INSTITUTO POLITÉCNICO NACIONAL

CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

LABORATORIO DE PROCESAMIENTO INTELIGENTE DE INFORMACIÓN GEOESPACIAL

"RECUPERACIÓN CONTROLADA DE INFORMACIÓN CUALITATIVA DESDE REPOSITORIOS DE DATOS"

TESIS

QUE PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

WALTER RENTERIA AGUALIMPIA

DIRECTOR DE TESIS:

DR. SERGUEI LEVACHKINE

México D. F. Diciembre, 2009

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

El desarrollo de la web semántica geoespacial requiere mecanismos de búsqueda que superen las comparaciones sintácticas y realicen análisis semántico, para recuperar información conceptualmente similar, esto permitiría reducir el riesgo de resultados vacíos cuando no hay correspondencia exacta entre la consulta y la información existente en los repositorios de datos. En este trabajo de investigación se presenta un modelo de recuperación de información, que integra un criterio semántico con criterios geoespaciales en un espacio no homogéneo; los criterios representan las dimensiones de este espacio, estas dimensiones son ponderadas de acuerdo a las preferencias de cada perfil de usuario. Para el proceso de integración se ha propuesto una expresión matemática que evalúa la relevancia de cada resultado. Se presenta un sistema que implementa el enfoque semántico geoespacial propuesto para recuperar información del dominio del turismo cultural, específicamente museos. Los resultados muestran las ventajas de integrar criterios semánticos y geoespaciales tomando en cuenta los perfiles de usuario, para ofrecer servicios personalizados.

ABSTRACT

The geospatial semantic web development requires search mechanisms to overcome the syntactic comparisons and perform semantic analysis, in order to retrieve information conceptually similar, this would allow reducing the risk of empty results when there is no exact correspondence between the query and the information available in data repositories. In this work, we describe an information retrieval model to integrate a semantic criterion with geospatial criteria in a not homogeneous space; the criteria represent the dimensions of this space, these dimensions are weighted in function of the user preferences or user profile; this integration uses an expression to evaluate the relevance of results. For the integration process is proposed a mathematical expression to evaluate the relevance of each result. We present a system that implements the geospatial semantic approach proposed to retrieve information from the domain of cultural tourism, museums specifically. The results depict the advantages of integrating geospatial and semantic criteria taking into account user profiles, offering more customized (personalized) services.

CAPÍTULO 1.

INTRODUCCIÓN

Y sabemos que a los que aman a Dios, todas las cosas les ayudan a bien…

Apóstol Pablo, Romanos 8: 28

A veces nuestro destino semeja un árbol frutal en invierno. ¿Quién pensaría que esas ramas reverdecerán y florecerán? Mas esperamos que así sea, y sabemos que así será.

Johann Wolfgang Von Goethe

1.1 Antecedentes

Actualmente en el campo de las Geociencias, el estudio de los sistemas de recuperación de información concentra sus esfuerzos en estudiar otras formas de representar el conocimiento, modificar la manera en que se almacenan y organizan los datos, así como la búsqueda de mecanismos que extraigan información controlando su precisión, para que correspondan de manera exacta o similar las respuestas arrojadas con las consultas que realizan los usuarios.

Dentro de este marco, en la actualidad las ontologías desempeñan un papel muy importante en las tareas de representación del conocimiento de un dominio, dotándolos de relaciones semánticas que enriquecen los trabajos de recuperación de la información en ellas contenidas. Cada vez más, estas nuevas formas de

representación toman mayor auge en el campo de los Sistemas de Información Geográfica (GIS), debido a que estas estructuras permiten modelar de manera formal el dominio geoespacial en el que se suscribe esta investigación.

En trabajos recientes se ha sugerido el uso de bases de datos espaciales por ser herramientas que permiten manejar y procesar la información geoespacial, describiendo los objetos espaciales que las forman a través de sus tres características principales: atributos, localización y topología. Los atributos proporcionan características de los objetos que describen lo que son. La localización, representada por la geometría del objeto y su ubicación espacial de acuerdo a un sistema de referencia, permite saber dónde está el objeto y qué espacio ocupa. Por último, la topología definida por medio de las relaciones espaciales entre los objetos, posibilita la interpretación semántica del contexto y la creación de jerarquías de elementos a través de sus relaciones.

El desarrollo de nuevas técnicas de recuperación de información ha permitido un significativo avance, posibilitando usar conceptos similares en lugar de términos sintácticamente iguales. La definición de los conceptos tiene lugar en el marco de las ontologías y en especial las jerarquías como caso particular de éstas, las cuales permiten organizar jerárquicamente los conceptos, definiendo propiedades y relaciones entre éstos. El uso de jerarquías como herramienta para representar el conocimiento, establece la necesidad de una nueva variable para ordenar los resultados que se obtienen después de una consulta, esta variable es denominada aquí como "*distancia" semántica*; de gran utilidad en el marco del presente campo de aplicación y los sistemas de recuperación de información espacial.

En el presente trabajo de investigación se usan estas estructuras jerárquicas como modelo conceptual, para organizar la información del dominio geográfico que

describe la ubicación de sitios culturales, se mide la similitud semántica entre los conceptos allí representados para identificar y recuperar los elementos que comparten propiedades similares, recuperando primero los más parecidos; controlando así el error y la precisión de las respuestas, que indican dónde se localiza un sitio cultural de acuerdo al tipo de lugar o patrimonio que preserva.

1.2. Planteamiento del problema

Es difícil localizar Puntos de Interés Cultural basado exclusivamente en la correspondencia exacta de resultados, más aún, cuando se toman en cuenta múltiples aspectos, tanto semánticos como geoespaciales, particularmente, cuando un turista ubicado en una de las estaciones del metro del centro histórico de la Ciudad de México desea buscar museos en esta región.

Dentro de este planteamiento surgen las siguientes preguntas de investigación:

- ¿Cómo evitar que los sistemas retornen resultados vacíos?
- ¿De qué manera se pueden flexibilizar los procesos de recuperación para traer resultados que van más allá de la coincidencia sintáctica?
- ¿Mediante qué estructuras de información es posible obtener resultados que por su similitud son relevantes para una consulta?
- ¿Cómo evaluar el grado de relevancia de cada uno de los resultados?
- ¿Cómo satisfacer las necesidades particulares de cada usuario?

1.3. Objetivos

En este apartado se describen los objetivos generales y particulares del presente trabajo de tesis.

1.3.1. Objetivo general

Diseñar e implementar un sistema para la recuperación de información contenida en repositorios de datos heterogéneos, permitiendo dar respuestas similares y con precisión controlada a consultas realizadas por usuarios, acerca de sitios turísticos culturales ubicados en el centro histórico de la Ciudad de México.

1.3.2. Objetivos específicos

- Diseñar y emplear jerarquías como estructura conceptual para representar y organizar información del dominio del turismo cultural.
- Desarrollar mecanismos de recuperación que den flexibilidad a las respuestas de las consultas sobre un Punto de Interés Cultural (CPI).
- Diseñar un modelo que tome en cuenta los resultados, los criterios y preferencias de los usuarios para controlar la precisión de las respuestas.
- Desarrollar una arquitectura para implementar las diferentes etapas del modelo de recuperación.

1.4. Motivación y justificación

De manera natural, lo que esperaría un turista cuando pregunta por un museo en particular, por ejemplo un *museo de pintura*; pero no hay ese tipo de museos "en el área cercana", entonces desearía que las respuestas que obtuviera sean: no hay exactamente ese museo, pero hay un *museo de escultura* a tan solo dos calles, o uno de *arte digital* a cinco calles, y así sucesivamente se mostraran los resultados más similares; incluso sería mejor, si las respuestas incluyeran una breve descripción del sitio.

En particular, las nuevas formas de representación del conocimiento junto con los mecanismos de recuperación de información, pretenden retornar a las consultas de usurario, resultados que van más allá de la simple coincidencia sintáctica; de allí que los principales problemas en este campo son:

- Representar el conocimiento en una forma "una estructura" donde un concepto se encuentre cerca "según una métrica" a otro, de acuerdo a qué tan parecidos son sus significados (similitud semántica) y no solamente por su igualdad sintáctica.

- Implementar sistemas de recuperación de información que den flexibilidad a las búsquedas, arrojando resultados similares según los significados de los conceptos buscados para evitar retornar resultados vacíos.

El caso de estudio de este trabajo está basado en los diferentes tipos de lugares de turismo cultural, particularmente en el área delimitada por el centro histórico de la Ciudad de México. La información de estos sitios y dependencias culturales están contenidas en una base de datos espacial, que sirve como repositorio de

datos para extraer y procesar la información cualitativa geoespacial con un enfoque semántico, que da flexibilidad al procesamiento de las consultas para controlar la precisión y el grado de similitud de las respuestas, marcando así un camino diferente para la recuperación de información que no puede ser resuelta con las aproximaciones tradicionales de los sistemas de búsqueda de información.

1.5. Beneficios Esperados

Con el desarrollo del presente trabajo se espera aportar a los sistemas actuales de recuperación de información dotándolos de características que flexibilicen la búsqueda y recuperación de información, en particular en el dominio del procesamiento semántico geoespacial, de tal manera que puedan realizarse consultas controlando la precisión con la que se recuperan los resultados, para dotar a los sistemas de recuperación de información de aspecto de la inteligencia artificial, en especial los que abordan el problema de retornar resultados que van más allá de la coincidencia sintáctica y abordan el área de la recuperación semántica.

1.6. Alcances y Límites

El caso de estudio de este trabajo está basado en los diferentes Puntos de Interés Cultural en especial los Museos, y particularmente en el área delimitada por el centro histórico de la ciudad de México. La información de estos sitios y dependencias culturales están contenidas en dos bases de datos, una espacial y otra conceptual, las cuales sirven como repositorios de datos para extraer y procesar la información cualitativa geoespacial con un enfoque semántico, que da flexibilidad al procesamiento de las consultas para controlar la precisión y el grado de similitud de las respuestas.

1.7. Hipótesis

Al Combinar el análisis semántico con el análisis geoespacial, para que se complementen mutuamente, se pueden obtener resultados más relevantes y más satisfactorios, es decir, resultados con la precisión de las operaciones geoespaciales, pero con la flexibilidad de los procedimientos del análisis de similitud semántica. Para determinar la importancia de un resultado, en esta investigación una de las hipótesis principales es que, en últimas quien más sabe lo que desea es el usuario, por ello él es quien define la relevancia a través de sus preferencias, y en consecuencia, la ponderación de los resultados.

1.8. Organización de la Tesis

Capítulo 1. En este capítulo se presenta una introducción general de la tesis, junto a las razones que motivan y justifican el presente trabajo, además se describe el planteamiento del problema, algunas preguntas de investigación que dirigen este trabajo, junto con los objetivos de la misma; por último se expone la hipótesis de la que se parte.

Capítulo 2. En este capítulo se hace un recorrido por el estado del arte en sistemas de Recuperación de Información, enfatizando la Información Geográfica, se citan varios trabajos que tienen relación con esta tesis. Se menciona la importancia del uso de jerarquías en diferentes áreas de investigación y las medidas de similitud semántica sobre diferentes tipos de estructuras conceptuales, enfocadas principalmente en jerarquías.

Capítulo 3. En este capítulo se profundiza en el marco teórico de la tesis, en el cual se describen conceptos relacionados con el estudio de estructuras de representación del conocimiento, así como la organización, almacenamiento y recuperación de información del dominio espacial con un enfoque semántico, y finalmente su relación y aplicación en estructuras jerárquicas como caso particular de las ontologías.

Capítulo 4. En el cuarto capítulo se describe la metodología propuesta a manera de un modelo de Recuperación de Información Semántico Geoespacial denominado ***GeoSeM*** (nombre que se ha formulado en Ingles - ***Geo****spatial* ***Se****mantic system to retrieve information from* ***M****useums*), y finalmente para evaluar dicho modelo, se presenta la arquitectura de un sistema de recuperación de información en el dominio del turismo cultural, empleando el soporte descrito en el marco teórico.

Capítulo 5. Este capítulo exponen los resultados obtenidos, de acuerdo con la metodología propuesta. Se lleva a cabo un análisis y discusión sobre los resultados, los cuales permiten evaluar la funcionalidad y bondades del modelo y del sistema propuesto.

Capítulo 6. En este capítulo se presentan las principales contribuciones y conclusiones obtenidas con el proceso de investigación llevado a cabo en este trabajo y se enuncian algunas propuestas como trabajo a futuro que permitirían expandir y mejorar el modelo propuesto.

En último lugar, se presentan las referencias a los trabajos que dieron soporte a este trabajo de investigación; así como los anexos, algunos apartes y detalles de implementación.

CAPÍTULO 2.

ESTADO DEL ARTE

No es reinventar la rueda, es hacerla andar.
No se puede imitar lo que se quiere crear.
Georges Braque

En este apartado se mencionan algunos de los trabajos, avances y tendencias que forman parte del contexto que engloba esta investigación. Sumado a esto, se describen trabajos que emplean estructuras conceptuales, es decir, ontologías y jerarquías como medio de representación y organización de la información. Finalmente, se enuncian algunas propuestas de recuperación semántica de información, así como trabajos que destacan la importancia y tendencia actual del uso de la similitud semántica en estructuras conceptuales, centradas participialmente en el dominio geoespacial.

2.1. Recuperación de información

La Recuperación de Información (IR – por sus siglas en inglés) basa sus procesos en los sistemas de información, ahora bien, al hablar de los sistema de información, es conveniente mencionar que, para que éstos den soporte a los procesos de recuperación se debe hablar de la cantidad y calidad de la información; pero un sistema de información no solo toma importancia por el volumen de información que contiene, sino también por la facilidad que proporciona el sistema para acceder a la información contenida, también por la representatividad y finalmente por el grado de satisfacción que la organización de esta información brinda al usuario. El enfoque de esta investigación va dirigido a la recuperación de información que, dada su similitud, se acerca en el mayor grado posible, a satisfacer las necesidades de los usuarios dentro del dominio geográfico.

La búsqueda y recuperación de la información se enfoca principalmente al proceso de interrogación (*queryng*) y al proceso de exploración (*browsing*); donde el primer proceso está dirigido por el usuario, introduciendo palabras clave que representan sus necesidades de información, para este caso el sistema retorna una serie de respuestas, generalmente ordenadas por relevancia, algunos ejemplos de sistemas que operan de esa manera en forma general son: Google, Yahoo!, entre otros.

En el proceso de recuperación por *browsing*, a diferencia del *queryng*, el usuario se encarga de explorar visualmente sin tener que expresar previamente cuáles son sus necesidades de información; es decir, el usuario reconoce lo que está buscando, sistemas de ejemplo que incluyen esta forma de recuperación en un modo general son e-bay, Mercadolibre, entre otros.

2.1.1. Recuperación de información geográfica

De la confluencia de los Sistemas de Información Geográfica (GIS – por sus siglas en inglés) con la Recuperación de Información, surge la Recuperación de Información Geográfica (GIR – por sus siglas en inglés). Esta incluye todas las áreas de investigación y componentes que tradicionalmente forman el núcleo de la IR, como son: los sistemas y motores de búsqueda, pero además con un énfasis en la información procedente de datos del dominio geográfico [BCJ05]. Se puede concluir entonces que, la recuperación de información geográfica se preocupa por la recuperación de información que involucra algún tipo de *percepción geoespacial*.

Aunque una de las principales tendencias en GIR es heredada de las áreas de procesamiento de lenguaje natural, al 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; no obstante, existe un amplio espectro de enfoques para resolver las tareas en GIR, los cuales van desde aproximaciones simples de recuperación de información sin indexación de términos geográficos, a modelos derivados de técnicas complejas de inteligencia artificial. Algunas de las técnicas usadas en la actualidad incluyen extracción de entidades geográficas, análisis semántico, bases de conocimiento geográfico (conjuntos de reglas, ontologías,…), adquisición de conocimiento, grafos y Análisis Formal de Conceptos (FCA – por sus siglas en inglés) y técnicas de expansión y flexibilización de consultas.

Con el objeto de unificar y establecer estándares, una estructura de organización de la información, es propuesta por ESRI, una de las compañías líderes en desarrollos de sistemas de información geográfica, su propuesta también señala el uso de jerarquías como estructura conceptual que facilita la recuperación de información, esta propuesta se observa en la Figura 2.1.

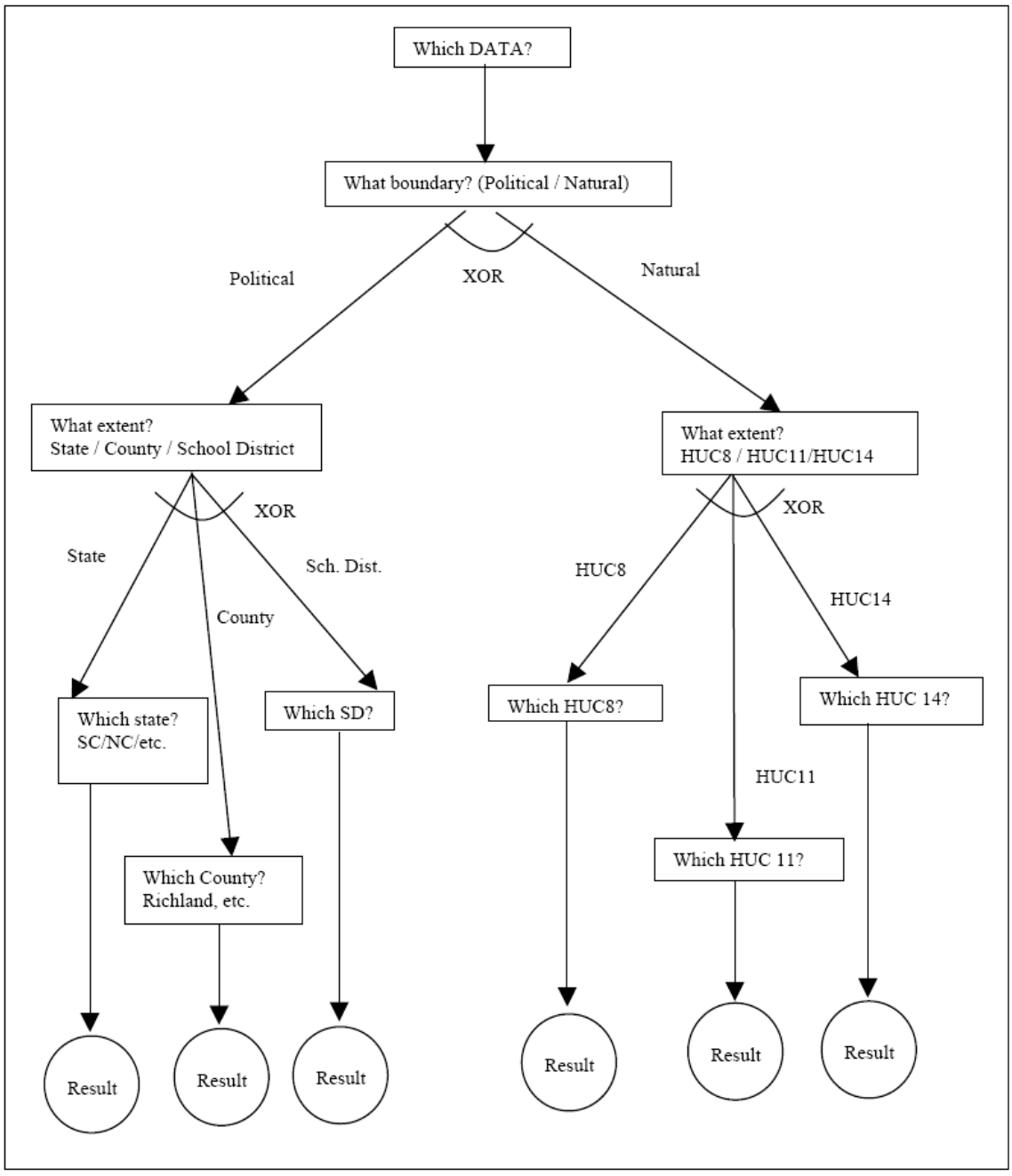

Figura. 2.1 Organización de la información del dominio geográfico (ESRI)

Aunque esta propuesta es una aproximación al problema de representación, deja de lado el problema de *análisis de similitud semántica* entre las entidades del mundo real que han sido abstraídas y representadas en su modelo.

2.2. Similitud semántica en estructuras conceptuales

Para hablar de similitud, hay que hablar de los primeros intentos y avances en formalización de éste término, los cuales se remontan a los años 60's y 70's con Shepard [She62] y luego con Tversky [Tve77]. Para determinar la similitud entre un objeto (entidad) y otro se parte de sus características comunes y diferentes, un ejemplo de la aplicación de este enfoque basado en las características de entidades espaciales es mostrado en [RER99].

Ahora bien, las entidades a comprar pueden formar parte de una entidad mayor que las contiene, es decir, una estructura de entidades, que de ahora en adelante se asumirá como estructura de conceptos o estructura conceptual, las estructuras conceptuales pueden ir desde complejos grafos a simples taxonomías o árboles, en la actualidad las estructuras conceptuales con mayor auge son las ontologías, y como se vio en el apartado anterior, en el dominio geográfico son muy usadas las jerarquías como caso particular de las ontologías, debido a que este tipo de estructuras conceptuales permiten representar de manera natural la información geográfica.

2.2.1. Medidas de similitud

El uso de las medidas de similitud es un área de investigación que ha sido explorado hace varios años. En el año de 1965 encontramos un trabajo de Rubenstein y Goodenough acerca de la similitud entre términos del idioma inglés, este trabajo está basado en la semejanza del contexto en que aparecen [RG65]. Para validar su trabajo se llevó a cabo un experimento donde dos grupos de personas, evaluaban la similitud entre 65 pares de sustantivos. La evaluación constaba de un valor numérico, en un rango que iba de 0.0 para describir

ausencia de semejanza, a un valor 4.0 que era indicio de existencia. Claramente este experimento fue una de las bases para que se siguieran realizando pruebas en el campo de las medidas de similitud.

Sin embargo, el experimento con el cual diversos autores se identifican, es el que realizaron Miller y Charles [MC91] en relación con su trabajo desarrollado en WordNet [Mil90]. En este experimento, un grupo de estudiantes fueron seleccionados para estimar la semejanza entre 30 pares de nombres. En este trabajo el grado de semejanza, al igual que en el trabajo de Rubenstein y Goodenough, se expresaba con un valor entre 0.0 y 4.0. A partir de este experimento se derivaron una serie de propuestas para medir la similitud entre elementos de estructura conceptuales taxonómicas. Aunque principalmente la mayoría de propuestas están orientadas bajo el marco de WordNet.

2.2.2. Similitud basada en probabilidad léxica

Uno de los algoritmos pioneros y más destacados es el propuesto en [Res93], el cual propone que la similitud entre dos conceptos c_1 y c_2 de una estructura taxonómica, puede ser obtenida mediante la expresión (1).

$$sim(c_1, c_2) = \max_{c \in S(c_1, c_2)} (-\log p(c)) \qquad (1)$$

En donde $S(c_1,c_2)$ representa el conjunto de conceptos de los cuales tanto c_1 como c_2 descienden.

Mientras que p(c) es la probabilidad del concepto *c*. El término concepto hace referencia al conjunto de términos que apuntan a una misma idea.

Ahora bien, para estimar la probabilidad de un concepto *c*, Resnik utiliza la frecuencia de aparición de los términos de ese concepto en el Brown Corpus of American English [FK82].

2.2.3 Medida de similitud sobre taxonomías

El uso de la probabilidad también es el fundamento de otros trabajos, por ejemplo el trabajo propuesto por Lin [Lin98], en el cual se obtiene de forma teórica la expresión (2) para el cálculo de la similitud entre dos conceptos c_1 y c_2 en una taxonomía como la de WordNet.

$$sim(c_1, c_2) = \frac{2 * \log p(c_3)}{\log p(c_1) + \log p(c_2)} \quad (2)$$

Donde c_3 es el nodo padre común a c_1 y c_2 directo, o más próximo.

Bajo este enfoque, Lin pretende dar una definición universal de similitud en términos de la teoría de la información. Para llegar a dicha definición Lin parte de tres premisas intuitivas.

1. La similitud entre A y B se relaciona con su entorno. Entre más características en común compartan, más similares son.

2. La similitud entre A y B está relacionada con las diferencias que existen entre ellos.

Entre más diferencias hayan, menos similares son.

3. La máxima similitud entre A y B se alcanza cuando A y B son idénticos.

2.2.4. Medida de similitud usando WordNet

Una propuesta más para medir la similitud es presentada por Leacock y Chodorow [LC98], estos dos autores consideran que la similitud entre c_1 y c_2 puede ser obtenida con la expresión (3).

$$sim(c_1, c_2) = -\log\left(\frac{len(c_1, c_2)}{2 * MAX}\right) \quad (3)$$

Donde, *len*(c_1,c_2) cuantifica el número de saltos entre c_1 y c_2, y *MAX* es el número de saltos entre el nodo raíz y los nodos hoja de la taxonomía.

De igual modo, otro destacado trabajo de investigación usando WordNet es el de Hirts y St-Onge [HSO98]. En este trabajo se emplean cadenas léxicas como representación de contexto para la detección y correlación de nodos entre conceptos. Se alejan un poco del enfoque probabilístico y cuantitativo (conteo de nodos entre conceptos), sustituyéndolo por la idea de cadenas léxicas para poder obtener la similitud.

Después de analizar destacados trabajos que han abordado parte del problema con el que nos enfrentamos en el presente trabajo de investigación, podemos observar la importancia de WordNet en la definición de las medidas de similitud, ya que gran número de las propuestas que existen en la actualidad van dirigidas al uso de estructuras taxonómicas.

2.2.5. Medidas de similitud más allá de la probabilidad

Dejando completamente de lado el enfoque probabilista, en [BT06] se plantea el problema que existe entre las medidas de similitud y se propone un algoritmo alterno. Los algoritmos encargados de medir la similitud semántica no pueden ser aplicados en taxonomías en las que no exista forma de estimar la probabilidad de un concepto *c*. Al observar el ejemplo de Resnik, en el que se estima la probabilidad de un concepto, por medio de la frecuencia de aparición de los términos asociados a este concepto dentro de un corpus significativo, como el Brown Corpus of American English, la probabilidad puede ser estimada aplicando la expresión (4).

$$\tilde{p}(c) = \frac{freq(c)}{N} \tag{4}$$

Donde $freq$(c) es el número de veces que aparecen en el *corpus* los nombres que corresponden al concepto *c* o a sus subconceptos, y N es el total de nombres presentes.

Esto nos lleva a un problema, ya que, si quisiéramos hacer uso de la expresión anterior para hacer un estimado de la probabilidad en taxonomías diferentes de WordNet, tendríamos que ver cómo medir la frecuencia de aparición de cada concepto. Lo que a su vez genera dos problemas: en primer lugar se está asumiendo que las instancias de la estructura conceptual son estáticas, es decir, no contemplan cambios conceptuales y en segundo lugar se deja de lado el problema de desbalance entre o desequilibrio de especificidad entre nodos, o sea, cuando la cantidad de instancias pertenecientes a una clase supera en gran número a la cantidad de instancias de la clase hermana.

Debido a estos problemas existe otra propuesta [BT06], en la cual se suponen dos conceptos c_0 y c_1, donde c_1 es subclase de c_0, lo cual no implica que c_1 sea hijo directo de c_0 en la ontología. Luego se discutirá el caso general de dos conceptos c_1 y c_2 que no son subclase uno del otro. Para el primer caso propuesto la medida de similitud es la expresada en (5).

$$ssim(c_0, c_1) = \frac{k + N_0}{k + N_0 + \log\left(\frac{E_0}{E_1}\right)} \tag{5}$$

Donde N_0, es el número de nodos que hay entre el nodo raíz y c_0, esto permite tener la noción de que descender un nivel en la ontología supone una mayor diferencia semántica en la parte alta de la ontología (la de los conceptos más genéricos) que en la parte baja (donde están los conceptos más específicos).

El cociente E_0/E_1 representa la relación entre la información que aporta c_0 y la información que aporta c_1. Si suponemos que el nodo raíz contiene el 100% de la información presente en la ontología, cada uno de los nodos inferiores contendrá una porción de esa información. La manera en que se modela el reparto de dicha información está relacionada con la estructura de la ontología.

