utilizada en el desarrollo del complemento para obtener los datos de las capas a las que se le realizará la tematización y su información asociada.

PostGis 1.5.1

“Es una extensión de base de datos espaciales de libre disposición, puede ayudar a responder preguntas que no podía responder mediante una simple base de datos relacional. Su conjunto de características iguala o supera las alternativas propietarias, lo que le permite crear consultas basadas en la localización y características con sólo unas pocas líneas de código SQL”. (Obe, 2011)

Algunas de las funciones espaciales con que cuenta el módulo PostGis se describen a continuación:

- Recuperación: obtienen propiedades y medidas de las geometrías.

- Comparación: comparan dos geometrías y obtienen información sobre su relación espacial.
- Generación: generan geometrías a partir de otros tipos de datos.

Visual Paradigm for UML 8.0

"Visual Paradigm es una herramienta multiplataforma que ayuda a generar diagramas para crear aplicaciones con calidad. Soporta múltiples usuarios trabajando en un mismo proyecto, además de facilitar el modelado de BD y el proceso de negocio" (Marañón, 2010). Esta es utilizada en la presente investigación ya que es una herramienta que permite realizar el modelado de todos los diagramas necesarios para la elaboración del complemento con una alta calidad, posibilita la generación de código a partir de los diagramas. La misma soporta el modelado de negocio, captura de requisitos, además de permitir el control de versiones y ser multiplataforma. Todas estas características conllevaron a que Visual Paradigm fuese seleccionada como herramienta CASE.

Entre sus principales características se destacan:

- Modelar procesos de negocio.
- Administrar requerimientos.
- Modelar visualmente el diseño lógico y físico de datos.

La misma tiene gran importancia en el desarrollo de la investigación porque aporta los modelos, la forma de administración y los requerimientos en el negocio, por lo que el autor asume el criterio del autor Marañón.

Lenguaje Unificado de Modelado (UML) 2.0

El Lenguaje Unificado de Modelado (Unified Modeling Language por sus siglas en inglés) es un lenguaje gráfico para visualizar, especificar, construir y documentar los artefactos de un sistema con gran cantidad de software. UML proporciona una forma estándar de escribir los planos de un sistema, cubriendo tanto las cosas conceptuales, tales como procesos del negocio y funciones del sistema, como las cosas concretas, tales como las clases escritas en un lenguaje de programación específico, esquemas de bases de datos y complementos software reutilizables (Grady y otros, 2000).

CartoWeb 3.0

CartoWeb es un software de publicación WebGIS construida en PHP sobre UMN MapServer que explota AJAX. Su característica más diferenciadora respecto a otros proyectos de clientes web ligeros sobre MapServer, es que ofrece un framework que ha sido diseñado con una arquitectura bastante modular y escalable, permitiendo separar la lógica de un servidor (cartoserver) encargado del diálogo con el servidor de mapas y la provisión de servicios a un cliente (cartoclient), cuya misión es acceder mediante SOAP a los servicios proporcionados por servidores cartoweb y renderizar de la manera apropiada la información hacia el cliente final (HTML, PDF, etc). Esto permite separar lógica y físicamente los clientes (cartoclient) de los servidores con múltiples configuraciones (N clientes - M servidores) y posibilidades de escalado. Funcionalmente presenta un abanico muy completo de características propias de un geoportal, con la posibilidad de ir añadiendo o desarrollando nuevos complementos, siendo esta otra de las fortalezas del sistema.

JavaScript

"JavaScript es un lenguaje de programación interpretado, por lo que no es necesario compilar los programas para ejecutarlos, o sea, los programas escritos con JavaScript se pueden probar directamente en cualquier navegador sin necesidad de procesos intermedios. Son utilizados principalmente para crear páginas web dinámicas" (Pérez, 2008). Este es utilizado en la presente investigación ya que el complemento que se desarrollará será integrado a la plataforma GeneSIG y esta hace uso de este lenguaje de programación del lado del cliente, además por sus ventajas para el desarrollo del complemento.

Algunas de las ventajas de este lenguaje de programación son las siguientes:

- Reducción de la carga del servidor.
- Gran usabilidad.
- Basado en programación orientado a objetos aunque no incorpora la creación de clases ni la herencia.

ExtJs 3.0

Es una biblioteca JavaScript que permite construir aplicaciones complejas además de flexibilizar el manejo de complementos de la página web como el DOM (Document Object Model por sus siglas en inglés), AJAX (Asynchronous JavaScript And XML por sus siglas en inglés) y DHTML (Dynamic HTML (HyperText Markup Language) por sus siglas en inglés). Brinda múltiples posibilidades para el trabajo con las validaciones y manejo de errores en el cliente. La personalización de temas de estilos es posible en su utilización y provee el trabajo con una amplia configuración e intenso trabajo con las hojas de estilo CSS (cascading style sheets por sus siglas en inglés). Tiene la gran ventaja de crear interfaces de usuario bastante funcionales.

Esta biblioteca incluye (Sánchez, 2008):

- Complementos de Interfaz de usuario (UI) de alto rendimiento y personalizables.
- Modelo de complementos extensibles.

Procesador de Hipertexto 5 (PHP 5)

“PHP es uno de los lenguajes más usados hoy en día para el desarrollo de páginas web dinámicas. Es de código abierto, adecuado para desarrollo web y que puede ser incrustado en HTML, o sea, que en un mismo archivo se podrá combinar código PHP con código HTML. Una de sus características más potentes es su soporte para gran cantidad de bases de datos. Entre su soporte pueden mencionarse MySQL, Oracle, PostgreSQL, entre otras” (Henst, 2001).

Algunas de sus ventajas son:

- Es un lenguaje multiplataforma.
- Completamente orientado al desarrollo de aplicaciones web dinámicas con acceso a información almacenada en una BD.
- Capacidad de conexión con la mayoría de los motores de BD que se utilizan en la actualidad.

Servidor de mapas MapServer 5.6.5

MapServer es un servidor de mapas, que fue concebido en un ambiente de desarrollo de código abierto para permitir construir aplicaciones que trabajen con información geoespacial, y que además tenga la cualidad de ser accesible a través de Internet (Antonio y otros, 2012). Según el sitio oficial MapServer, esta es una plataforma de código abierto para la publicación de los datos espaciales y aplicaciones de mapas interactivos para la web. Dentro de sus principales características se encuentran: generación automática de leyendas, utilización de datos en forma de mosaico, salida avanzada de cartografía, soporta scripts desarrollados en PHP, Java, C#, entre otros, soporta múltiples formatos raster y vectorial, se puede tener acceso a las características de los datos, puede funcionar sobre la mayor parte de versiones de Sistemas operativos: UNIX/LINUX, Microsoft Windows XP/NT/98/95 y hasta MacOS (MapServer, 2015).

Servidor de aplicaciones Apache 2.5.6

El Servidor Apache es uno de los más utilizados hoy en día en la red, es el software que gestiona los sitios web en Internet. Atiende las peticiones de los navegadores sirviendo las páginas y recursos que solicitan, restringe accesos, gestiona los errores, siendo un elemento crítico para el funcionamiento de la Web. Este servidor es utilizado principalmente para cargar y desplegar la aplicación que se esté desarrollando, además de que provee un alto grado de calidad y fortaleza para las implementaciones que utilizan el protocolo HTTP.

Entre sus principales características se destacan (Díaz, 2005):

- Multiplataforma.
- Es un servidor de web conforme al protocolo HTTP/1.1
- Extensible: gracias a ser modular se han desarrollado diversas extensiones entre las que destaca PHP, un lenguaje de programación del lado del servidor.

1.4.2 Herramientas propias del complemento

Para la realización de la propuesta de solución en la presente investigación se hará uso de otras herramientas que no forman parte de la plataforma GeneSIG, estas son Java y la herramienta Weka. Esta

última hará uso de la máquina virtual de Java para poder ejecutarse y poder darle solución al problema planteado en el presente trabajo.

Weka

Esta herramienta permite, a partir de ficheros de texto con el formato **.arff** (ver Fig. 5), utilizar distintos tipos de técnicas para extraer información. Es un paquete que contiene una colección de herramientas de visualización y algoritmos para análisis de datos y modelado predictivo, unidos a una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades. Esta se utiliza en diferentes áreas, en particular con finalidades docentes y de investigación. Algunas de las características de esta herramienta se definen a continuación (Witten, 1999):

- Está disponible libremente bajo la licencia pública general de GNU.
- Contiene una extensa colección de técnicas para preprocesamiento de datos y modelado.
- Soporta varias tareas estándares de MD, especialmente, preprocesamiento de datos, agrupamiento, clasificación, regresión, visualización, y selección.

La herramienta Weka será utilizada en el desarrollo del complemento como una biblioteca en el servidor ya que contiene los algoritmos requeridos para el análisis de datos, principalmente el K-Means, que es el que se va a utilizar para el desarrollo del complemento. Esta se ejecutará en el servidor pero como un servicio aparte al servicio web y al servicio de BD, o sea, se va a ejecutar en la máquina virtual de java y solo en el momento en que sea utilizado dicho complemento por lo que disminuiría la carga en los servidores web y de BD a la hora del procesamiento de los datos. A continuación se describe el proceso de cómo sería el funcionamiento de dicha herramienta para el funcionamiento del complemento en la plataforma GeneSIG:

Una vez que el cliente haya seleccionado la capa y los atributos de esta:

1. Mediante la función **fopen** de PHP, se crea el archivo identES**.arff**
2. El Weka lee este archivo y realiza el análisis haciendo uso del algoritmo K-Means.

3. Se crean los archivos **identESRandomize.arff** e **identESClusterizado.arff** siendo este último el que tiene el clúster al que pertenece cada elemento.

