Universidad de las Ciencias Informáticas.

Facultad 3

Título: Propuesta Arquitectónica de un Sistema de Recuperación de Información Geográfica para el motor de búsqueda Orión.

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas

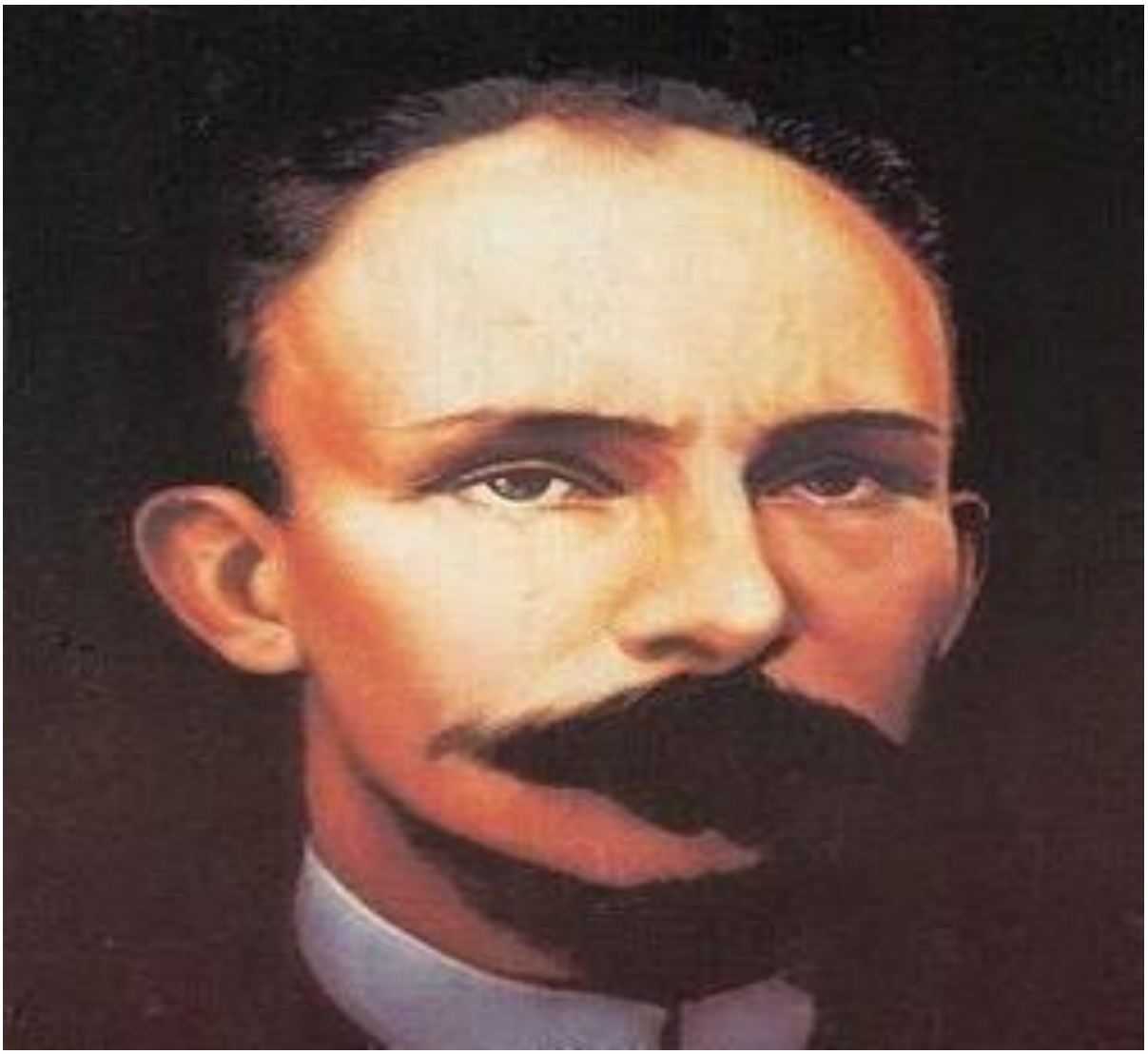
Autora: Anay Medina García

Tutores: Msc. Meylin Martínez Chong

Ing. Yusniel Hidalgo Delgado

Ciudad de la Habana, junio del 2011

"Año del 53 Aniversario del Triunfo de la Revolución"



"La fama es un premio justo de quien tiene el valor de sacrificar el grato sigilo de su persona a la idea que defiende".

José Martí

Yo, Anay Medina García declaro que soy la única autora de este trabajo y autorizo a la facultad3 de la Universidad de las Ciencias Informáticas a hacer uso de la misma en su beneficio. Para que así conste firmo el presente a los ____ días del mes de _____ del año _____.

Autora	Tutora	Tutor
Anay Medina García	Msc.Meylin Martínez Chong	Ing. Yusniel Hidalgo Delgado
_____	_____	_____
Firma de la Autora	Firma de la tutora	Firma del tutor

AGRADECIMIENTOS

A mis padres por apoyarme en estos cinco años de mi carrera, por educarme, por darme cariño, confianza.

A mis hermanas por darme consejos.

A mi novio Yordanis que lo quiero con la vida, por preocuparse por la terminación de mis estudios y por la admiración que siente por mí, por apoyarme en todas mis decisiones, por soportar mis malcriadeces.

A mí cuñado Kiomi y a mis suegros por darme confianza, especialmente a Oneida por levantarme cada día para venir para la universidad a hacerme cada vez mejor.

A mi tutor Yusniel por ayudarme mucho en la tesis, por ver confiado en mí.

A todas mis amigas especialmente del apto por enseñarme cada día algo nuevo y por ser sinceras conmigo.

Al Msc. David Silva Barrera por ayudarme a realizar la validación de la arquitectura.

A todos les agradezco, por haberme dado la oportunidad de graduarme.

DEDICATORIA

Dedico esta tesis a Fidel por haberme dado la oportunidad de ser una estudiante más de la Universidad de las Ciencias Informáticas.

A toda mi familia especialmente a mis padres Raúl y Victoria porque me han apoyado y se han preocupado por mi bienestar dentro y fuera de la universidad.

A mi novio al que extraño cada día que estoy en la universidad.

A mis amistades que me han apoyado en todo momento.

A mi tutor por ser un ejemplo a seguir.

En fin a la Vida por haberme dado la oportunidad de existir.

DATOS DE CONTACTOS

Msc. Meylin Martínez Chong

Graduada en la CUJAE como Ingeniera Informática en el año 2005. Actualmente se encuentra laborando en la Universidad de Ciencias Informáticas en el proyecto ERP en la línea de logística como jefa de línea.

Correo electrónico: meylin@.uci.cu

Ing. Yusniel Hidalgo Delgado

Graduado de Ingeniero en Ciencias Informáticas por la Universidad en Ciencias informáticas en el año 2010. Actualmente se encuentra laborando en la Universidad de Ciencias Informáticas en el proyecto ERP en la línea de logística como desarrollador de software.

Correo electrónico: yhdelgado@uci.cu

Resumen

Internet y la World Wide Web más conocido como www se han convertido en un enorme repositorio de información consultado diariamente por millones de usuarios. Además, otros repositorios de información, como las bases de datos documentales o las bibliotecas digitales, también han aumentado su popularidad considerablemente. Esto ha provocado que la Recuperación de Información se haya convertido en una de las áreas de investigación más importantes dentro de la informática.

Tanto los Sistemas de Información Geográfica como los Sistemas de Recuperación de Información han sido áreas de investigación muy importantes en las últimas décadas. Recientemente, un nuevo campo de investigación llamado Sistemas de Recuperación de Información Geográfica (SRIG) ha surgido fruto de la intersección de estas dos áreas.

Este trabajo tiene como objetivo proponer una solución arquitectónica para el desarrollo de un Sistema de Recuperación de Información Geográfica, específicamente para el motor de búsqueda Orión.

Para darle cumplimiento al objetivo planteado en la investigación se realizó el estado del arte sobre las Arquitecturas de los Sistemas de Recuperación de Información Geográfica. Se definieron las herramientas necesarias para la construcción de un Sistema de Recuperación de Información Geográficas basadas en software libre. Se validó la arquitectura propuesta.

Palabras Claves:

Arquitectura, Información geográfica, Recuperación de información.

Contenido

Contenido

INTRODUCCIÓN

Con el desarrollo de las Tecnologías de la Informática y las Comunicaciones TIC, y las capacidades de almacenamiento de la información digital, la cantidad de información generada en el planeta crece exponencialmente, a tal punto, que resulta tedioso encontrar información útil sobre una determinada materia.

Los Sistemas de Recuperación de Información, en lo adelante SRI, constituyen el mecanismo ideal para resolver este tipo de problemas. Estos permiten localizar y procesar la información de forma rápida y en forma automática. Son sistemas capaces de localizar cualquier contenido existente en la web, tales como textos, imágenes, videos, archivos de sonido, entre otros. En este sentido destacan los directorios temáticos, los motores de búsqueda o buscadores y los metabuscadores (1).