Ahora, si una clase *c* tiene una cantidad de información *E*, lo que vamos a suponer es que cada una de sus *n* subclases tendrá una cantidad de información *E/n*, es decir, que la información se va repartiendo equitativamente entre las clases hijas (*claro está, esto es un supuesto difícil de generalizar*). Con esto lo que se expresa es que, cuanto menor es el número de clases en que se divide una clase, mayor es el parecido entre estas subclases y la clase padre. Aquí el objetivo del diseño es organizar la información de forma manejable. Para ello se suele crear clases intermedias, evitando que una clase tenga un número muy elevado de subclases.

La agrupación de las subclases en clases intermedias se hace a partir de características cualitativas comunes, a veces definidas por el diseñador según el propósito.

La variable *k* permite ajustar el peso que se le asignan a las variables: profundidad de c_0 y tasa de información. Un aspecto negativo aquí, es que *k* debe ser obtenida empíricamente.

Después de realizar el cálculo de la similitud entre un concepto c_0 y unos de sus subconceptos c_1 en este algoritmo, lo siguiente es ver la manera de extender esto al caso de dos conceptos c_1 y c_2 que no son subclase uno del otro. Para ello, es necesario introducir el término distancia. La medida propuesta en la expresión (1) da lugar a un valor entre 0.0 (carencia de similitud) y 1.0 (máxima similitud, i.e exactitud). De esta manera, la distancia *dist( )* entre dos conceptos se puede definir a partir de su similitud semántica *ssim* como se muestra en (6).

$$dist(c_1, c_2) = \frac{1}{ssim(c_1, c_2)} - 1 \qquad (6)$$

Si $c_1 = c_2$ (i.e. el mismo concepto) entonces $ssim(c_1,c_2)=1$ *y* $dist(c_1,c_2)=0$.
Análogamente, si $ssim(c_1,c_2)=0$ (i.e. conceptos totalmente distintos), entonces la distancia conceptual entre ambos será infinita. Ahora si C es el conjunto de

conceptos que son superclase de c_1 y c_2, esta distancia se puede expresar como indica (7).

$$dist(c_1, c_2) = \min_{c_0 \in C} \left(dist(c_0, c_1) + dist(c_0, c_2) \right) \quad (7)$$

Derivada de esta expresión se puede obtener la expresión (8).

$$S_{12} = \max_{c_0 \in C} \frac{s_{01} * s_{02}}{s_{01} + s_{02} - s_{01} * s_{02}} \quad (8)$$

Donde S_{ij} es otra forma de representar *ssim*(c_1,c_2).

Hay que destacar que, este algoritmo no hace ninguna suposición sobre la taxonomía en la que se hacen los cálculos, se puede usar esta medida en cualquier dominio. Como consecuencia, no precisa de datos externos, puesto que los cálculos sólo dependen de la estructura conceptual.

2.2.6. Similitud semántica en jerarquías

Existen varias propuestas sobre cómo medir la similitud entre entidades contenidas en una estructura conceptual jerárquica, por ejemplo, Rada et al. proponen como distancia, simplemente el valor obtenido al contar la cantidad de conexiones entre un nodo (un concepto) y otro [RMBB89]. Ahora este proceso simple puede ser refinado estableciendo pesos a cada uno de los enlaces de la jerarquía como se proponen en [KK90]; en [GGW03] un estudio más exhaustivo es presentado, donde se propone un estudio de aquellos modelos que miden la similitud basados en la intersección de objetos, empleando una jerarquía que describa las relaciones entre los elementos que la componen. Al establecer los correspondientes valores de similitud entre entidades, es decir, conceptos en la jerarquía, se logra constituir un "*conocimiento semántico*" en la jerarquía como menciona Sarabia en [Sar08], el cual ayuda a identificar los objetos que tienen características en común, beneficiando así el mejoramiento de las medidas de

similitud. En [GGW03] consideran importante emplear una medida de similitud que sea "intuitiva" para el ser humano. En la Figura 2.2 se muestra una jerarquía que representa diferentes géneros de música.

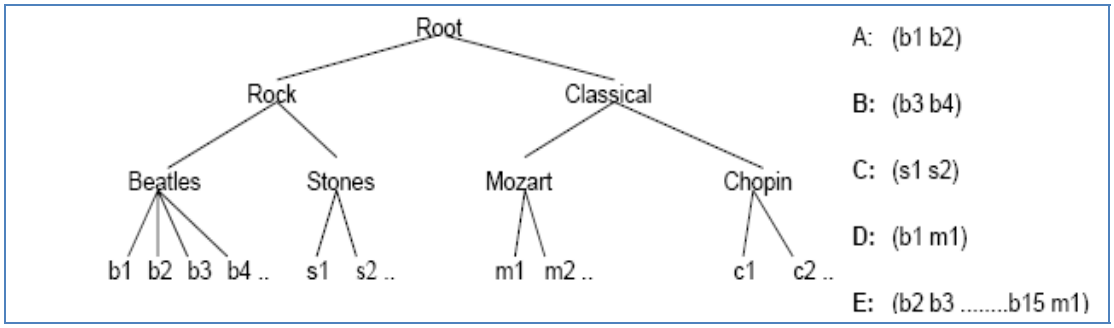

Figura 2.2. Jerarquía de CD`s de música empleado por Ganesan.

Partiendo de la jerarquía anterior, donde se suponen distintos usuarios (A, B, C, D, E) que compran un par de CD`s respectivamente, se pretende determinar qué usuario es más similar a otro dado el género de música que adquiere en sus compras.

En la jerarquía mostrada en la Figura 2.2 se observa que el cliente A compra dos CD´s de los Beatles (b1, b2), el cliente B adquiere (b3, b4), el cliente C compara (s1, s2) y el cliente D lleva (b1, m1). Puede verse que el cliente A y B son bastante similares ya que ellos adquirieron dos discos de los Beatles, mientras que el cliente A y C son menos similares, ya que a pesar de escuchar música rock prefieren diferentes agrupaciones musicales. En este sentido, Ganesan y Molina se enfocan al desarrollo de medidas de similitud que se acercan a los modos en que se supone, opera la intuición humana.

Las contribuciones que Ganesan y Molina aportan en el área de GIR, se pueden enunciar de manera resumida así:

- Introducen medidas de similitud que aprovechan la estructura natural jerárquica de un dominio, proporcionando resultados que son intuitivamente

más similares que los generados por las medidas tradicionales de similitud.

- Se consideran múltiples ocurrencias al medir la similitud entre elementos de la jerarquía.
- Se realizan comparaciones con las medidas de similitud que no explotan una jerarquía de dominio.

Las jerarquías a menudo son empleadas para estructurar la información y se han utilizado para la clasificación de texto, la recuperación de información, entre otras tareas donde la similitud juega un papel importante [FD95, SM98], particularmente para esta investigación nos permiten almacenar de manera práctica y natural información geográfica. En la Figura 2.3 se describe de forma general la evolución que han tenido los enfoques para medir la similitud, partiendo de su criterio.

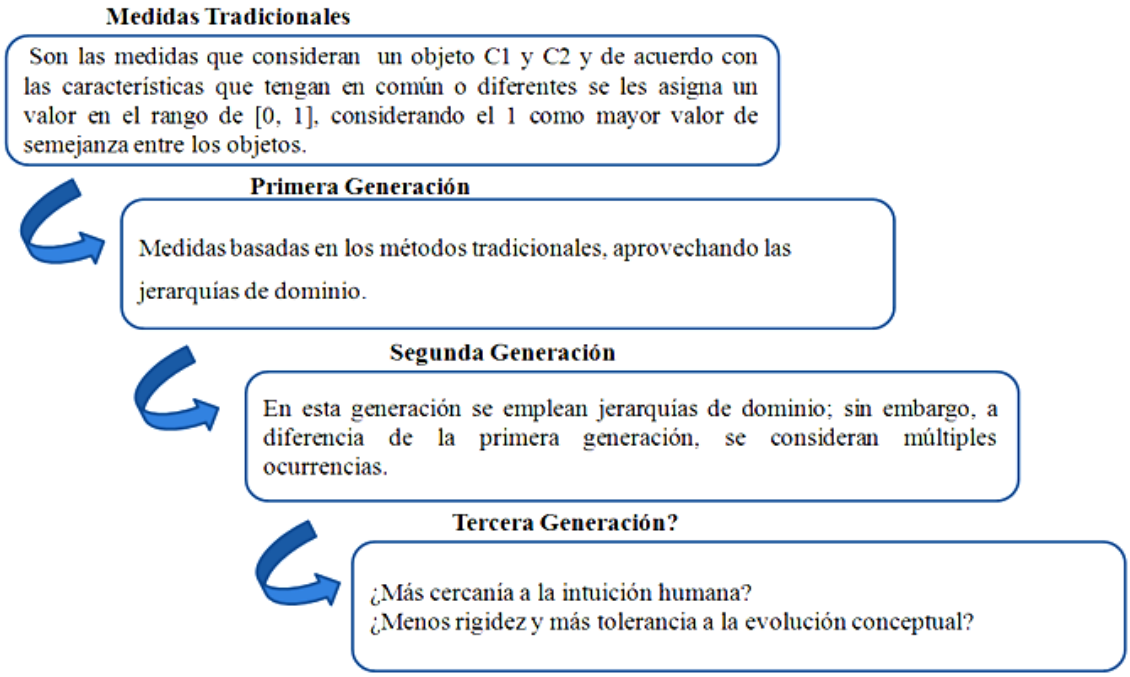

Figura 2.3. Evolución de las medidas de similitud según Ganesan y Molina

Para las medidas tradicionales no hay nociones universales de las características que deben ser tomadas en cuenta para evaluar qué tan similar es un objeto de

otro, por lo tanto, para este caso el concepto de *similitud* es necesariamente subjetivo. Este tipo de medidas son descritas en trabajos como [Rus79, SGM00, McG83, SB98, RIS+94].

2.3. Aplicaciones de GIR basadas en estructuras conceptuales

Está claro que las jerarquías son muy comunes en los sistemas de información, a tal punto que, se utilizan para organizar elementos por orden de importancia (según un criterio), o por explícitos como: tamaño, capacidad, tiempo, como por ejemplo en un sistema de archivos o en el organigrama de alguna empresa, en la estrategia de desarrollo, o en la estructura geográfica según la cobertura de una organización.

2.3.1. Jerarquías empleando SQL

En los sistemas tradicionales de almacenamiento de información, en especial las bases de datos, surge el problema de recuperar la información que por naturaleza está almacenada en forma de jerarquía, ya que no es posible en una sola consulta a la base de datos retornar toda o parte de la jerarquía. De tal manera que los esfuerzos por mejorar consultas sobre estructuras jerárquicas han traído innovaciones en los gestores de base de datos, como es el caso de SQL Server en sus últimas versiones, en las cuales admite un nuevo tipo de datos denominado *HierarchyID*, que ayudan a resolver algunos de los problemas al modelar y consultar información en una jerarquía. El tipo de datos *HierarchyID* permite construir relaciones entre los elementos de datos de una tabla y sobre todo representar una posición en una jerarquía.

2.3.1.1. Descripción de los datos HierarchyID

El tipo de datos *HierarchyID* es una representación compacta y binaria de una ruta de acceso ordenada [Teg08]. Resulta de utilidad siempre que se tenga que representar una relación anidada entre valores, en los que dicha relación se pueda expresar en una sintaxis de ruta de acceso ordenada.
Una ruta de acceso ordenada tiene el aspecto de una ruta de acceso de archivo, pero en lugar de usar los nombres de directorio y archivo, se usan valores numéricos. Como cualquier otra relación de elemento primario/secundario, todas las rutas de acceso ordenadas se deben delimitar mediante un nodo raíz. En SQL Server 2008, se usa un solo carácter (/) para representar textualmente el nodo raíz. Los elementos con la ruta de acceso ordenada suelen representarse mediante enteros, pero también se pueden usar valores decimales. La ruta de acceso ordenada debe terminar con otro carácter único (/).

Sin embargo, las rutas de acceso ordenadas no se almacenan en la base de datos como texto. Sino que se les aplica un algoritmo *hash* para obtener valores binarios, y se almacenan en las páginas de datos.

2.3.1.2. Restricciones de los HierarchyID

El principal inconveniente de los tipos de datos *HierarchyId*, está en que no representan un árbol automáticamente, es decir, cada fila almacena la información sobre la posición que tiene en el árbol, lo que no garantiza que se pueda generar un árbol al tomar todas las filas, ya que puede haberse borrado un nodo padre o puede existir más de un nodo con el mismo *HierarchyId, sumado a esta limitación,* se necesita controlar la concurrencia y unicidad de los *HierarchyId*, pues la base de datos no controla esos aspectos.
Finalmente, a diferencia de las referencias a otras entidades tradicionales (*foreign key*), los *HierarchyId* no mantienen la integridad referencial, lo que permitiría

eliminar un nodo que era padre de otros nodos sin que la base de datos genere un error, lo cual podría traer graves inconvenientes en estructuras conceptuales.

2.3.2. Fuentes heterogéneas

Cualquiera que haya requerido navegar en Internet se ha visto en la necesidad de buscar información, pero se ha encontrado con el problema que en la Web actual, el acceso a la información llega a convertirse en una tarea tediosa y frustrante debido a diversos problemas habituales en la tecnología actual de motores de búsqueda, el principal problema no es la falta de información, hoy los volúmenes de información son enormes y se piensa que crecerán a un ritmo cada vez más acelerado, uno de los problemas es que la información no está organizada de la manera en que los sistemas pueden procesarlas, otro problema es que quizá existe la información pero con otro nombre, o *incluso existe información similar* que, *con un pequeño margen* de error *podría satisfacer la búsqueda del usuario*. La Web Semántica es una Web extendida basada en el significado, en donde cualquier usuario en Internet podrá encontrar respuestas a sus consultas de forma más rápida y sencilla, aunque no siempre con perfecta exactitud.

Tim Berners Lee, en el artículo [BLH01] menciona cuatro componentes o características básicas necesarias para la evolución de la Web Semántica. Estos componentes son los siguientes: *expresar significado*, tener acceso a representaciones del conocimiento, *empleo de ontologías* y agentes artificiales.

2.3.2.1. Web semántica geoespacial

La *Web Semántica* es una extensión de la Web actual en la que es proporcionada la información con un *significado* bien definido, y se mejora la forma en la que las máquinas y las personas trabajan en conjunto [BLH01]. Otras definiciones de la Web Semántica se pueden encontrar en [Her03, W3C01].

La Web Semántica tiene como fin, superar las limitaciones de la Web que existe en la actualidad, mediante la introducción de descripciones explícitas de significado [BLH01]. Su visión está dirigida a la búsqueda de una Web más "*inteligente",* en la cual se pueda conseguir una comunicación efectiva entre computadoras, sus esfuerzos están orientados principalmente a la búsqueda de descripciones enriquecidas semánticamente para los datos que se encuentran en la Web, al más allá de los metadatos; con esto se esperan facilitar el camino para obtener mejores métodos de recuperación.

La Web Semántica Geoespacial busca dar un nuevo horizonte de trabajo al uso de la información espacial, se intenta incorporación la *semántica Geoespacial* en la Web tradicional. La idea primordial es poder elaborar ambientes donde puedan realizarse búsquedas para ubicar lugares geográficos dejando de usar sólo etiquetas o palabras clave, pero esta visión trae consigo varios retos, entre ellos se encuentran:

- Mejorar la representación de la información geográfica.
- Optimizar la integración de información.
- Añadir consultas que usen operadores espaciales.

Ahora bien, existen tres componentes que engloban a la Web Semántica Geoespacial, los cuales se pueden agrupar de la siguiente manera:

Componente geoespacial: Mapas, Objetos geográficos Relaciones geográficas Sistemas de referencia.
Componente Web: Interoperabilidad entre tecnologías, Servicios compartidos
Componente semántica: Razonamiento automático e inferencia

En función de estos componentes es posible definir algunas de las aplicaciones con mayor necesidad actualmente, estas son:

- Plataformas de búsqueda empleando componentes de tipo geográfico.
- Descubrimiento de conocimiento a través de relaciones semánticas, espaciales y temporales.
- Seguimiento al comportamiento de los usuarios de la Web Geoespacial.

El principal interés en la *semántica geoespacial* está centrado en mejorar los resultados de las búsquedas y la interoperabilidad entre datos espaciales. El UCGIS (University Consortium for Geographic Information Science – por sus siglas en inglés) ha tomado el desafío de profundizar la investigación de la web semántica geoespacial. Existen varios autores que han tomado parte en esta área, como es el caso de Egenhofer. En [Ege02], se describe claramente la necesidad de la semántica en los datos para realizar búsquedas en la web y obtener resultados más cercanos a la realidad del dominio.

El consorcio Universitario para la Ciencia de la Información Geoespacial (UCGIS) ha identificado a la Web Semántica Geoespacial como una prioridad en investigaciones recientes. El desarrollo de las ontologías espaciales ha sido identificado como un compromiso de investigación a largo plazo por el consorcio. Las principales áreas de interés están enfocadas a la recuperación de información y la interoperabilidad de la información espacial.
Para finalizar, existen otra manera de abordar la semántica de un dominio, basado en las teorías de redes semánticas, en las cuales se distinguen tres categorías: redes is-a, grafos conceptuales, y redes de marcos. En conclusión hay que mencionar que este tipo de estructuras tienen un alto poder de expresividad igual o mayor que la lógica de predicados de primer orden, de acuerdo con el análisis realizado en [Sow00, SR92], donde se puede consultar más detalles.

2.3.3. Sistemas e implementaciones actuales

GeoCLEF: Dentro del área de GIR existen avances cuya principal orientación es el procesamiento de información contenida en documentos, bajo este enfoque

surgen algunas propuestas reunidas en el Foro Geo-CLEF, para la evaluación de sistemas de recuperación de información geográfica multilenguaje, cuya tarea más importantes es la de búsqueda de información geográfica y la proposición de nuevas sub-tareas como el análisis de consultas (query parsing), para identificar aspectos geográficos en una consulta y además, la búsqueda geográfica de imágenes. Estas tareas están dirigidas por dos procesos, uno monolingüe donde se utiliza el mismo idioma tanto para las consultas como para las colecciones y otro bilingüe, que incluye traducción, puesto que el idioma de la consulta es distinto al de la colección utilizada. Un ejemplo de arquitecturas típicas para sistemas que emplean este enfoque se presenta en [PGGU08], como se ilustra en la Figura 2.4

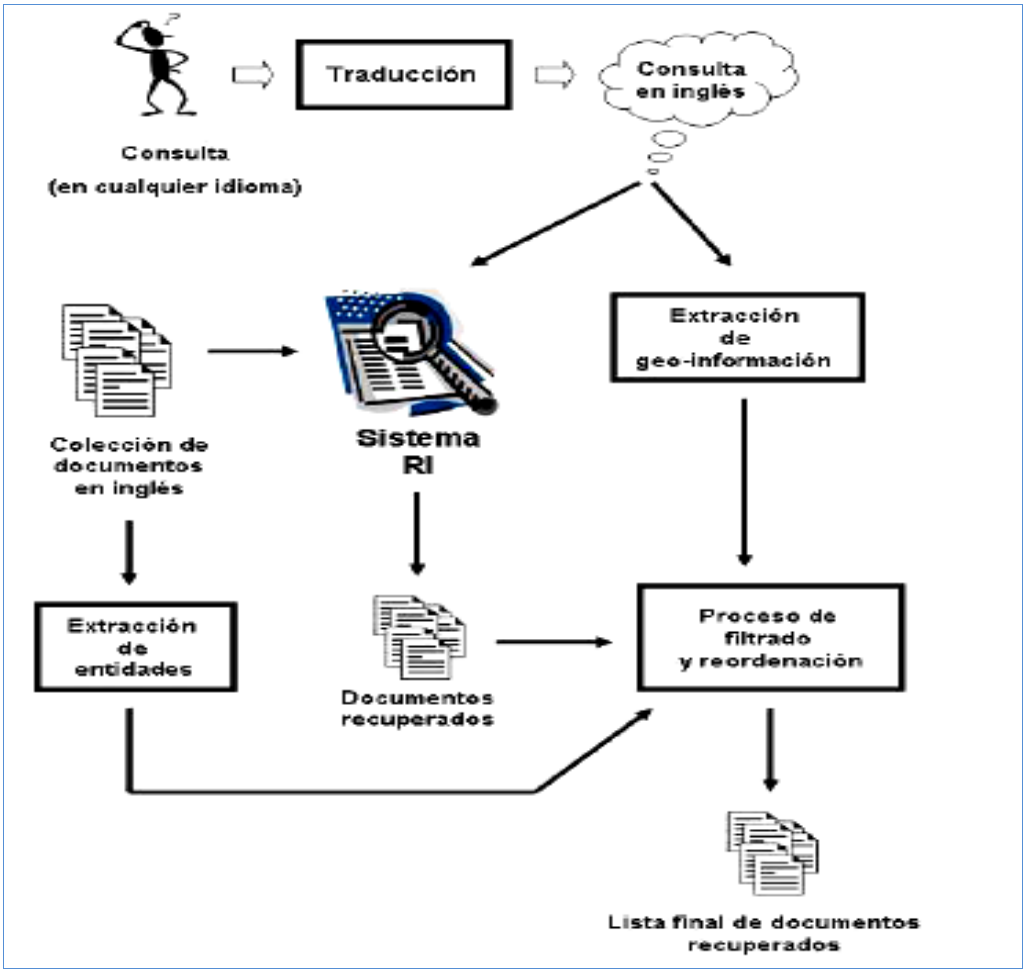

Figura 2.4 Ejemplo de una arquitectura basada en el enfoque de procesamiento de texto (GeoUJA)

Wikimapia: llamada a veces la Wikipedia geográfica, es una aplicación tipo mashup, es decir, un recurso en línea que combina otros recursos y servicio, en este caso mapas de Google con un sistema wiki, permitiendo a los usuarios añadir información en forma de notas a cualquier región o localidad del planeta, en ese sentido, la información a recuperar obedece a la coincidencia de etiquetas con la consulta, ejemplos de búsqueda en este sistema se muestran en la Figura 2.5.

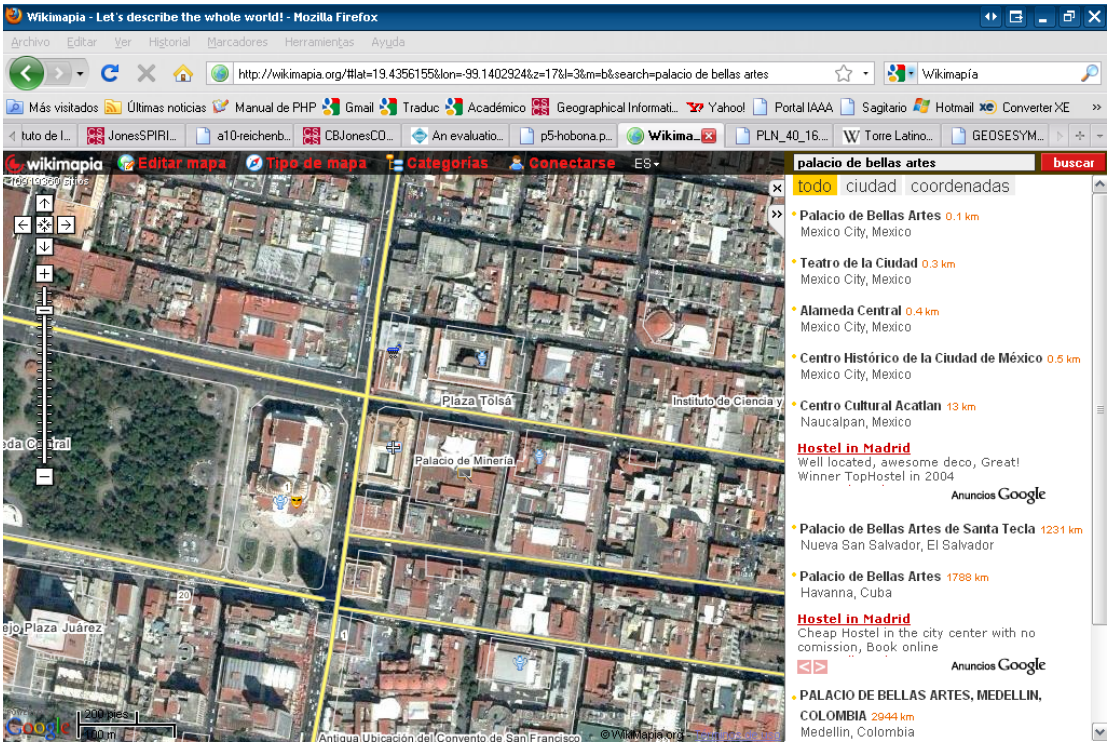

Figura 2.5 Ejemplo de búsqueda de un sitio turístico en Wikimapia

Al consultar por el museo Palacio de Bellas Artes, se observa que los resultados arrojados se ordenan de acuerdo a la cercanía en forma descendente, si hay coincidencia exacta entonces aparecerá primero aquel sitio que coincida de manera exacta y por defecto seguirán aquellos sitios que están a menor distancia, sin importar características semánticas de cada sitio, sino que obedece a una categorización; esto queda más claro si en lugar de buscar un sitio en particular, consultamos una categoría en específica, i.e. museos, donde definitivamente los resultados que aparecen solo obedecen a coincidencias sintácticas, y aunque la

búsqueda se realiza en la misma área del ejemplo anterior, no aparece el *Palacio de Bellas Artes*, que es un museo en la región de análisis actual, ver la Figura 2.6.

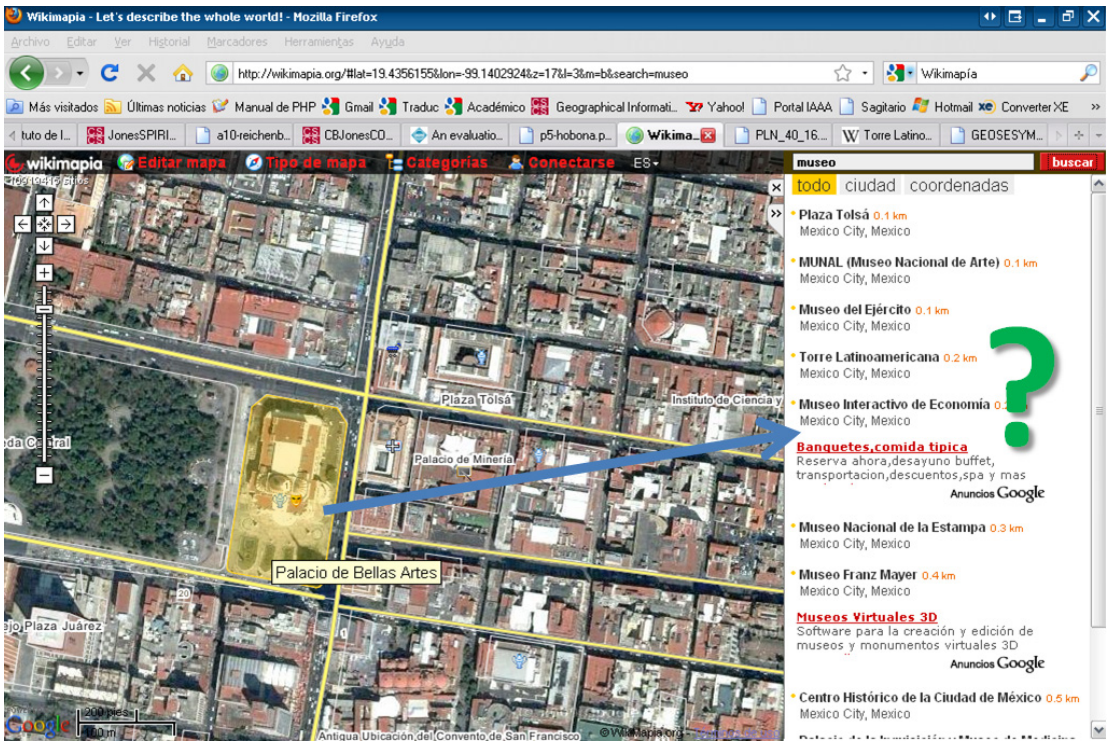

Figura 2.6 Problemas en la búsqueda por categoría de sitios turístico en Wikimapia.