4. Finalmente es capturado el archivo clusterizado por PHP, mostrándose al final la tematización en el mapa.

```
@relation 'identES'

@attribute id real
@attribute latitud real
@attribute longitud real
@attribute a1 real
@attribute a2 real
@attribute a3 real

@data
5881,-73.1509170532227,19.5046291351318,0.8,0.9,1.1
5882,179.059997558594,17.576000213623,0.7,1,1.1
5883,-75.4236297607422,19.8858509063721,1.8,2,1.6
5884,-75.2725296020508,19.9730491638184,1.2,1.8,0.9
5885,-77.6104431152344,18.9706592559814,1,1,0.9
5886,-84.6658630371094,18.3964290618896,0.6,0.4,0.3
5887,-73.8149032592773,19.3126392364502,0.2,0.2,0.5
5888,-76.2477035522461,19.8377799987793,2.2,2.3,1.6
5889,-72.9610977172852,8.50965595245361,0.9,0.4,1
5890,-72.9829788208008,20.0667991638184,0.3,0.3,0.6
5891,-75.0912933349609,19.8223991394043,2.4,2.9,2
5892,-77.8362426757813,18.1136493682861,0.8,1.2,0.8
5893,-74.9535675048828,20.675609588623,0.8,0.6,0.6
5894,-75.5888519287109,19.8265495300293,2.5,2.5,2.5
5895,-75.7083129882813,20.6790294647217,1,0.6,0.6
5896,-76.0389633178711,19.7805995941162,2.3,2.9,2.2
5897,-73.9990463256836,20.1798496246338,1.1,1.1,1.1
5898,-75.5743637084961,19.7380294799805,1.6,1.6,2.2
5899,-75.474479675293,19.9176902770996,2.1,2,2.6
```

Fig. 5: Muestra del fichero identES.arff generado.

```
@relation 'identESRandomize'

@attribute id numeric
@attribute latitud numeric
@attribute longitud numeric
@attribute a1 numeric
@attribute a2 numeric
@attribute a3 numeric

@data

8407,-76.156052,20.01417,2.9,2.3,3.2
11469,-75.341507,19.914009,2.1,1.8,1.6
9344,-73.252098,20.368,1,0.9,1
8038,-75.680939,20.02491,2.6,3.2,2.3
11530,-75.571274,19.667839,1.7,1.1,2
7727,-75.534813,19.684191,2.9,3.5,2.7
11599,-75.471039,19.774639,2.4,2.2,2.2
7119,-73.58905,20.40686,0.8,0.8,1.2
6555,-75.287514,19.79904,2.1,2.6,2.2
6213,-77.291043,19.577118,1.2,0.9,0.8
9533,-80.33844,22.015289,0.2,0.1,0.2
10709,-76.085251,19.93033,1.8,2.3,1.4
6120,-75.378304,19.78791,2.6,2.3,2.7
11200,-75.648262,19.647209,1,0.9,0.9
9320,-69.289871,19.685711,0.6,0.1,0.4
```

Fig. 6: Muestra del fichero identESRandomize.arff generado.

```
@relation 'identESClusterizado'

@attribute id numeric
@attribute latitud numeric
@attribute longitud numeric
@attribute a1 numeric
@attribute a2 numeric
@attribute a3 numeric
@attribute cluster {cluster1,cluster2,cluster3}

@data

8407,-76.156052,20.01417,2.9,2.3,3.2,cluster1
11469,-75.341507,19.914009,2.1,1.8,1.6,cluster3
9344,-73.252098,20.368,1,0.9,1,cluster2
8038,-75.680939,20.02491,2.6,3.2,2.3,cluster1
11530,-75.571274,19.667839,1.7,1.1,2,cluster3
7727,-75.534813,19.684191,2.9,3.5,2.7,cluster1
11599,-75.471039,19.774639,2.4,2.2,2.2,cluster1
7119,-73.58905,20.40686,0.8,0.8,1.2,cluster2
6555,-75.287514,19.79904,2.1,2.6,2.2,cluster1
6213,-77.291043,19.577118,1.2,0.9,0.8,cluster2
9533,-80.33844,22.015289,0.2,0.1,0.2,cluster2
10709,-76.085251,19.93033,1.8,2.3,1.4,cluster3
6120,-75.378304,19.78791,2.6,2.3,2.7,cluster1
11200,-75.648262,19.647209,1,0.9,0.9,cluster2
9320,-69.289871,19.685711,0.6,0.1,0.4,cluster2
```

Fig. 7: Muestra del fichero identESClusterizado.arff generado.

1.4.3 Metodologías a utilizar

La MD es una disciplina que ha crecido enormemente en los últimos años. Las organizaciones han comprendido que los grandes volúmenes de datos que residen en sus sistemas pueden ser analizados y explotados para obtener nuevo conocimiento a partir de los mismos. MD, es el proceso de extraer conocimiento útil, comprensible y novedoso de grandes volúmenes de datos, para ello se hace necesario contar con una guía que oriente la investigación por un buen camino, estas guías tienen una serie de pasos a cumplir comúnmente llamadas metodologías. Las metodologías permiten llevar a cabo el proceso de MD de forma sistemática. Ayudan a las organizaciones a entender el proceso de descubrimiento de

conocimiento y proveen una guía para la planificación y ejecución de los proyectos. Las principales metodologías de la MD son SEMMA y CRISP-DM (Moine y otros, 2011).

"La metodología es la ciencia que enseña a dirigir determinados procesos de manera eficiente y eficaz para alcanzar los resultados deseados y tiene como objetivo dar la estrategia a seguir en el proceso" (Zayas, 1995). Las metodologías de desarrollo de software conforman un enfoque para estructurar una aplicación con el objetivo de lograr productos de alta calidad. Están formadas por un conjunto de procedimientos, reglas, técnicas y herramientas que guían a los desarrolladores en el proceso de cómo crear un software. Las herramientas son importantes en el desarrollo de un software, una buena selección y utilización de las mismas permite desarrollar las aplicaciones en tiempo y con una alta calidad, cumpliendo siempre con las necesidades del cliente. Para el desarrollo del complemento propuesto en la presente investigación se hará uso, principalmente, de las herramientas utilizadas por la plataforma GeneSIG para garantizar una integración factible.

CRISP-DM (Cross-Industry Standard Process for Data Mining)

CRISP–DM, creada por el grupo de empresas NCR1 y Daimer Chrysler2 en el año 2000, es actualmente la guía de referencia más utilizada en el desarrollo de proyectos de MD. Estructura el proceso en seis fases de forma cíclica: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implantación (Chapman, 2000). Cada fase es descompuesta en varias tareas generales de segundo nivel. Las tareas generales se proyectan a tareas específicas, pero en ningún momento se propone como realizarlas. Es decir, CRISP-DM establece un conjunto de tareas y actividades para cada fase del proyecto pero no especifica cómo llevarlas a cabo. A continuación se muestra la relación de cada una de estas fases.

[1] La corporación NCR es una TIC especializada en soluciones para la venta al por menor y la industria financiera. La oficina central se encuentra ubicada en Estados Unidos. Fue fundada por John Henry Patterson en 1884.
[2] Daimler AG es una importante empresa alemana, principalmente dedicada a la industria del automóvil. Fundada en 1998 por Gottlieb Daimler, Carl Benz.

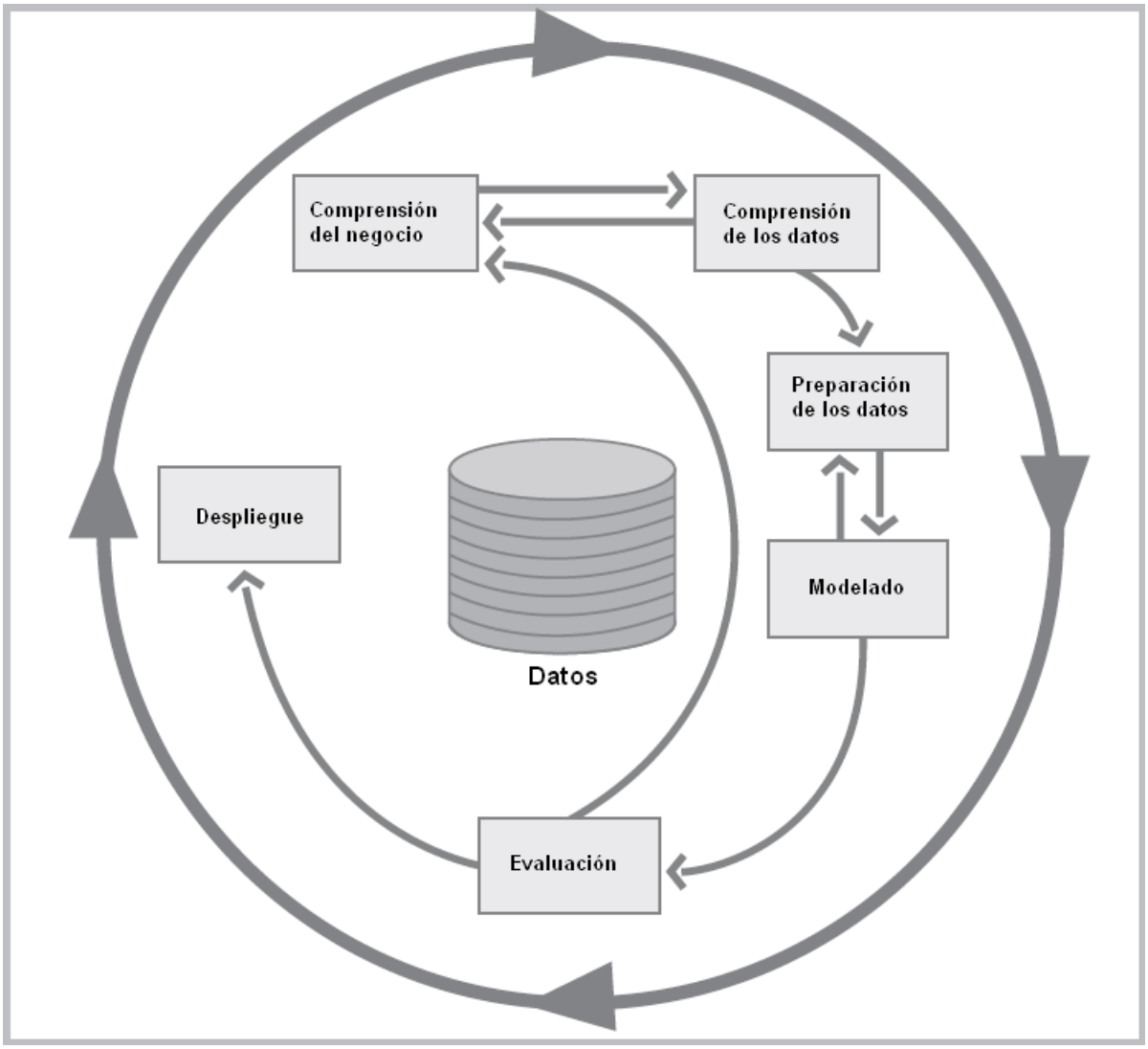

Fig. 8: Fases de la metodología CRISP-DM (IBM, 2012).