En ocasiones resulta complejo realizar una búsqueda sobre una determinada materia, debido a que los sistemas que poseen estas informaciones no cuentan con servicios que permitan a los usuarios visualizar la localización geográfica del contenido que desea obtener.

En el año 2010, en la Universidad de las Ciencias Informáticas, se comenzó el desarrollo del motor de búsqueda Orión. Este motor de búsqueda permite recuperar información existente en la web interna de una empresa o universidad, sin embargo, no tiene definida la arquitectura para la construcción de un Sistema de Recuperación de Información Geográfica que aporte a la toma de decisiones dentro o fuera de la institución.

Atendiendo a la **situación problemática** descrita anteriormente se enuncia el siguiente **problema a resolver:** ¿Cómo contribuir a la búsqueda de información a partir del análisis de la información geoespacial contenida en el motor de búsqueda Orión?

El **objeto de estudio** se define como: Arquitecturas de los Sistemas de Recuperación de Información Geográfica y el **campo de acción** se define como: la arquitectura de un Sistema de Recuperación de Información Geográfica en el motor de búsqueda Orión.

Para darle solución al problema descrito, se ha planteado el siguiente **objetivo general:**

Proponer una solución arquitectónica para el desarrollo de un Sistema de Recuperación de Información Geográfica.

Del cual se desglosan los siguientes **objetivos específicos**:

- ✓ Estudiar el estado del arte sobre el objeto de estudio que se investiga para realizar una valoración crítica y tomar posición al respecto.
- ✓ Definir las herramientas necesarias para la construcción de un Sistema de Recuperación de Información Geográficas basadas en software libre.
- ✓ Validar la arquitectura propuesta.

Para dar cumplimiento a los objetivos específicos planteados con anterioridad se definen las siguientes **tareas de investigación:**

- ✓ Realización del Diseño Metodológico de la investigación para identificar el problema a resolver y los objetivos de la investigación.
- ✓ Confección del cronograma de ejecución de la investigación para planificar la ejecución de la misma.
- ✓ Evaluación del contenido de la información obtenida sobre el tema que se investiga para establecer un diagnóstico de las tendencias actuales y tomar posición al respecto.
- ✓ Caracterización de las principales arquitecturas utilizadas en el desarrollo de los sistemas de recuperación de información geográfica para tener identificadas cuales son las arquitecturas más usadas a nivel mundial.
- ✓ Comparación de las principales arquitecturas utilizadas en el desarrollo de los sistemas de recuperación de información geográfica para seleccionar lo mejor de cada una de ellas.
- ✓ Elaboración de la propuesta de la posible arquitectura a utilizar en la construcción de un sistema de recuperación de información geográfica basado en software libre.
- ✓ Caracterización de las principales herramientas basadas en software libre para la construcción de un Sistema de Recuperación de Información Geográfica.
- ✓ Valoración crítica de los principales resultados obtenidos durante la investigación para establecer conclusiones finales.

Idea a defender

Si se define una correcta solución arquitectónica para un sistema de recuperación de información geográfica, entonces se podrá incorporar una herramienta que permita a los usuarios buscar información tanto espacial como textual en el motor de búsqueda Orión.

Resultados esperados

Obtención de una propuesta arquitectónica para el desarrollo de un Sistema de Recuperación de Información Geográfica para el motor de búsqueda Orión.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA

Introducción

En este capítulo se exponen una serie de criterios valorativos sobre las principales tendencias, técnicas, tecnologías y herramientas que son utilizados en la construcción de una propuesta arquitectónica para los Sistemas de Recuperación de Información Geográfica. También se describen algunas aplicaciones de estos sistemas a nivel mundial.

Sistemas de Información Geográfica

Son varios los autores que han intentado dar una definición general de lo que son los Sistemas de Información Geográfica (SIG). (Burrough, 1988) los define como un conjunto de herramientas potentes para recoger, almacenar, recuperar, transformar y mostrar datos espaciales del mundo real para unos propósitos particulares. En este trabajo se adoptará el concepto dado por el Environmental Systems Research Institute Inc (ESRI, 1995) que es la principal empresa que comercializa este tipo de herramientas informáticas, los define como un conjunto organizado de hardware, software y datos geográficos, diseñados para manipular, analizar y mostrar todo tipo de información referenciada geográficamente con el fin de resolver problemas complejos de planificación y gestión (2).

También puede definirse como un modelo de una parte de la realidad referido a un sistema de coordenadas terrestre y construido para satisfacer necesidades concretas de información. En un sentido más genérico, los SIG son herramientas que permiten a los usuarios crear consultas interactivas, analizar la información espacial, editar datos, mapas y presentar los resultados de todas estas operaciones.

Los Sistemas de Información Geográfica han constituido durante los últimos veinte años una de las herramientas de trabajo más importantes para investigadores, analistas y planificadores; en todas sus actividades tienen como insumo el manejo de la información relacionada con diversos niveles de agregación espacial o territorial, lo cual está creando la necesidad de que estos usuarios de información espacial conozcan acerca de esta tecnología. Aunque los SIG tienen gran capacidad de análisis, estos no pueden existir por sí mismos, deben tener una organización, personal y equipamiento responsable para su implementación y sostenimiento, adicionalmente este debe cumplir un objetivo y estar garantizados los recursos para su mantenimiento.

Técnicas utilizadas en los SIG

En los sistemas de información geográfica las técnicas que se emplean para la representación de datos en los SIG son los métodos Raster y Vectorial y para la creación de datos las herramientas de Diseño Asistido por Ordenadores (sus siglas en inglés, CAD).

Herramientas CAD

Se utilizan especialmente para crear diseños y planos de construcción tanto de manufactura como obras de infraestructura, estos sistemas no requieren de componentes relacionales ni de herramientas de análisis, las herramientas CAD actualmente se han ampliado como soporte para mapas, pero tienen utilidad limitada para analizar y soportar bases de datos geográficas grandes. Tiene como objetivo producir un dibujo de un objeto, una casa, el esquema de una red viaria, etc.

SIG Vectoriales

El modelo vectorial es una estructura de datos utilizada para almacenar datos geográficos. Esta forma de expresión espacial implica la utilización de los tres tipos de elementos espaciales, de carácter geométrico, en que pueden ser interpretados los objetos geográficos: puntos, líneas y polígonos. Los atributos temáticos, que corresponden a las unidades espaciales, se manejan, habitualmente, desde tablas, sujetas al concepto de base relacional (3).

SIG Raster

El modelo Raster es un método para el almacenamiento, el procesado y la visualización de datos geográficos. Su forma de proceder es dividir la zona de afección de la base de datos en una retícula o malla regular de pequeñas celdas (píxeles) y atribuir un valor numérico a cada celda como representación de su valor temático. Dado que la malla es regular, el tamaño del píxel es constante y se conoce la posición en coordenadas del centro de una de las celdas, se puede decir que todos los píxeles están geográficamente referenciados. Para tener una descripción precisa de los objetos geográficos contenidos en la base de datos el tamaño del píxel debe ser reducido en función de la escala, lo que dotará a la malla de una resolución alta; sin embargo, a mayor número de filas y columnas en la malla, mayor va a ser el esfuerzo en el proceso de captura de la información y mayor costo computacional al momento de procesarla (4).

El modelo de datos Raster es útil para describir objetos geográficos con límites difusos, como por ejemplo la dispersión de una nube de contaminantes, o los niveles de contaminación de un acuífero subterráneo, donde los contornos no son absolutamente nítidos; en esos casos, el modelo Raster es más apropiado que el vectorial.

Los dos modelos de SIG no son excluyentes, ya que sus ventajas e inconvenientes se complementan, siendo necesario frecuentemente trabajar con ambos modelos de datos. Este hecho ha propiciado que la mayoría de los SIG dispongan de ambas naturalezas, aunque normalmente con el predominio de una de ellas. A continuación se presenta una tabla que muestra los beneficios e inconvenientes de los modelos mencionados anteriormente.

Modelo	**Ventajas**	**Inconvenientes**
Vectorial	1. Estructura de datos más compacta con ficheros menos voluminosos. 2. Topología mejor definida con mayor capacidad de análisis. 3. Más adecuado para la representación de datos bien definidos como: ríos, carreteras, etc.	1. Estructura de datos más compleja. 2. Mayor dificultad de proceso en operaciones de superposición. 3. Insuficiente representación en caso de alta variabilidad espacial. 4. Gran dificultad en el tratamiento de imágenes digitales. 5. Dificultad de aprendizaje y complejidad de manejo
Raster	1. Estructura de datos más sencilla. 2. Operaciones de análisis sencillas y potentes. 3. Mejor presentación de la variabilidad espacial y de elementos poco definidos. 4. Gran capacidad para el	1. Estructura de datos menos compacta con grandes ficheros de datos. 2. Peor presentación grafica de resultado. 3. Relaciones topológicas más difíciles de representar. 4. Limitaciones de resolución

	tratamiento de imágenes digitales. 5. Más facilidad en el aprendizaje y uso.	como consecuencia de su relación con el volumen de almacenamiento. 5. Facilidad en el tratamiento de imágenes digitales.