Ahora, al buscar sistemas o aplicaciones para recuperación de información en México, se encuentran pocos resultados, los más cercanos son por ejemplo aplicaciones que más bien se quedan al nivel de GIS, por ejemplo:

Sigmetropoli2025: esta aplicación Web permite localizar un lugar pero a partir de su dirección, más no a partir de categorías y mucho menos de aspectos semánticos, un ejemplo de búsqueda en este sistema es mostrado en la Figura 2.7. Algunas dificultades de este sistema radican en el hecho que es poco probable que los turistas conozcan la dirección de todos los lugares a donde desean ir, no obstante se cita esta aplicación porque es de las pocas que hay en el ámbito local, además, poseen una representación icónica que se aproxima a una categorización,

en la imagen se puede observar que el *Museo del Palacio de Bellas Artes* es catalogado como un sitio cultural, representado con el ícono de una estatua.

Figura 2.7 Ejemplo de búsqueda en el sitio de SIG Metrópoli 2025 [SIGM].

Xaqui: es otro sitio Web con un enfoque similar al anterior, no obstante este permite buscar sitios de acuerdo a categorías, por ejemplo: museos; los resultados se pueden delimitar de acuerdo a un área geográfica, pero la lista de resultados no obedece algún ordenamiento semántico o algún criterio geoespacial, sino un orden alfabético, como puede observarse en la Figura 2.8, este sistema también utiliza íconos sobre el mapa para representar categorías de sitios de interés: museos, restaurantes, hospitales, entre otros.

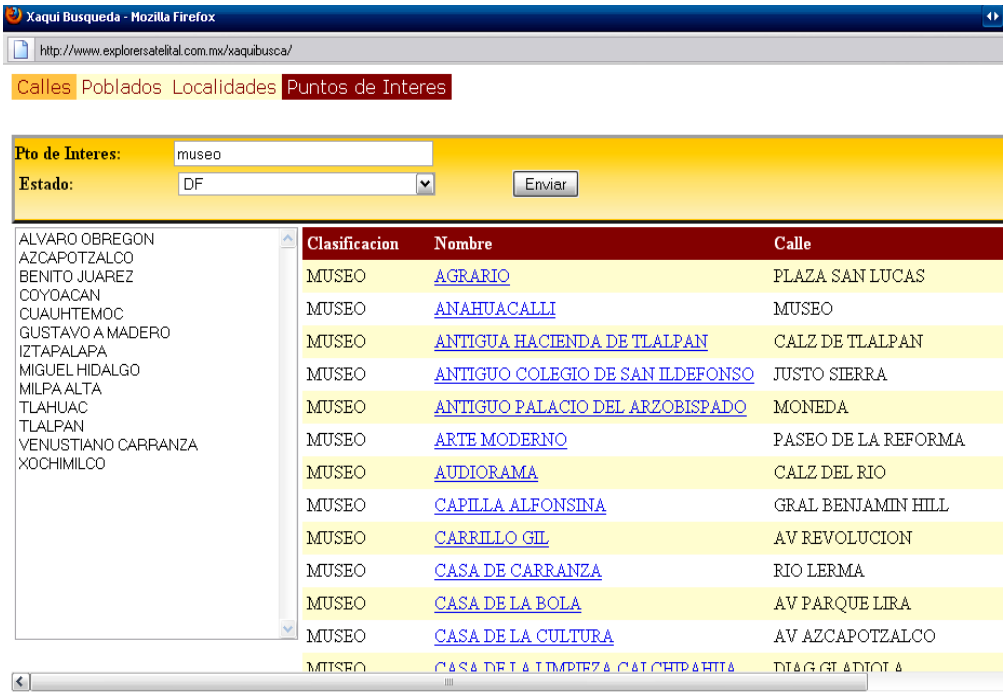

Figura 2.8 Ejemplo de búsqueda en el sitio xaqui [XAQ].

GeoRel: es un proyecto suizo en dispositivos móviles para evaluar la relevancia geográfica, que incluye aspectos, tales como: el espacial (dónde), temporalidad (cuándo), tópicos de objetos (qué, cuál) y motivacional del usuario. Ver Figura 2.9

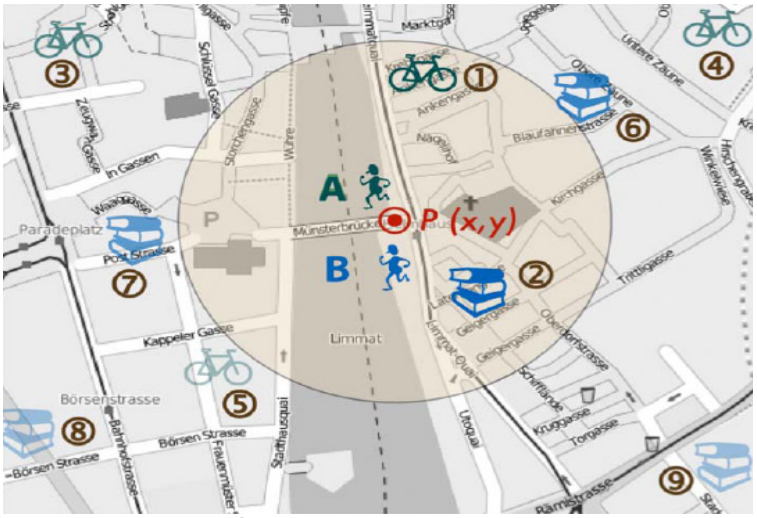

Figura 2.9 Ejemplo del análisis de proximidad espacial empleado en GeoRel.

Parte de su propuesta se centra en el análisis de proximidad espacial, empleando buffers, la propuesta se puede consultar en [REI09].

2.3.3.1. GIR con enfoque semántico geoespacial

***iRank* CIC-IPN:** en [MAT09] se propone abordar la recuperación y ponderación de información geográfica desde repositorios no estructurados orientada por ontologías, asumiendo que los procesos de recuperación pueden mejorarse si se integran las relaciones geográficas y topológicas asociadas a los objetos, extraídas desde fuentes de información heterogéneas, para luego ponderar los resultados de carácter topológico, geográfico y semántico mediante un método denominado *iRank*, que a su vez retoma ideas del uso de medidas de similitud en estructuras conceptuales. La propuesta para evaluar la relevancia de una entidad, en este caso un documento D_G, dada una consulta Q_G es calculado mediante la expresión (9).

$$RelInt = \frac{RelCon(Cq, Cd) + RelGeo(Gq, Gd) + RelTopoly(Tq, Td)}{Número\ de\ fuentes\ geográficas} \quad (9)$$

STORM University of Newcastle: en [HF05] es propuesto el modelo de relevancia ontológica espacio temporal (STORM por sus siglas en inglés), el cual usa una rejilla tridimensional para representar el grado de importancia de documentos geoespaciales en función de criterios espaciales, temporales y similitud semántica. Este enfoque se presenta como una interfaz interactiva para visualización tridimensional que puede complementar los sistemas de evaluación de relevancia basados en listas.

OASIS-Cardiff University: Christopher Jones et al [JAT01], proponen una forma de recuperar información basada en integración de medidas, su propuesta se basa en crear una distancia híbrida, que combina una medida de distancia jerárquica con la distancia euclidiana, las cuales son ponderadas por unos pesos pre-establecidos para luego sumarse. La expresión para evaluar la relevancia finalmente es la sumatoria de la distancia conceptual (jerárquica), la distancia temática y la distancia geométrica. Una desventaja que se observa es que, los autores establecen un valor para cada peso y sencillamente lo multiplican por los

valores de cada criterio, como una especie de promedio, sin especificar el porqué de esos valores en particular. La distancia jerarquía es calculada mediante (10).

$$\mathrm{HD}(q,c) = \sum_{x \in \{q.\mathrm{PartOf} - c.\mathrm{PartOf}\}} \frac{\alpha}{\mathrm{Lx}} + \sum_{y \in \{q.\mathrm{PartOf} - c.\mathrm{PartOf}\}} \frac{\beta}{\mathrm{Ly}} + \sum_{z \in \{q,c\}} \frac{\gamma}{\mathrm{Lz}} \quad (10)$$

Donde *Lx, Ly* y *Lz* son los niveles jerárquicos de los lugares individuales dentro de sus respectivas jerarquías. α, β y γ son pesos que proporcionan control para la aplicación de la medida propuesta. Además, se emplea la expresión (11) para calcular la distancia temática

$$TD(a,b) = \frac{C_{a,x_1}}{L_{x_1}} + \frac{C_{x_1,x_2}}{L_{x_2}} + \frac{C_{x_2,x_n}}{L_{x_n}} + \ldots + \frac{C_{x_n,b}}{L_b} \quad (11)$$

Donde cada razón en la expresión, corresponde a un enlace en la ruta más corta entre a y b. Los valores L_i representan el nivel de términos *i*, mientras que $C_{j,k}$ representan los pesos de los enlaces entre los términos *j* y k.

Finalmente las medias son combinadas mediante la expresión (12).

$$TSD(q,c) = w_e ED(q,c) + w_h HD(q,c) \quad (12)$$

Donde W_e y W_h son los pesos para la distancia Euclidiana y la distancia jerárquica.

Finalmente existen trabajos que ordenan los resultados según la relevancia de los atributos no presentes en la consulta, mas detalles en [Chu06], y por otro lado, existen enfoques derivados de las ciencias sociales que usan Escalamiento Multidimensional (MDS por sus siglas en ingles) para explorar la similitud en datos con múltiples atributos, generando resultados de carácter interpretativo [BG05].

2.4. Comentarios finales

Para resumir, se presenta en la Figura 2.10 un diagrama que engloba los puntos esenciales que se mencionan en esta revisión del estado del arte. Por una parte se describen los trabajos relacionados con las modalidades de recuperación de información, por otra parte se examinan algunas medidas para evaluar la similitud

entre entidades, enfocándose en aquellas que están organizadas en estructuras conceptuales, además se enfatiza en las estructuras conceptuales jerárquicas, debido al rol que juegan las jerarquías para poder medir la similitud semántica entre conceptos del dominio geográfico. En este sentido son descritos algunos trabajos que han servido como base para el desarrollo de nuevas investigaciones en este campo. Finalmente se mencionan algunos trabajos y aplicaciones que, o bien son sistemas para recuperación de información geográfica o además de ello, incluyen aproximaciones semánticas para la recuperación, centrándonos en la manera como integran diferentes criterios de búsqueda y cómo evalúan la relevancia de los resultados.

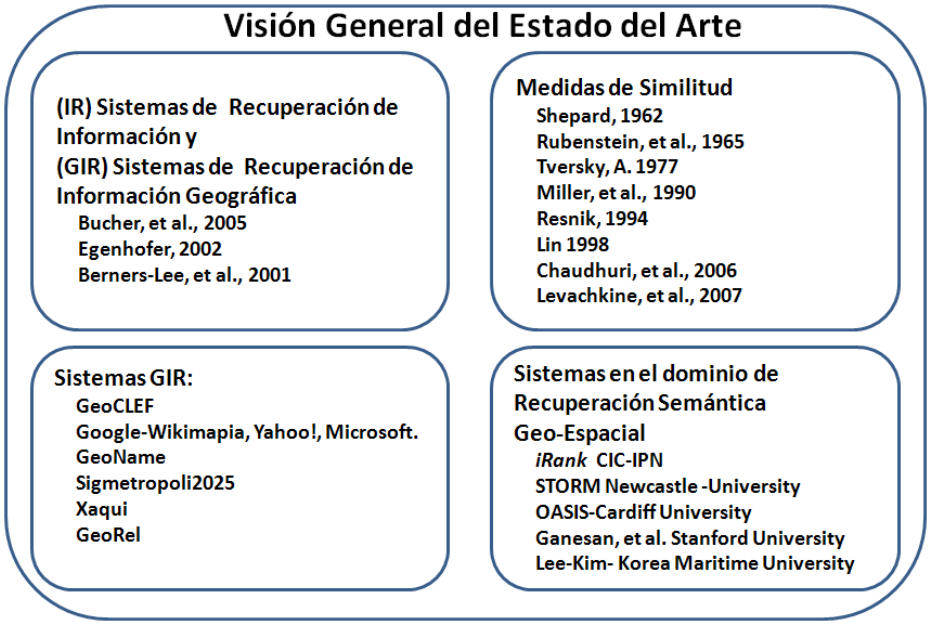

Figura 2.10. Esquema general del Estado del Arte.

El uso de medidas de similitud semántica toma cada vez mayor importancia en los sistemas de recuperación de información, pero es necesario adaptar esas medidas para que operen de manera adecuada en el dominio geográfico, permitiendo abarcar aquellos problemas donde no basta la comparación sintáctica, sino otras características que permiten recuperar resultados que por sus atributos o propiedades similares los hacen relevantes.

CAPÍTULO 3.
MARCO TEÓRICO

Poca ciencia aleja muchas veces de Dios, y mucha ciencia conduce siempre a ÉL.
Sir Francis Bacon

En este capítulo se presentan aquellos conceptos que dan soporte a la metodología propuesta en esta investigación, además se definen los conceptos y términos que se utilizarán. En especial, se describirán los aspectos relacionados con la recuperación de información geográfica desde fuentes de datos geográficas. También, se explican conceptos involucrados en los procesos de extracción, búsqueda y ponderación de datos. Para finalizar, se muestran las definiciones bajo las cuales se abordará el cálculo de similitud en jerarquías, apoyados en la *teoría de confusión*.

Este capítulo se organizará de la siguiente forma: en la primera sección se profundiza en los conceptos propios de IR y GIR, a continuación se tratan aspectos de la información geográfica y espacial, luego se explican las fuentes de datos geográficos, posteriormente se describe en detalle cada una de las fuentes de datos, se abordan las ontologías y su relación con GIR, seguidamente se habla del

tipo de consultas espaciales y geográficas de manera amplia, para pasar al análisis e indexado espacial. Finalmente en las últimas secciones se sientas las bases para el uso de las jerarquías y la *teoría de confusión*, y se concluye con una discusión y comentarios finales.

3.1 Fundamentos de la recuperación de información geográfica

Partiendo de la introducción presentada en la sección (2.1) donde se muestra la recuperación de información geográfica GIR (por sus siglas en ingles) como la encargada de proporcionar accesos a fuentes de información geográfica y espacial, incluyendo tanto, procesos de ponderación, almacenamiento e indexado [SJAB04]. GIR es un tema interdisciplinario de reciente nacimiento, pero de rápido desarrollo académico y comercial. En donde se procesan datos por medio de alguna noción de la relevancia geográfica en la información. La investigación actual enfrenta varios retos como se menciona en [Mat09], de entre los cuales se destacan:

- La extracción de términos geográficos de datos estructurados, que se complica aún más para datos no estructurados.
- La identificación y eliminación de ambigüedades en los procedimientos de extracción.
- Formas de almacenamiento eficiente de ubicaciones y sus relaciones.
- Desarrollo de motores de búsqueda y algoritmos para aprovechar las características de la información geográfica.
- La combinación de relevancia geográfica y contextual; así como, técnicas para permitir al usuario interactuar y explorar los resultados de consultas para sistemas GIR.

En IR el modelo subyacente que proporciona acceso a los datos es probabilista, se interesa en cuestiones subjetivas e indeterminadas; de tal forma que un documento es relevante (a un cierto grado) para el usuario y su consulta. Mientras que la recuperación de datos es determinista con respecto a operaciones de recuperación, o sea, si un documento cumple las condiciones especificadas en una consulta, entonces éste es por definición "relevante".
Los GIR contemplan ambos enfoques: recuperación determinista y recuperación probabilística.

En [LP04, Lar07] son presentas algunas definiciones, entre las cuales se dice que un GIR se enfoca en los problemas de encontrar fuentes de información que están relacionadas con ubicaciones geográficas. Debido a que la mayoría de motores de búsqueda tratan la terminología geográfica al igual que otra terminología, lo cual trae como consecuencia que se recuperen documentos irrelevantes. Actualmente, GIR trabaja en los siguientes retos:

• Existen muchos lugares con el mismo nombre (documentos que se refieren al lugar equivocado son recuperados).
• Existen muchos usos para nombres de lugares. Por ejemplo, se usan nombres de personas y de organizaciones para referirse a un lugar.
• Existen consultas que incluyen preposiciones espaciales, tales como cerca, afuera, que requieren consideraciones especiales en el ámbito geográfico.

Una de las metas en GIR es la de proporcionar acceso a fuentes de información georeferenciada. Por lo tanto, se puede considerar a la recuperación de información geográfica como una especialización de la recuperación de información clásica. Esto incluye todas las áreas que tradicionalmente conforman el núcleo de la investigación en IR, con especial énfasis en el indexado y recuperación de información orientada espacial y geográficamente. Estos procesos se realizan de

manera común a través de consultas espaciales (*Spatial Query*) o de consultas geográficas.

En la literatura computacional han existido distinciones entre IR y recuperación de datos (*data retrieval*) este último término está asociado con sistemas manejadores de bases de datos (SMDB). En la Tabla 3.1 se muestra una comparación de estos dos enfoques. El estudio de sus características permitirá establecer la diferencia entre estos.

Tabla 3.1 Espectro de la Recuperación de información Vs. Recuperación de datos.

Recuperación de Información	**Recuperación de Datos**
Enfoque Probabilista	Enfoque Determinista
La extracción se hace desde contenidos	Se toman elementos completos
No es necesaria coincidencia exacta	Debe haber coincidencia exacta
Admite consultas en lenguaje natural	Consultas en lenguaje estructurado
Resultados ordenados en función de su relevancia	Resultados obedecen a la coincidencia, sin un orden

Entonces, al hablar de recuperación de información geográfica, se hace alusión a ambos modelos de recuperación, el determinista (e.g. recuperar todos los datos que corresponden a una coordenada en particular) y probabilista (e.g. encontrar sitios de interés relacionados en determinada área).

De esta manera, la recuperación de información requiere de indexados para garantizar un acceso eficiente a grandes volúmenes de información en bases de datos. La mayoría de los sistemas IR obtienen cada índice a partir del contenido de

los elementos a ser indexados. Que puede obedecer a palabras clave, extracción por inferencia u otras técnicas de análisis inteligente para asignación de índices.

En la recuperación actual de elementos de una BD, los algoritmos utilizados para establecer coincidencias (*matching*) de una consulta con los elementos del índice (o el contenido de una base de datos) están basados en modelos de recuperación tradicional. Los modelos de recuperación de información guían a una clase de algoritmos de recuperación que pueden involucrar el cálculo de probabilidades y métodos de inferencia estadísticos, hasta enfoques basados en modelos que usan el espacio de documentos [Sal89].

El principal enfoque de estos modelos es encontrar el conjunto de coincidencias potenciales entre una *consulta* y un documento, donde la ponderación se basa en parámetros que miden el *grado de coincidencia*, de tal forma que las "*mejores*" coincidencias (de acuerdo a alguna medida) reciben mayor peso.

Como se explicó, los métodos de recuperación de datos son deterministas, lo que implica que requieren *coincidencia* exacta entre lo que solicita en la *consulta* y lo que contiene la BD. Por esta razón, de manera general la lógica *Booleana* es utilizada en el procesamiento de los lenguajes basados en consultas, como también en los catálogos en línea. Es conveniente resaltar que, en GIR no se descarta el uso de métodos de coincidencia determinista aproximada, parcial y precisa para el procesamiento de *consultas geográficas* [Lar07].

Las consultas en sistemas de IR son generalmente expresadas como enunciados en lenguaje natural de acuerdo a las necesidades de los usuarios que buscan información. Estas *consultas* pueden ser imprecisas, incluso ambiguas. Mientras que en la recuperación de datos la *consulta* es típicamente expresada en algún tipo de lenguaje estructurado cuya sintaxis es precisa y preestablecida.

De forma general, los resultados de una búsqueda son presentados en un orden ponderado, donde el criterio de ponderación se basa en el "*grado de coincidencia*" entre la *consulta* y el elemento de la base de datos.

3.2 Información geográfica y espacial

A pesar de que los términos geográfico, espacial y geoespacial, se usan frecuentemente de manera sinónima, no son iguales; aunque existe una estrecha relación entre ellos, no obstante hacen alusión a aspectos diferentes.

El término *geoespacial* ha sido diversamente definido, es un término ampliamente utilizado para englobar una ubicación espacial[1], los métodos analíticos y datos geográficos o terrestres. Igualmente el término también es muy usado en conjunción con GIS y Geomática. Muchos sistemas de información geográfica usan el término análisis geoespacial para hacer alusión a operaciones basadas en datos vectoriales: superposición de capas, análisis de proximidad, entre otras.

Lo anterior refleja la amplitud de la información geoespacial y su difícil extracción cuando no está estructurada para hacer un procesamiento semántico, tal es el caso de muchas fuentes de información actualmente (entre ellos el abordado aquí), así que es difícil explotar el conocimiento de estos dominios.

3.2.1 Datos geográficos

En este trabajo se usa el término objeto o dato geográfico para hace referencia a *características* del ámbito de los GIS. Las características que definen a un objeto

[1] Open Geospatial Consortium (OGC), www.opengis.org

geográfico, de acuerdo a la OGC son: debe existir un único objeto geográfico relacionado con cada entidad del mundo real (el globo terráqueo). Por "objeto geográfico" (OG) se hace referencia a la unidad esencial de datos con la cual es conformada una base de datos geográfica.

- Por "entidad del mundo real" se asume cualquier cosa o fenómeno del cual se tengan almacenados datos geo-referenciados.
- Los objetos tienen un único y persistente ID. A veces usado para localizarlo.
- Los OG tienen descripciones, las cuales no tienen restricciones únicas.
- Los OG se componen de uno o más atributos (propiedades con nombre, tipo, valor, u otra descripción opcional). Los atributos pueden ser nulos. La semántica de un atributo requiere que sea identificado unívocamente por su nombre. Debe ser posible preguntar por un atributo o por nombre.

Un OG es una combinación de una coordenada geométrica y de un sistema de referencia de coordenadas. En general, un OG es un conjunto de puntos geométricos, representado por posiciones directas, las cuales guardan las su ubicación en un sistema de referencia de coordenadas [OGC].

Por otra parte, los datos geográficos son mucho más que imágenes digitales de mapas. Ya que describen el entorno permitiendo por ejemplo, planificar el crecimiento de ciudades, anticipar y reaccionar ante ciertos eventos naturales o artificiales, servicios de emergencia, además son fundamentales para análisis de la dinámica de muchos otros dominios como el ambiental, climatológico, patrones delictivos, entre otros. Comúnmente los datos geográficos se procesan con un GIS, que a su vez generan como uno de sus productos de información mapas [Mat08].

En México se producen estos datos geográficos en forma de cartografía digital generalmente en formato *raster o vectorial*.

3.2.2 Datos espaciales

Los datos espaciales se refieren a entidades o fenómenos que cumplen los siguientes características básicas:

• Tienen posición absoluta: sobre un sistema de coordenadas (x, y, z)
• Tienen una posición relativa con respecto a otros elementos (topología)
• Tienen una figura geométrica que los representa (punto, línea, polígono)
• Tienen atributos que lo describen (características propias)

En conclusión, se puede decir que todos aquellos datos con alguna forma de referenciación con coordenadas terrestres (i.e. geo) son datos geográficos, que en forma general pueden llegar a constituir datos geo-espaciales.

3.3 Ontologías como fuente semántica de datos

Las ontologías en el área computacional, tienen hoy un significativo auge. Pero, ¿qué son las ontologías en el campo de la computación, de dónde vienen?, Para comenzar se presentan algunas definiciones generales, que hoy son casi consenso.

• Una Ontología es una especificación explícita de una conceptualización [Gru95].
• "Una Ontología es el estudio metafísico de la naturaleza del ser y su existencia" [Wor].
• "Campo de la metafísica que investiga y explica la naturaleza, las características y las relaciones esenciales de todos los seres como tales, o los principios y las causas de ser" [Gua95].

Entonces, una Ontología es una descripción formal de conceptos *consensada* en *un dominio de discurso*, bajo un marco teórico que especifica un *vocabulario*.

Dicho vocabulario define entidades, clases, propiedades o atributos, funciones o procesos y relaciones entre estos componentes.

Se debe señalar que las ontologías toman un papel clave en la resolución de interoperabilidad semántica entre sistemas de información, por esta razón, su uso dentro del contexto computacional aporta avances a los ya conocidos metadatos.

3.3.1 Ontologías como conceptualización compartida

Una de las primeras y más citadas aportaciones a este campo de las ontologías es la realizada por Thomas Gruber a mediados de los 90's, al presentarlas como un medio para reutilizar y compartir el conocimiento entre aplicaciones [Gru95]. La definición propuesta expresa que: "*Una Ontología es una especificación explicita, formal de una conceptualización compartida*". De donde se realizan las siguientes observaciones: Una *conceptualización* se refiere a un modelo abstracto de algún fenómeno en el mundo, el cual identifica los conceptos relevantes de dicho fenómeno. *Explícito* significa que los tipos de conceptos utilizados, y las restricciones en su utilización, están claramente definidos. *Formal* se refiere al hecho de que la Ontología debe ser legible y procesable por los programas (*machine-readable*). Mientras que *Compartido* significa que la Ontología surge a partir de un *consenso* entre muchas partes.

Según esta definición una base de datos puede ser considerada una Ontología. Ya que es una representación abstracta de un fenómeno del mundo real, también es explícita y procesable por computadora. Sin embargo, surge la pregunta: ¿representa un consenso, es decir un conocimiento compartido?, los esquemas de base de datos son típicamente desarrollados para uno o un conjunto limitado de aplicaciones. Mientras que una Ontología debe ser un acuerdo entre muchas partes acerca de un dominio.

3.3.2 Ontologías, sus tipos y la lógica de operación.

Otro aporte significativo sobre ontologías, es el de Nicola Guarino [Gua95]. Realiza una introducción formal de los principios básicos para ingeniería del conocimiento, explorando las relaciones entre Ontología y representación de conocimiento.

Dado que una ontología es una especificación de una conceptualización compartida, los expertos en el dominio, los usuarios y diseñadores necesitan estar en común acuerdo en los conceptos especificados, para que la ontología sea útil.

Ya que es difícil lograr tal acuerdo, resulta útil dividir el conocimiento en capas para diferentes ontologías basándose en su nivel de generalidad, así no es necesario que todos estén de acuerdo con todas las ontologías. Aunque sí debe haber consenso en las ontologías de nivel superior (las de mayor generalidad conceptual), las cuales son el punto de partida para ontologías de dominio específico y de aplicación [Mee99]; por lo anterior se sugiriere que el conocimiento se establezca por lo menos en tres capas: la del lenguaje, la del dominio y la de aplicación, es decir, un orden de generalización descendente.

En el trabajo [Gua95], Guarino identifica igualmente tres capas de conocimiento correspondientes a cuatro diferentes tipos de Ontologías, basándose en sus niveles de generalidad, denominadas como sigue:

- Ontologías de Alto Nivel o Genéricas: éstas describen conceptos más generales. En relación con los Sistemas de Información, estas ontologías describirían conceptos básicos. Por ejemplo, WordNet [Wor] y Cyc [Len95].
- Ontologías de Dominio: Describen un vocabulario relacionado con un dominio genérico. Por ejemplo, podría ser una descripción de datos y entidades relacionados con el sensado remota con un ambiente urbano.

- Ontologías de Tareas o de Técnicas básicas: Describen una tarea, actividad o artefacto. Por ejemplo, la evaluación de la contaminación sonora en ambientes urbanos o la descripción de características generales de componentes, procesos o funciones.
- Ontologías de Aplicación: Describen conceptos que dependen tanto de un dominio específico como de una tarea específica y, generalmente son una especialización de ambas. Fonseca propone que este tipo de ontologías nazcan a partir combinaciones de ontologías de nivel superior [Fon00]. Ellas representan las necesidades de usuarios relacionados con una aplicación específica.

En [McG83] se presenta una clasificación bastante completa, puede verse como un espectro de ontologías de acuerdo al grado de expresividad como muestra la Figura 3.1.