Comprensión del negocio: esta fase inicial se centra en la comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial, y luego convertir este conocimiento en una definición del problema de minería de datos, y un plan preliminar diseñado para alcanzar los objetivos.

Comprensión de los datos: una vez establecidas las bases de proyecto es necesario comprender los datos con los que se cuenta. En esta fase se realiza la recogida, exploración, descripción y validación de los datos.

Preparación de los datos: esta fase cubre todas las actividades para construir el conjunto de datos. Estas tareas son ejecutadas en múltiples oportunidades y sin orden. Las tareas incluyen selección y transformación de tablas, registros y atributos y limpieza de datos para su posterior modelado.

Modelado: en esta fase se seleccionan y aplican varias técnicas de modelado y se calibran los parámetros para obtener óptimos resultados.

Evaluación: en esta fase se realiza la evaluación de los resultados según los objetivos del negocio establecidos en la primera fase. Se realiza una revisión del proceso y luego se establecen los pasos y acciones que se realizarán a continuación. En esta fase se pueden encontrar nuevas necesidades que obliguen al proceso a volver a alguna de las fases anteriores.

Despliegue: en esta fase se realiza el despliegue, generación del informe final y la revisión del proyecto.

PRODESOFT 1.5

La metodología de desarrollo que guiará todo el proceso de desarrollo de software en la presente investigación será Proceso de Desarrollo de Software (PRODESOFT), diseñado por la UCID (Unidad de Compatibilización e Integración de Software para la Defensa) actual XETID (Empresa de Tecnologías de la Información para la Defensa). Su modelo de desarrollo de software describe la secuencia de actividades de alto nivel para la construcción y desarrollo de soluciones con una combinación entre los modelos basado en Componente, el Iterativo y el Incremental.

El desarrollo de un producto de software va unido a un ciclo de vida compuesto por una serie de fases que comprenden todas las actividades, desde el momento en que surge la idea de crear un nuevo producto de software, hasta aquel en que el producto deja definitivamente de ser utilizado por el último de sus usuarios. El ciclo de vida proporciona el marco de referencia básico para dirigir el proyecto, independientemente del trabajo específico involucrado en cada fase. Este se descompone en el tiempo en cinco fases secuenciales (Figura 5) que son: Inicio, Modelación, Construcción, Explotación Experimental, Despliegue. En las primeras fases las iteraciones son realizadas con mayor énfasis en la determinación del alcance del proyecto, planificación, identificación y descripción de requisitos y la creación de la línea base de la arquitectura, luego a medida que va avanzando el proyecto el énfasis de las iteraciones cambia

centrándose más en el análisis, diseño, implementación y pruebas del complemento (SOFTWARE-DEFENSA, 2012).

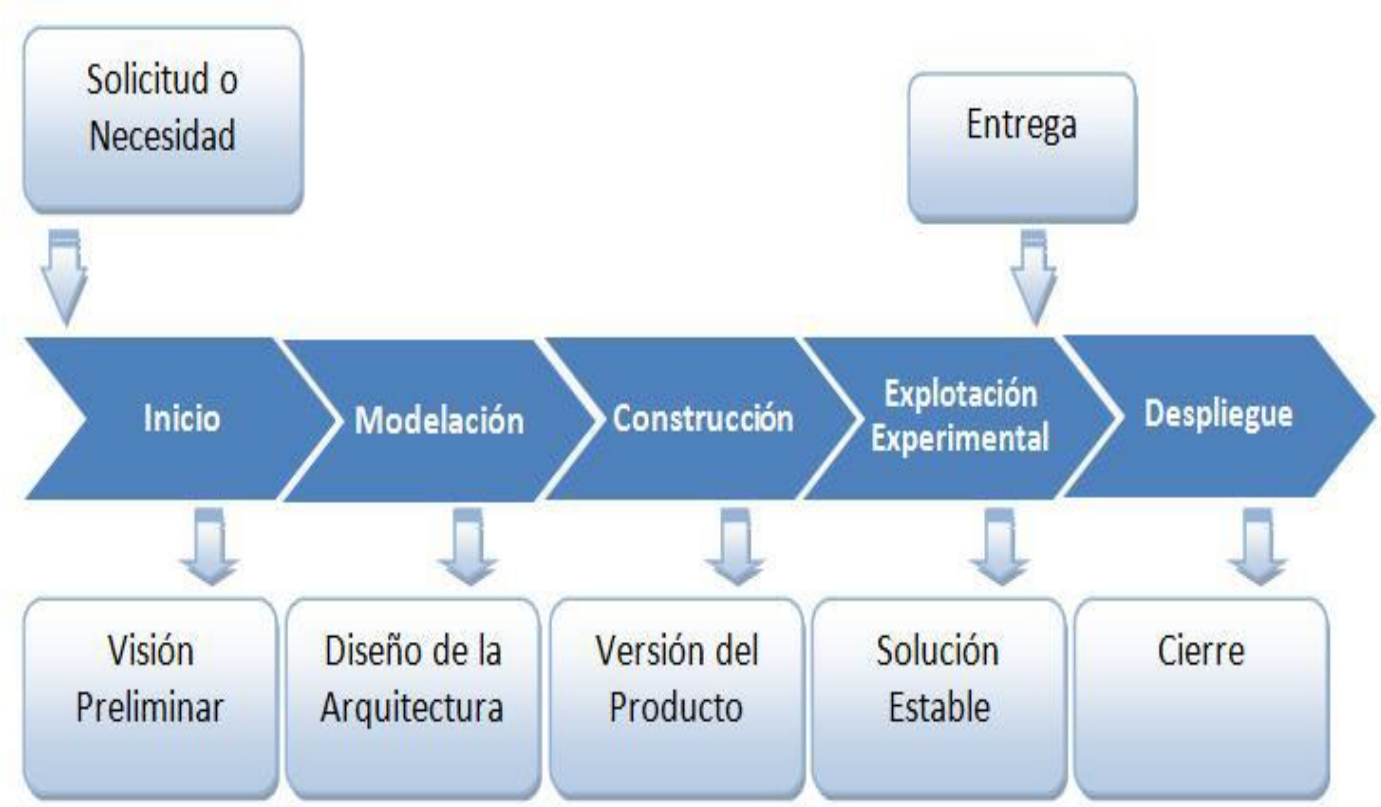

Fig. 9: Etapas del ciclo de vida de PRODESOFT (Lissa Curvelo Oliva y otros, 2012)

1.4.4 Integración de las metodologías

Luego de analizadas las principales características de cada una de las metodologías anteriormente planteadas se procede a la integración de las mismas para seguir una misma estrategia de desarrollo de software. Es por ello que se define a continuación cómo será el proceso de desarrollo con ambas metodologías.

- Se integrarán las fases de “Comprensión del Negocio” y “Comprensión de los Datos” de la metodología de MD CRISP-DM con la etapa de “Inicio” de la metodología PRODESOFT.

- Se integrarán las fases de “Preparación de los datos” y “Modelado” de la metodología de MD CRISP-DM con la etapa de “Modelación” de la metodología PRODESOFT.
- Se integrará la etapa de “Evaluación” de la metodología de MD CRISP-DM con las etapas de “Construcción” y “Explotación Experimental” de PROFESOFT.
- Se integrará la etapa de Despliegue de la metodología de MD CRISP-DM con la etapa de Despliegue de la metodología PRODESOFT.

Una vez integradas ambas metodologías la propuesta quedaría de la siguiente manera:

1. Fase de Inicio para la comprensión del negocio y los datos.
2. Fase de Preparación y Modelación de los datos.
3. Fase de Construcción y Evaluación del sistema.
4. Fase de Despliegue.

1.5 Conclusiones del capítulo 1

El análisis realizado permitió seleccionar las herramientas necesarias para el desarrollo de la investigación. El uso de las metodologías de MD y las metodologías de desarrollo de software como una misma línea de desarrollo permitió el desarrollo de la aplicación y su correcto funcionamiento tanto para la realización de la MD como para el proceso de identificación de elementos similares. La caracterización de principales técnicas y algoritmos de identificación de elementos similares, permitieron la selección del algoritmo K-Means para darle solución al problema planteado, así como el uso de una serie de herramientas propias de la plataforma como por ejemplo: PostgreSQL 8.4, PostGis 1.5.1, CartoWeb 3.0, ExtJs 3.0, Visual Paradigm for UML 8.0, Lenguaje Unificado de Modelado (UML) 2.0, PHP 5, MapServer 5.6.5, Apache 2.5.6 y como herramientas propias para el desarrollo del complemento, la herramienta Weka.