Tabla 1. Comparación entre los modelos Raster y Vectorial.

Componentes arquitectónicos de un SIG

Los Sistemas de Información Geográfica están compuestos de los siguientes aspectos (5).

- ✓ Hardware
- ✓ Software
- ✓ Datos
- ✓ Recursos Humanos
- ✓ Procedimientos

Hardware

Hardware es la computadora en la que opera el SIG. Actualmente, un SIG corre en un amplio rango de tipos de hardware, desde servers de computadoras centralizados hasta computadoras desktop utilizadas en configuraciones individuales o de red.

Una organización requiere de hardware suficientemente específico para cumplir las necesidades de la aplicación. Algunas cosas a considerar incluyen: velocidad, costo, soporte, administración, escalabilidad y seguridad.

Software

Los programas SIG proveen las herramientas y funcionalidades necesarias para almacenar, analizar y mostrar información geográfica, los componentes principales del software SIG son:

- ✓ Sistema de manejo de base de datos.
- ✓ Una interfaz gráfica de usuarios para el fácil acceso a las herramientas.

- ✓ Herramientas para la captura y manejo de la información geográfica.
- ✓ Herramientas para el soporte de consultas, análisis y visualización de datos geográficos.

Datos

El componente más importante de un SIG son los datos. Primero y principalmente se requiere de buenos datos de base. Lograr esto frecuentemente absorberá el 60-80% del presupuesto de implementación de un SIG. Asimismo, recolectar buenos datos de base es un proceso largo, que frecuentemente demora el desarrollo de productos que pueden utilizarse para justificar la inversión. La mayoría de los SIG emplean dos bases de datos, una para crear y mantener la información y la otra para ayudar a organizar y manejar los datos.

Recursos Humanos

La tecnología de SIG es de valor limitado y se necesita de un personal para el manejo del sistema. Los usuarios de SIG varían desde especialistas técnicos, que diseñan y mantienen el sistema, hasta aquellos que lo utilizan para ayudar a realizar sus tareas diarias.

Procedimientos

Para que un SIG tenga una implementación exitosa debe basarse en un buen diseño y reglas de actividad definidas, que son los modelos y las prácticas operativas exclusivas en cada organización.

Las modernas tecnologías SIG trabajan con información digital, para la cual existen varios métodos utilizados en la creación de datos digitales. El método más utilizado es la digitalización, donde a partir de un mapa impreso se transfiere a un medio digital por el empleo de un programa CAD con capacidades de georeferenciación.

Estándares

La definición de estándares, si bien es muy importante en muchas áreas sobre todo en el ámbito de las ingenierías, cobra una especial relevancia en el campo de los sistemas de información geográfica debido a su aplicación en muchas otras áreas por ejemplo en la arquitectura, la ingeniería civil, la gestión medioambiental, etc. Sin estos estándares cada desarrollador enfocaría los sistemas desde su propio punto de vista muy condicionado por el campo de aplicación del SIG específico. En el año 1994 se creó el Consorcio

Geoespacial (sus siglas en inglés, OGC) con el objetivo de fomentar la interoperabilidad entre las herramientas SIG que se desarrollaban.

Además del OGC, la Organización Internacional de Estandarización (ISO, del inglés Internacional Organization for Standardization) por medio del ISO/TC 211 (*ISO Technical Committee 211*) también ha participado activamente en el proceso de estandarización. Aunque originalmente estas dos organizaciones trabajaban de manera independiente, hoy en día trabajan de manera conjunta y con propósitos bien definidos y diferenciados. Mientras que el ISO/TC 211 trabaja en estándares estáticos, abstractos y de alto nivel, el OGC se centra en estándares orientados a la industria y la tecnología. Este esfuerzo colaborativo ha hecho posible que muchas organizaciones públicas estén participando en el desarrollo de infraestructuras de datos espaciales para compartir su información espacial.

La estandarización es de gran importancia por dos motivos. En primer lugar, porque de esta surgió una arquitectura para los SIG en la que se puede basar a la hora de definir la arquitectura para un sistema de recuperación de información geográfica. En segundo lugar, porque existen muchos puntos en común entre ambos campos, lo que nos permite utilizar en la arquitectura servicios estandarizados en el campo de los SIG (6).

Aplicaciones de los Sistemas de Información Geográfica

En la mayoría de los sectores los SIG pueden ser utilizados como una herramienta de ayuda a la gestión y toma de decisiones, algunos de ellos como:

Web

Los SIG son de gran uso en la web ya que al utilizar estos sistemas, los usuarios pueden acceder a través de la web para buscar todo tipo de bases de datos, archivos de textos, mapas; sin tener que preocuparse por su localización física en la red.

Infraestructuras

La elaboración de mapas, así como la posibilidad de elaborar otro diferente tipo de consulta, ya sea gráfica o alfanumérica, son las funciones más comunes para estos sistemas, también son utilizados en trabajos de ingeniería, inventarios, planificación de redes, gestión de mantenimiento, entre otros.

Empresas y negocios

El uso de los SIG para las aplicaciones comerciales es obvio, debido a que es de crucial importancia para cualquier negocio conocer dónde se encuentran los mercados potenciales.

Las aplicaciones operacionales incluyen, el uso de las funcionalidades de los SIG para supervisar la provisión de productos en una red de distribución. Para este propósito, los SIG serán utilizados para apoyar actividades diarias de rutina. Las aplicaciones tácticas proporcionan información requerida para la toma de decisiones.

Hoy día, mucho esfuerzo es dedicado a aplicar los SIG a los problemas tácticos. Es esencial integrar bases de datos internas y externas, para un óptimo proceso de toma de decisiones.

Sistemas de Recuperación de Información

A lo largo de estos años, son mucho los autores que han podido dar sus propios conceptos referentes a los Sistemas de Recuperación de Información. Los SRI se ocupan de encontrar documentos de naturaleza no estructurada (texto) que se encuentren en grandes colecciones normalmente almacenadas de forma digital para satisfacer necesidades de información.

Modelos Matemáticos de los Sistemas de Recuperación de Información

De forma simplificada, podemos considerar un modelo como un método para representar tanto documentos como consultas en SRI y comparar la similitud de esas representaciones. Para ello, los modelos de recuperación tienen que proporcionar de manera implícita o explícita una definición de relevancia.

Aunque no son los únicos, existen tres modelos muy utilizados para la recuperación de información, estos son: el modelo booleano, el modelo probabilístico y el modelo espacio vectorial. Estos modelos representan tanto los documentos como las consultas mediante un conjunto de términos clave que se emplean como términos de indexación.

Modelo Probabilístico

El modelo probabilístico define la recuperación como un proceso de clasificación, de tal forma que para cada consulta existen dos clases: la clase de los documentos relevantes y la clase de los documentos no relevantes. Por tanto, dado un documento *D* y una consulta determinada se puede calcular con qué probabilidad pertenece el documento a cada una de las clases. Si la probabilidad de que pertenezca a la clase de los documentos

relevantes es mayor que la probabilidad de que pertenezca a la clase de los no relevantes (es decir, $P(R|D) > P(NR|D)$) el documento será relevante para la consulta (7).

La probabilidad de que un documento sea relevante para una consulta dada, depende de la forma en que se representan los documentos en el sistema. Una vez calculada la probabilidad de los documentos relevantes para la consulta dada, son ordenados descendentemente y mostrados al usuario (8).

Modelo Booleano

El modelo booleano es uno de los modelos de recuperación de información más sencillos. Está basado en la teoría de conjuntos y el álgebra de Boole. Las consultas se formulan como expresiones lógicas que combinan mediante operadores del álgebra de Boole (por ejemplo, AND, OR, NOT) los términos clave de búsqueda. La idea en la que está basado este modelo es que un término clave puede estar presente o ausente en un documento y, por tanto, sólo serán relevantes aquellos documentos que contengan los términos clave que se indica en la consulta (6).

La indización se realiza asociando un peso binario a cada término del índice: 0 si el término no aparece en el documento y 1 si aparece al menos una sola vez. El grado de similitud entre un documento y una consulta también será binario y un documento será relevante cuando su grado de similitud sea igual a 1, de lo contrario, el documento no tendrá ninguna relevancia en cuanto a la consulta.

La principal ventaja de este modelo es su sencillez y su eficiencia. Por el contrario, su principal inconveniente es la dificultad para formular exactamente las necesidades de información de los usuarios empleando el álgebra de Boole y términos clave.

Modelo Espacio Vectorial

La idea fundamental en la que se basa el modelo vectorial es considerar que tanto la importancia de un término clave con respecto a un documento, como los documentos y las consultas se pueden representar como un vector en un espacio de alta dimensión. Por tanto, para evaluar la similitud entre un documento y una consulta; es decir, para obtener la relevancia del documento con respecto a la consulta, simplemente hay que realizar una comparación de los vectores que los representan. Uno de los métodos más habituales para comparar el grado de similitud es calcular el coseno del ángulo que forman ambos vectores. Cuanto más se parezcan los vectores, más próximo a cero grados será el ángulo que forman y, en consecuencia, más se aproximará a uno el coseno de ese ángulo. Esta forma de comparar los vectores se conoce como la fórmula del coseno (6).