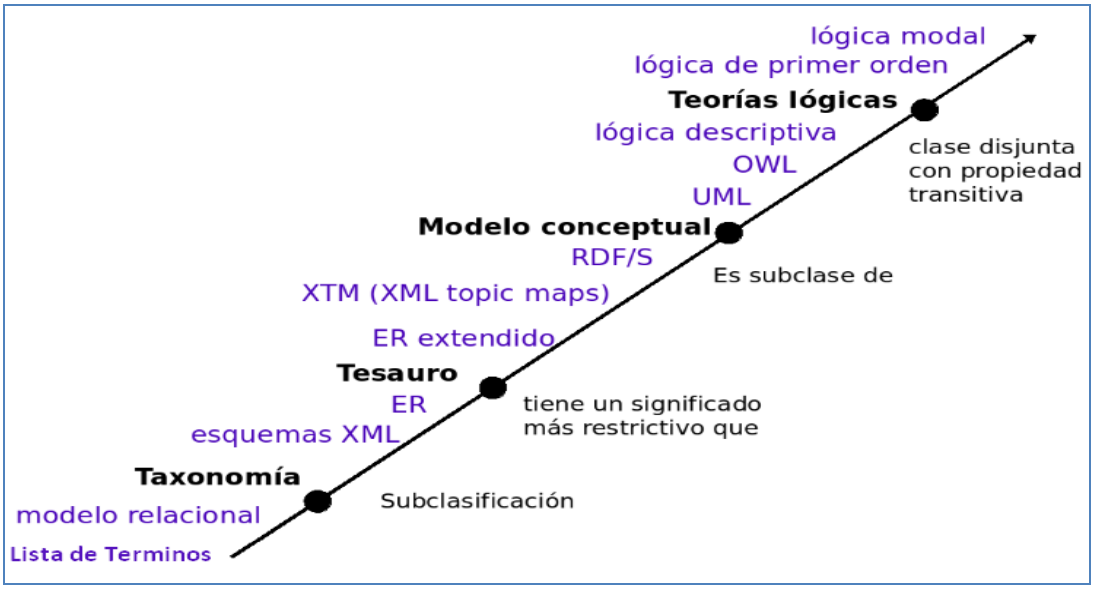

Figura 3.1 Espectro de Ontologías.

Sumando a lo anterior, otra clasificación es propuesta por Van Heijst [HSW96]:

- Ontologías terminológicas: Especifican los términos que son usados para representar el conocimiento en el universo del discurso. Suelen ser usadas para unificar vocabulario en un campo determinado.

- Ontologías de información: Especifican la estructura de almacenamiento de bases de datos. Ofrecen un marco para el almacenamiento estandarizado de información.
- Ontologías de modelado de conocimiento: Especifican conceptualizaciones del conocimiento. Contienen una rica estructura interna y suelen estar ajustadas al uso particular del conocimiento que describen.

Ahora bien, en términos prácticos, se observa que el diseño y desarrollo de una Ontología contempla los siguientes pasos generales:
- Definir clases del dominio.
- Conformar una jerarquía de clases (superclase- subclases).
- Definir atributos (slots).
- Describir los valores permitidos para los atributos.
- Finalmente poblar o rellenar los valores de los slots.

3.4 Consultas en geoinformática

Para hablar de las consultas en el área de información de carácter geoespacial, hay que abordarlo desde las consultas geográficas y consultas espaciales, lo que implica consultar a una base de datos indexada espacialmente. Dicha indexación se efectúa con base en las relaciones entre elementos particulares de la base de datos y dentro de un sistema de coordenadas particular. Las consultas espaciales o búsquedas espaciales (*spatial querying*) son un término más amplio.

Una *consulta* espacial puede ser definida como aquella que involucre relaciones espaciales (*intersección, contenido, con límites, adyacencia, proximidad*) de entidades geométricamente definidas y localizadas en el espacio, sin considerar la naturaleza del sistema de coordenadas [DFMP92].

Una *consulta geográfica*, también requiere de lo espacial, pero a través de sistemas de coordenadas, las cuales están bien definidas y delimitadas en relación con su correspondencia al mundo real (i.e espacio de coordenadas terrestres). Como se señala en [Fra91], existen características de los datos geográficos que requieren estructuras de datos y métodos de acceso espacial.

En general, las relaciones geográficas en los sistemas de coordenadas impuestos en el mundo real son relaciones geométricas [Lar07]. Dentro de un marco de trabajo geométrico, donde la dirección y distancia pueden ser medidas en una escala continua, muchos tipos de relaciones entre objetos definidos dentro de este espacio pueden ser definidos usando la geometría.

3.4.1 Lenguajes de consulta.

Dado los sistemas de almacenamiento, los cuales posibilitan mantener las estructuras conceptuales (para nuestro caso las ontologías) en repositorios de datos), es conveniente definir cómo acceder a la información contenida en estos sistemas, típicamente descritos en RDF. Una de las herramientas más usadas hoy, es el API de Jena [Jen07], desarrollado en Java por Hewlett-Packard. Jena a su vez está compuesta por un conjunto de herramientas, que dan soporte al desarrollo de ontologías tales como RDFS, DAML+OIL y OWL, mediante las cuales se pueden definir OntoClases y OntoPropiedades, las cuales dan sustento al trabajo desarrollado en esta investigación, especialmente a uno de los productos de investigación de este trabajo, como lo es la API de confusión V1.0.

Ahora, un excelente complemento para efectuar las consultas sobre RDF y RDFS, es el lenguaje RQL (RDF Query Language – por sus siglas en ingles) el cual nos permite navegar por los grafos que hay en el modelo RDF, a través de

mecanismos de posicionamiento en los nodos del modelo que, satisfacen semánticamente una consulta (i.e a través de las relaciones y propiedades entre clases e instancias). Este marco da origen a RDQL y SPARQL.

Para RDQL, la estructura de grafos que contiene un RDF, permite especificar un patrón de una tripleta de la siguiente forma: un sujeto que cumple un predicado sobre un objeto, de tal forma que, este nivel de expresividad logra recuperar todas aquellas tripletas <sujeto, predicado, objeto> o parte de las mismas dentro del grafo en el modelo RDF, que cumplen dicho patrón.

SPARQL es propuesto por la W3C [Spa07], como una alternativa más flexible aún, de tal forma que es posible extraer información de manera más variada, desde grafos o sub-grafos RDF hasta URIs, incluso generando nuevos grafos a partir de la información de los grafos RDF consultados.

3.5 Análisis espacial y procesamiento de datos

El análisis espacial, es con frecuencia referido como modelado. Se refiere al análisis de fenómenos distribuidos en el espacio y que poseen dimensiones físicas (la ubicación de, cercano a, o la orientación de un objeto respecto a otro; relación con un área de un mapa referenciado o relacionado a una ubicación especifica en la superficie del planeta). Además de esto, el análisis espacial a veces es visto como un modelo de interpretación de resultados útiles para evaluación de disponibilidad, capacidad, estimación o predicción.

Enfocado a GIS, se puede decir que existen cuatro tipos de análisis espacial: *análisis por sobreposición* y *contigüidad*, *análisis de superficies*, *análisis lineal*, y

análisis raster. Este incluye funciones GIS como sobreposición topológica, generación de buffers, y modelado de redes como cita [Mat09].

3.5.1 Semántica espacial

Aunque no existe un consenso universal para establecer una definición formal y que en consecuencia sea aceptada por la comunidad científica en general. Se intenta describir al afirmar que la semántica espacial se encarga del significado espacial y/o geográfico de los datos. Por ejemplo, el nombre de un objeto geográfico y sus posibles representaciones (fotos, datos vectoriales, imágenes raster, etc). Así como las relaciones que tienen con otros objetos (si están cerca en términos de distancia o tiempo, si comparten un punto, si una carretera los conecta, etc.)

Por otra parte, están las denotaciones que tienen esos datos, es decir la semántica se describe a partir de metadatos con respecto a descripciones, atributos y aspectos espaciales de los objetos geográficos. También, es comúnmente citada la estadística que hasta el 80% de la información está referenciada espacialmente en alguna forma.

Sin embargo, los aspectos espaciales de mucha de esa información no son utilizables para la computación debido a su inescrutable semántica (no está expresada o es expresada informalmente). Además de las limitaciones de los razonadores y lenguajes formales para representar la semántica espacial. Ya que actualmente, descubrir, procesar, analizar y visualizar datos geográficos, aun requiere expertos que puedan entender el significado de los datos espaciales y que puedan cumplir con dichas tareas de forma inteligente.

3.6 Jerarquías y Teoría de Confusión

A continuación se describen los conceptos de jerarquías que conforman los fundamentos para definir la *Teoría de Confusión*. En [GAL04, LG04] se formulan para el concepto de jerarquías las siguientes definiciones:

Definición (*Conjunto de elementos).* Un conjunto *E* cuyos elementos se definen explícitamente.

Ejemplo: {rojo, azul, verde, naranja, amarillo}.

Definición (*Conjunto ordenado).* Un conjunto de elementos cuyos valores están ordenados por la relación *< ("menor que")*

Ejemplo: {muy frío, frío, templado, tibio, caliente, muy caliente}.

Definición (*Cubrimiento)*. *K* es un cubrimiento del conjunto *E* si *K* es un conjunto de subconjuntos $e_i \subset E$, tal que $\cup ie_i = E$

Cada elemento de *E* está en algún subconjunto $e_i \in k$. Si *K* no es un cubrimiento de *E*, se puede hacer añadiendo un nuevo e_j, denominado "otros", que contiene todos los elementos restantes de *E* que no pertenecen a ningún e_i previo.

Definición (*Conjunto exclusivo*). *K* es un conjunto exclusivo si $e_i \cap e_j = \varnothing$ para cada $e_i, e_j \in k$

Sus elementos son mutuamente exclusivos. Si *K* no es un conjunto exclusivo, se puede hacer reemplazando cada par que se traslape $e_i, e_j \in k$ con tres: $e_i - e_j, e_j - e_i, e_i \cap e_j$.

Definición (*Partición).* *P* es una partición del conjunto *E,* si P es cubrimiento y conjunto exclusivo de *E.*

Definición (*Variable cualitativa).* Una variable uni-valuada que toma valores simbólicos.

Su valor no puede ser un conjunto. Por simbólico se entiende que es cualitativo, opuesto a numérico, o variables cuantitativas.

Definición (Valor Simbólico). Un valor simbólico *v* representa al conjunto *E*, descrito por $v \alpha E$, si *v* puede ser considerado un nombre o una representación de *E*.

Definición (Jerarquía). Para un elemento del conjunto *E*, una jerarquía *H* de *E* es otro conjunto de elementos, donde cada e_i es un valor simbólico que representa a cualquier elemento de *E* o una partición; y $\cup_t \left\{ \rho_t \middle| \varepsilon_t \propto \rho_t \right\} = E$ (La unión de todos los conjuntos representados por los e_i de *E*).

Suponga una **Jerarquía H1,** Para un conjunto *E* = {Canadá, USA, México, Cuba, Puerto_Rico, Jamaica, Guatemala, Honduras, Costa_Rica}, la jerarquía H1 es {Norte_America, Islas_Caribe, América_Central}, donde Norte_América $\propto$ {Canadá, USA, México}; Islas_Caribe $\propto$ {Islas_Hispano-parlantes, English_Speaking_Island}; Islas_Hispano-parlantes $\propto$ {Cuba, Puerto_Rico}; English_Speaking_Island $\propto$ {Jamaica}; América_Central $\propto$ {Guatemala, Honduras, Costa_Rica}.

Las jerarquías permiten comparar valores cualitativos que pertenecen a ella.

Definición (*Variable Jerárquica)* Una variable jerárquica es una variable cualitativa cuyos valores pertenecen a la jerarquía.

Ejemplo: lugar_de_origen toma valores de H1.

Definición Se utilizará la siguiente notación: a) **padre_de** (*v*). En un árbol que representa una jerarquía, el nodo **padre_de** un nodo es aquel del cual desciende; b) los **hijos_de** (v) son los valores que cuelgan de *v*. Los nodos con el mismo padre son **hermanos**; c) **abuelo de**, **hermano_de, tío_de**, **ascendentes**, **descendientes**... son definidos, cuando ellos existen; d) La **raíz** es el nodo que no tiene padre.

3.6.1 Jerarquía simple, ordenada, porcentual y mixta

Una jerarquía describe la estructura de valores cualitativos en un conjunto *E*. Son definidas las siguientes jerarquías:

Definición (*Jerarquía Simple)*: Una jerarquía simple (normal) es un árbol con raíz *E* y si un nodo tiene un hijo, éste forma una partición del padre.

Una jerarquía simple describe una jerarquía donde *E* es un conjunto (así sus elementos no se repiten ni son ordenados).

Ejemplo: ser viviente {animal {mamífero, pez, reptil, otro animal}, planta {árbol, otra planta}}.

Definición (*Jerarquía Ordenada)*: En una jerarquía ordenada, los nodos de algunas particiones obedecen una relación de orden.

Ejemplo objeto {diminuto, pequeño, mediando, grande}

Definición (*Jerarquía Porcentual)*: En una jerarquía porcentual, el tamaño de cada conjunto es conocido.

Ejemplo: Continente Americano (740M) {Norteamérica (430M) {USA(300M), Canadá(30M), México(100M)} América Central (10M), Sudamérica(300M)}.

Definición (*Jerarquía Mixta)*: Una jerarquía mixta combina los tres casos anteriores Para cada uno de los tipos de jerarquías se define *conf(r,s)* como el error de usar un valor *r* en vez de s.

3.7 Teoría de Confusión

La *teoría de confusión* consiste en modelar la similitud entre objetos de una o de diversas jerarquías. Se basa en una medida asimétrica y dependiente del contexto denominada *confusión* (de usar un valor cualitativo en vez del valor esperado o correcto) [GAL04]. El *término "confusión"* se definió para diferenciar este enfoque

de otros que se orientan a medir distancias entre conceptos (i.e., simetría, medidas independientes del contexto, cercanía, similitud).

En este caso, la asimetría de la *confusión* se da por su definición y su dependencia al contexto por la estructura jerárquica. El concepto de *confusión* permite definir la cercanía a la que un objeto cumple un predicado como también, derivar otras operaciones y propiedades entre valores jerárquicos.

3.7.1 Confusión de usar r en vez de s en jerarquías simples

Definición. Si *r*, *s* $\in H$, entonces la *confusión* de usar *r* en vez de *s*, que se escribe como

conf(*r*, *s*), es:

- *conf* (*r*, *r*) = *conf* (*r*, *s*) = 0, donde *s* es cualquier ascendente de *r*; (regla 1)
- *conf* (*r*, *s*) = 1 + *conf* (*r*, padre_de(*s*)) . (regla 2)

Para medir *conf*, se cuentan los enlaces *descendentes* de *r* a *s*, al valor reemplazado, este valor se suele normalizar al dividirlo por la profundidad d H

3.7.2 Confusión de usar r en vez de s, en jerarquías ordenadas

Definición. Para jerarquías simples compuestas por conjuntos ordenados, la *confusión* de usar *r* en vez de *s*, *conf'* (r, s), se define como:

- *conf'* (*r, r*) = *conf* (*r*, cualquier ascendente de *r*) = 0;
- Si *r* y *s* son hermanos distintos, *conf'* (r, s) = 1 Si el padre no está en un conjunto ordenado; entonces, *conf'* (r, s) es igual a la distancia relativa de *r*

a *s* es y a la vez es igual el número de pasos requeridos para llegar de *r* a *s* en el orden definido, dividido entre la cardinalidad-1 del padre;

(regla 3)

• *conf'* (*r*, *s*) = 1 + *conf* (*r*, padre_de(*s*)).

Esto es como *conf* para las *jerarquías constituidas por conjuntos*, excepto que allí el error entre dos hermanos es 1, y aquí es un número ≤ 1.

Por ejemplo: Temperatura = {congelado, frío, normal, tibio, caliente, ardiendo}; en esta lista *conf'* (congelado, frío) = 1/5, mientras que la *conf'* (congelado, ardiendo) = 5/5 = 1.

3.7.3 Confusión de usar r en vez de s, en jerarquías porcentuales

Considerando la jerarquía *H* (de un elemento del conjunto *E*) pero compuesta por un conjunto desordenado en vez de un conjunto ordenado.

Definición. Para conjuntos desordenados, la *confusión* de usar *r* en lugar de s, *conf''* (*r*, *s*), es:

• *conf''* (*r*, *r*) = *conf''* (*r*, *s*) = 0, cuando *s* es cualquier ascendente de *r*;

• *conf''* (*r*, *s*) = 1 – proporción relativa de *s* en *r*.

(regla 4)

3.7.4 Confusión de usar r en vez de s, en jerarquías mixtas

Definición. Para calcular *conf''*(*r*, *s*) *en una jerarquía mixta*:

• Aplicar (regla 1) para la ruta *ascendente* de *r* a *s*;

• En la ruta descendente, usar (regla 3) en vez de (regla 2), si *p* es un conjunto ordenado; o usar (regla 4) en vez de (regla 2), cuando los tamaños de *p* y *q* son conocidos. Es decir, usar (regla 4) para las jerarquías porcentuales en lugar de (regla 2). Esta definición es consistente y reduce

las definiciones previas de jerarquías porcentuales, simples, ordenadas y mixtas.

3.7.5 Valores obtenidos dada una confusión.

Definición. Un valor *u* es igual al valor *v*, dentro de una *confusión* dada ε, escrita $u =_{\varepsilon} v$, si $conf(u, v) \leq \varepsilon$ (Esto significa que el valor *u* puede ser usado en vez de *v*, con un error ε).

Ejemplo: Si v = limón (ver Figura 3.2), entonces (a) el conjunto de valores es igual a v con *confusión* 0 es {limón}; (b) el conjunto de valores iguales a v con *confusión* 1 es {cítrico, limón}; (c) el conjunto de valores iguales a v con *confusión* 2 es {planta, cítrico, pino, limón}.

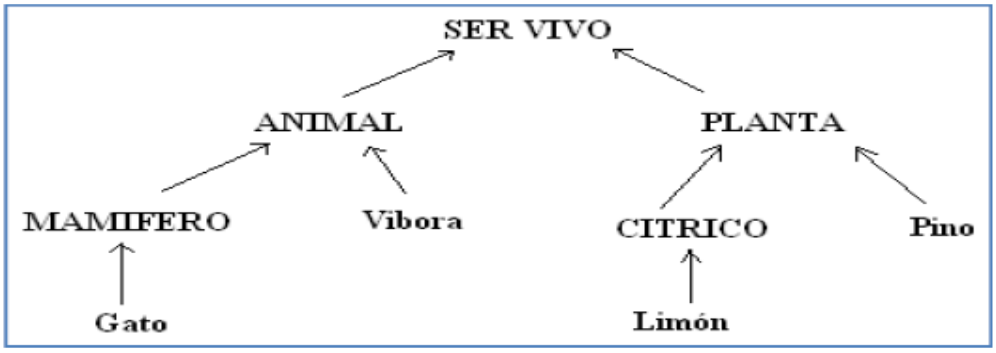

Figura 3.2 Jerarquía para un ser vivo

3.7.6 Objetos idénticos, muy similares, algo similares

Los objetos son entidades descritas por pares *k* (propiedades, valores), que en la notación se hará referencia como pares (variable, valor) [LG07]. En bases de datos son denominados como pares (relación, atributo). Un objeto o con *k* pares (variable, valor) es escrito como (o (v_1 a_1) (v_2 a_2)... (v_k a_k)).

Se desea estimar el error de utilizar el objeto o' en vez del objeto o. Para un objeto o con *k* (tal vez jerárquica) variables v_1, v_2 ,..., v_k y valores a1 , a_2 ,..., a_k, se dice

que otro objeto o' con las mismas variables v_1, v_2 ,..., v_k pero con valores a'_1, a'_2 ,..., a'_k, se presentan las siguientes definiciones:

Definición *o´* es *idéntico* a *o*, if $a'_i = a_i$ para todas $1 \leq i \leq k$. Los valores correspondientes son idénticos.

Si todo lo que se sabe sobre *o y o´* son los valores de las variables v_1 , v_2 ,..., v_k, y ambos objetos tienen valores idénticos, se puede decir que "con todo lo que se sabe", *o* y *o´* son el mismo.

Definición *o´* es *sustituto* para *o*, si $conf(a'_i, a_i) = 0$ para todo $1 \leq i \leq k$.

No se trata de una *confusión* entre el valor del atributo de *o´* y el valor correspondiente de *o*. Se puede usar *o´* en vez del *o* solicitado con $conf = 0$.

Definición *o´* es *muy similar* a *o*, si $\sum_i conf(a'_i, a_i) = 1$.

Definición *o´* es *similar* a *o*, si $\sum_i conf(a'_i, a_i) = 2$.

Definición *o´* es *algo similar* a *o*, si $\sum_i conf(a'_i, a_i) = 3$.

Definición En general, *o´* es $similar_n$ a *o*, si $\sum_i conf(a'_i, a_i) = n$.

3.8 Comentarios finales

Se describieron los fundamentos de la recuperación de información, tomando especial cuidado del enfoque geoespacial. Se expuso los elementos considerados hoy, como pilares para la representación del conocimiento en estructuras conceptuales procesables por máquinas, y finalmente se presentaron los cimientos teóricos de la *teoría de confusión*, los cuales darán sustento a una de las principales fases de la propuesta metodológica.

El objetivo es exponer las razones que sustentan el uso de un mecanismo integrador para recuperar información más allá de las coincidencias sintácticas, dirigido más bien, por la similitud semántica de los datos en el dominio del turismo cultural, organizado en estructuras jerárquicas como las que aquí se presentaron.

Finalmente se reconoce que, aunque la similitud semántica no es un asunto trivial, más aún cuando se busca que las máquinas la procesen, en este apartado se estableció el marco común de trabajo para determinar en qué medida son similares dos entidades de una estructura conceptual jerárquica, que condensa el conocimiento de un dominio en particular, de esta manera las entidades con mayor similitud semántica, son las que produzcan menor confusión, o sea, menor error al ser usadas como alternativas en lugar de la entidad originalmente consultada, esto inmediatamente conlleva a pensar que, cuando un usuario realice una consulta preguntando por una entidad, es posible ofrecer como resultados, otras entidades semánticamente similares, en adición a la entidad buscada.

De tal forma que, si la entidad consultada existe, se recuperará y se añadirán otras respuestas semejantes, cumpliendo así con el nivel actual en el estado del arte de IR, pero si no se cuenta con la entidad consultada, entonces se evitarán los resultados vacíos, regresando resultados estrechamente relacionados con la consulta debido a su similitud semántica, con lo cual se logra contribuir al estado del arte de los sistemas de recuperación de información geográfica, principalmente en el dominio del turismo cultural.

CAPÍTULO 4.

METODOLOGÍA

El éxito no se logra sólo con cualidades especiales. Es sobre todo un trabajo de constancia, de método y de organización.
J.P. Sergent
El hombre encuentra a Dios detrás de cada puerta que la ciencia logra abrir.
Albert Einstein

En este capítulo se describe la metodología desarrollada en el presente trabajo de investigación, metodología compuesta por tres grandes bloques: la conceptualización del dominio, el análisis multi-criterio y la parte de procesamiento e interacción con el usuario. Dichos bloques están constituidos a su vez por módulos, destacando en ellos, el diseño de la base de datos conceptual, la base de datos geográfica, el análisis semántico y el análisis geoespacial. Finalmente se describe el modulo que procesa las consultas, y la integración de resultados. La metodología propuesta se presenta como un modelo para la recuperación de información, el cual es expandible y aplicable a otros dominios. En este trabajo se aborda la recuperación de información del dominio del turismo cultural, pretendiendo encontrar Puntos de Interés Cultural (CPI por sus siglas en Ingles) particularmente museos.

Antes de describir en detalle cada uno de los pasos de la metodología, se describen los escenarios a los que esta busca dar solución:

- El Usuario ubicado en una estación del Sistema de Transporte Colectivo Metro, dentro del área del Centro histórico de la Ciudad de México (ejemplo: Estación Metro "Bellas Artes"), consulta al sistema por un *Museo de Pintura*.

- El sistema busca "*Museo de pintura*" en su base de datos de acuerdo a una estructura conceptual jerárquica previamente construida, con el fin de obtener los resultados exactos y/o más similares.

  Ahora bien, existen dos posibilidades o escenarios generales:

  - ***(a)*** Que haya uno o más museos que satisfagan la consulta.
  - ***(b)*** Que no hayan museos que cumplan con la consulta

  Bajo el escenario ***(a)*** "que un sí haya algún *Museo de Pintura* en el área del centro histórico, no hay inconveniente alguno, debido a que es el caso clásico que pueden resolver los sistemas manejadores de bases de datos convencionales, puesto que hay coincidencia exacta de la consulta con algún registro de una base de datos.

  El escenario ***(b)*** "que no haya algún *museo de pintura*", es donde surgen los problemas, es a partir de este punto donde colapsan o fallan los sistemas actuales de recuperación, porque en su base de datos no encuentran registros que satisfagan *exactamente* la consulta, y no retornan resultados al usuario.

Este último escenario es el más interesante, debido a que desde el punto de vista de recuperación de información, se aborda un problema abierto para los sistemas tradicionales de procesamiento de consultas, los cuales simplemente basan su filosofía de trabajo en comparar palabra con palabra y/o carácter con carácter. Ahora, dentro de este escenario pueden presentarse varios casos, a saber:

1- Que simplemente no hayan Sitios de Interés CPI *iguales* a lo que busca el usuario, en tal caso, la peor decisión es la que toman los sistemas actuales de recuperación, porque no retornan resultados, es decir, se quedan en "silencio". La situación anterior es común en los buscadores, es el típico "*no match found*" o en el caso de sistemas GIS "*Your search PLACE did not match any places*".

   Es decir, piense en lo siguiente: como dueño de una empresa usted no preferiría que un usuario se vaya insatisfecho porque Usted no tiene exactamente lo que él busca, Usted le recomendaría algo bastante similar a lo que él desea. Por ejemplo, si un turista quiere visitar un *museo de pintura*, el sistema debe devolverle como resultados museos donde se exhiban Murales, acuarelas, Pintura al Oleo y en general museos de pintura a pesar que el nombre del museo no diga textualmente que es un *museo de pintura*.

   Incluso se deben considerar como resultados, *museos de escultura*, *arquitectura* y *arte sacro*, puesto que estos museos hacen parte del mismo tipo de museos, i.e. *museos de las Bellas Artes*, luego el sistema debería retornar museos de la categoría más próxima semánticamente, o sea: *museos de artes decorativas*, *artes*

aplicadas, y finalmente el sistema debería regresar museos menos similares como: *museos de historia, antropología, etnografía* y *folklore*, y así sucesivamente, como ilustración de lo antes dicho se puede observar la Figura 4.1, donde los museos que satisfacen la consulta, son: *Museo del Palacio de Bellas Artes* y *Museo José Luis Cuevas*, ambos tienen igual valor de *confusión* "ε=0", o sea, son los más similares al museo consultado (como se explicó en los apartados 3.6 y 3.7) i.e ambos son *museos de Pintura*.