CAPÍTULO 2: PRESENTACIÓN DE LA SOLUCIÓN PROPUESTA.

Introducción:

En el presente capítulo se describen las características propuestas para la elaboración del complemento. Se definen los requisitos tanto funcionales como los no funcionales del mismo, de manera que resuman todo el proceso de análisis de la investigación. También se generan los artefactos del diseño que sustentarán la implementación.

2.1 Modelo de Negocio

La modelación de procesos de negocio permite realizar una exploración del dominio del problema, con el fin de lograr comprensión por parte del equipo de desarrollo de los procesos que se realizan actualmente en la entidad y la relación que existe entre estos. De esta forma se van determinando necesidades operacionales, así como restricciones que presenta la entidad, obteniéndose finalmente un entendimiento del negocio para dar paso a la fase inicial del sistema (SOFTWARE-DEFENSA, 2012).

2.2 Modelo de Dominio.

El modelo del dominio muestra clases conceptuales significativas en un dominio del problema, es el artefacto más importante que se crea durante el análisis orientado a objetos. Un modelo del dominio es una representación de las clases conceptuales del mundo real, no de complementos software. No se trata de un conjunto de diagramas que describen clases software, u objetos software con responsabilidades.

Diagrama de clases del Modelo de Dominio

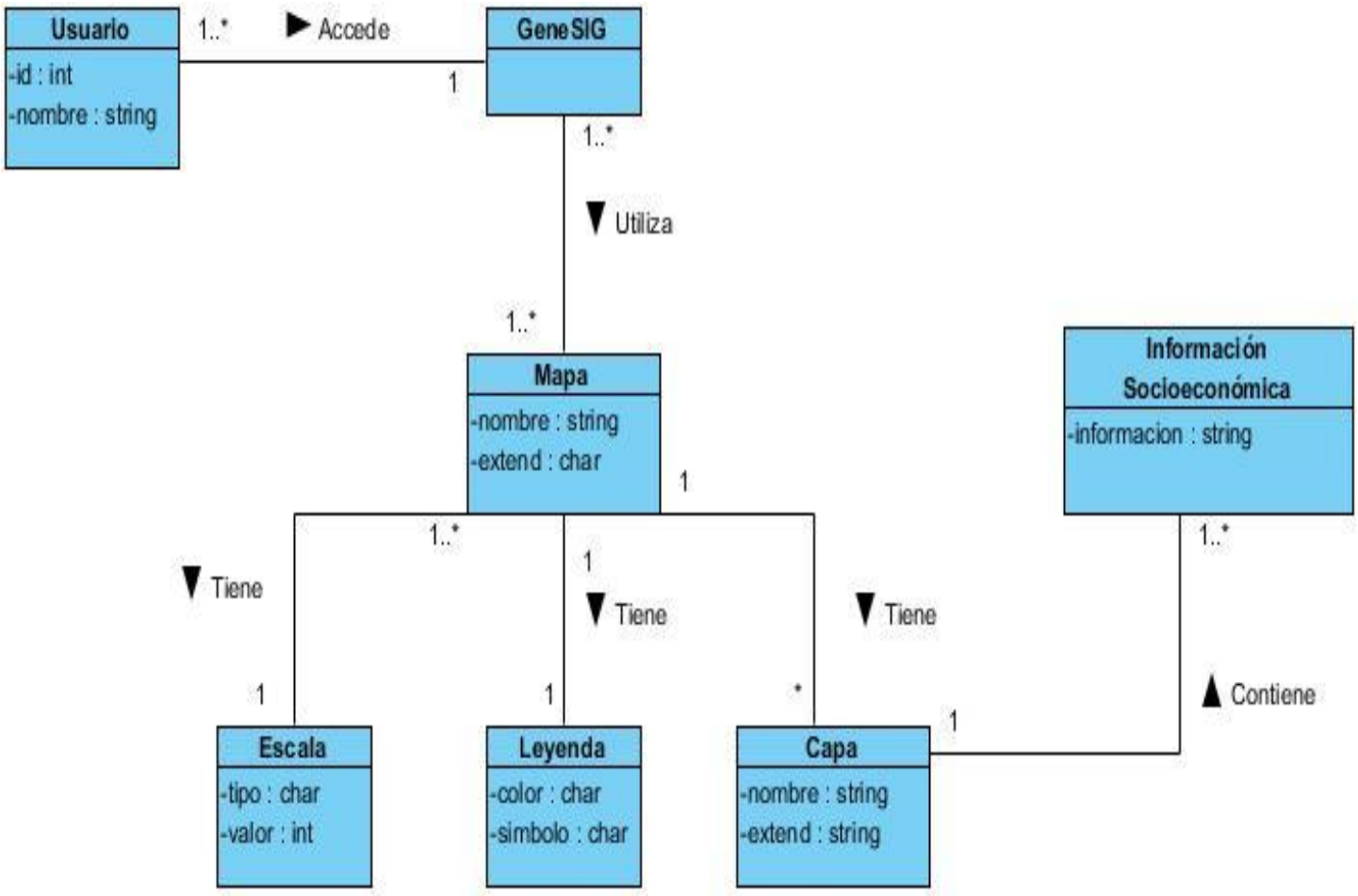

Fig. 10: Diagrama de clases de Modelo de Dominio.

2.2.1 Definición de las clases del Modelo de Dominio

Usuario: Se refiere a la persona que tiene los privilegios de interactuar con las funcionalidades del sistema.

Mapa: Cartografía de una localización referenciada geográficamente la cual posee escalas, capas y leyenda.

Escala: Es la porción respecto a la realidad a la que están representados los diferentes objetos en un mapa.

Capa: Agrupación de un grupo de entidades de un mismo tipo en un mapa.

Leyenda: Es una lista donde se explica el significado de los distintos símbolos utilizados en un mapa.

Información socioeconómica: Es un conjunto organizado de datos procesados referentes a los distintos elementos que están presentes en un mapa.

2.3 Levantamiento de Requisitos:

Un requisito, es una descripción de las necesidades que surgen respecto a un producto determinado. El principal objetivo al definir estos es identificar las necesidades para el correcto funcionamiento del producto. La definición y descripción de los mismos, ayuda a que exista una mayor comunicación entre el cliente y el equipo de desarrollo. Existen varias clasificaciones de requisitos, en la presente investigación se describirán los requisitos funcionales y los no funcionales.

2.4.1 Requisitos Funcionales:

Los requisitos funcionales son condiciones que el sistema debe cumplir. Para el desarrollo de la solución propuesta, el complemento debe cumplir con los siguientes requerimientos o requisitos funcionales:

RF 1: Selección de la capa

Este requisito le permite al usuario que seleccione una capa de la BD en la cual se identificarán los elementos similares.

RF 2: Selección de los datos de la capa.

Este requisito le permite al usuario que una vez que seleccione una capa, se muestre toda la información referente a esta capa en la BD para que el usuario seleccione los valores con los cuales quiere conformar la matriz que será utilizada posteriormente en el análisis.

RF 3: Confeccionar la matriz de datos.

Este requisito permite obtener la matriz de datos según los datos asociados a los atributos seleccionados previamente.

RF 4: Pre-procesar los datos.

Este requisito permite pre-procesar los datos previamente seleccionados para asegurarse de que todos los valores sean numéricos y que sean valores válidos, entiéndase por valores inválidos todos aquellos que no sean numéricos o sean nulos, además se le da la posibilidad de modificar los valores con problemas por la media en ese atributo, cero o simplemente eliminar el elemento para que no sea tenido en cuenta en el procesamiento.

RF 4.1: Buscar errores en la matriz de datos.

Este requisito permite detectar cualquier tipo de error que puedan tener los datos en la matriz de datos.

RF 4.2: Modificar los datos que contengan errores en la matriz de datos.

Este requisito permite modificar los datos en caso de que se tenga valores inválidos, en este caso se completarán dándole valores a este campo con 0, eliminando esa fila o columna donde se encuentra el elemento o dándole un valor según la media teniendo en cuenta esa fila o esa columna donde se encuentre el elemento.

RF 5: Crear el archivo con extensión .arff.

Este requisito es el que crea el archivo con la extensión **.arff** a partir de la matriz de datos, para que puedan ser reconocidos por Weka.

RF 6: Mandar a ejecutar el análisis.

Este requisito es el que manda a ejecutar al Weka el análisis para que realice el agrupamiento mediante el algoritmo K-Means.

RF 7: Mostrar la información referente al resultado de la tematización.

Este requisito va a mostrar la información referente al resultado de la tematización y del análisis realizado por el Weka.

RF 8: Visualizar el mapa con los datos previamente seleccionados.

Este requisito permite obtener el mapa con todos sus elementos previamente seleccionados por el cliente según el resultado del análisis de los datos que devuelve el Weka.

2.4.2 Requisitos No Funcionales:

Los requisitos no funcionales son restricciones de los servicios y cualidades brindadas por el sistema que surgen en función de las necesidades del usuario. Estas propiedades están en función de brindar al producto, usabilidad, rapidez y confiabilidad.

RNF 1: Rendimiento

- El tiempo de respuesta de la velocidad de procesamiento y actualización de la información estará dado por 2 razones fundamentales:

    1. El tiempo que demore el Weka en realizar el análisis.

    2. El tiempo que demore GeneSIG en construir el mapa tematizado.

RNF 2: Interfaz

- Se hará uso de los tooltip para indicarle al usuario lo que debe de hacer en cada funcionalidad de la aplicación.

- Las alertas de errores deberán ser lo más informativas posibles para que el usuario entienda cuál es el error y pueda corregirlo.

RNF 3: Software

- El servidor debe tener instalado Java 1.6 o superior para poder utilizar la herramienta Weka a través de la máquina virtual de Java.

RNF 4: Restricciones de Diseño

- El producto debe de diseñarse sobre una arquitectura cliente-servidor.