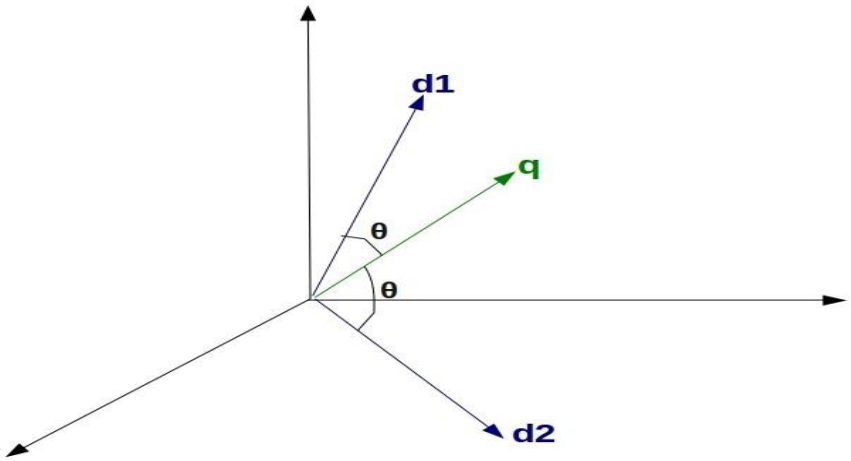


Figura1. Representación matemática del Modelo Espacio Vectorial.

Relación de los Sistemas de Recuperación de Información con otras Ciencias

La ciencia de la información, la informática, la documentación, la lógica, la inteligencia artificial, entre otras son algunas de las disciplinas que en mayor o menor medida contribuyen a la investigación teórica y aplicada de la recuperación de información.

Documentación

La recuperación de información dentro del proceso documental ha sido una de las tareas de más importancia desde el punto de vista del usuario, esta se encarga de informar al receptor sobre dónde se halla la información que necesita para generar nuevo conocimiento, contribuyendo de este modo al avance científico en particular y al progreso en general.

Muchos son los autores como Walker, Harter y Lancaster que han podido dar definiciones generales del término recuperación de información dentro del contexto de la ciencia de la documentación y todos han podido concluir que es la aplicación del conjunto de técnicas, métodos, y actividades para buscar, localizar y recuperar de una manera eficiente en los diversos SRI la información relevante que requiere el usuario, y satisfacer de esta forma su necesidad de información (9).

Sistemas Operativos y Redes

Los sistemas operativos y las redes han contribuido a que los ordenadores sean cada vez más potentes, rápidos, baratos y a que se hayan convertido en una herramienta personal

de trabajo utilizada en mucho ocasiones para la búsqueda y recuperación de información, lo que facilita el proceso de comunicación y difusión de la información.

Informática

Desde un principio, la recuperación de información ha estado ligada a la ciencia de la informática no sólo por el uso de ordenadores y de las tecnologías de la información como herramientas de trabajo, sino también porque una parte importante de la investigación ha estado orientado al diseño de mejores sistemas de recuperación de información. Para Baeza-Yates, el problema de la recuperación de información desde el punto de vista de la informática consiste principalmente en diseñar y construir índices eficientes, para el procesamiento de las consultas de los usuarios con un alto rendimiento, y en desarrollar algoritmos de rango que mejoren la calidad de los resultados obtenidos (10).

Sistemas de Recuperación de Información Geográficas

Los Sistemas de Recuperación de Información Geográfica (SRIG) han aparecido recientemente con la intersección de dos áreas más consolidadas como son: los sistemas de recuperación de información y los sistemas de información geográfica. Esta intersección tiene como objetivo que los usuarios puedan realizar consultas sobre colecciones de documentos teniendo en cuenta tanto su ámbito textual como su ámbito espacial.

El objetivo principal de los sistemas de recuperación de información geográfica, es definir estructuras de indexación y técnicas para almacenar y recuperar documentos de manera eficiente empleando tanto las referencias textuales como las referencias geográficas contenidas en el texto (11).

Un aspecto fundamental en un SRIG es la detección de las entidades geográficas presentes tanto en el texto de la colección como en la consulta de usuario.

Arquitectura de un SRI

Un sistema de recuperación de información se define como el proceso donde se accede a una información previamente almacenada, mediante herramientas informáticas que permiten establecer ecuaciones de búsquedas específicas (12).

Generalmente, todos los sistemas de recuperación de información comparten una misma arquitectura, la cual se detalla a continuación (13).

Interfaz: un usuario con necesidades de información bien definidas, interactúa con la interfaz del sistema, mediante la cual introduce las consultas al mismo. La interfaz puede

estar basada en una interfaz web, una interfaz de escritorio o ambas.

Sistema de Formulación de Consultas: realiza un preprocesamiento de las consultas trasladando las consultas hechas en el lenguaje natural a consultas entendibles por los sistemas de información.

Mecanismo de evaluación de consultas: compara los documentos representados en el sistema de información con la consulta reprocesada para obtener un subconjunto de documentos relevantes que satisfagan la consultas por el usuario, ordenados estos de acuerdo a un criterio de relevancia.

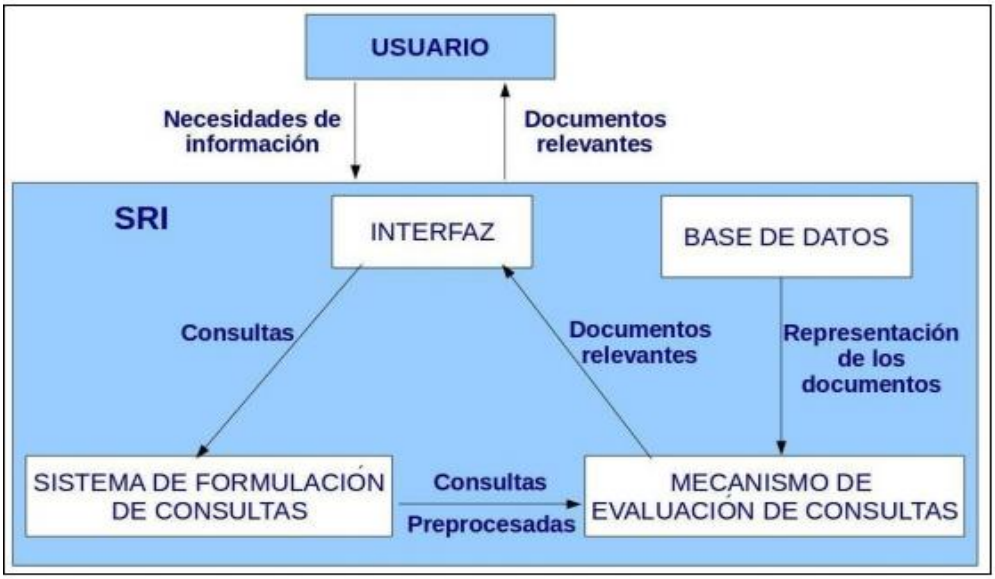


Figura2. Arquitectura de un Sistema de Recuperación de Información.

Herramientas utilizadas en la creación de los SRIG

AJAX

AJAX[1] es una técnica de desarrollo web para crear aplicaciones interactivas. Mantiene una comunicación asíncrona con el servidor en segundo plano. Esto significa aumentar la interactividad, velocidad y usabilidad en el sistema (11).

[1] AJAX: Asynchronous JavaScript And XML- JavaScript asíncrono y XML

Servidor de Mapas

Le permiten al usuario la máxima interacción con la información geográfica. Un servidor de mapas funciona enviando a petición del cliente, desde su navegador de internet, una serie de páginas HTML, con una cartografía asociada en formato de imagen. Las primeras versiones de servidores de mapas sólo permitían realizar funciones básicas de visualización y consultas alfanuméricas simples. En las versiones más recientes es posible realizar funciones mucho más avanzadas.
Como ejemplo de servidores de mapas se encuentran MapServer y GeoServer.

MapServer

MapServer es un servidor de desarrollo de código abierto para construir aplicaciones de internet espaciales. Permite crear "imágenes de mapas geográficos": mapas que pueden dirigir a los usuarios hacia el contenido y es multiplataforma.
Este servidor puede ser invocado desde páginas web para generar de forma dinámica imágenes en los formatos más habituales para la publicación en la web como gif y png. Soporta gran cantidad de formatos tanto vectoriales como raster (14).

GeoServer

Servidor de código abierto escrito en Java. Pretende operar como un nodo a través de una infraestructura de datos espaciales libre y abierta para ofrecer datos geoespaciales (6).