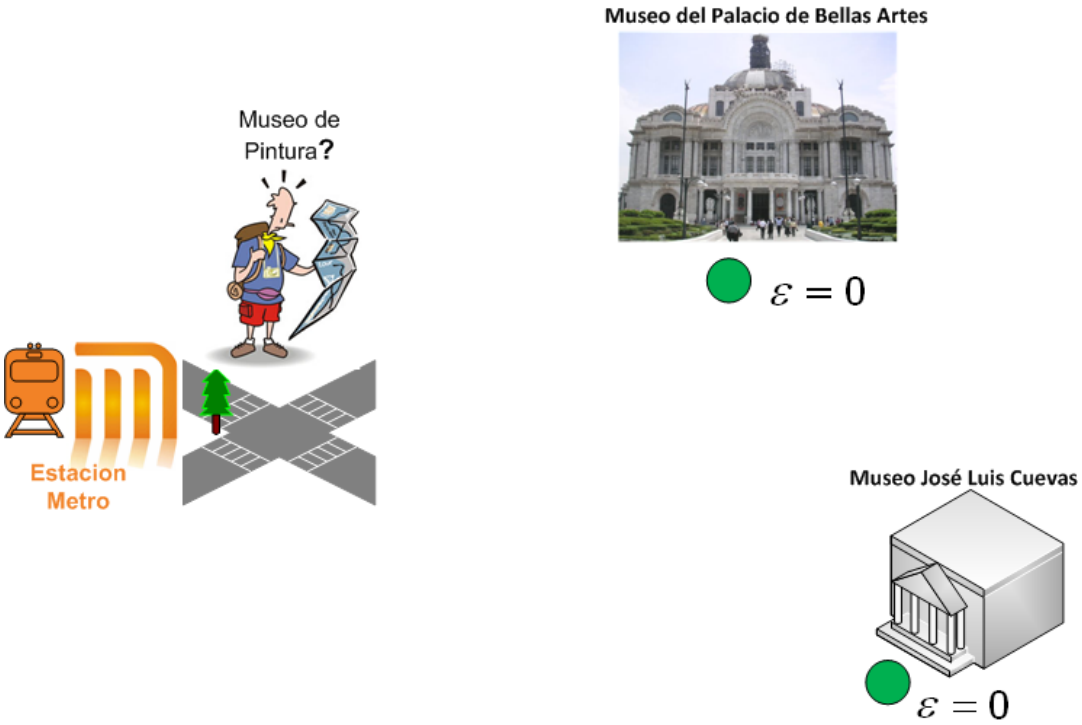

Figura 4.1 Ejemplo de dos Sitios Turísticos Culturales (Museos) similares.

2- Otro escenario posible, resulta cuando hay museos *bastante parecidos* entre sí con lo que busca el usuario, pero con diferentes distancias desde donde se encuentra el usuario, bajo esta situación, el sistema debe estar en capacidad de recomendar entre varios sitios con igual valor de *confusión,* el museo que esté más cercano, de tal manera que aparezca primero aquel museo que está ubicado a unos pasos de donde él se encuentra y luego los más distantes. Con el

objetivo de ilustrar este escenario se presenta la Figura 4.2, donde el CPI "*Museo Palacio de Bellas Artes*" se encuentra a menor distancia que el CPI "*Museo José Luis Cuevas*".

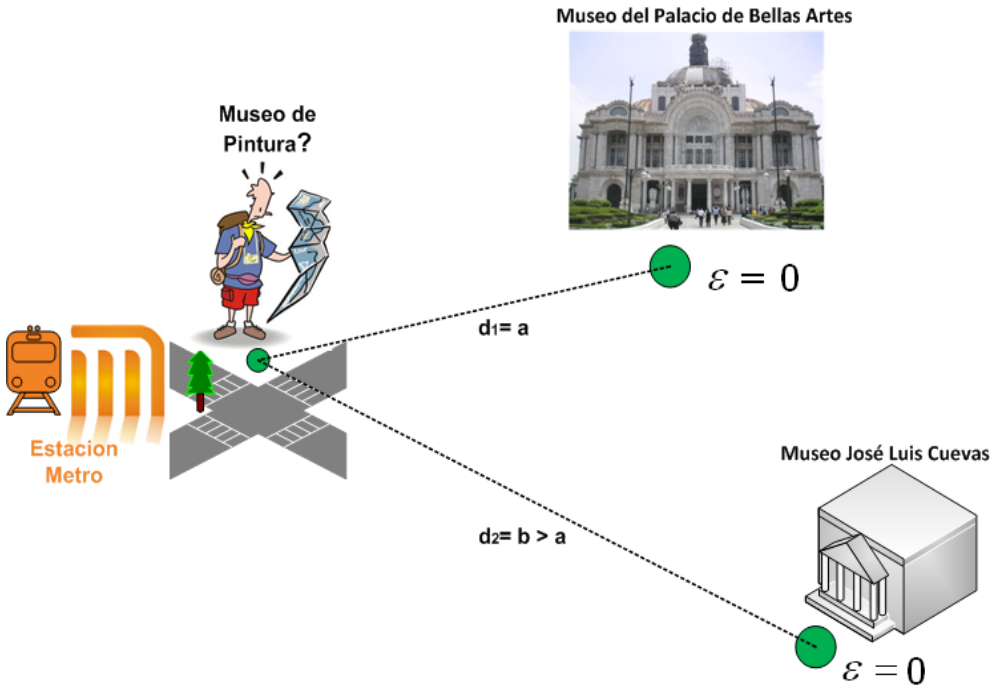

Figura 4.2 Ejemplo de dos Sitios Turísticos Culturales con igual similitud pero con diferentes distancias

3- Finalmente, es importante considerar el siguiente escenario, que hayan Puntos de Interés CPI_1 y CPI_2 similares al buscado, distantes el uno del otro, pero alrededor del CPI_2 hay mayor cantidad de Puntos de Interés que en las proximidades de CPI_1, los cuales podrían interesarle al usuario, puesto que podría visitar estos otros CPI cercanos haciendo más enriquecedor su recorrido, por lo que se le debería dar mayor prioridad en el ordenamiento de los resultados, si así lo señalara o prefiriera el usuario.

Bajo este análisis, podría ser útil considerar en la fase de diseño de la base de datos espacial, un atributo de los Sitios de Interés que contemple la cantidad de sitios de interés en un área próxima o

vecina (por ejemplo, a una distancia no mayor a 5 calles), esta cantidad de Puntos de Interés Cultural se almacenará en una variable llamada *densidad* σ, de tal manera que su valor, refleje la *densidad cultural*, es decir, cuántos sitios turísticos culturales están cercanos del CPI al que se desea visitar, en la Figura 4.3 se representa gráficamente esta situación, donde el CPI "*Museo Palacio de Bellas Artes*" tiene una densidad σ=2 y el CPI "*Museo José Luis Cuevas*" tiene una densidad σ=5.

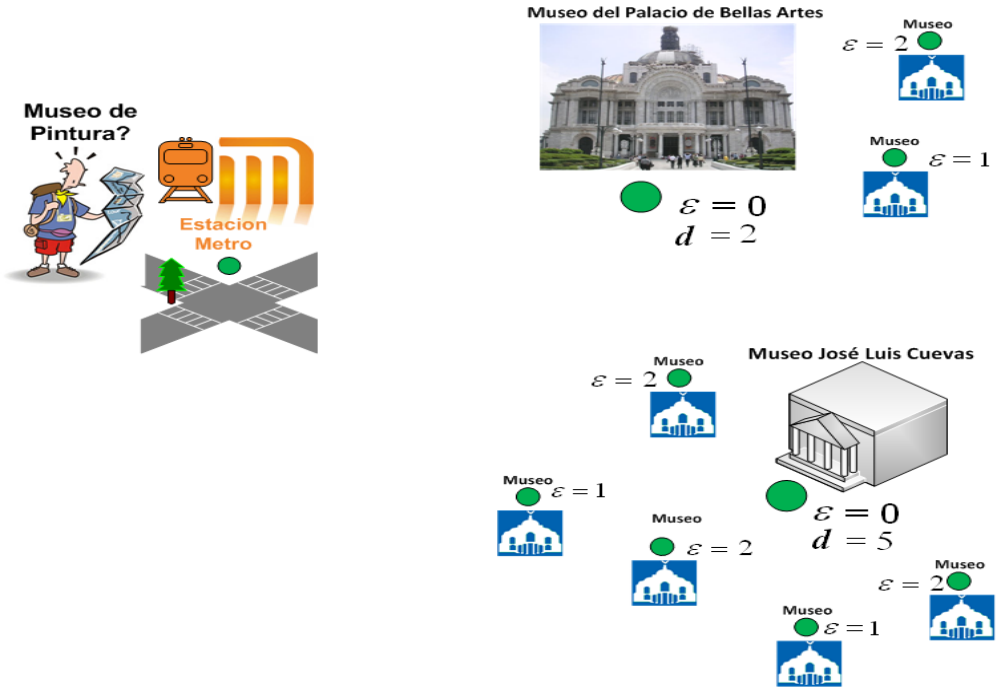

Figura 4.3 Ejemplo del valor de densidad σ, cantidad de museos circundantes

Dentro de estos escenarios hay que hacer hincapié en que, cada apartado descrito implica buscar alguna *forma de ordenar* los resultados.

Ahora bien, teniendo claro el dominio alrededor del cual se mueve esta investigación, plantearemos la metodología a seguir, esta es ilustrada a través del diagrama de la Figura 4.4.

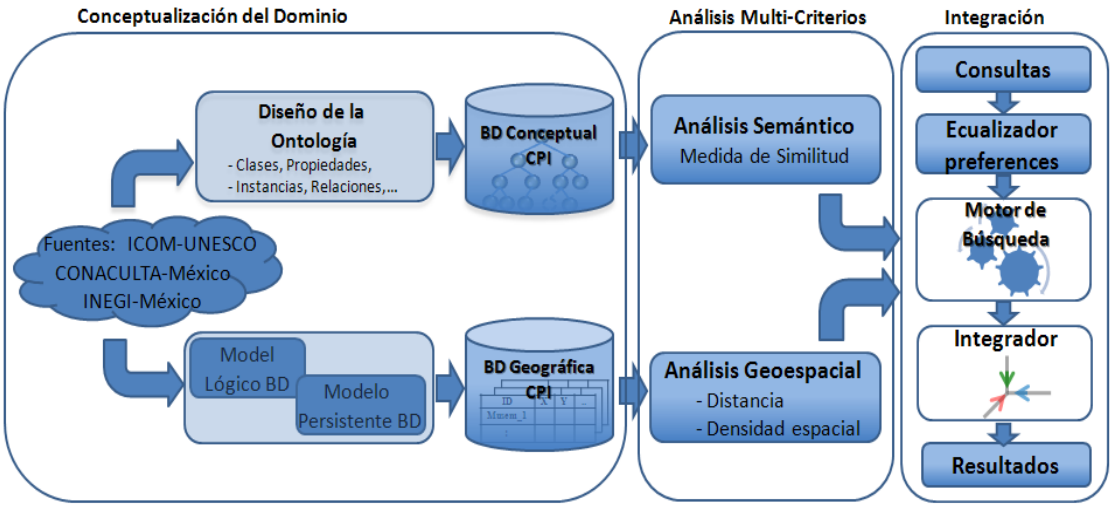

Figura 4.4 Metodología propuesta

En este diagrama se describen los pasos seguidos para aproximarnos al problema de localizar CPI iguales o similares a los que busca un usuario, el objetivo principal es flexibilizar los procesos de búsqueda y satisfacer en diferentes grados una consulta; para lo cual es necesario dotar los datos de un nivel de semántica que permita encontrar otra información semánticamente similar a la buscada y contenida en el repositorio, la metodología está dividida en tres grandes etapas, a saber:

Conceptualización del Dominio, Análisis Multi-criterios e Integración, que a su vez, se componen de módulos menores, como son: recopilación de información de las principales fuentes internacionales y nacionales del dominio del turismo cultural, particularmente para museos, diseño de la estructura conceptual u ontología y diseño de la base de datos geográfica, posteriormente los módulos de análisis semántico y geoespacial y finalmente los módulos de la última etapa: adquisición de la consulta, ecualizador de preferencias, motor de búsqueda, el integrador de criterios y referencias y presentación de resultados.

Desglosemos ahora las fases o etapas de la propuesta metodológica como se describe a continuación.

4.1. Conceptualización del Dominio

Con esta etapa se busca abstraer una problemática o situación del mundo real, esta primera fase del proceso permite modelar y captar las características más importantes de un problema y su entorno, para lograrlo se desglosa en varios procedimientos: *Diseño de la Ontología* y *Diseño de la base de datos geográfica*, que a su vez se subdivide en el diseño de un *modelo lógico* y la implementación del *modelo persistente.*

4.1.1 Diseño de la base de datos conceptual

Como se plantea en [GP04], diseñar una Base de Datos Conceptual (BDC), para este caso una jerarquía, requiere plantear una forma de entender y describir un dominio de manera formal para que las computadoras la puedan procesar, y especificada de manera explícita mediante un lenguaje. Esta Ontología surge del consenso de un grupo y a su vez dirigida a un grupo. En la fase de diseño de esta ontología se definen las clases, subclases, conceptos, relaciones, propiedades e instancias de la estructura que contiene la semántica del dominio. En este trabajo se ha determinado usar jerarquías como caso particular de las ontologías, ya que con éstas se logra captar la semántica del dominio de los CPI que deseamos modelar. No obstante la metodología no se limita a jerarquías, sino que contempla las ontologías y posteriormente el uso de alguna medida de similitud que tenga la capacidad de operar sobre este tipo de estructuras. El resultado del diseño puede verse en la Figura 4.6 y una ampliación de la clase *Arte,* se presenta en la Figura 4.5. Estas jerarquías han sido desarrolladas en el ambiente *protégé V3.4* [Pro08].

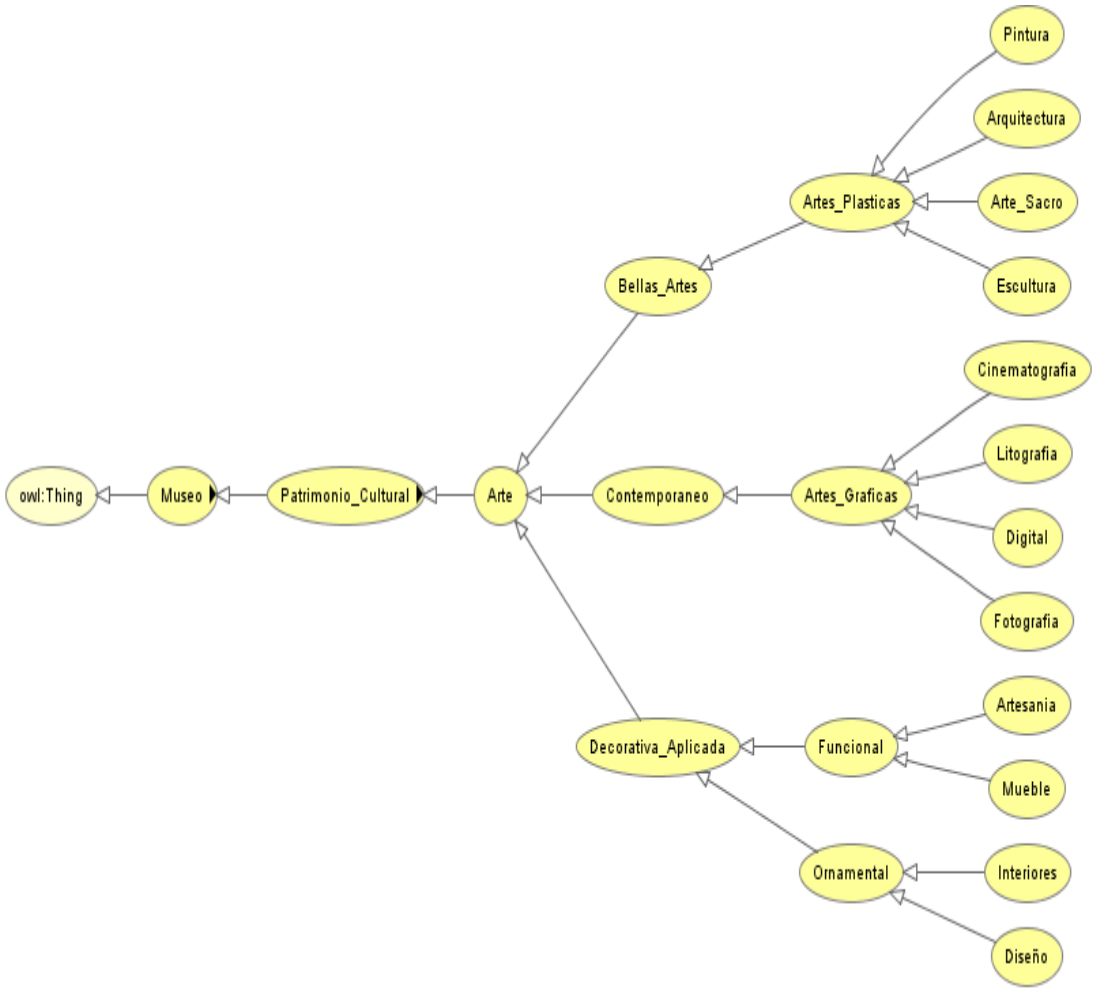

Figura 4.5 Vista detallada de sub-jerarquía de Arte

El modelo o conceptualización semántica debe estar contenido en algún repositorio de datos para que sea reutilizable, en esta metodología se plantea el uso de una Base de Datos Conceptual, de la cual posteriormente se podrá hacer análisis semántico. La información para el diseño de la ontología se basa en los datos provistos por el International Council of Museums (ICOM) [ICO01] y la clasificación de museos que este organismo propone, junto con la información provista por las principales organizaciones y autoridades en esta materia, como son: el Consejo Nacional para la Cultura y las Artes (CONACULTA) y la Secretaría de Turismo, según las particularidades y la riqueza cultural del país.

Figura 4.6 Diseño de Ontología para Sitios Turísticos Culturales (Museos)

4.1.2 Diseño de la base de datos geográfica

El proceso de conceptualización también se extiende a la fase de *Diseño de un Modelo Lógico* y luego uno *Persistente* de la BD, el cual contendrá información de todos los museos que se podrán recuperar de manera semántica; esta información contendrá detalles geográficos y de interés cultural de cada uno de los sitios, para ser consultados por el usuario.

Las fuentes de la base de datos geográfica son obtenidas de los diccionarios de datos geográficos que provee el Instituto Nacional de Estadística, Geografía e Informática (INEGI) [INE05], complementados con la información que suministra la Secretaria de Cultura del Gobierno del Distrito Federal 2009 [SCG09], la Secretaria de Turismo de la Ciudad de México [CPI09] y el Consejo Nacional para la Cultura y las Artes (CONACULTA).

4.1.2.1 Diseño lógico de la base de datos

Dado el análisis conceptual donde se definieron las propiedades de los clases y las relaciones entre estas, es posible analizar los datos disponibles e identificar redundancias en los mismos para conocer qué tipo de información finalmente se almacenará, este análisis permite establecer el modelo formal para la base de datos geoespacial, para lo cual se propone usar el modelo de bases de datos relacionales, a continuación se presenta en la Figura 4.7 un diagrama que ilustra de manera general el diseño lógico para la base de datos Geoespacial.

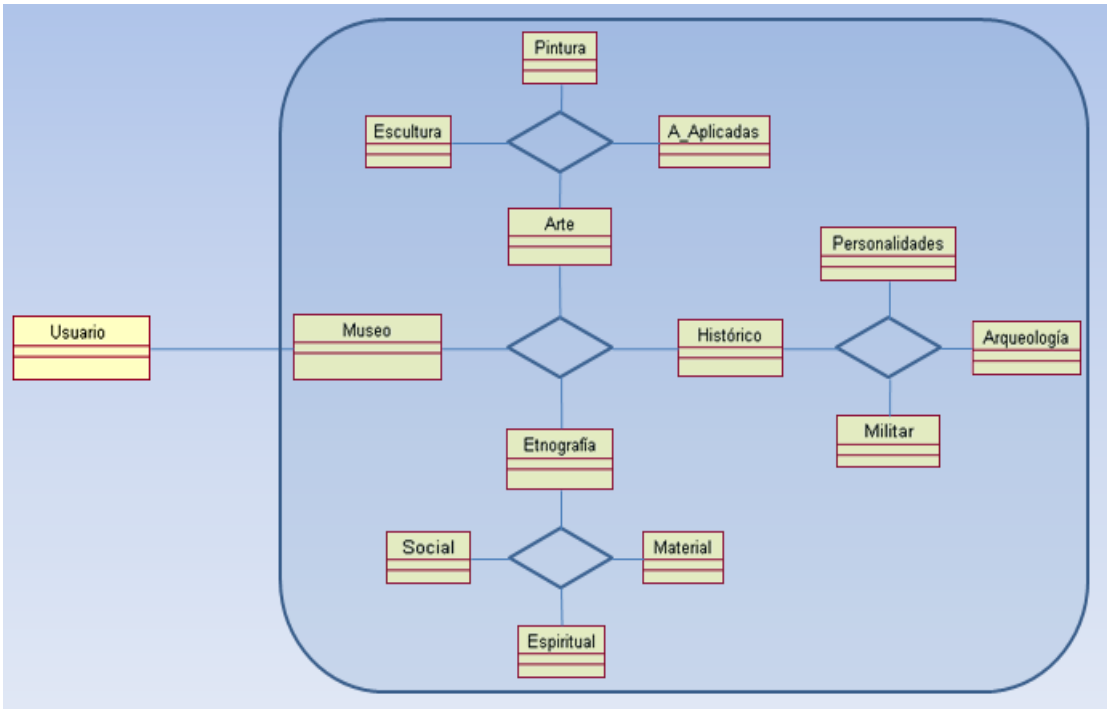

Figura 4.7 Diseño lógico de la base de datos para Museos

4.1.2.2 Diseño persistente de la base de datos

Hay que subrayar que tanto el esquema o diseño anterior como el diseño físico o persistente son descripciones de los mismos datos, pero en el Diseño lógico el nivel de abstracción es más alto; realmente los únicos datos que existen están a nivel físico, por eso se le denomina persistente, por que se almacenan en un dispositivo como sería el caso de archivos en un disco. Para esta metodología, toda la información geoespacial de los sitios turísticos se reflejará como registros en tablas de la base de datos relacional administrada mediante el gestor Postgre/PostGis.

Ahora, con la conceptualización del dominio ya definida, es posible pasar a la fase de procesamiento de los datos.

4.2. Análisis semántico

En esta fase se toman en cuenta los fundamentos abordados en el marco teórico descrito en el capítulo tres, donde se propuso el uso de la *Teoría de confusión* [LG07] para controlar la *precisión* con la que se responde a una consulta del usuario, es decir, con qué grado de similitud semántica se satisface un consulta, si hay varias respuestas que satisfacen una consulta, ¿cuál de ellas produce menos error?, o incluso, si no hay resultados que satisfagan la consulta, ¿será posible encontrar algún resultado que bajo un pequeño margen de error satisfaga la petición?, la respuesta a estas preguntas, es afirmativa.

Es posible regresar una entidad parecida conceptualmente con la entidad buscada a través de *confusión*, i.e mediante el valor de similitud que regresa *confusión; se puede recuperar* una clase de la estructura conceptual parecida a la buscada. Este análisis intenta dotar al modelo de recuperación de cierta flexibilidad, que no poseen los actuales buscadores de información, sobre todo cuando no hay resultados que correspondan exactamente con la consulta.

Esta fase hace uso de la base de datos conceptual para hacer análisis de similitud semántica entre conceptos de la *jerarquía de Museos*, para dar de manera flexible, respuesta a las consultas del usuario.

La perspectiva semántica se alcanza en la medida que los resultados retornados se ordenarán de acuerdo a, qué tanto se parecen conceptualmente y en qué medida satisfacen la consulta; ya que el valor de *confusión* permite medir la distancia conceptual entre conceptos que se encuentran representados en la jerarquía de clases de museos, de manera que los conceptos identifican a las clases, las cuales a su vez representan nodos, y los

nodos más cercanos al nodo buscado, contendrán aquellas instancias que más se parecen a la consulta del usuario y por tanto la satisfacen.

4.3. Análisis Geoespacial

La aproximación al dominio del problema donde se desarrolla este trabajo, posibilita la integración del análisis semántico con el análisis geoespacial. Parte de la propuesta de este trabajo está dirigida a complementar y enriquecer la medida de similitud en variables cualitativas propuesta por Levachkine y Guzmán-Arenas [LG04], ya que al aplicar esta medida en jerarquías, no es posible *diferenciar* y para este caso en particular, *ordenar* conceptos que pertenecen a una misma subclase, siempre y cuando esta subclase no presente un orden natural; la propuesta que se plantea en este trabajo es una aproximación a ese problema; pero centrado especialmente en el *dominio geoespacial*.

Se formula el uso de dos criterios para reordenar o mejor aún, complementar los resultados que presenten el mismo valor *confusión*, el primer criterio consistirá en ordenar los resultados de acuerdo a la *distancia euclidiana,* es decir, qué Puntos de Interés Cultural (CPI) están a menor distancia del usuario. El segundo criterio consiste en ordenar los resultados en función de un valor σ, el cual determina la cantidad de museos que se encuentran en la *vecindad* del museo analizado (i.e. *densidad cultural*); de esta manera el *análisis geoespacial* complementa al *análisis semántico* y son mutuamente enriquecidos.

Con los resultados del análisis semántico es posible encontrar todos aquellos CPI que por su semántica, se parecen más a lo que busca el usuario,

obteniendo resultados como los que se muestran en la Tabla 4.1 cuando se consulta por un *Museo de Pintura*; pero se observa que no hay diferenciación u orden entre resultados con el mismo valor de similitud.

Se requiere entonces alguna manera de reordenar los datos que presentan el mismo valor, es decir, un procedimiento que haga lo que ilustra la Tabla 4.1 en la cuarta columna, para darle prioridad a los resultados según el análisis geoespacial.

Tabla 4.1 Refinar "*confusión*" usando otros criterios

Valor de similitud *conf*(r, Museo de Pintura) = ε, i.e. valor del error	Resultados	Subclase de	Reordenar por otro criterio
0	Museo del Palacio de Bellas Artes	Pintura	
0	Museo Nacional de Arte	Pintura	
0	Museo Mural Diego Rivera	Pintura	
0	Salón de la Plástica Mexicana II	Pintura	
1	Museo Nacional de San Carlos	Escultura	
1	Museo Franz Mayer	Arte decorativo	
1	Laboratorio de Arte Alameda	Arte Digital	
4	Museo Nacional de las Culturas	Etnología	
4	Museo de la Ciudad de México	Etnología	

4.4. Criterios de búsqueda

Dado que se tienen claro el camino de análisis, ahora se especificará cómo se procede con los datos. En primer lugar, la base de datos conceptual se encarga de almacenar toda la información que se ha conceptualizado para dotar de semántica la información del dominio, dicha información está representada en una estructura conceptual, que en este caso es una jerarquía como caso particular de las ontologías, con la descripción de sus clases, sub-clases y las relaciones entre las mismas; en este diseño se ha utilizado la

relación *es-un* "*is-a*", para denotar la asociación entre una clase, sus sub-clases y sus instancias o nodos hoja.

La recuperación de las instancias de manera semántica, permite extraer de la base de datos geoespacial, sus propiedades geográficas y aquellas propiedades que las describen en el ámbito de Sitios Turísticos Culturales. Ahora con las propiedades geográficas es posible enriquecer el análisis semántico; para profundizar en esto, la siguiente sección detalla cómo se lleva a cabo el procesamiento e integración de los datos, en este sentido es conveniente tomar como punto de partida el conjunto de escenarios presentados al inicio de este apartado, los cuales describen y delimitan el alcance de este trabajo.

Cada escenario implica buscar alguna *forma de ordenar* los resultados, pero a la vez el inconveniente principal surge a la hora de integrar cada uno de estos resultados obtenidos a través de diferentes criterios; de allí que una de las principales propuestas de este trabajo es, diseñar un modelo que integre y ordene los resultados de una consulta sobre los diferentes Sitios Turísticos Culturales, de acuerdo a los siguientes criterios que dan sustento a la metodología aquí propuesta:

1) La similitud semántica del CPI consultado con los CPI contenidos en la base de datos espacial.
2) La distancia desde cada CPI hasta la estación del metro donde está el usuario, que luego podría ser modificada a cualquier lugar.
3) La cantidad de Puntos de Interés Cultural en el área próxima del CPI que se desea visitar.

4.4.1 Criterio de similitud

Para abordar el primer criterio de ordenamiento, se ha propuesto en este trabajo el uso de la *teoría de confusión* [LG07], con el fin de determinar los Sitios Turísticos Culturales más similares semánticamente al CPI consultado por el usuario.

Al observar el siguiente ejemplo se puede entender cómo trabaja *confusión* aplicada sobre la jerarquía que se ilustra la Figura 4.8.