RNF 5: Software

- La PC cliente debe de tener sistema operativo GNU/Linux o Windows instalado y un navegador web.
- La PC servidor debe tener sistema operativo GNU/Linux Debian 6
- La PC servidor debe tener instalada el Gestor de Base de Datos PostgreSQL 8.4 con su extensión a PostGis para el almacenamiento de los datos espaciales.
- La PC servidor debe tener Apache con el módulo para PHP 5.3 o superior.
- La PC servidor debe tener instalado OpenJDK 6 o superior.

RNF 6: Licencia

El complemento será distribuido bajo licencia libre de código abierto GPL (General Public License), permitiéndose el uso y modificación del mismo.

2.4 Arquitectura

Según Pressman: “La arquitectura de software alude a la estructura general del software y las formas en que la estructura proporciona una integridad conceptual para un sistema. En su forma más simple, la arquitectura es la estructura u organización de los complementos del programa (módulos), la manera en

que estos complementos interactúan, y la estructura de datos que utilizan los complementos. En un sentido más amplio, sin embargo, los complementos pueden generalizarse para representar elementos importantes del sistema y sus interacciones" (Pressman, 2005).

Se concibe al mismo complemento como un complemento en sí a insertarse en la plataforma GeneSIG, por lo cual asume su distribución en paquetes estándar que garantiza su correcta integración concentrando el núcleo del sistema en el complemento **js** donde se implementa la lógica del negocio. Está basado en el principio de estandarización para el desarrollo de software estableciendo una política de desarrollo orientada a la reutilización de complementos.

La presente investigación presentará una arquitectura basada en componente. Esta permite alcanzar un mayor nivel de reutilización de software, aún en contextos distintos a aquellos para los que fue diseñado. Permite que las pruebas sean ejecutadas probando cada uno de los complementos antes de probar el conjunto completo de complementos ensamblados. Cuando existe un débil acoplamiento entre complementos, el desarrollador es libre de actualizar y/o agregar complementos según sea necesario, sin afectar otras partes del sistema. Dado que un complemento puede ser construido y luego mejorado continuamente, la calidad de una aplicación basada en componente se elevará con el paso del tiempo (Curvelo, 2012). A continuación se muestra el diagrama de complemento.

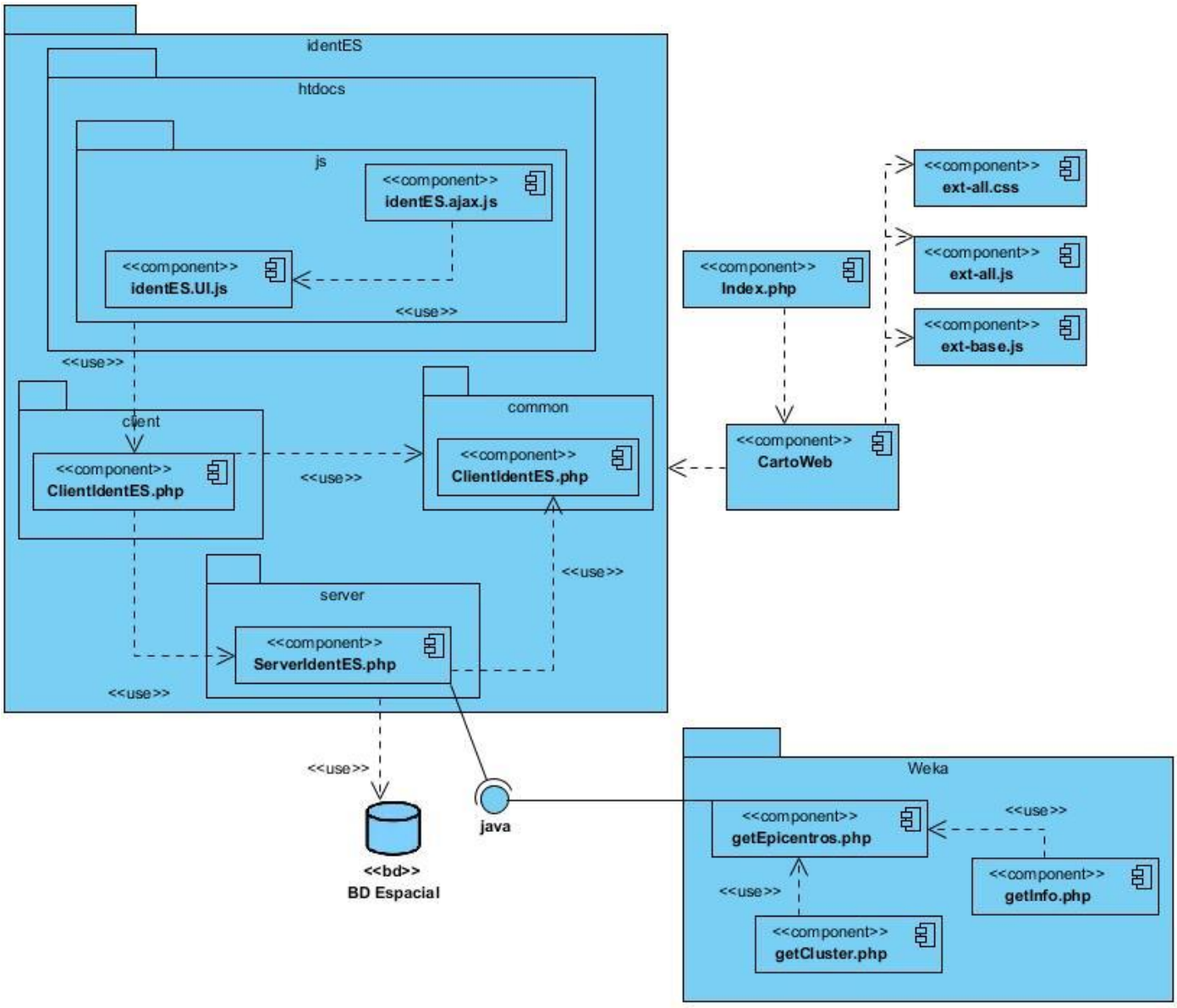

Fig. 11 Diagrama de componente.

2.5 Diagrama de Clases del Diseño

El diagrama de clases del diseño (Fig. 7) describe gráficamente las especificaciones de las clases de software y de las interfaces en una aplicación. Usualmente contiene la siguiente información:

- Clases, asociaciones y atributos.
- Interfaces
- Métodos
- Información sobre los tipos de atributos
- Dependencia

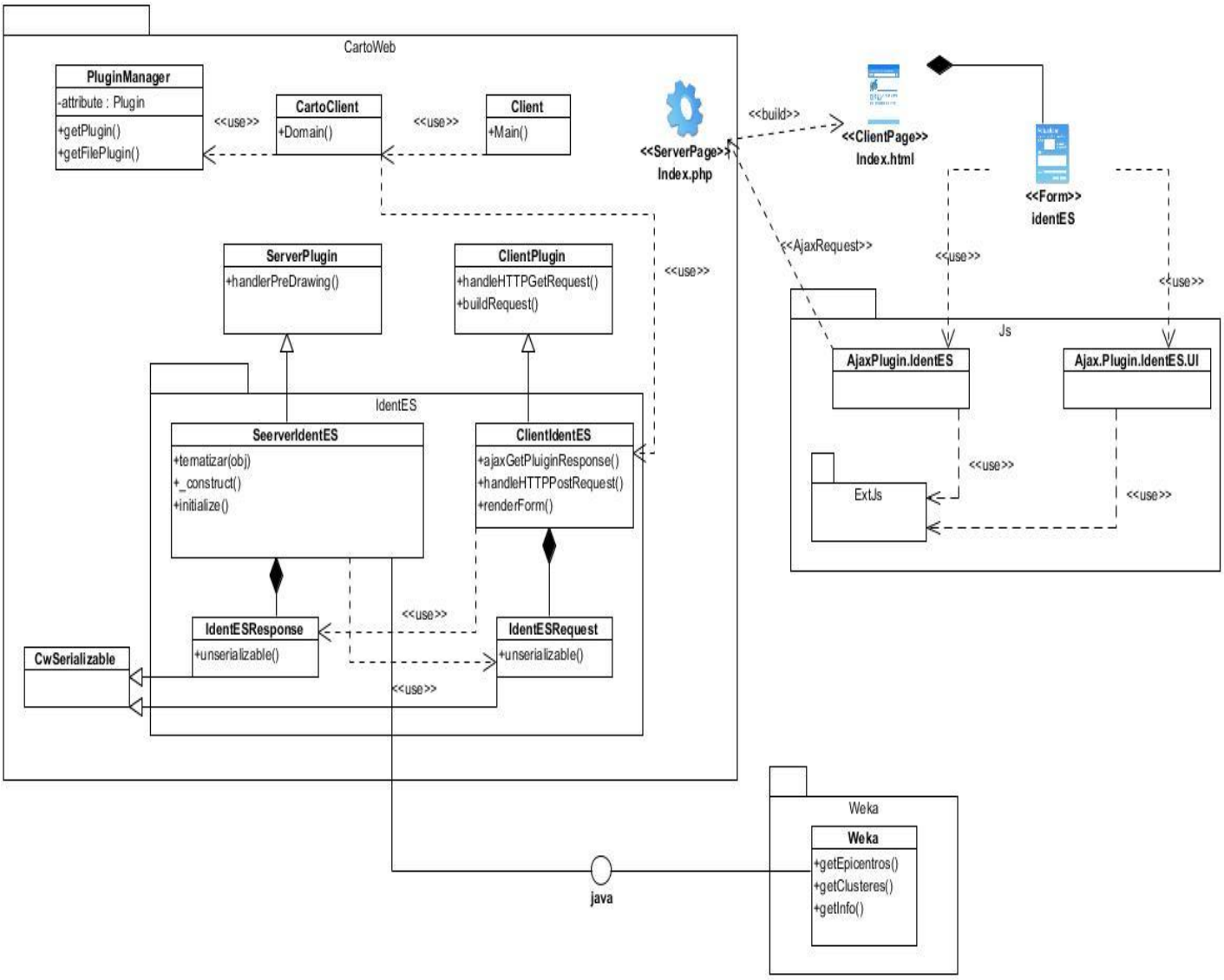

Fig.12: Diagrama de Clase del Diseño.