Funciones que permiten realizar los servidores de mapas

- ✓ Visualización: para alejar o acercar los elementos cartográficos.
- ✓ Identificación de atributos alfanuméricos en cada elemento cartográfico.
- ✓ Consultas de atributos alfanuméricos: sencillas, como la búsqueda de topónimos o más complejas, con operadores booleanos.
- ✓ Conexión de bases de datos locales a la base de datos remota del servidor de mapas, de cara a la creación de mapas temáticos con datos alfanuméricos propios.
- ✓ Cálculo de rutas óptimas para la navegación de vehículos. Capacidad de imprimir mapas manteniendo la escala.

Open Street Map (OSM)

OSM es un proyecto colaborativo para crear mapas libres y editables. A partir de los datos de este proyecto se pueden producir mapas de carreteras, mapas náuticos, etc. También este proyecto se aplica para el cálculo de las rutas óptimas para vehículos y peatones.
A medida que el proyecto ha ido madurando y su base de datos ha mejorado rápidamente en calidad y cobertura, ha ido surgiendo a su alrededor todo un ecosistema de herramientas informáticas y servicios, convirtiéndose en una fuente de datos factible para determinados proyectos complejos que hacen uso de estos datos de una forma creativa, productiva o inesperada.

DIGMAP

DIGMAP quiere decir “Descubrir nuestro mundo pasado a través de mapas digitalizados”.
Es un proyecto investigativo que propone el desarrollo de soluciones para bibliotecas digitales. DIGMAP tiene el propósito de convertirse en la principal fuente internacional y servicio de referencia para mapas antiguos y bibliografía relacionada (15).
El proyecto consiste en un conjunto de servicios disponibles en Internet, desarrollados con software libre que pueden ser reutilizados por otros servicios. Desarrolla ambientes sofisticados de navegación y búsqueda para los usuarios, como también servicios de interoperabilidad con sistemas de información externos.
Este proyecto propone características avanzadas para la indexación automática de mapas históricos (16).

Herramientas para el trabajo con Bases de Datos

PostgreSQL

Servidor de base de datos objeto- relacional, no necesita de licencias de software, es multiplataforma.
Los sistemas de bases de datos relacionales soportan un modelo de datos que consisten en una colección de relaciones con nombre, que contienen atributos de un tipo específico (17).
PostgreSQL utiliza un modelo cliente-servidor y usa multiprocesos para garantizar la estabilidad del sistema. Un fallo en uno de los procesos no afectará el resto y el sistema continuará funcionando (18).
Tiene soporte total para transacciones, vistas, procedimientos almacenados y almacenamiento de objetos de gran tamaño. Consume bastante recursos y carga de

mejor forma el sistema, presenta gran potencia y flexibilidad. Tiene mejor soporte para vistas y procedimientos almacenados en el servidor, además tiene ciertas características orientadas a objetos. Se destaca en ejecutar consultas complejas, subconsultas y joins de gran tamaño. Permite la duplicación de bases de datos maestras en múltiples sitios de réplica y cuenta con funcionalidades de compatibilidad para ayudar en la transición desde otros sistemas menos compatibles como SQL.

PostgreSQL es la mejor opción en sistemas en las que la consistencia de datos sea fundamental con información realmente importante como por ejemplo los bancos, pese a su mayor lentitud.

PostGIS

Extensión al sistema de base de datos objeto-relacional PostgreSQL. Creado como un proyecto de investigación de tecnologías de bases de datos espaciales.

PostGIS habilita el servidor PostgreSQL, permitiendo que sea utilizado como un servidor de datos espacial para los sistemas de información geográfica (19).

Con PostGIS se puede usar todos los objetos que aparecen en la especificación OpenGIS como puntos, líneas, polígonos, multilíneas, multipuntos, y colecciones geométricas.

Lenguaje de Programación que utilizan los SRIG

Lenguaje R

Es un lenguaje de programación interpretado como Java y no compilado como C, lo cual significa que los comandos escritos en el teclado son ejecutados directamente sin necesidad de construir ejecutables (20).

Proporciona un amplio abanico de herramientas estadísticas (algoritmos de clasificación y agrupamiento) y gráficas. R se distribuye gratuitamente bajo los términos de *General Public Licence (GNU)*. El código fuente está disponible en varias formas, principalmente en C (21).

Está disponible para los sistemas operativos Windows, Unix, GNU/Linux, entre otros. R incluye muchas funciones que pueden ser utilizados para la lectura, visualización y el análisis de datos espaciales. Se enfoca en los datos espaciales, donde las observaciones pueden ser identificadas como localizaciones geográficas y donde se pueden adicionar información referente a algunos lugares, si la ubicación se registra con cuidado.

El lenguaje R permite al usuario, programar bucles para analizar conjuntos sucesivos de datos. También es posible combinar en un solo programa diferentes funciones

estadísticas para realizar análisis más complejos. R está orientado a objeto lo que significa que las variables, datos, funciones y resultados se guardan en la memoria activa del computador en forma de objetos con un nombre específico (20).

Librerías y Motores de búsqueda utilizadas para crear SRIG

RMAP

La biblioteca o paquete RMAP utiliza objetos de R para el acceso general a las fuentes de datos geográficos. Estas fuentes de datos pueden ser simples objetos de investigación, o potencialmente objetos que obtienen sus datos de bases de datos o servidores.

Proj4

Proj4 es una librería utilizada por multitud de aplicaciones SIG como PosGiS, MapServer. Las proyecciones cartográficas se manejan utilizando la biblioteca Proj4. No presenta ningún tipo de páginas web personalizadas.

Geospatial Data Abstraction Library (GDAL)

GDAL es una librería de software para la lectura y escritura de formatos de datos geoespaciales. Presenta un único modelo abstracto de datos que utiliza para acceder a todos los formatos que soporta y es multiplataforma. Contiene una variedad de utilidades en línea de comando para la traducción y el proceso de datos geoespaciales. Utiliza el software MapServer. GDAL provee una serie de métodos y rutinas a los desarrolladores de software, para la manipulación y el procesamiento de datos geoespaciales, como por ejemplo conversión de datos en formato raster, reproyección de imágenes, redimensionamiento de imágenes, conversión de datos en formato vectorial, reproyección de datos vectoriales, entre otros.

Como biblioteca presenta un único modelo de datos abstractos para la aplicación de llamadas para todos los formatos soportados (22).

Alejandria Digital Library (ADL)

ADL ofrece un fácil acceso a las colecciones de materiales de referencias geográficas. La colección ADL proporciona capacidades de búsqueda de los materiales geográficamente referenciados. La colección ADL contiene información que apoya la ciencia básica, incluyendo la Tierra y las ciencias sociales. La colección ADL ofrece búsqueda geoespaciales como el mecanismo de búsqueda principal. Esta colección contiene fotografías, modelos de elevación digital, gráficos digitales raster, los mapas de imágenes de satélite y el mundo. ADL ofrece a los clientes HTML tener acceso a sus colecciones y

nomencladores, y herramientas específicas de gestión de la información (23).

Apache Lucene

Librería de código abierto utilizada para la creación de sistemas de recuperación de información, escrito en Java. Permite a los desarrolladores (Java, .Net, C++) integrar funciones de indexación y búsquedas de información textual dentro de sus proyectos. Durante muchos años ha sido la librería gratuita más utilizada para recuperar información. Lucene siendo una librería, no es una aplicación finalizada lista para ser utilizada, como lo pudiera ser un programa buscador de archivos o un buscador de sitios Web (24).

Básicamente Lucene se divide en dos grandes partes, la indexación y la búsqueda. En la indexación Lucene crea un índice sobre el que luego realiza la búsqueda, y el cual construye mediante el análisis de la información que le llega. En el proceso de búsqueda Lucene usa el índice que ha construido anteriormente para devolver la información que coincida (25).

En el centro de la arquitectura lógica de Lucene se encuentra el concepto de Documento que contiene campos de texto. Esta flexibilidad permite a Lucene ser independiente del formato del fichero. Es útil para cualquier aplicación que requiera indexado y búsqueda de texto completo (26).

Solr

Es una plataforma de búsquedas basada en Apache Lucene, que funciona como un "servidor de búsquedas". Sus principales características incluyen búsquedas de texto completo, resaltado de resultados y manejo de documentos (como Word y PDF). Solr es escalable, permitiendo realizar búsquedas distribuidas y replicación de índices, y actualmente se está usando en muchos de los sitios más grandes de Internet (27).

Está escrito en Java, pero se puede usar en cualquier lenguaje, simplemente usando las peticiones GET para realizar búsquedas en el índice, y POST para agregar documentos. Fácil de configurar y usar. La principal característica de Solr es su API estilo REST[2], ya que en vez de usar drivers para comunicarse con Solr se puede realizar peticiones HTTP y obtener resultados en XML[3] o JSON[4]. Además, los XML son legibles por personas, y los JSON se pueden usar para realizar pruebas (28).

Como la interfaz principal de Solr es web, necesita un contenedor de servlets. Este

[2] REST: Representational State Transfer- Transferencia de Estado Representacional
[3] XML: Extensible Markup Language- Lenguaje de etiquetado extensible
[4] JSON: JavaScript Object Notation - Notación de Objetos de JavaScript

servidor de búsqueda tiene varios servlets que exponen distintos servicios, como UPDATE o SELECT (29).