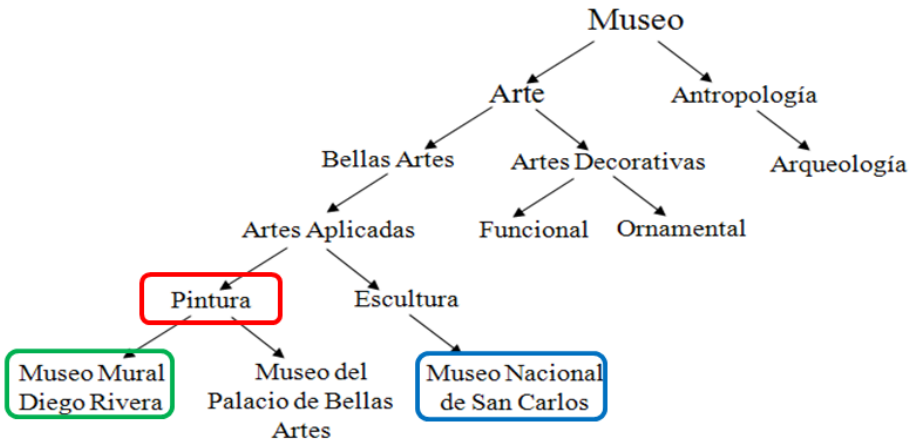

Figura 4.8 Ejemplo general del uso de confusión en la jerarquía de Museos.

Supongamos que un usuario ubicado en la estación del metro Balderas desea visitar un *Museo de Pintura*, el sistema hace la búsqueda en la jerarquía, y calcula la *confusión* de usar cualquier nodo en lugar de usar *Museo de Pintura*, ejemplo:

conf (*Museo Mural Diego Rivera* , *Pintura*) = 0, ¿porqué?

Recordemos las reglas de *confusión* del apartado 3.7.1:

- *conf* (*r*, *r*) = *conf (r, algún_ancestro_de (r))* = 0
- *conf* (*r*, *s*) = 1 + *conf* (*r*, padre_de(*s*))

De la jerarquía se puede observar que:

Pintura = *Algún_ancestro_de*(*Museo Mural Diego Rivera*).
Entonces:

conf (*Museo Mural Diego Rivera* , *Pintura*) =
conf (*Museo Mural Diego Rivera* , *Algún_ancestro_de*(*Museo Mural Diego Rivera*)) = 0

Visto de otra forma, *Museo Mural Diego Rivera* es una instancia especializada de *Museos de Pintura*.

Otro ejemplo se presenta al usar *Museo Nacional de San Carlos* en lugar de *Museo de Pintura*, y se procede como sigue:
conf (*Museo Nacional de San Carlos*, *Pintura*) = 1, ¿porqué?, analicemos:

conf (*Museo Nacional de San Carlos*, *Pintura*) =
conf (*Museo Nacional de San Carlos*, *Pintura*) = 1 + *conf* (*Museo Nacional de San Carlos*, *Padre_de*(*Pintura*))
De la jerarquía se sabe que: *Artes Aplicadas* = *Padre_de*(*Pintura*) entonces al reemplazar se obtiene:

conf (*Museo Nacional de San Carlos*, *Pintura*) = 1 + *conf* (*Museo Nacional de San Carlos*, *Artes Aplicadas*)
De la jerarquía se sabe que: *Artes Aplicadas* = *algún_ancestro_de*(*Museo Nacional de San Carlos*)

conf (*Museo Nacional de San Carlos*, *Pintura*) = 1 + *conf* (*Museo Nacional de San Carlos*,
algún_ancestro_de(*Museo Nacional de San Carlos*))

conf (*Museo Nacional de San Carlos*, *Pintura*) = 1 + 0

conf (*Museo Nacional de San Carlos*, *Pintura*) = 1

El valor de *confusión* es uno, ya que el nodo *Pintura* es hermano del nodo *Escultura*, y este es padre del nodo *Museo Nacional de San Carlos.*

Igual se procede con todos los nodos de la jerarquía, para más detalles ver la Tabla 4.2. que presenta los resultados generales, donde el valor en cada una

de las celdas es el valor al calcular *conf*(***r***,***s***) sin normalizar aún, valor que representa el error de usar un concepto en lugar otro, es decir, si se consulta por ***s*** pero se tiene ***r***, qué tan similares son semánticamente; de tal manera que el valor cero indica que el concepto encontrado es exactamente igual al buscado o un caso especializado del mismo, y por otro lado, a medida que este valor crece indica que hay mayor error al usar el concepto ***r*** en lugar del buscado ***s***.

Desde un punto de vista conceptual, un valor mayor implica que los conceptos son menos similares semánticamente, esto quiere decir que, en una búsqueda los resultados que *mejor satisfacen* una consulta son los *más cercanos a cero*.

Tabla 4.2 Ampliación de los ejemplos al calcular *confusión*.

s →

r ↓

	Punto de Interés Cultural	Museo	Patrimonio Cultural	Arte	Bellas Artes	Artes Gráficas	Decorativas_Aplicadas	Pintura	Escultura	Arte Sacro	Arte Digital
Punto de Interés Cultural	0	1	2	3	4	4	4	5	5	5	5
Museo	0	0	1	2	3	3	3	4	4	4	4
Patrimonio Cultural	0	0	0	1	2	2	2	3	3	3	3
Arte	0	0	0	0	1	1	1	2	2	2	2
Bellas Artes	0	0	0	0	0	1	1	1	1	1	2
Artes Gráficas	0	0	0	0	1	0	1	2	2	2	1
Decorativas_Aplicadas	0	0	0	0	1	1	0	2	2	2	2
Pintura	0	0	0	0	0	1	1	0	1	1	2
Escultura	0	0	0	0	0	1	1	1	0	1	2
Arte Sacro	0	0	0	0	1	1	0	2	2	0	2
Arte Digital	0	0	0	0	1	0	1	2	2	2	0
Antropología	0	0	0	1	2	2	2	3	3	3	3
Arqueología	0	0	0	1	2	2	2	3	3	3	3
Etnología	0	0	0	1	2	2	2	3	3	3	3
Museo Nacional de Arte	0	0	0	0	0	1	1	0	1	2	2
Museo del Palacio de Bellas Artes	0	0	0	0	0	1	1	0	1	2	2
Museo Mural Diego Rivera	0	0	0	0	0	1	1	0	1	2	2
Museo de la SHCP	0	0	0	0	0	1	1	1	0	2	2
Museo Nacional de San Carlos	0	0	0	0	0	1	1	1	0	2	2

Los resultados al aplicar *confusión conf*(**r,s**), permite ordenar los resultados como se muestra en la Tabla 4.3.

Tabla 4.3 Aplicando *confusión* para buscar *Museos de Pintura* o similares

Valor de *confusión* ε *conf*(r, Museo de Pintura)	Instancias resultado de la consulta	Subclase de
0	Museo del Palacio Bellas Artes	Pintura
0	Museo José Luís Cuevas	Pintura
0	Museo Nacional de Arte	Pintura
0	Museo Mural Diego Rivera	Pintura
1	Museos de la SHCP	Escultura
1	Salón de la Plástica Mexicana II	Escultura
1	Museo Nacional de las Culturas	Escultura
1	Museo Nacional de Arquitectura	Arquitectura

4.4.2 Criterio de distancia.

Dentro del marco del análisis semántico, el anterior criterio contempla la situación donde hay coincidencia exacta de la consulta del usuario con la información contenida en la base de datos conceptual, pero qué sucede si solo hay resultados similares, esto es, por ejemplo si hay varios resultados con $\varepsilon = 1$, allí se deben reordenar los resultados de otra manera, por ejemplo, bajo un segundo criterio, para esto se utiliza el análisis geoespacial, donde se propone refinar según la distancia que hay que recorrer para desplazarse desde la estación del metro donde se encuentra el turista hasta el Sitio Turístico Cultural. Ahora aclaremos esto, continuando con el ejemplo de la sección anterior, observe que el CPI llamado *Museo Mural Diego Rivera* tiene el mismo valor de *confusión* que el *Museo José Luis Cuevas*, pero ¿cuál debe aparecer primero en los resultados?, es claro que debe aparecer primero aquel CPI para el cual hay que recorrer menos distancia desde la estación del Metro Balderas (lugar donde se encuentra ubicado el usuario), obsérvese la Figura 4.9 que ilustra la situación descrita.

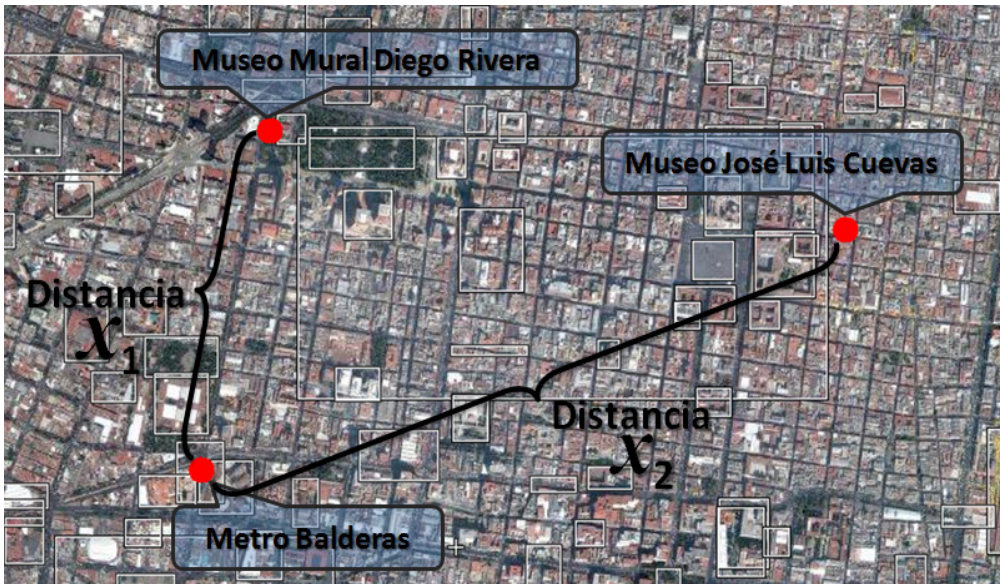

Figura 4.9 Ejemplo de uso del criterio de ordenamiento por distancia

Como se observa el *Museo José Luis Cuevas* se encuentra más lejos que el *Museo Mural Diego Rivera*, por tanto debe aparecer primero y así para todos los casos donde varios museos tengan el mismo valor de *confusión*, debe listarse primero el que se encuentra a menor distancia si se usa este criterio, y así lo prefiera el usuario. Bajo este análisis los resultados menos distantes quedarían reordenada como muestra la Tabla 4.4.

Tabla 4.4 Ejemplo usando el criterio de distancia para ordenar los resultados de sub-jerarquía de Museos de Artes

Valor de *confusión* ε	Instancias resultado de la consulta	Latitud	Longitud	Distancia (m)
2	Museo de Arte Popular	19°26'2"N	99°8'46"W	**850.706848**
0	Museo Mural Diego Rivera	19°26'10"N	99°8'48"W	**1070.78347**
0	Museo del Palacio Bellas Artes	19°26'7"N	99°8'28"W	**1254.81337**
2	Museo de Franz Mayer	19°26'14"N	99°8'35"W	**1318.80051**
0	Museo Nacional de Arte	19°26'10"N	99°8'21"W	**1460.75133**
2	Museo de la Caricatura	19°26'11"N	99°8'9"W	**1742.32082**
2	Museo Ex Teresa Arte Actual	19°26'2"N	99°7'50"W	**2083.90548**
3	Museo Templo Mayor	19°26'5"N	99°7'49"W	**2147.46933**
0	Museo José Luis cuevas	19°26'0"N	99°7'44"W	**2225.36977**

4.4.3 Criterio de Densidad.

Otra manera de abordar la situación de resultados con el mismo valor de similitud arrojados por el análisis semántico, es a través del criterio de *densidad*, el cual permite ordenar los resultados en función de la cantidad de museos que haya en las proximidades del lugar al que se desea visitar, lo que se pretende es que aparezca primero el museo con mayor cantidad de Puntos de Interés Cultural dentro del área más próxima circundante, por ejemplo, un área con radio menor a seis calles; este proceso es conocido en sistemas GIS con el nombre de análisis de proximidad, y en especial la operación de *buffering*, que permite determinar la relación de proximidad entre características de objetos geográficos.

Con el fin de ampliar los conceptos, obsérvese el siguiente ejemplo: cuando el usuario consulte por un *Museo de Pintura* el sistema regresará museos de pintura ordenados de acuerdo a su valor de densidad $\boldsymbol{d}$, que por ejemplo, para el caso del *Museo Palacio de Bellas artes* su valor de densidad $\boldsymbol{d}$ es 9, porque los museos en su vecindad o dentro de un área con radio menor a cinco calles son:

{*Museo Nacional de Arte, Museo Franz Mayer, Museo Nacional de la Estampa, Palacio Postal, Palacio de Iturbide, Palacio de Minería, Centro Cultural de la SHCP, Museo del Ejército y Fuerza Aérea, Museo Mural Diego Rivera*}

Igualmente se calcula para cada uno de los Sitios Turísticos Culturales que son similares al buscado por el usuario, según lo indica el análisis semántico mediante el cálculo de *confusión*, para obtener entonces la Tabla 4.5.

Con el fin de ilustrar la manera de calcular del valor de densidad σ para el "*Museo Palacio de Bellas Artes*" se muestra la Figura 4.10.

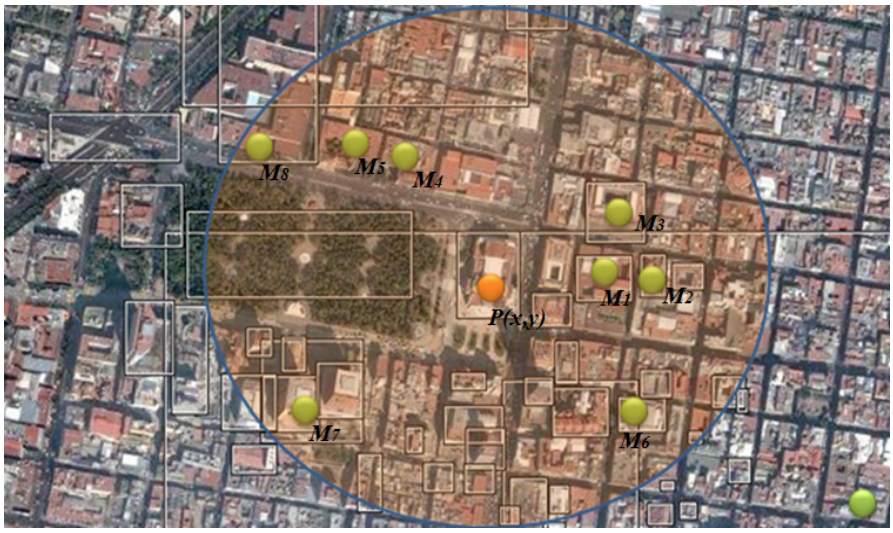

Figura. 4.10 Ejemplo de uso del criterio de densidad.

Tabla 4.5 Ejemplo de uso del criterio densidad para ordenar los resultados.

Valor de *confusión*	Instancias resultado de la consulta	Densidad σ	Museos Vecinos
3	**Museo Templo Mayor**	**22**	***Museo de la Caricatura, Museo José Luis cuevas, museo mexicano del diseño, Museo Nacional de las Culturas, Museo Ex-Teresa Arte Actual, Antiguo Palacio del Arzobispado, Museo de la Autonomía Universitaria, Museo de Comida Mexicana, Centro Cultural de España en México, Antiguo Colegio de san Ildefonso, Museo de la luz, Museo Puro Corazón, Museo Salón de la Plástica Mexicana II, Museo de la SEP, Museo SHCP, Museos Casa de la Imprenta, Museo Pinacoteca Virreinal de San Diego, Museo Cocina Mexicana, Museo Galería Cocina Duque de Herdez, Museo Recuerdo de la Academia Mexicana de la Lengua, Museo Serfín de Indumentaria Indígena, Museo del Calzado***
0	**Museo José Luis Cuevas**	**20**	***Museo Nacional de las Culturas, Museo Ex-Teresa Arte Actual, Antiguo Palacio del Arzobispado, Museo Templo Mayor, Museo de la Autonomía Universitaria, Museo de Comida Mexicana, Centro Cultural de España en México, Antiguo Colegio de san Ildefonso, Museo de la luz, Museo Puro Corazón, Museo Salón de la Plástica Mexicana II, Museo de la SEP, Museo SHCP, Museos Casa de la Imprenta, Museo Pinacoteca Virreinal de San Diego, Museo Cocina Mexicana, Museo Galería Cocina Duque de Herdez, Museo Recuerdo de la Academia Mexicana de la Lengua, Museo Serfín de Indumentaria Indígena, Museo del Calzado***
0	**Museo Nacional de Arte**	**12**	***Museo del Palacio Bellas Artes, Museo Interactivo de Economía, Museo Franz Mayer, Museo Nacional de la Estampa, Museo Palacio Postal, Laboratorio Arte Alameda, Palacio de Iturbide, Palacio de Minería, Museo del Ejército y Fuerza Aérea, Museo Mural Diego Rivera, Museo Manuel Tolsá, Museo Palacio Cultural Banamex***
0	**Museo del Palacio Bellas Artes**	**9**	***Museo Nacional de Arte, Museo Franz Mayer, Museo Nacional de la Estampa, Palacio Postal, Palacio de Iturbide, Palacio de Minería, Centro Cultural de la SHCP, Museo del Ejército y Fuerza Aérea, Museo Mural Diego Rivera***
0	**Museo Mural Diego Rivera**	**9**	***MUNAL, Museo Franz Mayer, Museo Nacional de la Estampa, Museo Palacio Postal, Laboratorio Arte Alameda, Palacio de Iturbide, Palacio de Minería, Museo del Ejército y Fuerza Aérea, M. P. Bellas Artes.***

4.5 Espacio para la integración de criterios

Dado que se tienen definidos los criterios que harán parte del sistema de recuperación de información, es necesario, establecer la forma como se integrarán estos criterios, para tal fin se propone un espacio en analogía con los espacios métricos multidimensionales, este espacio lo llamaremos: *espacio para la integración de criterios*.

El principal objetivo de este espacio es servir como mecanismo para combinar el análisis semántico con el análisis geoespacial, ya que si estos se complementan mutuamente, se obtienen resultados más satisfactorios, es decir, resultados con la *precisión* y *flexibilidad* de los procedimientos de análisis de *similitud semántica* y con la *refinación* de las *operaciones geoespaciales*. Nosotros creemos que esta es una aproximación a la forma como los humanos hacen recomendaciones en función de varios criterios simultáneamente.

El espacio para la integración de criterios que estamos planteando debe cumplir las siguientes propiedades y definiciones:

(i) Cada dimensión es usada para mapear un criterio de recuperación.

(ii) Cada criterio o dimensión es independiente de los otros criterios.

(iii) Cada dimensión es configurada de tal forma que los valores más cercanos a cero "al origen", son los más apropiados para satisfacer una consulta, como se ilustra en la Figura 4.11.

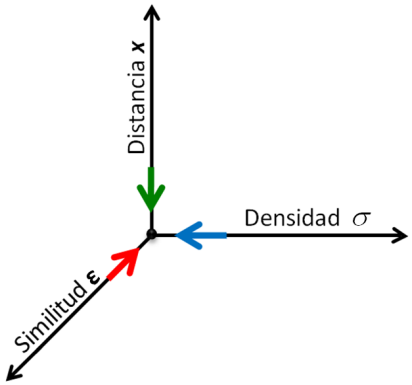

Figura. 4.11 Valores más cercanos a cero satisfacen mejor la consulta del usuario

De esta forma, un museo M_i que satisface una consulta Q_a será representado como una tripleta de valores $M_i(\varepsilon,\sigma,x)$ en el espacio. Con las siguientes definiciones:

El valor de similitud ε es definido así: si $\varepsilon \to 0$ entonces el museo M_i es muy similar al buscado, si $\varepsilon = 0$ entonces el museo satisface exactamente a la consulta.

El valor de distancia x es definido como la *distancia euclidiana* entre las coordenadas geográficas del usuario (x_u, y_u) y las coordenadas del museo M_i que es objeto de análisis (x_m, y_m). si $x \to 0$ entonces el museo M_i está muy próximo al usuario.

El valor d' es definido como el número de museos localizados en las vecindades del museo M_i, es decir, la cantidad de museos dentro de un área *A* con radio *r*, radio con el que a su vez se normaliza cada distancia *x*. Para satisfacer la propiedad (***iii***) se requiere la siguiente transformación $\sigma = 1 - \dfrac{\sigma'}{\max\left(Density_m\right)}$, si $\sigma \to 0$ entonces el museo M_i tiene alta *densidad cultural*, o sea, cuenta con un elevado número de museos en su vecindad. La *densidad* es un parámetro a definir, mediante el análisis de proximidad geoespacial, para determinar el buffer a usar; este incluso podría estar en función de la distancia que el usuario esté dispuesto a caminar, 5, 8, 10 calles. La función *max( )* regresa el valor de densidad máxima entre el conjunto de valores de densidades de los museos denominado $Density_m$.

(iv) Cada dimensión tienen un grado de relevancia, determinado por un peso w_i, estos pesos son provistos por el usuario cuando configura el nivel de importancia que él asigna a cada criterio en su consulta.

El conjunto de pesos w_i será llamado perfil del usuario, este perfil es formado mediante la configuración que el usuario proporciona al indicar sus preferencias. Hay que notar que, el usuario no tiene que ser un experto para configurar los pesos, solo tiene que "ecualizar sus preferencias", i.e. atenuar o dar mayor prioridad a los criterios de acurdo a sus gustos o necesidades.

De esta manera, el ecualizador de preferencias es el componente del modelo que captura las necesidades o deseos del usuario, este proceso es menos tedioso y complejo que en los sistemas actuales, ya que el usuario solo debe responder a la pregunta:

¿Qué tan importante es para Usted cada criterio?

Es decir, el usuario debe indicar el nivel de importancia que él otorga a cada criterio según sus preferencias, conformando así un perfil como se muestra en la Figura 4.12.

Figura. 4.12 El nivel de importancia de cada criterio constituye el peso para cada dimensión.

(v) Dada una consulta Q_a sobre un museo M_i, su tripleta de valores $M_i(\varepsilon,\sigma,x)$, será representado por un punto en el espacio de integración.

(*vi*) El valor de relevancia R, de cada resultado (i.e. cada punto en el *espacio para la integración de criterios*) es calculado como señala la expresión propuesta en (13).

$$R_w(c_i, w_i) = 1 - \left(\frac{\sum_{i=1}^{n} (c_i)^{w_i}}{n} \right) \qquad (13)$$

Donde c_i son los criterios de recuperación, w_i los pesos asignados por el usuario a cada criterio mediante el ecualizador de preferencias y n es el número de criterios a utilizar, en este caso n=3.

Está expresión permite determinar los resultados que más se aproximan o satisfacen la consulta.

De tal manera que si $R_w \to 1$ entonces el resultado satisface mejor la consulta, un ejemplo del espacio para la integración se muestra en la Figura 4.13.

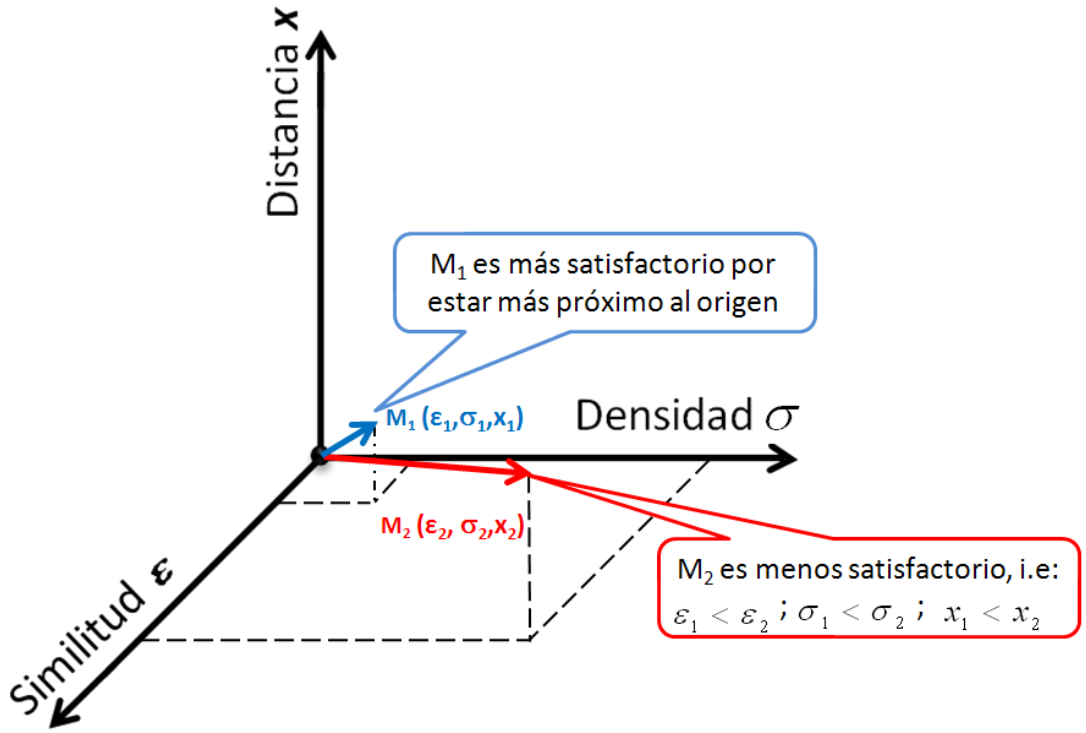

Figura. 4.13. Un museo con valores más próximos al origen satisface mejor la consulta, en este caso M_1 es mejor que M_2.

4.6. Presentación de resultados

Esta fase final contempla la visualización de los resultados y otras características adicionales de los Sitios Turísticos Culturales recuperados según la consulta del usuario; ésta es solo la interfaz para interacción con el usuario, por lo tanto es de notar que no se pretende desarrollar una nueva herramienta de desplegado de mapas, de las cuales ya hay muchas enmarcadas en el desarrollo de Sistemas de Información Geográfica (GIS por sus siglas en Ingles). El enfoque de este trabajo se centra en hacer más flexibles las búsquedas y controlar la precisión con que se recuperan los resultados de una consulta, ya sea en sistemas de búsquedas tradicionales o en GIS. Esto se logra empleando estructuras de información que dotan de semántica a los sistemas de recuperación de información. La presentación de los resultados contempla mostrarlos ordenados de manera semántica y geoespacialmente (según los criterios ya definidos en el apartado 4.5), junto con la ubicación de los museos sobre un mapa digital y otro detalles de índole informativa.

4.6.1 Visualización de resultados

La interfaz que permite ingresar las consultas también se encarga de presentar los resultados ya ordenados según el análisis semántico, pero tomando como principio la tendencia de los sistemas orientados a la Web 2.0. Las cuales presentan más intercambio de información con el usuario, la propuesta para integrar los resultados del análisis semántico con el geoespacial, es permitir que sea el usuario quien decida cual enfoque del análisis semántico geoespacial utilizar y en qué medida, esto en función de lo que más le interesa.

4.7 Arquitectura de la aplicación

Con el desarrollo de las secciones anteriores es posible plantear una arquitectura que materialice cada una de las fases de la metodología, permitiendo así, descender en el nivel de abstracción de desarrollo del sistema hacia el prototipo o modelo funcional. Ahora se describen los componentes de la arquitectura que representa nuestra propuesta de solución, como se ilustra en la Figura 4.14.