2.6 Patrones de Diseño

Los patrones de diseño son el soporte de las soluciones a problemas comunes en el desarrollo de un software. Se refieren al diseño de interacción o interfaces que permiten establecer una estructura común ante problemas semejantes.

2.6.1 GRASP

Los Patrones Generales de Software para Asignar Responsabilidades GRASP (General Responsibility Assignment Software Pattern por sus siglas en inglés), describen los principios fundamentales de la asignación de responsabilidades a objetos, expresados en forma de patrones. (Grosso, 2011)

Para obtener un buen diseño se aplicaron los patrones Experto, Creador, Controlador, Alta Cohesión y Bajo Acoplamiento. A continuación se describe cómo estos patrones solucionan los problemas de diseño en aplicaciones orientadas a objeto realizando una breve descripción de cada uno de ellos.

Patrones GRASP utilizados

- **Experto:** asigna responsabilidad a las clases expertas con la información necesaria para llevar a cabo una tarea determinada. Este se evidencia en las clases "*ServerIdentES*" y "*ClientIdentES*".

- **Creador:** crea una nueva instancia en la clase que tiene la información necesaria para realizar la creación del objeto, almacena o maneja varias instancias de la clase y contiene o agrega la clase. Este se aplica en las clases "*ServerIdentES*" y "*ClientIdentES*" con este fin.

- **Controlador:** Representa una especie de fachada en la capa del dominio para la capa de la interfaz. O sea, puede evidenciarse en la forma en que la plataforma maneja las peticiones del usuario enviándolas al cliente del complemento para que les pueda dar solución, en este caso sería "*ClientPlugin*".

- **Alta Cohesión:** asigna responsabilidades de manera que la información que almacena una clase sea coherente y esté relacionada con la clase. Constituye una premisa de GeneSIG donde cada complemento realiza una labor única dentro del sistema y auto-identificable creando un sistema relativamente fácil de mantener, entender y reutilizar.

- **Bajo Acoplamiento:** diseñado con el objetivo de tener las clases lo menos ligadas entre sí. De tal forma que en caso de producirse una modificación en alguna de ellas, se tenga la mínima repercusión posible en el resto de las clases, para potenciar la reutilización, y disminuir la dependencia entre las clases. Constituye otra de las premisas de GeneSIG garantizando que cada

complemento dentro de la estructura del complemento no dependan unos de otros a no ser estrictamente necesario.

2.6.2 GOF

Los patrones GOF (Gang Of Four por sus siglas en inglés) deben su nombre al grupo compuesto por sus creadores: Erich Gamma, Richard Helm, Ralph Jhonson y John Vlisides quienes en su publicación "Design Patterns" describen 23 patrones de diseño de gran utilidad y aplicables en el diseño de un sistema.

Estos patrones ofrecen las siguientes ventajas (Pressman, 2005):

- Proporcionan elementos reusables en el diseño de sistemas de software.
- Efectividad comprobada en la resolución de problemas similares.
- Formalizan un vocabulario común entre diseñadores.
- Facilitan el aprendizaje de las nuevas generaciones de diseñadores y desarrolladores utilizando conocimiento ya existente.
- Estandarizan el diseño.

Patrón GOF utilizado

Fachada (Facade): el patrón Fachada es del tipo estructural. Este se aplica para proporcionar una interfaz simple en un subsistema complejo. Crea una única clase, simplificando el acceso a diversas clases, para que los programadores utilicen esta para interactuar con todas las clases. Este se evidencia en la clase *"Ajax.Plugin.IdentES.UI".*

2.7 Conclusiones del capítulo 2

La modelación y el diseño son una parte esencial en el desarrollo de una solución informática, que le permite a los desarrolladores una mejor comprensión del trabajo a realizar y disminuye considerablemente el tiempo de desarrollo. Cada artefacto que se generó contribuyó en el desarrollo del complemento permitiendo que los resultados se ajustaran a la propuesta de solución. El modelo del dominio generado

sirvió de entrada para saber cómo es el funcionamiento del uso de los complementos de la plataforma GeneSIG. Para un mejor desarrollo del sistema se identificaron los requerimientos tanto funcionales como los no funcionales. Contribuyendo a una mejor comprensión de las características de la propuesta de solución, el diagrama de clase elaborado mostró las principales clases así como las relaciones entre estas. El uso de los patrones de diseño GRASP y GOF permitieron probar el correcto funcionamiento de cada complemento, clase o método o implementado, lo que permitió fijar las bases para la siguiente fase de desarrollo: implementación y prueba.

CAPÍTULO 3: IMPLEMENTACIÓN Y PRUEBA

Introducción

El contenido de este capítulo aborda la implementación y validación del sistema propuesto, donde se garantizó el cumplimiento a los requisitos definidos en el capítulo anterior, utilizando estándares de codificación para facilitar la comprensión del código. Se especifican los métodos y técnicas de prueba a realizar para verificar el correcto funcionamiento del complemento con el objetivo de identificar y corregir fallos cometidos durante el desarrollo de los casos de prueba.

3.1 Diagrama de Paquetes

Un diagrama de paquetes muestra como un sistema está dividido en agrupaciones lógicas mostrando las dependencias entre esas agrupaciones. Dado que normalmente un paquete está pensado como un directorio, los diagramas de paquetes suministran una descomposición de la jerarquía lógica de un sistema. Los paquetes están normalmente organizados para maximizar la coherencia interna dentro de cada paquete y minimizar el acoplamiento externo entre estos. Se propone de esta forma, el siguiente diagrama de paquetes, para el complemento a desarrollar.

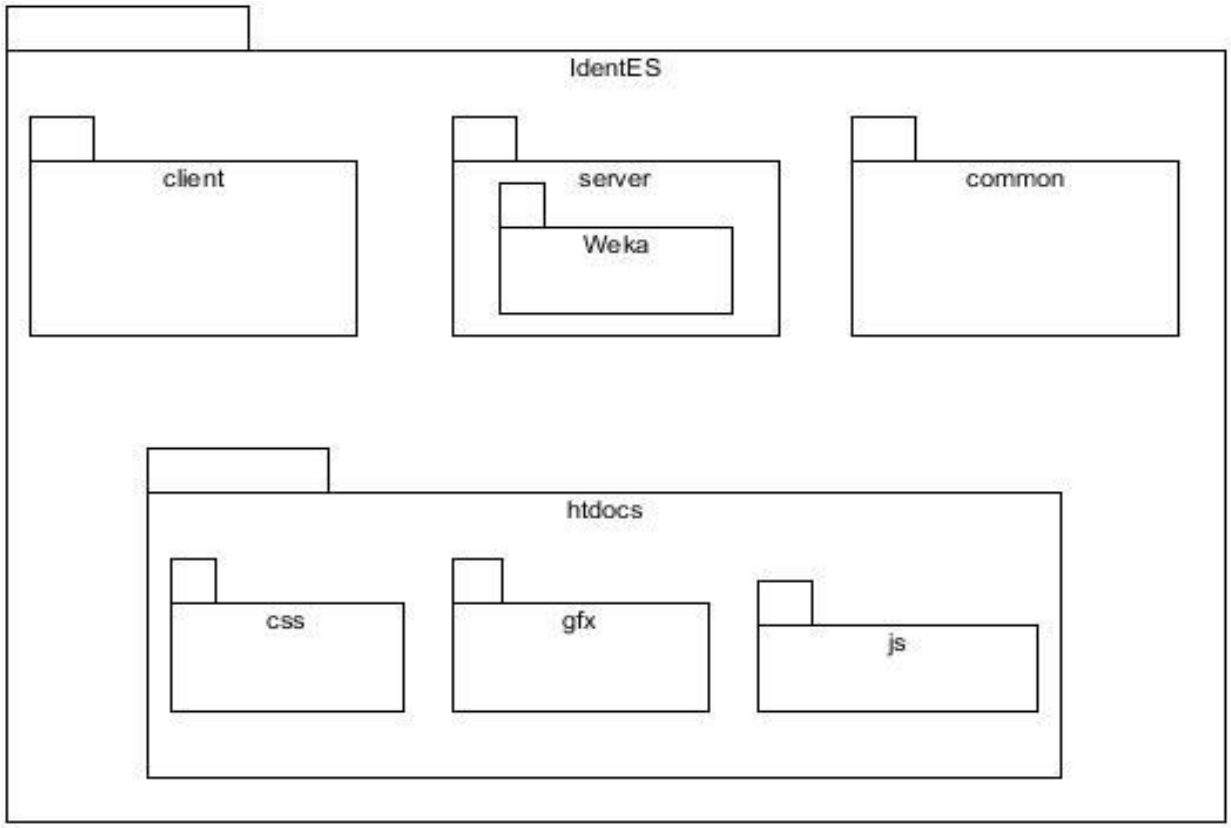

Fig.13: Diagrama de Paquete.

3.1.1 Descripción del diagrama de paquetes.

Client: Contiene todos los complementos (de extensión php) del complemento que implementan el módulo CartoClient del CartoWeb.

Server: Contiene todos los complementos (de extensión php) del complemento que implementan el módulo CartoServer del CartoWeb.

Common: Contiene todos los complementos que implementan aquellas clases que servirán de puente para la comunicación entre client.php y server.php.

Htdocs: Se define como el directorio que contiene los elementos que serán publicados por el servidor de aplicaciones web.