SOLR utiliza la biblioteca de Java Lucene, para la indexación de texto completo y de búsqueda.

También posee las siguientes propiedades:

- ✓ La replicación es independiente del sistema operativo para que pueda reproducir el mismo índice en una plataforma Windows, así como en Linux.
- ✓ Los índices se puede configurar para operar en sincronía: Si alguno de los índices se modifica, las demás copias de dicho índice se actualiza automáticamente.
- ✓ SOLR apoya la búsqueda de indexación distribuida por la realización de diversas máquinas y la fusión de los resultados.
- ✓ SOLR puede indexar y realizar búsquedas al mismo tiempo.
- ✓ Cuenta con un esquema basado en XML para la gestión de campos indexados.

Conclusiones parciales

En este capítulo se presentaron los conceptos necesarios para el desarrollo de este trabajo. Se realizó un estudio en el que se demuestra que la arquitectura que comparten todos los sistemas de recuperación de información está compuesta por una interfaz, un sistema de formulación de consultas y un mecanismo de evaluación de consultas. Se definieron como herramientas para la construcción de un SRIG, pare la creación de aplicaciones iterativas se utilizó AJAX, Solr como servidor de búsqueda de información, para la presentación de la información geográfica el servidor de mapas MapServer y para el trabajar con la información tanto textual como espacial el sistema de base de datos PostgreSQL y la extensión a este sistema PosGIS.

CAPÍTULO 2: ARQUITECTURA PROPUESTA Y VALIDACION

Introducción

En este capítulo se describe la arquitectura lógica que se propone para el desarrollo de sistemas de recuperación de información geográfica y los componentes por los que está formada. En primer lugar, se expone una definición de lo que es una arquitectura de software. También se hace referencia en este capítulo a proyectos y concursos que utilizan los SRIG para recuperar cualquier tipo de información tanto espacial como textual. Se presenta el funcionamiento de la arquitectura mediante la descripción de la interacción de las distintas tecnologías y se muestra la interacción entre las herramientas seleccionadas en el capítulo anterior. Por último se realiza una validación de la arquitectura lógica que se propone.

Arquitectura de software

La arquitectura de software es la organización fundamental de un sistema encarnada en sus componentes, las relaciones entre ellos y el ambiente y los principios que orientan su diseño y evolución (30).

Conjunto de patrones y abstracciones coherentes que proporcionan el marco de referencia necesario para guiar la construcción del software para un sistema de información.

La principal motivación para utilizar una arquitectura de software es que una buena arquitectura es como un buen diseño, si el tamaño y la complejidad del sistema aumenta la estructura del sistema se va a hacer más importante que la selección del algoritmo o la estructura de datos apropiada. Un buen diseño de la estructura que soportará el sistema traerá como consecuencia un mejor comportamiento y funcionalidad del mismo. Sin embargo un mal diseño arquitectónico puede anonadar la ejecución del mejor de los algoritmos. Además de que un diseño pobre del algoritmo y la estructura de datos puede dejar cojo una buena arquitectura. Una buena arquitectura de software ayuda a identificar puntos de cómputo y cuellos de botella en los flujos de datos, así que esto permite describir las características del algoritmo y la estructura de datos adecuados para el mejor funcionamiento del sistema (31).

Soluciones actuales que utilizan los Sistemas de Recuperación de Información Geográfica

SPIRIT es un proyecto desarrollado para la localización de nombres de lugar. Este

proyecto propone trabajar con *Text-First* y *Geo-First*, para la indexación.

Primero se emplea un índice para filtrar los documentos (el índice invertido en el Text-First y el índice espacial en el Geo-First). El conjunto de documentos resultante es ordenado por sus identificadores y posteriormente filtrado usando los índices mencionados anteriormente (11).

En SPIRIT, cada referencia geográfica mencionada en el texto de un documento se almacena en su cuadro delimitador; es decir, el referente de cada documento está formado por varios cuadros delimitadores.
Uno de los proyectos pioneros, anterior incluso al proyecto SPIRIT, en cuanto a la georeferenciación de documentos contenidos en bibliotecas digitales es GIPSY (*Georeferenced Information Processing System*). En este proyecto, las referencias geográficas encontradas en los textos se convierten en representaciones geométricas (por ejemplo, en polígonos) y a todas esas representaciones se les asigna un valor de ponderación en función de propiedades derivadas del contenido de los documentos (por ejemplo, de la frecuencia de aparición del término). Luego se agregan los polígonos construyendo representaciones topográficas en tres dimensiones teniendo en cuenta los factores de ponderación y se establece un umbral para determinar la elevación mínima que determina que el área es relevante (6).

A nivel mundial se realizan congresos y foros de evaluación GeoCLEF. La tarea más importante definida en GeoCLEF es la de buscar información geográfica. GeoCLEF no solo evalúa sistemas de búsqueda de información geográfica, sino que también propone nuevas subtareas, como la de análisis de consultas, cuyo objetivo es identificar aspectos geográficos en una consulta (32).

Arquitectura Propuesta

Interfaz

El sistema debe contar con dos interfaces de usuario diferentes: una interfaz de usuario de administración, destinada a realizar el mantenimiento de la estructura de indexación y del sistema en general, y la interfaz de usuario de consulta, empleada por los usuarios para realizar consultas al sistema e interactuar con los resultados (6).

Modelos de recuperación de información

Los modelos a utilizar son el booleano, probabilístico y vectorial, mencionados en el capítulo 1.

Servicio de Gazetteer

Es un diccionario geográfico que contiene, además de nombres de lugar, otros nombres alternativos, poblaciones, localizaciones de lugares y otra información relacionada con el lugar. También se conoce como un recurso externo para obtener información geográfica.

Servicio de Ontología del Espacio geográfico

Especificación explícita y formal de una conceptualización compartida. Este servicio proporciona un vocabulario de clases y relaciones para describir un ámbito determinado y es utilizado en la creación de un índice espacial.

Servidores de Mapas

Permite generar imágenes png, jpeg, gif; que representan los mapas definidos a partir de la información geográfica disponible en las fuentes de datos a las que tiene acceso el servicio.

Mecanismo de evaluación de consultas

Se encarga de recibir las consultas y de resolverlas empleando para ello la estructura de indexación (6).

Construcción del índice

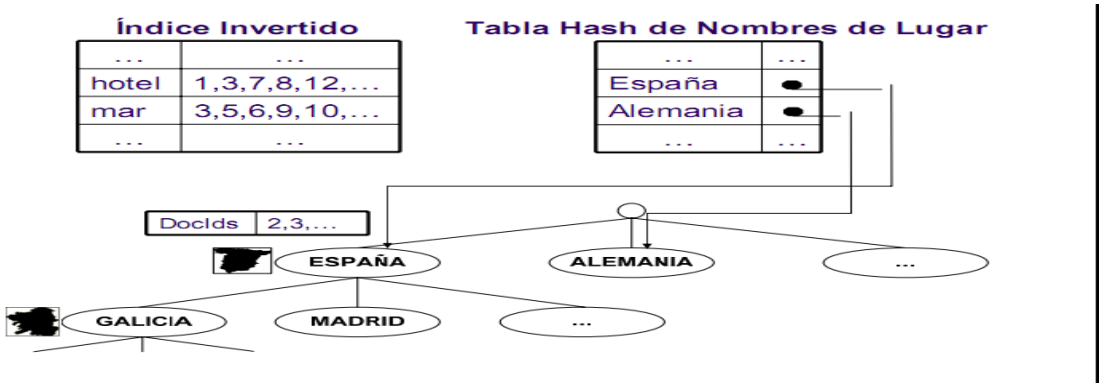


Figura 3. Estructuras para la construcción del índice.

Se emplean dos estructuras auxiliares. Primero, una tabla hash de nombre de lugares que almacena para cada nombre de lugar su posición en la estructura de indexación.

La segunda estructura auxiliar es un índice invertido, esta estructura asocia a cada palabra en el texto (organizado como un vocabulario) una lista de puntos a los documentos donde aparecen. El principal inconveniente de esta técnica es que ignora por completo las referencias geográficas. Los nombres de lugar son considerados simplemente como palabras (11).

Procesamiento del Lenguaje Natural (PLN)

Es una técnica que utiliza para la reducción de grupos de nombres, un extractor de raíces (en inglés stemmers), para eliminar las palabras vacías (stopwords) una lista de palabras sin contenido semántico y un Reconocedor de Entidades (*Named Entity Recognizer*, *NER*) para detectar y reconocer posibles entidades en cualquier texto.

Extracción de palabras claves

Para reducir el número de palabras claves, se pueden emplear las técnicas clásicas de recuperación de información mencionadas con anterioridad.