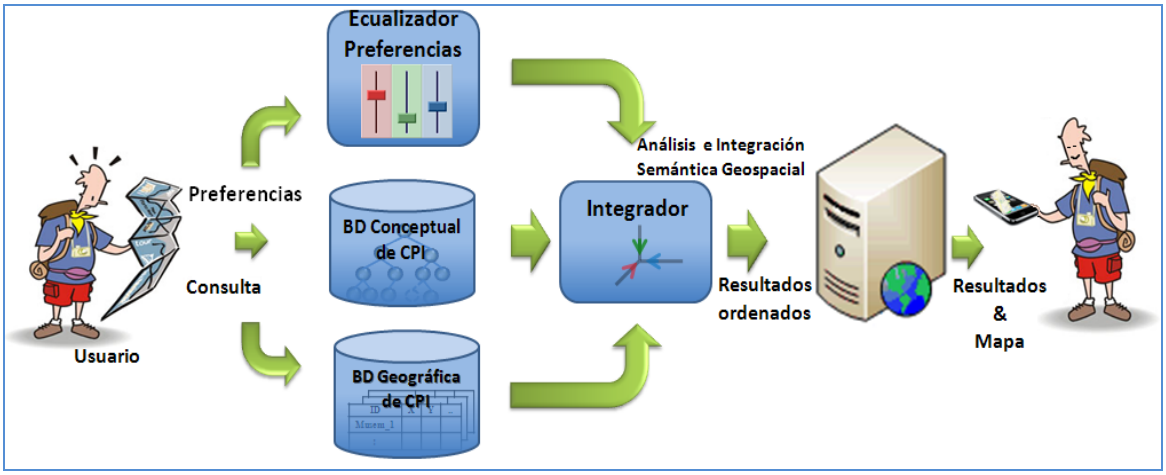

Figura. 4.14 Arquitectura de la Aplicación

Las consultas de usuario, son capturadas desde una interfaz Web, desarrollada en PHP 5.0, la consulta junto con las preferencias del usuario pasan a conformar la entrada del sistema. Una vez capturada la consulta, se realiza un proceso de búsqueda en los repositorios de datos. El repositorio conceptual o archivo OWL, permite encontrar un museo y los museos semánticamente más similares, mientras que la descripción y cada una de las propiedades de los museos, se encuentran almacenada en el repositorio de datos geográfico. Ahora bien, ¿por qué dos repositorios?, las razones de tener los dos repositorios son:

- La separación aporta flexibilidad al sistema, es decir, cuando se desee

agregar o eliminar un museo, solo bastará con modificar un registro de la base de datos geográfica, sabiendo: a qué clase pertenece este museo, características como su ubicación geográfica, etc.

- La separación de la semántica del problema de los datos crudos, esto permite abstraer la información del tipo conceptual, para realizar procesos ulteriores de razonamiento e inferencia sobre información que, muchas veces es imperceptible en bases de datos convencionales.

Cada museo en la base de datos geográfica, se refleja en la base de datos conceptual a través de un cambo denominado "subclase", dicho campo permite saber a qué subclase de la ontología pertenece la instancia en cuestión, es decir, el registro en la base de datos geográfica es un nodo que apunta a algún nodo hoja de la ontología (repositorio de datos conceptual).

Posteriormente se realiza el análisis semántico geoespacial de los museos, según los criterios definidos y explicados al principio, este análisis es procesado junto con las preferencias de usuario en el módulo de integración; esta integración producirá para cada resultado un valor de relevancia que permitirá conformar una tabla ordenada.

La tabla de resultados ordenados tendrá un campo adjunto denominado geometría de los datos, esta geometría es adjuntada mediante el gestor PostgreSQL/PostGis. Finalmente, el servidor de mapas GeoServer toma la tabla de resultados y con su geometría adjunta produce una nueva capa, la cual será sobrepuesta a una capa del Distrito Federal, en particular la zona del centro histórico.

4.8. Comentarios Finales

El desarrollo de esta metodología ofrece una propuesta orientada a flexibilizar la búsqueda de información y controlar la precisión con que se recuperan resultados, englobando dos grandes análisis: el *Semántico* y el *Geoespacial*, mostrando cómo se enriquecen mutuamente i.e se combinan para complementarse y ofrecer mejores respuestas, ya que la rigidez del enfoque clásico-numérico es flexibilizado a través del componente semántico, pero al mismo tiempo, las debilidades que presenta el análisis semántico, son reforzadas o refinadas, mediante el uso de criterios del dominio geoespacial, los, que actúan como mecanismos de filtrado o mejor aún, mecanismos para refinar entre entidades conceptualmente iguales, pero que por su ubicación en el espacio (corta distancia u otro criterio espacial), pueden representar alternativas más interesantes.

Esta propuesta metodológica se materializa en una arquitectura que ayuda a evitar resultados vacíos, dado que si no hay correspondencia exacta de la consulta con los registros de la base de datos geográfica (i.e los nombres de los museos), entonces se vale de la información de la base de datos conceptual, para recuperar museos que por sus temáticas son los más semejantes; ofreciendo así, resultados semánticamente similares (museos más parecidos al consultado), y geoespacialmente más próximos (análisis espacial de distancia) y culturalmente más interesantes (mayor probabilidad de encontrar lo que busca un turista en un área con mayor densidad de museos).

Adicionalmente, toda esta riqueza y flexibilidad obtenida al integrar el componente semántico con el geoespacial, es aprovechado por el usuario de manera personalizada, ya que él tiene la oportunidad de decidir qué es lo que más le interesa, además de la satisfacción de nunca ver resultados vacíos, se encuentra

con un servicio que toma en cuenta sus preferencias, permitiéndole explorar todo el universo de posibilidades a través del ecualizador de preferencias, lo cual muestra un modelo de recuperación de información en el dominio geográfico, que se acerca a la personalización de servicios, vinculando al mismo tiempo la semántica de los dominios históricamente más importantes para la economía México, como es el turismo cultural.

Finalmente, debe quedar claro que, esta propuesta metodológica, es fácilmente aplicable a otros dominios o subdominios, principalmente aquellos donde es necesario valerse de la semántica presente en el dominio para complementar el enfoque de recuperación geoespacial y viceversa, y lograr que los sistemas adquieran ese nivel de inteligencia que les permita ofrecer alternativas o incluso recomendar opciones adicionalmente a las que busca un usuario.

CAPÍTULO 5.

PRUEBAS Y RESULTADOS

Hemos aprendido a volar como los pájaros, a nadar como los peces; pero no hemos aprendido el sencillo arte de vivir como hermanos.
Martin Luther King

Lo verdadero es siempre sencillo, pero solemos llegar a ello por el camino más complicado.
George Sand

En este capítulo, se presentan los resultados obtenidos con el modelo de recuperación e integración propuesto. Además la implementación de un sistema Web que retorna resultados a consultas sobre museos, basándose en la propuesta metodológica de este trabajo. El sistema toma en cuenta los criterios de recuperación descritos en la sección anterior, junto con las preferencias del usuario, a través del denominado ecualizador de preferencias.

Se destaca que, para cada consulta, se recupera un conjunto de resultados ordenados en función del grado de relevancia. Los resultados son Puntos de Interés Cultural (CPI), en particular Museos representados por una tripleta de atributos que, según sus valores, se ajustan en mayor o menor medida al museo que busca el usuario. Además, se describen algunas características de cada uno de los resultados.

5.1 Discusiones sobre los principales componentes del modelo

Antes de proceder, se presenta una amplia discusión y evaluación de cada uno de los criterios empleados, con el fin de establecer un marco común de interpretación de cada una de las pruebas y resultados.

5.1.1 Evaluación de la similitud semántica del modelo

Para abordar esta sección, se han planteado unas preguntas de investigación que servirán para guiar al lector a manera de discusión. Para entender el modo de operación de esta parte del modelo, las preguntas son:

¿Es realmente semántico el enfoque usado en este modelo?
¿Verdaderamente evita resultados vacíos?
¿Qué tan cierto es que se logra controlar la precisión del conjunto de resultados basado en su similitud semántica?

Para abordar estas preguntas se inicia destacando que, como parte de los productos de esta tesis, se desarrolló un API para el lenguaje de programación JAVA, con la cual se implementa la medida de similitud propuesta en la *teoría de confusión* [LG04], esta API está abierta a la comunidad científica del Centro de Investigación en Computación y al Instituto Politécnico Nacional, previa autorización de sus autores.

Con esta API, denominada API de *confusión V.1.0* se pueden obtener resultados como: dada una consulta, en la cual se pregunta por un museo de pintura, y no habiendo *Museos de Pintura*, entonces los resultados serán *Museos de Escultura*, de

Arquitectura y *Arte Sacro* que, según la ICOM [ICO01] son museos de *las Bellas Artes*.

Se ilustran ahora algunos ejemplos que corroboran su funcionamiento. En la Tabla 5.1 se muestran los resultados que regresa la API de *confusión V.1.0* al consultar por un *Museo de Pintura*.

Tabla 5.1. Resultados retornados por la API de *confusión V.1.0* al consultar *Museo de Pintura*

Clase	ε	Clase	ε	Clase	ε	Clase	ε
Pintura	0	Maya	4	Tejido	4	Moneda	4
Escultura	1	Madera	4	Manufactura	4	Bovicultura	4
Arquitectura	1	Policía	4	Lengua	4	Zapoteca	4
Arte Sacro	1	Música	4	Danza	4	Óptica	4
Fotografía	3	Mecánica	4	Colombina	4	Biográfico	4
Diseño	3	Agricultura	4	Indumentaria	4	Ejército	4
Litografía	3	Electricidad	4	Mercado	4	Vida Extinta	5
Cinematografía	3	Marina	4	Comida	4	Alimenticia	5
Interiores	3	Silvicultura	4	Tradicional	4	Suelo	5
Mueble	3	Azteca	4	Fiesta	4	Ambiente	5
Artesanía	3	Cunicultura	4	Astronomía	4	Entomología	5
Digital	3	Tolteca	4	Cerámica	4	Silvestre	5
Meteorología	4	Porcicultura	4	Hidrografía	4	Mineralogía	5
Comunidad	4	Medios	4	Literatura	4	Humana	5
Acuicultura	4	Educación	4	Oceanografía	4	Reserva	5
Acontecimiento	4	Avicultura	4	Fuerza Aérea	4	Medicinal	5
Bibliográfico	4	Capricultura	4	Ciencia Jurídica	4	Fauna Nacional	5
Chichimeca	4	Papel	4	Totoneca	4	Biodiversidad	5
Alternativa	4	Naturista	4	Transporte	4	Ecoparque	5
Cestería	4	Olmeca	4	Mundial	4	Flora Nacional	5

Ahora, si alguien pregunta por un *Museo de Etnología* (i.e. un museo cuyo principal interés es el estudio y preservación de la historia de las costumbres y tradiciones de un pueblo a través de sus muestras culturales) y no hay museos denominados como tal, entonces el sistema regresa *Museos de Danza, Música, Indumentaria, Comida, entre otros* (i.e. museos que poseen muestras culturales especializadas de

una sociedad).

Los resultados que regresa la API de *confusión V.1.0* al consultar por un *Museo de Etnología* se presentan en la Tabla 5.2.

Tabla 5.2. Resultados retornados por la API de *confusión V.1.0* al consultar *Museo de Etnología*

Clase	ε	Clase	ε	Clase	ε	Clase	ε
Danza	0	Tolteca	1	Marina	2	Hidrografía	2
Lengua	0	Totonaca	1	Cunicultura	2	Litografía	2
Indumentaria	0	Colombina	1	Arquitectura	2	Digital	2
Música	0	Olmeca	1	Capricultura	2	Acuicultura	2
Literatura	0	Medios	2	Avicultura	2	Escultura	2
Comida	0	Policía	2	Electricidad	2	Fuerza Aérea	2
Fiesta	0	Acontecimiento	2	Arte Sacro	2	Silvestre	3
Educación	0	Naturista	2	Agricultura	2	Humana	3
Mercado	0	Porcicultura	2	Bovicultura	2	Mineralogía	3
Cestería	0	Tradicional	2	Fotografía	2	Reserva	3
Tejido	0	Bibliográfico	2	Manufactura	2	Flora Nacional	3
Cerámica	0	Mueble	2	Artesanía	2	Ambiente	3
Papel	0	Alternativa	2	Diseño	2	Vida Extinta	3
Madera	0	Oceanografía	2	Moneda	2	Fauna Nacional	3
Ciencia Jurídica	0	Meteorología	2	Ejercito	2	Suelo	3
Zapoteca	1	Comunidad	2	Interiores	2	Entomología	3
Azteca	1	Astronomía	2	Cinematografía	2	Ecoparque	3
Chichimeca	1	Pintura	2	Transporte	2	Biodiversidad	3
Maya	1	Mecánica	2	Óptica	2	Medicinal	3
Mundial	1	Silvicultura	2	Biográfico	2	Alimenticia	3

Como se ha visto en los ejemplos anteriores, a partir de la similitud semántica, los resultados que se obtienen son los más parecidos a lo que se busca, es decir, la distancia en el significado de un resultado con respecto al buscado es la mínima en la estructura conceptual donde están representados los conceptos del dominio del turismo cultural, en particular museos; es así como se evitan los resultados vacíos.
Al analizar la lista de resultados se observa que, a medida que se desciende en la lista, los resultados son menos parecidos a lo que inicialmente se buscó. De esta

manera es posible tener los resultados ordenados de acuerdo con los resultados que presentan mayor similitud, y entonces, controlar la precisión del conjunto de resultados basado en su similitud semántica de los mismos.

5.1.2. Evaluación de la medida de densidad cultural

Es importante abordar este apartado analizando los datos que suministra el Instituto Nacional de Estadística, Geografía e Informática, INEGI y el Consejo Nacional para la Cultura y las Artes (CONACULTA) de México; en sus datos se puede observar la distribución de museos por entidad federativa de todo el país, como se ilustra en la Figura 5.1.

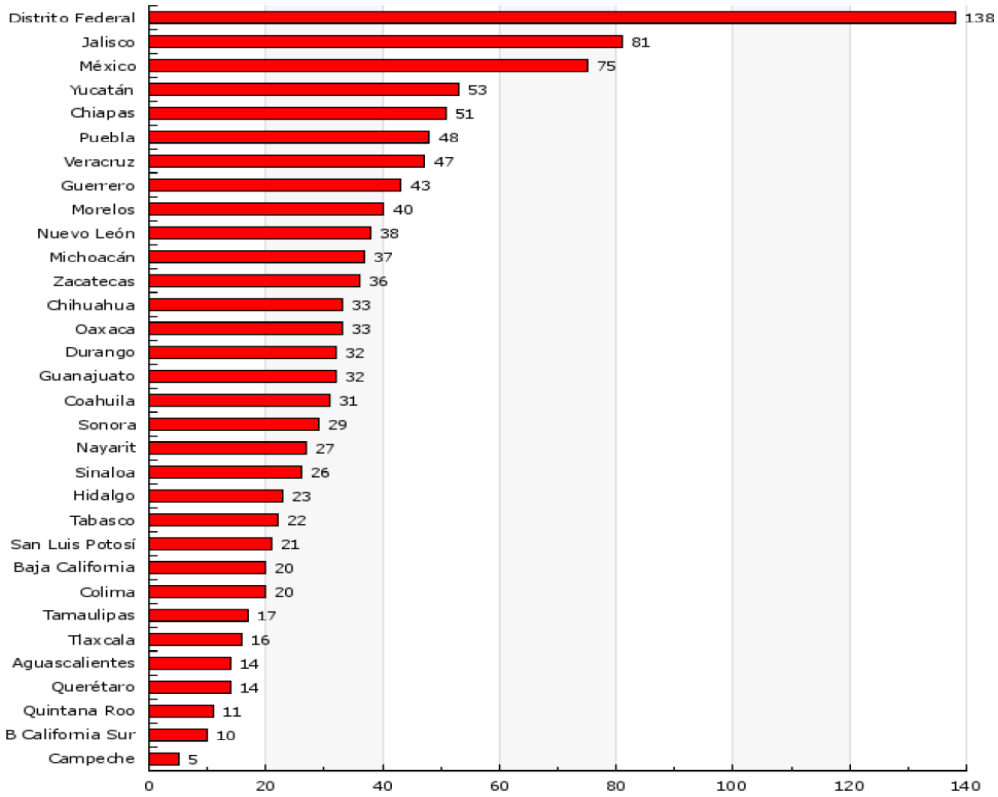

Figura 5.1 Distribución de museos por entidad federativa, 2007[1].

En la figura se puede apreciar que la concentración de museos más alta está en el Distrito Federal. Después de indagar se encontró que, dentro del Distrito Federal la

[1] Fuente: CNCA, Sistema de Información Cultural,

zona con mayor concentración de museos es el centro histórico de la ciudad, con aproximadamente 50 museos.

Si esto se mira desde un punto de vista estadístico, se puede afirmar que gran parte de la información que alguien quisiera obtener acerca de la cultura, se encuentra muestreada o bien representada en los museos de esta zona.

Ahora, esto da cabida para plantear la siguiente hipótesis:

Un turista desearía ir donde hay más de aquello que él está buscando, es decir, hay mayor probabilidad de satisfacer su necesidad en un área con mayor concentración cultural, que en un área desprovista de Puntos de Interés Cultural (CPI).

De acuerdo a lo anterior, se plantea la siguiente pregunta con fines de ilustrar la situación aquí descrita.

Suponga que un estudiante de doctorado desea conocer una jovencita, para entablar con ella una relación de amistad y posiblemente de matrimonio, Usted, ¿dónde le recomendaría que hiciera su búsqueda?, o ¿qué oportunidad tendría de hallar una novia aquí en el CIC?, donde la concentración de población femenina estudiantil es poca, o por el contrario, si se dirige a una escuela de ciencias sociales donde la densidad de población femenina es elevada; en resumen, se le recomendaría al estudiante de doctorado que visitara una escuela de ciencias sociales, pues allí tendría mayor probabilidad de encontrar la mujer de sus sueños.

Bajo este planteamiento, ¿dónde hay mayor probabilidad que un turista satisfaga sus necesidades culturales? ¿En un sector donde hay mayor concentración de museos o en uno donde hay poca concentración de museos? i.e. densidad cultural. En este trabajo se asume que hay mayor probabilidad que el turista encuentre lo

http://sic.conaculta.gob.mx/consulta2/ (agosto de 2007).

que busca donde hay mayor densidad cultural, particularmente museos.

Lo anterior lleva a pensar que, si alguien desea conocer mucho sobre la cultura de un país o una región a través de las visitas a museos, uno de los lugares más apropiados es la Ciudad de México.

5.1.3 Evaluación de la medida de distancia

Realmente no es simple medir la distancia entre dos lugares de una ciudad, debido a que existen muchos factores a considerar a la hora de establecer todos los aspectos que intervienen para ir de un punto A hacia un punto B, porque en las grandes ciudades como la Ciudad de México, se requiere más tiempo y recorrido en ir de A hacia B y no así en la dirección contraria; por ejemplo, los puentes peatonales permite acortar camino a los peatones pero no a los que se desplazan en autos. Por otro lado, las distancias se convierten en no triviales porque, si usted va en automóvil debe considerar además del recorrido en auto, la existencia y disponibilidad de estacionamiento, en adición al trayecto que debe caminar después de estacionar su automóvil, y no sería muy grato que el estacionamiento esté muy retirado de su lugar de destino, principalmente si se poseen limitaciones físicas. Estos y otros factores hacen que la distancia desde un punto a otro deje de ser simétrica y algo trivial. Solamente para mostrar el modelo propuesto en este trabajo, la distancia se calcula a través de la distancia Euclidiana, no obstante el modelo admite utilizar una métrica más ajustada a la realidad como se acaba de describir, esta distancia es apropiada en el sentido que, el área del centro histórico del D.F. con respecto al globo terráqueo. Representa distancias cortas, que se pueden aproximar de un semi-arco terrestre a una línea.

Una medida un poco más ajustada a la realidad, es la distancia denominada city block, también llamada distancia de Manhattan, cuyas propiedades son:

Para $x, y, z \in \mathbb{R}^2$ donde x, y, z representan coordenadas geográficas,

se verifica que:

1. $d(x,y)=0$ si, y silo si $x=y$ "axioma de separación"
2. $d(x,y)\geq 0$ "resultado en los reales"
3. $d(x,y)=d(y,x)$ "**axioma de simetría**"
4. $d(x,y)\leq d(x,z)+d(z,y)$ "desigualdad triangular"

No obstante, como se observa en la tercera propiedad, esta distancia aún asumen simetría, es decir, no importa en qué sentido se desplace una persona dentro de una ciudad, la distancia será siempre igual.

Pero un aspecto que hay que resaltar es que, cualquiera que sea la medida de distancia usada, habrá usuarios para los cuales no importará mucho el recorrido que tengan que hacer para llegar al lugar turístico que buscan; pero sin duda alguna, habrá usuarios para los cuales la distancia será un factor trascendental, entonces más que desear, necesitarán al realizar una consulta, que los resultados sean, además de ser similares conceptualmente, sean aquellos que estén menos distantes de su ubicación.

5.1.4 Evaluación del ecualizador de preferencias

Esta sección se aborda mediante la dirección de las siguientes preguntas:
¿Realmente es importante incluir un ecualizador en el modelo?
Esta pregunta se aborda solamente mencionando dos frases:

- *Los usuarios desean servicios personalizados.*
- *Las empresas no desean perder usuarios, desean satisfacer sus preferencias.*

Otra pregunta que sirve para analizar la necesidad del ecualizador, es:

¿Cómo hacen la ponderación los sistemas actuales?

- Muchos de los sistemas hasta hoy, han basado la manera como el usuario interactúa con el sistema a través de formularios o en su versión más simple, listas de chequeo (*check List box*), ejemplo *iGoogle, Yahoo,* entre otros.

- En otros sistemas el diseñador o el "experto" define el nivel de importancia de cada criterio, pero de esta manera se asume que el experto conoce, incluso, más que el propio usuario, lo que éste busca, y entonces en lugar de servicios personalizados, se ofrecen "servicios masificados", en otras palabras, el experto es una especie de oráculo, o simplemente promedia los gustos o preferencias de la población, asumiendo que el usuario X también deseará lo que el resto de la población quiere.

- En otros sistemas el usuario se ve forzado a ingresar valores numéricos, pero realmente no sabe qué significa cada valor, y en el caso que algún usuario esté más o menos familiarizado con los valores, para otros usuarios no será muy intuitivo ingresar valores numéricos. Como resultados de esta investigación se ha encontrado que se requieren otras formas menos complejas de captar sus preferencias, por eso se sugiere el uso del ecualizador de preferencias, en analogía a la manera como una persona elige y configura sus preferencias musicales mediante las frecuencias y niveles de sonido que quiere escuchar, pues quien más sabe lo que desea escuchar es el propio usuario.

¿Qué cosas no se logran con el ecualizador?

Es importante aclarar que el ecualizador solo se encarga de captar el nivel de importancia que cada usuario le asigna a cada criterio, de tal manera que bajo la misma consulta, los resultados serán diferentes puesto que, por ejemplo, si un usuario asigna mayor importancia a la distancia al momento de "ecualizar", entonces los resultados que aparecerán, estarán concentrados en las proximidades de la estación que haya especificado, es decir, su ubicación inicial. Pero si otro

usuario le otorga preferencia a la similitud, y le resta importancia a la distancia, quiere decir que el usuario prefiere sitios exactamente iguales o muy parecidos a lo que él está buscando, sin importar que tenga que desplazarse más.

5.1.5 Evaluación y discusión de la expresión de relevancia

Inicialmente se propuso la expresión (14) en el apartado para evaluar la relevancia.

$$R(c_i, w_i) = 1 - \left(\sum_{i=1}^{n} \left(c_i . \left(1 - w_i\right) \right)^P \right)^{\frac{1}{p}} \quad (14)$$

Donde c_i representa los criterios, w_i los pesos para cada criterio y p un valor que determina el tipo de relación entre criterios.

Esta expresión arrojó buenos resultados, pero no era suficientemente discriminatoria, es decir, dado un museo $M_i(\varepsilon, \sigma, x)$ con su tripleta de valores: ***ε***, ***σ*** y ***x***, en ocasiones se tenían resultados aparentemente similares, i.e. un museo $M_1(0.7, 0.2, 0.8)$ y otro museo $M_2(0.4, 0.2, 0.7)$, e incluso un museo $M_3(0.14, 0.2, 0.9)$, con ese conjunto de valores, se requiere una expresión que dada la ponderación del usuario, aleje del origen, los museos menos relevantes, en otras palabras, si uno de los valores de un museo, lo sitúan cerca al origen, y este valor corresponde al criterio que el usuario estableció mayor importancia, entonces este museo debe ser el más importante y los otros menos relevantes, por lo tanto la expresión debe hacer más relevante un museo $M_i(\varepsilon, \sigma, x)$, según el nivel de importancia que cada usuario establece para cada criterio, sin tomar en cuenta los otros criterios aunque tengan valores parecidos; la propuesta es la mencionada en la expresión (13) del capítulo anterior 4.6, para efectos de recordarla, se enuncia nuevamente:

$$R_w(c_i, w_i) = 1 - \left(\frac{\sum_{i=1}^{n} (c_i)^{w_i}}{n} \right)$$

Ahora, esta última expresión, solucionó este inconveniente, ya que se logró obtener resultados que, dadas las preferencias de usuario, se alejan del origen del espacio para la integración de criterios aquellos museos que, de acuerdo con su tripleta de valores (i.e $\boldsymbol{\varepsilon}$, σ y $\boldsymbol{x}$,) no se ajustan a lo que busca el usuario. Esto tiene mayor explicación en las propiedades que se definieron en el espacio para integrar los criterios (ver sección 4.6), en otras palabras esta expresión tiene la ventaja que, aquellos resultados que no son relevantes, son alejados de manera exponencial del origen, dado que con pequeñas variaciones en c_i, la razón de crecimiento es alta, según los valores w_i capturados en el ecualizador, así los resultados relevantes, son aquellos que su tripleta de valores simultáneamente están más cerca del origen.

Para comparar las dos expresiones se realizaron varios experimentos, observando el mismo comportamiento presentado en los resultados de la consulta *Qa* en las Tablas 5.3 y 5.4.

Qa = Museo de Arqueología cerca al metro Zócalo.

En los resultados de las tablas solo se observan coincidencias del primer resultado: *Museo Templo Mayor*, pero al analizar la Tabla 5.3 se observa que, museos que deberían aparecer en los primeros lugares de la lista, i.e *museos de Arqueología*, ni siquiera aparecen; e.g. *Museo de la SEP*, que es un *museo de Arqueología* poco distante del metro Zócalo (a solo 445 metros), no aparece en la lista de resultados obtenidos mediante la expresión (14), mientras que en los resultados de la Tabla 5.4, obtenidos mediante la expresión (13), aparece en primer lugar un museo de *Arqueología Mexicana*: *Museo Templo Mayor*, en segundo lugar, un museo de *Arqueología Mundial*: *Museo Nacional de las Culturas* y en tercer lugar aparece un museo de *arqueología Local*: *Museo de la SEP* y posteriormente aparecen *museos de Etnología* que, junto con los *museos de Arqueología* hacen parte de la clase: *museos de Antropología*.

Tabla 5.3 Resultados de la consulta Qa con la expresión de integración (13) **Wε=1,Wσ=0,Wx=1**

Museo	R(ε,Wε, σ,Wσ, x,Wx)
Museo del Templo Mayor	0.93
Museo de la Ciudad de México	0.82
Museo de la Autonomía Universitaria	0.80
Museo Nacional de las Culturas	0.78
Museo Ex Teresa Arte Actual	0.77
Museo José Luis Cuevas	0.67
Museo de la Luz	0.63
Museo Nacional de Arte	0.60
Museo Franz Mayer	0.58

Tabla 5.4 Resultados de la consulta *Qa* con la expresión de integración (14) **Wε=1,Wσ=0,Wx=1**

Museo	R_W(ε,Wε, σ,Wσ, x,Wx)
Museo del Templo Mayor	0.57
Museo Nacional de las Culturas	0.52
Museo SEP	0.50
Museo Puro Corazón	0.50
Museo Galería Cocina Duque de Herdez	0.49
Museo Cocina Mexicana	0.47
Museo Recuerdo de la Academia Mexicana de la Lengua	0.46
Museo Introductor a la Catedral Metropolitana	0.45
Museo Mexicano de Diseño	0.45
Museo Numismático	0.44
Museo Serfín de Indumentaria Indígena	0.43
Museo SHCP	0.43

Al observar la Figura 5.2 es posible apreciar la distribución de los datos según el tipo de cultura que preservan, coleccionan o difunden en cada museo. Esta gráfica permite ver cómo es la concentración de museos en la zona del centro histórico.