Gfx: Contiene las imágenes que serán utilizadas en el complemento.

Css: Contiene las hojas de estilo que serán utilizadas en el complemento.

Js: Contiene todos los complementos (de extensión js) del complemento que implementan el módulo cliente en JavaScript del CartoWeb.

3.2 Diagrama de Despliegue

El propósito del modelo de despliegue es capturar la configuración de los elementos de procesamiento y las conexiones entre estos elementos en el sistema. El modelo consiste en uno o más nodos, dispositivos, y conectores entre estos. El modelo de despliegue también mapea procesos dentro de estos elementos de procesamiento, permitiendo la distribución del comportamiento a través de los nodos que son representados. A continuación se muestra el diagrama de despliegue modelado para la aplicación a desarrollar:

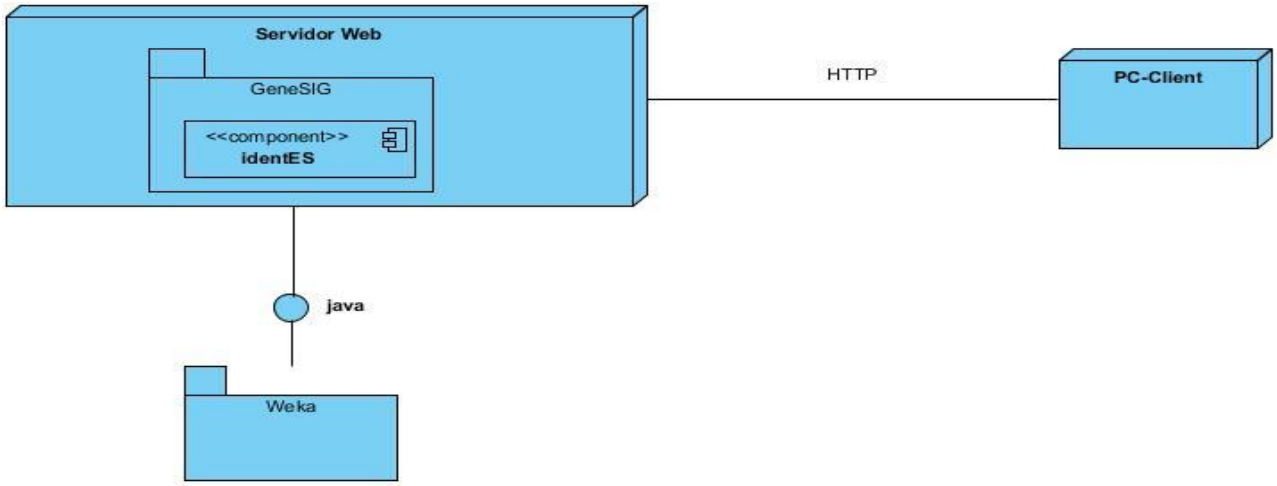

Fig. 14: Diagrama de Despliegue.

Servidor Web: Es el nodo donde se encuentra la plataforma GeneSIG.

identES: Complemento para la plataforma GeneSIG para la identificación de elementos similares.

PC Cliente: nodo desde el cual los usuarios se conectan para hacer sus peticiones (HTTP) a GeneSIG y de esta forma pueden interactuar con el complemento según las acciones deseadas.

Weka: Herramienta que se ejecutará una vez que el cliente desee utilizar el complemento identES.

3.3 Pruebas

La prueba de software consiste en ejecutar el programa con la intención de descubrir errores previos a la entrega al usuario final. Tiene como objetivo encontrar y documentar los defectos que pueden afectar la calidad del software. Las pruebas se aplican durante todo el ciclo de desarrollo del software para diferentes objetivos y en distintos niveles de trabajo (Pressman, 2005).

Una estrategia de prueba describe el enfoque y los objetivos generales de las actividades de prueba, así como consideraciones especiales afectadas por requerimientos de recursos o que tengan implicaciones en la planificación. La estrategia de prueba definida consiste en la realización del método de caja negra por la técnica de partición de equivalencia.

3.3.1 Pruebas de Caja Negra

Las pruebas de caja negra se realizan sobre la interfaz, la técnica seleccionada para implementarla es la de partición de equivalencia, esta técnica trata cada parámetro como un modelo algebraico donde unos datos son equivalentes a otros. Logra reducir un rango amplio de posibles valores reales a un conjunto reducido de clases de equivalencia, entonces es suficiente probar un caso de cada clase, pues los demás datos de la misma clase son equivalentes.

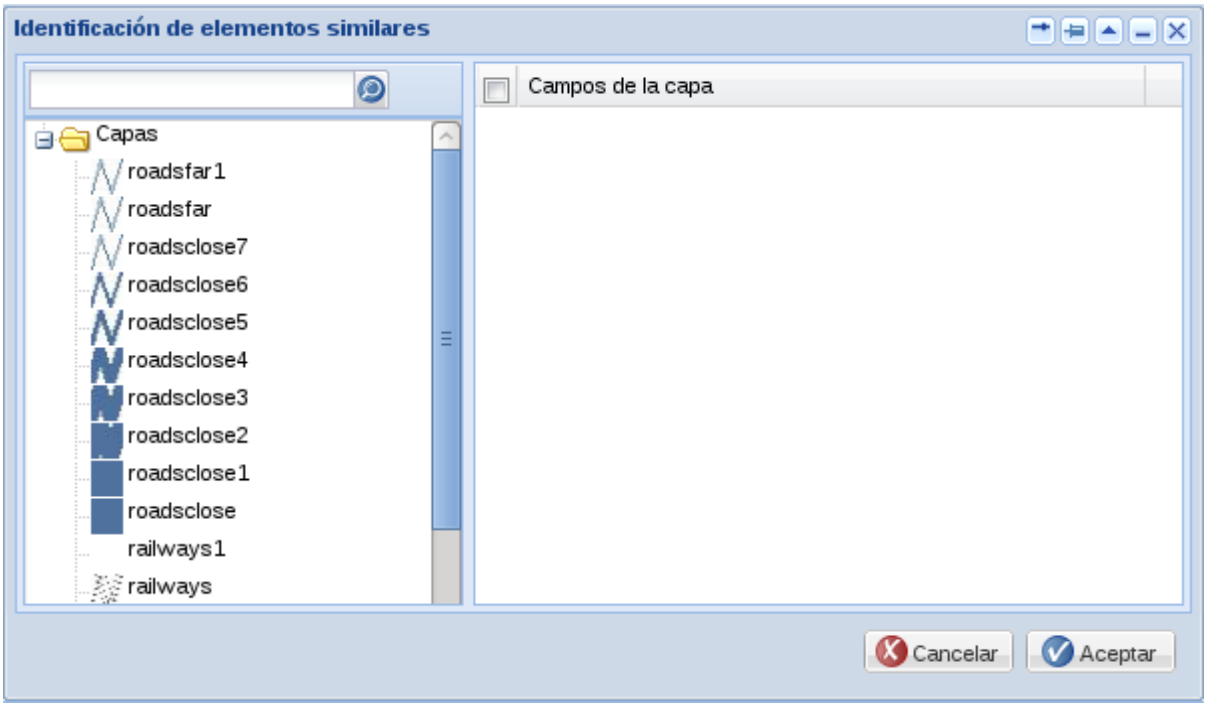

Fig. 15: Interfaz gráfica para la selección de la capa.

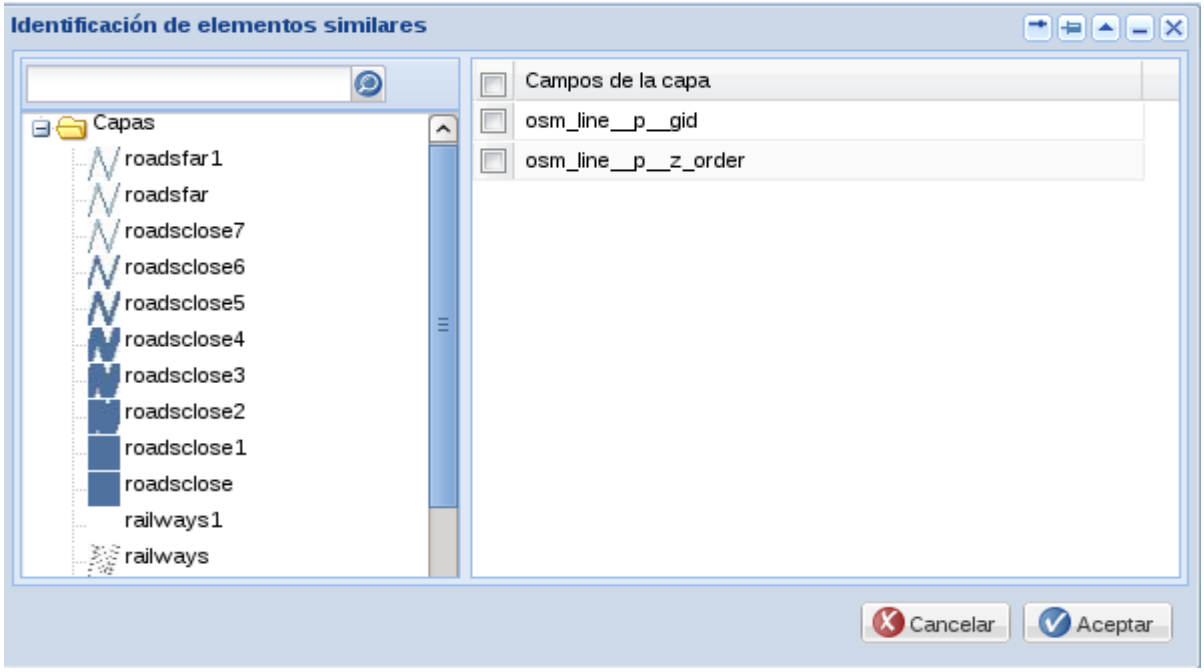

Fig. 16: Interfaz gráfica para la selección de los elementos de la capa.