Estructura de Indexación

El componente principal de la estructura de indexación es un árbol compuesto por nodos que representan nombres de lugar. Estos nodos están interconectados por medio de relaciones de contenido. En cada nodo se almacena: (i) la palabra clave (un nombre de lugar), (ii) las referencias geográficas asociadas con el nombre de lugar, (iii) un delimitador de rectángulo mínimo de la geometría que representa ese lugar, (iv) una lista con los identificadores de los documentos que incluyen referencias geográficas a ese lugar y (v) una lista de nodos hijos que estén geográficamente contenidos en ese nodo. Cada nodo emplea un R-tree (árbol balanceado) para mejorar el rendimiento a la hora de acceder a los nodos hijos. R-tree en uno de los índices espaciales clásicos más empleados para mejorar la eficiencia en sistemas de información geográfica.

Bases de datos

Deben existir dos bases de datos, una geográfica y otra documental, en la primera para almacenar todas las informaciones de forma geográfica por ejemplo mapas y en la segunda para almacenar cualquier tipo de información de forma textual.

En la siguiente figura se muestra una propuesta arquitectónica de un sistema de recuperación de información geográfica. La arquitectura se divide en tres capas: el flujo de trabajo para la construcción del índice, los servicios de procesado y las interfaces de usuario. La capa inferior se puede definir de forma simplificada como un proceso que recibe como entrada una colección de documentos y produce como resultado una estructura de indexación. El primer paso de este flujo de trabajo es la tarea Extracción de palabras clave donde todos los documentos son analizados y se extraen las palabras clave del texto. En esta tarea se preparan los documentos para su procesado en el resto del flujo de trabajo del sistema y se almacenan en base de datos para que puedan ser consultados posteriormente. El resultado final del proceso, además de la estructura de indexación, contiene información almacenada en las bases de datos documental y geográfica que permite mejorar la calidad de los resultados presentados a los usuarios.

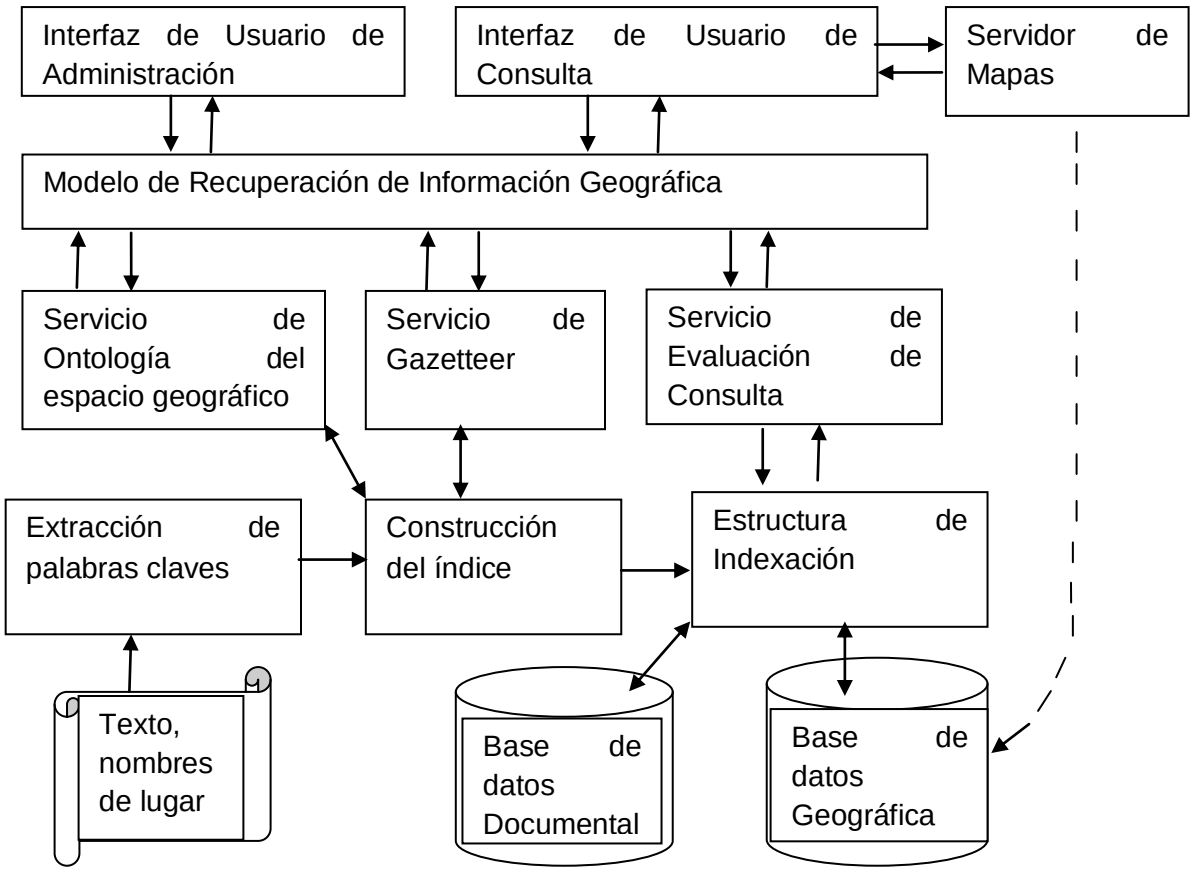


Figura 4. Arquitectura Lógica Propuesta.

En la derecha se puede observar el servicio de evaluación de consultas, que es el

componente que se encarga de recibir consultas y emplea la estructura de indexación para resolverlas.

Para la construcción del índice se utiliza el servicio de ontología del espacio geográfico y el servidor de gazetteer. Por encima de estos servicios se sitúa un módulo de recuperación de información geográfica encargado de coordinar la tarea efectuada por cada servicio en respuesta a las peticiones del usuario. La capa superior de la arquitectura muestra las diferentes interfaces de usuario mencionadas anteriormente.

La interfaz de usuario de consulta tiene un acceso directo a un servidor de mapas. El servidor de mapas tiene dos objetivos fundamentales. El primer objetivo es proporcionar los mapas interactivos para la interfaz de usuario de consulta, esto permite que el usuario pueda navegar por la cartografía que proporciona el sistema buscando zonas de interés para su consulta. El segundo objetivo es la creación de representaciones cartográficas de los resultados de las consultas (6).

Para la construcción del índice se emplean dos estructuras auxiliares. Mantener separados el índice textual del índice espacial tiene muchas ventajas. En primer lugar, todas las consultas textuales pueden ser procesadas de manera eficiente por el índice invertido y todas las consultas espaciales pueden ser procesadas de manera eficiente por el índice espacial. Además, el sistema soporta consultas que combinen aspectos textuales con espaciales. Así mismo se pueden manejar de manera independiente las actualizaciones en cada uno de los índices, esto hace que se puedan añadir o eliminar datos de forma sencilla. Finalmente, se pueden aplicar optimizaciones específicas a cada estructura de indexación de manera individual.

Tipos de consultas soportadas

La característica más importante de una estructura de indexación es el tipo de consultas que pueden resolver. Los siguientes tipos de consultas son relevantes en un sistema de recuperación de información geográfica:

- ✓ Consultas puramente espaciales. Un ejemplo de este tipo de consultas es recuperar todos los documentos que se refieran a la siguiente área geográfica". El área geográfica en la consulta puede ser un punto, una ventana de consulta, o incluso un objeto complejo como un polígono.

- ✓ Consultas puramente textuales. Estas son consultas del tipo recuperar todos los documentos donde aparezcan las palabras hotel y mar".

Para realizar una consulta el usuario interactúa con el sistema mediante la interfaz de usuario de consulta, esta interfaz le permite al usuario indicar las palabras claves a buscar en la consulta, y también navegar por el mapa para visualizar el área de interés para la consulta. Esta navegación se presenta mediante la petición de mapa que, aunque aparece una única vez, se realiza para cada movimiento en el mapa, para cada ajuste del nivel zoom. Las peticiones al servidor de mapas provocan que este componente tenga que recuperar la información geográfica necesaria de las fuentes de datos.

Una vez que el usuario tiene seleccionada el área de interés en el mapa puede realizar una consulta. Esta consulta la recibe el módulo recuperación de información geográfica, que es el encargado de coordinar todos los servicios que ofrece el sistema y conoce en qué componente debe delegar la petición. El componente encargado de resolver la consulta como se mencionaba anteriormente es el servicio de evaluación de consulta, este servicio delega en la estructura de indexación la resolución de consultas. Luego la estructura de indexación divide la consulta en un componente textual para resolver las consultas textuales y un componente espacial, que resuelve la parte espacial de la estructura. Una vez obtenidos ambos resultados los combina en una única lista de resultados. Esta es la lista que devuelve como resultado el servicio de evaluación de consulta, que es el encargado de poner los elementos de la lista en un orden según la importancia que representa el documento para el usuario de acuerdo a la consulta que realizó. Finalmente esta lista ordenada llega al módulo de recuperación de información geográfica que se lo envía al usuario para que pueda analizarla e interactuar con ella. El usuario puede consultar cualquiera de los documentos incluidos en la lista mediante el módulo de recuperación de información geográfica. Dicho módulo se encarga de recuperar el documento solicitado de la fuente de datos donde se encuentra almacenado. Además el usuario puede representar los documentos georeferenciados en el mapa mediante peticiones al servidor de mapas. Estos dos tipos de peticiones se realizan de forma asíncrona. Esto es posible debido a que en la interfaz de usuario se emplea la tecnología AJAX.

A continuación se muestra una figura de cómo interactúan las herramientas seleccionadas en el capítulo 2 teniendo en cuenta la arquitectura propuesta en la figura mostrada anteriormente.

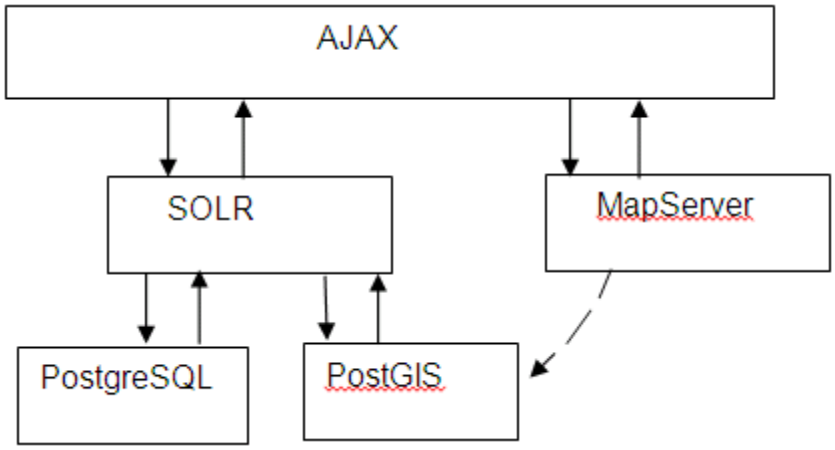


Figura 5. Interacción entre herramientas.

Validación de la Arquitectura

Para validar la arquitectura lógica propuesta se realizarán entrevistas a varios integrantes del **Departamento de Geoinformática.**

El departamento de Geoinformática surge en enero de 2010 como parte del centro de desarrollo "Geoinformática y Señales Digitales".

Este centro tiene la **misión** de desarrollar productos, servicios y soluciones informáticas en el campo de la Geoinformática, contribuyendo a la formación integral de profesionales que respondan a las necesidades del progreso científico-técnico y socioeconómico permitiendo un posicionamiento en el mercado nacional.

Resultados de las Encuestas

Después de realizar las encuestas, la propuesta arquitectónica fue aceptada, por el Ing. Romanuel Ramón Artunéz, el Ing.Yuniel Eliades Proeza, el Ing. Odiel Estrada Molina y por el Msc. David Silva Barrera, donde este último firmó un acta aceptando la arquitectura lógica propuesta.

En los anexos se encuentra un Acta de Aceptación, donde el Msc. David Silva Barrera realizó un resumen aceptando la arquitectura lógica de un Sistema de Recuperación de Información Geográfica, para el motor de búsqueda Orión.

Conclusiones parciales

En este capítulo se expuso la solución arquitectónica propuesta para el desarrollo de un sistema de recuperación de información geográfica. Su composición separada por componentes para una mejor relación entre las tecnologías que integran esta arquitectura. El aspecto más importante de esta arquitectura, es la comunicación entre las capas. Se obtuvo una interacción entre las herramientas seleccionadas para la construcción de un SRIG mencionadas en el capítulo anterior. Por último para validar la arquitectura propuesta, se realizaron entrevista a varios integrantes del Centro de Geoinformática, de ésta forma se validó la arquitectura propuesta. Además uno de los integrantes de Geoinformática, firmó un acta aceptando la arquitectura lógica de un Sistema de Recuperación de Información Geográfica presentada en este capítulo.

CONCLUSIONES

Con la investigación realizada se propuso una arquitectura para los Sistemas de Recuperación de Información Geográfica para resolver las necesidades de que el motor de búsqueda pueda contar con una arquitectura que permita a los usuarios realizar consultas tanto textuales como espaciales. Se describió la solución propuesta basándose en las principales definiciones arquitectónicas descritas en la Fundamentación Teórica. Para esta solución se seleccionó el trabajo con el servidor de búsqueda Solr, para la presentación de la información geográfica MapServer, para el acceso a los datos el sistema de base de datos PostgreSQL y la extensión a este sistema PosGIS. Para validar la arquitectura propuesta se realizaron encuestas, donde varios especialistas del departamento de Geoinformática estuvieron en total acuerdo con la arquitectura lógica propuesta en éste trabajo.

RECOMENDACIONES

Se recomienda desarrollar un Sistema de Recuperación de Información Geográfica que ayude a los usuarios a buscar información tanto textual como geográfica en la web. Se recomienda además desarrollar un servicio de ontología del espacio geográfico y un servicio gazetteer para un mejor funcionamiento del sistema mencionado anteriormente.

REFERENCIAS BIBLIOGRAFICAS

1. **Delgado, Yusniel Hidalgo.** *Orión, un motor de búsqueda para la web de la UCI.* 2010.

2. **Téllez, Esperanza Ayuga.** *S.I.G. Definiciones Básicas.* 2008.

3. Cursos Abiertos de la UNED. [En línea] //ocw.innova.uned.es.

4. **Escobar, Dr F.** Introducción a los SIG. [En línea] //www.sli.unimelb.edu.au/gisweb/.

5. SIG . [En línea] //www.fcagr.unr.edu.ar/mdt/GTS/Zonaedu/GIS3htm.htm.

6. **Naveiras, Diego Seco.** *Técnicas de indexación y recuperación de documentos utilizando referencias geográficas y textuales* . 2009.

7. **Jones, S.E. Robertson y S.K.** *Relevance weighting of search terms. Journal of the American Society for Information Science.*

8. **Comeche, Juan. A Martínez.** *Los modelos clásicos de recuperación de información y su vigencia.* 2008.

9. **José Antonio Salvador Oliván, Rosario Arquero Avilés.** *Una aproximación al concepto de recuperación de información en el marco de la ciencia de la documentación.*

10. **Sergio Gómez F, Diego Avella, Lorena Lobo Leguizamón.** Programa de Sistemas de Información: Tecnología de Redes . [En línea] //recuperaciondinformacion.blogspot.com.

11. **Miguel R. Luaces, Jose R. Paramá, Oscar Pedreira y Diego Seco.** *Una estructura de indexación para la recuperación de documentos con referencias geográficas.*

12. **Pinto, María.** Búsqueda y Recuperación de Información. [En línea] 2004. //www.mariapinto.es/e-coms/recu_infor.htm#ri1.

13. **Herrera, Antonio .G López.** Modelos de Sistemas de Recuperación de Información Lingüística Difusa. [En línea] 2006.

14. **Roldán, Méndez.** *El servidor de mapas MapServer, una solución recomendada para la representación de Información Geoespacial* .

15. **Santos, Filipe.** DIGMAP. 2007. //www.digmap.eu/doku.php?id=wiki:project_flyer:spanish.

16. **Borbinha, José.** DIGMAP. [En línea] //www.digmap.eu.

17. **Lockhart, Thomas.** *Tutorial de PostgreSQL.*

18. **Martínez, Rafael.** PostgreSQL. [En línea] www.postgresql.org.es .

19. PostGIS. [En línea] 2011. //postgis.refractions.net.

20. **Paradis, Enmanuel.** *R para Principiantes.* 2003.

21. El lenguaje R y su entorno grafico . [En línea] www.casado-d.org/edu/instalarR.html.

22. Geospatial Data Abstraction Library (GDAL). [En línea] //www.freegis.org.

23. Alejandria Digital Library. [En línea] //www.alexandria.ucsb.edu/.

24. **Vadim Paz Madrid Gorelov, Ángel F. Zazo, Carlos G. Figuerola, José Luis Alonso Berrocal.** *Librerías Lucene y dotLucene para Recuperación de Información. Estudio y desarrollo de casos prácticos.* 2007.

25. **Plunchete.** Introducción a Apache Lucene (Java). [En línea] //es.debugmodeon.com.

26. **Torres, Victor Antonio.** Hipasec Sistemas. [En línea] //www.hispasec.com/unaaldia/3993.

27. **Seta, Leonardo De.** Apache Solr: una introducción . [En línea] //www.dosideas.com.

28. **Suárez, Jose Manuel Sánchez.** Introduccion a Apache Solar. [En línea] //www.adictosaltrabajo.com/tutoriales.

29. Apache Solr. [En línea] //lucene.apache.org/solr/.

30. **Reynoso, Carlos Billy.** *Introducción a la Arquitectura de Software.* 2004.

31. **Rodriguez, Jose Antolio Cobo.** *Línea Base Arquitectónica para el Polo Sistemas Tributarios y de Aduanas.* 2008.

32. **José Manuel Perea Ortega, Miguel Angel García Cumbreras, Manuel García Vega, L. Alfonso Ureña López.** *Sistemas de Recuperación de Información Geográfica multilingües en CLEF* .

ANEXOS