Tabla 2: Descripción de la secuencia.

Nombre de la sección	*Escenario de la sección*	*Descripción de la funcionalidad*
SC 1: Seleccionar datos de una capa	*SC 1.0: Selecciona datos de una capa satisfactoriamente.*	*Se le muestra toda la información referente a ese dato*
	SC 1.1: El actor selecciona la opción cancelar	*Redirecciona a otra interfaz*
	SC 1.2: No selecciona ningún dato	*No se le habilita ningún panel*

Con la realización de este método de prueba se demostró que las funciones del software son operativas y que las salidas del mismo eran las esperadas.

La prueba de caja negra no es una alternativa a las técnicas de prueba de caja blanca. Más bien se trata de un enfoque complementario que intenta descubrir diferentes tipos de errores que los métodos de caja blanca (Pressman, 2005).

3.4 Conclusiones del capítulo 3

El diagrama de despliegue permitió comprender la distribución física del sistema, usando el protocolo HTTP para la comunicación entre el cliente y el servidor. Mediante la aplicación de pruebas de Caja Negra mediante la técnica de Participación Equivalente, se pudieron corregir las principales no conformidades que lograron satisfacer los requisitos funcionales identificados.

Conclusiones Generales

Como resultado de la investigación se desarrolló el análisis, diseño, implementación y pruebas de un complemento para la plataforma GeneSIG que permite identificar los elementos geoespaciales similares entre sí con múltiples atributos, dada su información socioeconómica. Arribando a las siguientes conclusiones:

- El análisis de las tendencias actuales en el proceso de identificación de elementos geoespaciales similares permitió la selección del algoritmo K-Means para el desarrollo del complemento.
- El uso de las metodologías permitieron guiar el proceso de desarrollo de software y de MD.
- La implementación del complemento permitió dar cumplimiento a los requisitos funcionales identificados y brindarle una nueva funcionalidad a la plataforma GeneSIG haciendo uso de una herramienta externa.
- Con las pruebas realizadas se validó el complemento detectando los errores del mismo ante de integrarlo a la plataforma GeneSIG.

Recomendaciones

Con el objetivo de mejorar los resultados obtenidos durante el desarrollo de la presente investigación se proponen las siguientes recomendaciones para futuras investigaciones:

1. Estudiar con mayor profundidad la herramienta Weka para que otros complementos de la plataforma GeneSIG haciendo uso de esta, puedan brindar mejores soluciones.

Referencias

A. Hinneburg, D. A. Keim. 1998. *An efficient Approach to Clustering in Large Multimedia Databases with Noise.* 1998.

Ager, Ingenieros. 2003. Datos espaciales en los SIG. 2003.

Agudelo, Denisse Cangrejo Aljure y Juan Gabriel. 2011. *Minería de Datos Espaciales.* 2011.

Andritsos, Periklis. 2002. *Data clustering techniques. .* Toronto : s.n., 2002.

Atkinson, A. Coffey. 2005. *Estrategias Complementarias de Investigación.* 2005.

Barrera, J. Hurtado de. 2000. *Metodología de la investigación holística.* Caracas : s.n., 2000.

Cuadras, Carles M. 1989. *Distancias Estadísticas.* Barcelona : s.n., 1989.

Date, C. J. 2001. *Introducción a los Sistemas de bases de Datos.* 2001.

Díaz, Álvaro. 2005. *Graphical User Interface (GUI) para el programa servidor de mapas.* 2005.

Díez., José Luis. 2003. *Técnicas de agrupamiento para identificación y control por modelos locales.* Valencia : s.n., 2003.

Elizondo, José Jaime Esqueda. 2002. *Matlab e Interfaces Gráficas.* 2002.

Española, Real Academia. 2014. *Real Academia.* 2014.

Espinosa, Roberto. 2010. *Teoría de Data Mining.* 2010.

ESRI. 1998. *Base de Datos Geográficas.* 1998.

Flores, Gil. 1994. *Análisis de Datos Cualitativos. Aplicaciones a la Investigación Educativa.* Barcelona : PPU, 1994.

García, Alfonso Pérez. 1993. *ESTADISTICA APLICADA CON SAS.* 1993.

Gastelú, Carlos Arturo Torres. 2011. *Servicio y localización espacial.* Veracruz : s.n., 2011.

Gilfillan, Ian. 2000. *La Biblia de MySQL.* 2000.

Graeff, B. 2001. *Querying Raster Data Structures –Probabilistic and non-probabilistic approaches on knowledge based template matching methods.* Budapest : s.n., 2001.

Gran, Aurea. *Distancias Estadísticas y Escalado Multidimensional (Análisis de Coordenadas Principales).*

Grosso, Andrés. 2011. *Prácticas de Software.* 2011.

Grupo Implantación, Soporte y Tecnologías. 2015. *Centro de Desarrollo de Geoinformática y Señales Digitales (GEYSED).* La Habana : s.n., 2015.

Güting, R. H. 1994. *An Introduction to Spatial Database Systems.* Alemania : s.n., 1994.

I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. 2006. *Rapid Prototyping for Complex Data Mining Tasks.* 2006.

IBM. 2012. *Manual CRISP-DM de IBM SPSS Modeler.* 2012.

J. Han, M Kamber. 2001. *Data Mining: Concepts and Techniques.* Estados Unidos : s.n., 2001.

J. Zilberstein, H. Valdés. 1999. *Aprendizaje escolar, diagnóstico y calidad educativa.* México : Ediciones CEIDE, 1999.

Juan Miguel Moine, Ana Silvia Haedo, Silvia Gordillo. 2011. *Estudio comparativo de metodologías para minería de.* 2011.

JcomplementoVicenteño, Francisco González y Carlos Álvaro. 2013. *Inteligencia Geoespacial.* 2013.

Kolatch, E. 2001. *Clustering Algorithms for Spatial Databases.* 2001.

Krasnov, M. Kiseliov, A. Makarenko, G. Shikin, E. 2005. *Curso de Matemática Superior para Ingenieros.* La Habana, Cuba : Editorial Félix Varela, 2005.

Kruse, Robert L. 2014. *Estructura de Datos y Diseño de Programas.* 2014.

Leonard Kaufman, Peter J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis.* 1990.

Lissa Curvelo Oliva, Laura S. Ortega Retureta, Yoanna Columbié Cisnero, Yudelky González Milán. 2012. *RODESOFT PDS UCID.* 2012.

MapServer. 2015. *MapServer 7.0.0-beta1 .* 2015.

Marañón Rodríguez, Hermes Alex, Duharte Montero, Pedro David. 2010. *Planificación y Diseño del Portal para la Comunidad Técnica Cubana de PostgreSQL.* La Habana : s.n., 2010.

Marin Ester, Hans-Peter Kriegel, Jörg Sander. 2001. *Geographic Data Mining and Knowledge Discovery, chapter Algorithms and Applications for Spatial Data Mining.* 2001.

Microsoft. 2014. *Conceptos Básicos sobre Bases de Datos.* 2014.

Minotta, Enrique Alberto Hurtado. 2006. *Estadística Descriptiva y Analítica.* 2006.

NCGIA. 2015. *Centro Nacional de Información Geográfica y Análisis .* 2015.

P. Chapman, J. Clinton, Keber. 2000. *CRISP-DM 1.0 Step by step guide.* 2000.

Parr, O. 2000. *Data Mining Cookbook Modeling Data for Marketing, Risk, and Customer Relationship Management.* 2000.

Parra, Hugo García Mancilla y JcomplementoMatus. 2013. *Estadística descriptiva e inferencial I.* 2013.

Peñuelas, Rodríguez. 2005. *Metodología de Investigación.* México : s.n., 2005.

Pérez, Javier Eguíluz. 2008. *Introducción a JavaScript.* 2008.

Plata, María Teresa Escobedo Portillo y Jorge A. Salas. 2007. *Mahalanobis y las aplicaciones de su distancia estadística.* 2007.

Pressman, Roger. 2005. *Ingeniería de Software. Un enfoque práctico.* 2005.

R. Hernández, C. Fernández, P. Baptista. 2003. *Metodología de la Investigación.* México : s.n., 2003.

R. O. Duda, P. E. Hart, D. Stork. 2001. *Pattern Classification",.* 2001.

Regina Obe, Leo Hsu. 2011. *PostGIS in Action .* USA : s.n., 2011.

Ritzman, Lee J. Krajewski y Larry P. 2000. *Administración de operaciones: estrategia y análisis.* México : s.n., 2000.

Rodaisy Sánchez Rodríguez, Abella Pérez y Loren. 2009. *Sistema para la gestión de la información de profesores. Desarrollo del Módulo Extensión.* 2009.

Rodríguez Gómez, Gregorio y otros. 1999. *Metodología de la Investigación cualitativa.* 1999.

Rouse, Margaret. 2012. *Análisis de Datos.* 2012.

S., Christian Van Der Henst. 2001. *¿Qué es el PHP?* 2001.

Sánchez, Juan Eladio. 2008. *ExtJS lo bueno, lo malo y lo feo.* 2008.

Shalini S. Singh, N. C. Chauhan. 2011. *K-means v/s K-medoids: A Comparative Study.* Gujarat : s.n., 2011.

Ubuntu. 2012. *Guía Documentada para Ubuntu.* 2012.

Wiley, John. 2008. *Wiley Encyclopedia of Computer Science and Engineering.* 2008.

Witten, Ian H. 1999. *Weka: Practical machine learning tools and with Java implementations.* 1999.

Zayas, C. Álvarez de. 1999. *La escuela en la vida.* La Habana : Pueblo y Edcación, 1999.

Zhexue Huang, Michael K. 1990. *A fuzzy k-modes algorithm for clustering categorical data.* 1990.