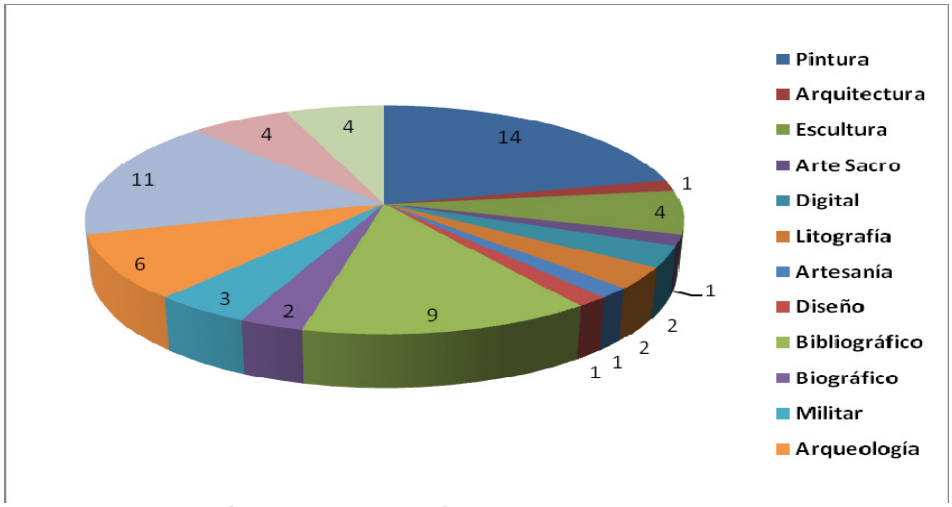

Figura. 5.2 Distribución de museos según el tipo de bienes culturales que preservan.

Las bondades de la expresión (13) "$\mathbf{R_w}$" son resaltadas al compararla con la expresión tradicional (14) "**R**", a través de la Figura 5.3, donde se ilustran los resultados obtenidos al realizar la consulta de *Museo de Arqueología* tomando como punto de partida la estación del metro Zócalo. Es posible ver que dada la consulta *Qa*, de los 6 museos de arqueología se recuperaron 5, es decir 83.3%, mientras que a través de la expresión tradicional solo se recuperaron 2, es decir 33.3%, de igual manera sucede con los museos semánticamente más similares, es decir, con los *museos de Etnología*, que junto a los *museos de Arqueología* hacen parte de la clase de *museos de Antropología*; con la expresión tradicional R solo se recuperó 1 de los 11 museos de etnología en el área, pero con expresión (13) "$\mathbf{R_w}$", se recuperaron 6, es decir un 54.5%.

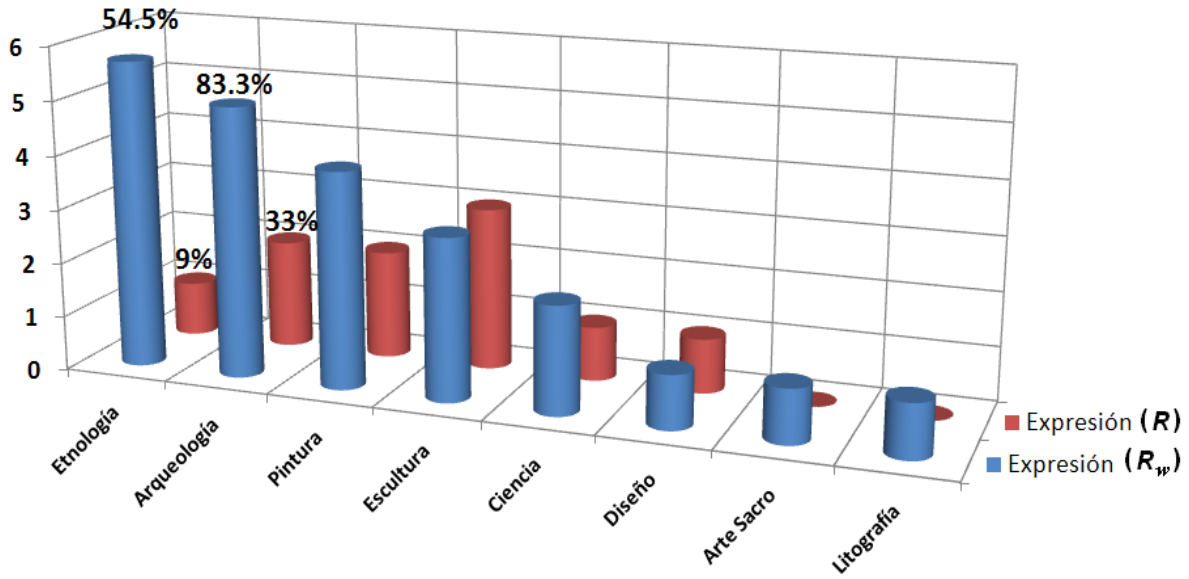

Figura. 5.3 Comparación de museos recuperados por la expresión **R** y $\mathbf{R_w}$.

5.2 Evaluación del sistema

Esta sección se dividirá en cuatro partes, las tres primeras dedicadas a analizar la forma como se procesan las consultas, y la última para evaluar resultados que aparentan ser errores, pero bajo la dirección de esta investigación no lo son, por el contrario, reflejan flexibilidad frente a la manera como actualmente se procesa información en los sistemas de recuperación.

5.2.1 Detalles de la interfaz

Se presenta el resultado final de la implementación de un sistema bajo los conceptos desarrollados con el modelo de recuperación e integración propuesto. En la Figura 5.4 se ilustra la interfaz.

Figura. 5.4 Interfaz del sistema implementado en esta tesis.

- La consulta es capturada a través de dos campos, en el primero se indica el museo o tipo de museo buscado, en el segundo la ubicación del usuario, esto se ilustra en la Figura 5.5

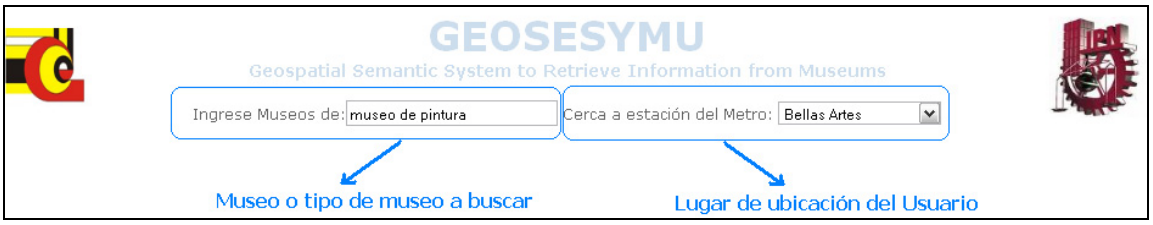

Figura. 5.5 Captura de la consulta

- El ecualizador de preferencias permite especificar el rol que juega cada criterio, es decir, el nivel de importancia para cada usuario, como se muestra en la Figura 5.6.

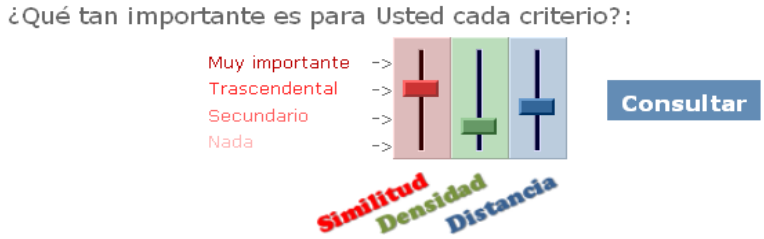

Figura. 5.6 Uso del ecualizador de preferencias.

- En el lado izquierdo aparece el listado de resultados ordenados de acuerdo con el grado de relevancia descrito en el modelo (capítulo 4), al hacer clic sobre un resultado el mapa es re-centrado y ampliado para mostrar ese museo en particular, y por otro lado, en la parte inferior del mapa aparece toda la descripción del museo seleccionado, estos aspectos se pueden observar en las Figuras 5.7, 5.8. y 5.9.

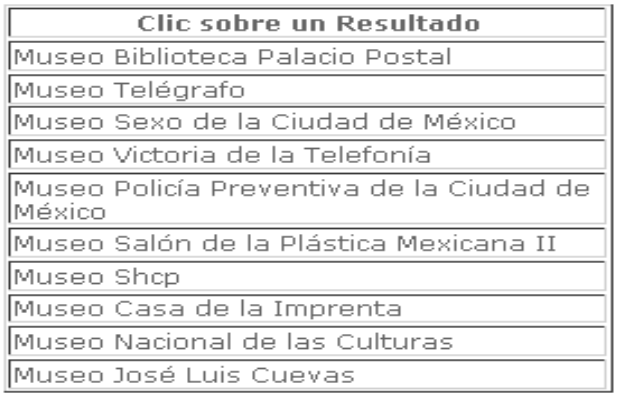

Figura. 5.7 Lista de resultados ordenados.

Figura. 5.8 Ampliación y centrado del mapa en el museo seleccionado.

Figura. 5.9 Información adicional del Museo seleccionado.

Bajo el marco de discusión de la sección 5.1, se puede entonces abordar las pruebas al sistema. A continuación se describe en la Tabla 5.5 un conjunto de ejemplos de consultas que se pueden realizar en el sistema implementado.

Tabla 5.5 Ejemplos del tipo de consultas soportadas

Ejemplos de consultas utilizadas en esta tesis
Q1= Museo cerca al Metro Hidalgo
Q2= Museo de Pintura cerca al Metro Bellas Artes
Q3= Museo de Arqueología cerca al Metro Zócalo
Q4= Museo de Historia cerca al Metro Allende
Q5= Museo de Ciencia cerca al Metro Pino Suárez
Q6= Museo de Arte cerca al Metro Salto del Agua
Q7= Museo de Bellas Artes cerca al Metro Garibaldi
Q8= Museo Nacional de Arte al Metro Pino Suárez
Q9= Museo del Templo Mayor cerca al Metro Zócalo
Q10= Museo de Etnología cerca al Metro Balderas
Q11= Museo Indumentaria cerca al Metro Isabel la Católica
Q12= MUNAL cerca al Metro Hidalgo

Se debe aclarar que cada una de las consultas retornará resultados en función de las preferencias que el usuario "ecualice", definidas en el ecualizador de preferencias, como se ilustra en la Figura 5.10

Figura. 5.10 Configuración de preferencias por parte del usuario

Después de contar con la consulta y con la ecualización que el usuario realiza, el sistema procesa la consulta de manera semántica y geoespacial como se describió en los apartados (4.1 y 4.2).

5.2.2 Búsqueda por instancia

Un ejemplo de este tipo de consultas es la consulta Q8 y Q9, en este caso el ecualizador debe estar configurado para que regrese resultados con exactitud, es decir, la máxima similitud, como se muestra en la Figura 5.11, un ejemplo es ilustrado en la Figura 5.12, donde al buscar Museo Palacio de Bellas Artes, con la configuración señalada en el ecualizador, efectivamente aparece en primer lugar del listado el *Museo Palacio de Bellas Artes* y luego museos muy similares (i.e. *Museos de Pintura*), sumado a esto, al hacer clic sobre dichos museos, se puede observar su posición en el mapa y algunos datos del mismo, como su horario, pagina Web, teléfono, costo de entrada, entre otros.

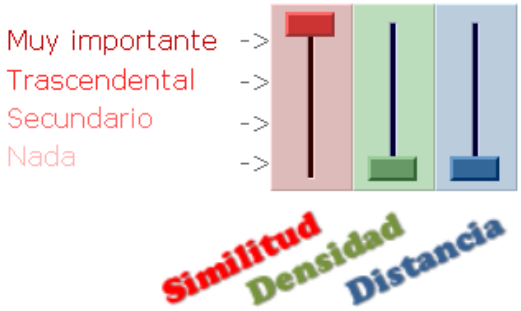

Figura. 5.11 Configuración que regresa resultados con exactitud, es decir, la máxima similitud

Figura. 5.12 Resultados al buscar Museo Palacio de Bellas Artes con la configuración de máxima similitud

5.2.3 Búsqueda por el alias de un museo

La situación es igual que en el caso anterior, solo que aquí se parte del hecho que, algunos museos son más conocidos por un sobrenombre, alias o algunas siglas especiales, ejemplo de esta situación de búsqueda es MUNAL "Museo Nacional de Arte". Si el usuario busca MUNAL, los resultados regresados se pueden observar en la Figura 5.13.

Figura. 5.13 Resultados al buscar el alias de un museo con configuración de máxima similitud

5.2.4 Búsqueda por tipo de museo

En esta forma de búsqueda puede suceder que:

El usuario desee visitar una clase de museos, y no está interesado en otros aspectos; entonces el ecualizador debe estar configurado como se mostró en la Figura 5.11, bajo esas preferencias los resultados obtenidos aparecen en la Figura 5.14, en la cual los museos están ordenados en función de la similitud semántica con el tipo de museo buscado, de tal manera que se observa en primer lugar: Museo Mural Diego Rivera, Museo Palacio Cultural Banamex, Museo Palacio de Bellas Artes y posteriormente otros museos también de pintura. Se puede ver que en el listado no se hace distinción respecto a la distancia entre la ubicación de cada museo y la ubicación del usuario, debido a que este criterio se "anuló", no se tomó en cuenta en el ecualizador.

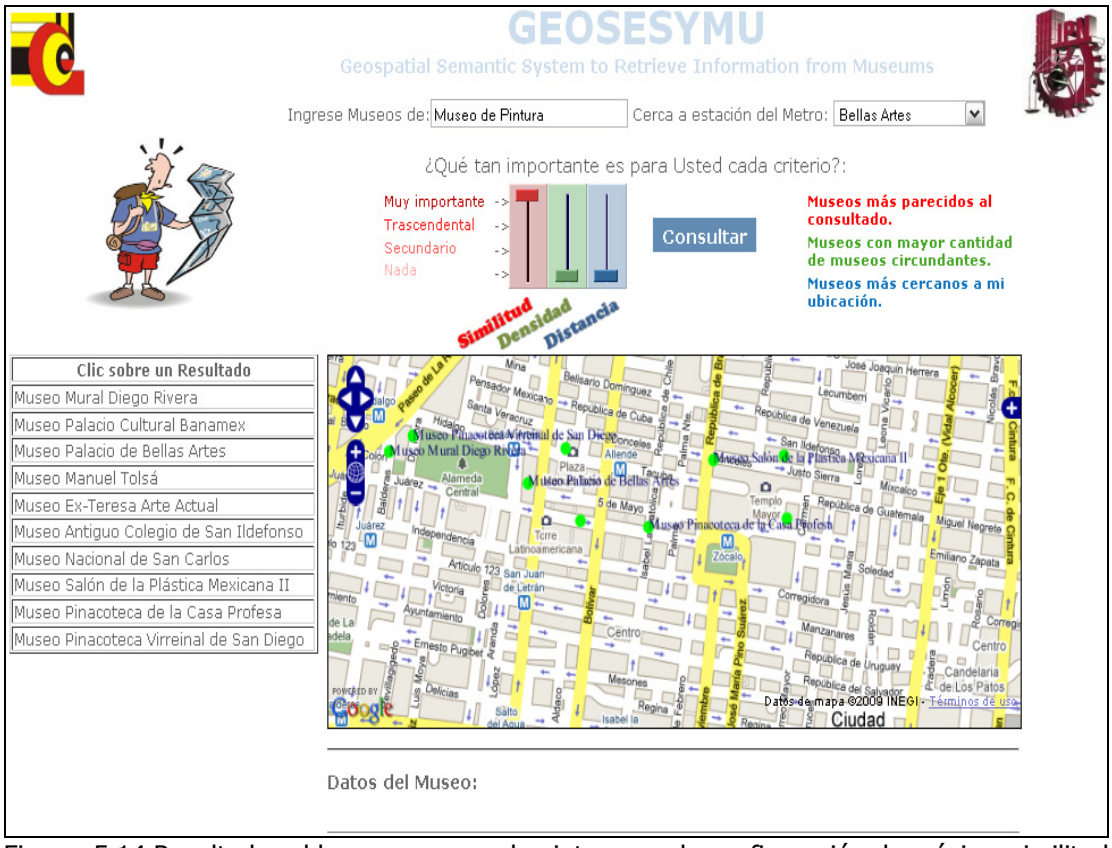

Figura. 5.14 Resultados al buscar museos de pintura con la configuración de máxima similitud

Otro ejemplo similar se presenta al buscar un *Museo de Fotografía*, se observa en la Figura 5.15 que en primer lugar está el *Museo Laboratorio Arte Alameda*, luego el *Museo Centro de la Imagen*, los cuales son museos de *Arte Digital (i.e categoría hermana de la clase Fotografía),* seguidamente el *Museo de la Caricatura* y el *Museo Nacional del la Estampa*, estos son *Museos de Litografía*, los museos de *Arte Digital* y los *Museos de Litografía*, son subclase de las *Artes Gráficas*; posteriormente aparecen museos de otras artes: *Diseño*, *Arquitectura*, *Arte Sacro*, entre otras, que son museos menos similares, pero que junto con los museos *Arte Digital*, hacen parte de una misma clase de museos, la clase de *Arte*.

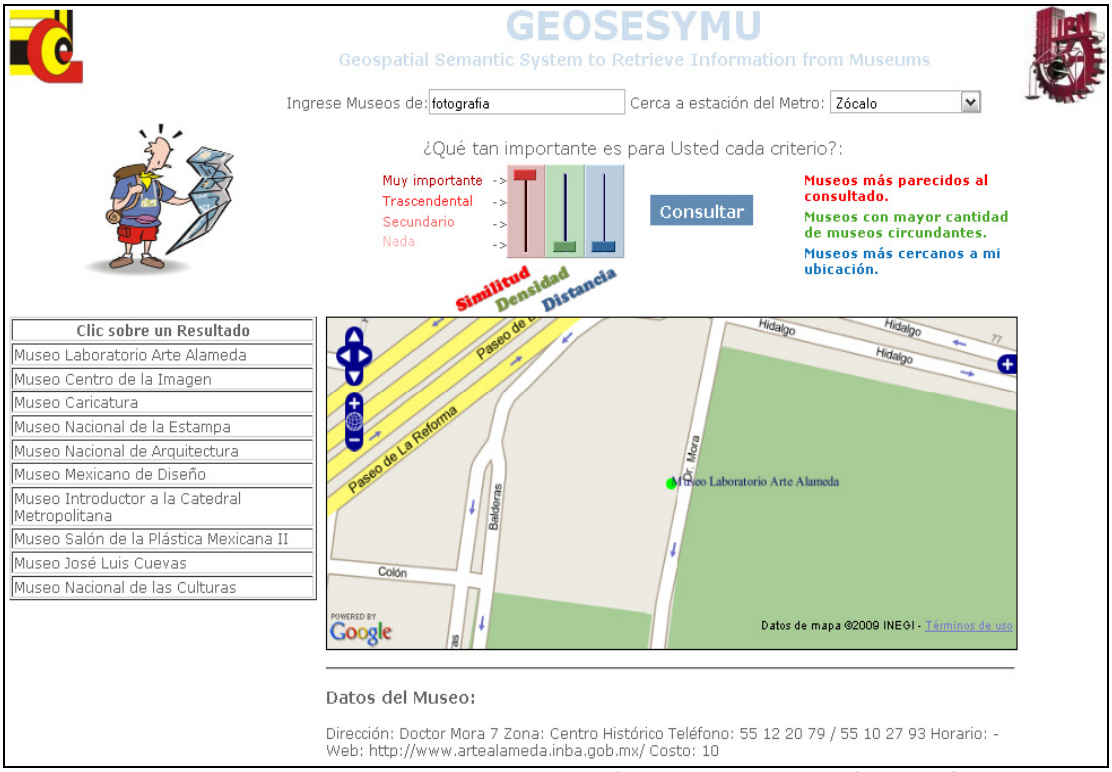

Figura. 5.15 Resultados al buscar museos de fotografía con la configuración de máxima similitud.

Pero además de que al usuario le interese una clase de museos, puede solicitar que sean museos cercanos; entonces el ecualizador estará configurado para tomar en cuenta tanto la similitud como la distancia, ver Figura 5.16. Los resultados obtenidos en función de esa configuración se observan en la Figura 5.17, donde se puede apreciar que, primero aparecen museos del mismo tipo, pero que a su vez, se encuentran más cerca del usuario en relación a los demás museos del mismo tipo.

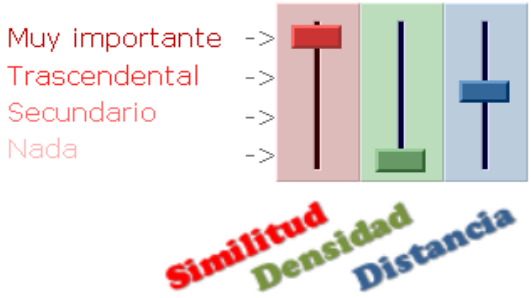

Figura. 5.16 Configuración que regresa resultados con mayor similitud, pero simultáneamente los más cercanos.

Figura. 5.17 Resultados al buscar museos de pintura con la configuración de alta similitud y sitios bastante próximos.

Ahora bien, si el usuario desea visitar un tipo de museo, y no está muy seguro de lo que realmente busca, hay mayor "chance", mayor probabilidad de encontrar lo que busca en un área con mayor cantidad de museos, entonces el ecualizador debe ser configurado como se ve en la Figura 5.18,

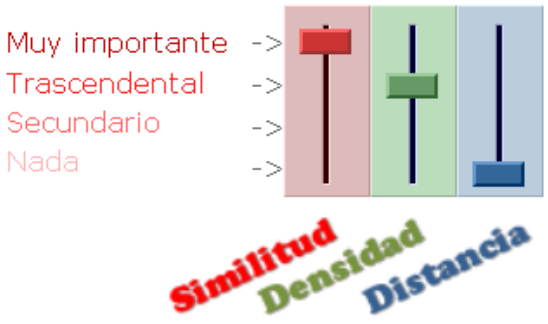

Figura. 5.18 Configuración que regresa resultados con mayor similitud, pero simultáneamente los de mayor densidad.

Con lo cual los resultados obtenidos bajo estas preferencias se observan en la Figura 5.19, donde aparecen primero los museos del tipo solicitado, i.e. museos de pintura, museos tales como: *Museo Pinacoteca de la Casa Profesa*, *Museo Salón de la Plástica Mexicana II*, *Museo Manuel Tolsá*, los cuales simultáneamente tienen mayor densidad cultural; por supuesto es el resultado esperado.

Figura. 5.19 Resultados al buscar museos de pintura con la configuración de alta similitud y alta densidad, sin importar la distancia del usuario hasta cada museo.

5.2.5 Aparentes errores

Al observar el resultado de algunas consultas podría pensarse que son errores, no obstante, en esta sección se explicará que, antes que errores, son resultados que reflejan la flexibilidad y potencialidad del sistema:

Si por ejemplo, hay un usuario ubicado en el metro Bellas Artes, y este usuario tiene limitaciones físicas, o simplemente ese día no está dispuesto a desplazarse mucho, y realiza la siguiente consulta: *Museo Ex-Teresa Arte Actual*, con la configuración de preferencias señalada en la Figura 5.20, se observarán los siguientes resultados.

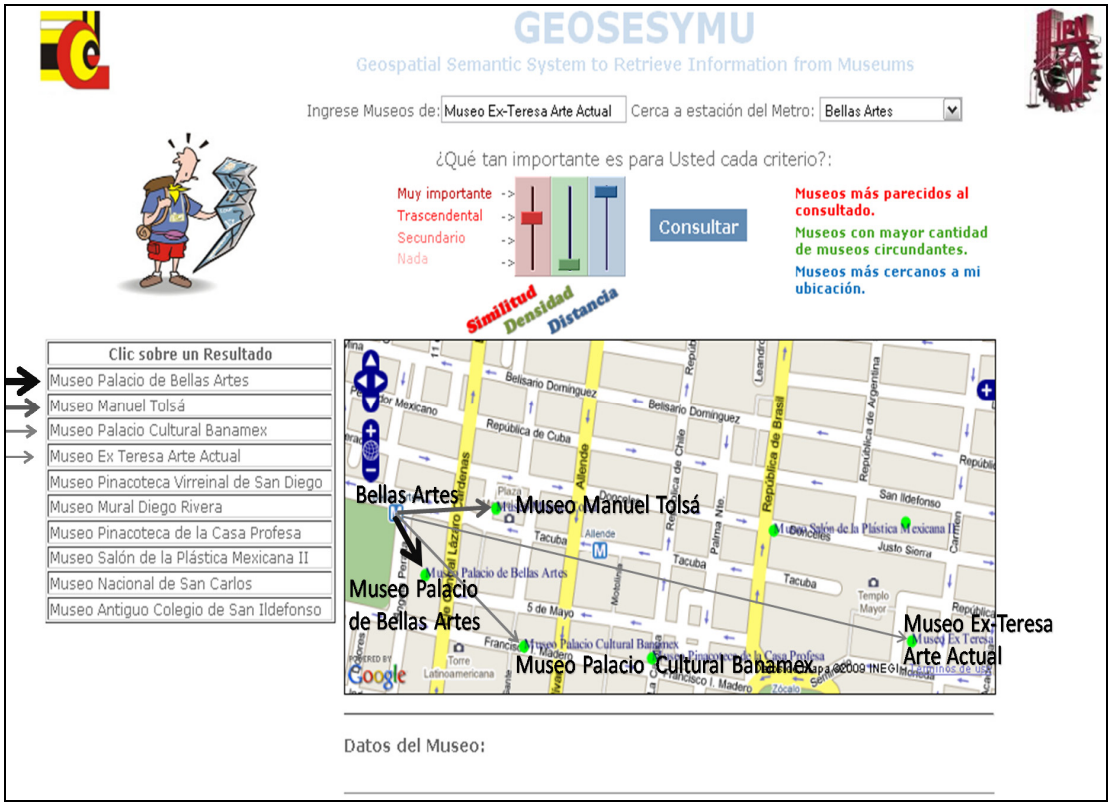

Figura. 5.20 Resultados al buscar: Museo Ex-Teresa Arte Actual desde la estación Bellas Artes con la configuración que regresa resultados con alta similitud, sin importar la densidad, pero que sean los museos menos distantes.

En estos resultados el sistema procede de la siguiente manera, dado que analiza la semántica, se detecta que el usuario está buscando un museo de pintura, pero al hacer el análisis geoespacial, descubre que el usuario está buscando un museo muy cerca a su ubicación, entonces el sistema recomienda los *museos más similares* al consultado y *simultáneamente los más cercanos,* tomando como *factor primordial la distancia*, no obstante también le proporciona el museo que originalmente buscó, pero con menor relevancia. En la Figura 5.20 se ilustra este caso, donde se ofrece como primera alternativa un museo muy similar: *Museo Palacio Bellas Artes* porque solo está a escasos 100 metros, pero el *Museo Ex-Teresa Arte Actual* se encuentra a casi 2000 metros de distancia.

Para contemplar las potencialidades de este enfoque piense en la siguiente

situación:

Si en lugar de preguntar por un museo en particular, el usuario pregunta por un tipo de museo que esté muy cerca, pero resulta que *no hay exactamente* de ese *tipo de museo en las proximidades*, pero mediante el análisis semántico se encuentra que hay *museos muy similares* y simultáneamente *muy cercano*, entonces recomienda estos museos, véase la Figura 5.21. un ejemplo claro se observa al consultar por un Museo del Ejército; el sistema recomienda el *Museo Policía Preventiva de la Ciudad de México* por ser de la misma clase y estar a unos 200 metros, seguido por el *Museo Ejército y Fuerza Aérea Mexicanos* que está mucho más lejano.

Figura. 5.21 Resultados cuando no hay museos sintácticamente iguales pero sí, muy similares y muy próximos, ejemplo 1.

Finalmente puede darse el caso que, definitivamente no hayan museos de la clase que se está buscando, entonces el sistema regresa los más similares situados a la menor distancia. Por ejemplo, si un usuario ubicado en la estación del metro zócalo busca museos de Electricidad, los resultados son los de la Figura 5.22, donde

aparece primero el Museo de la Luz, debido a que los museos de *Óptica*, *Electricidad*, *Mecánica* y *Astronomía*, son similares semánticamente, i.e. son subclase de la clase *Física* y a su vez de la clase *Ciencia*.

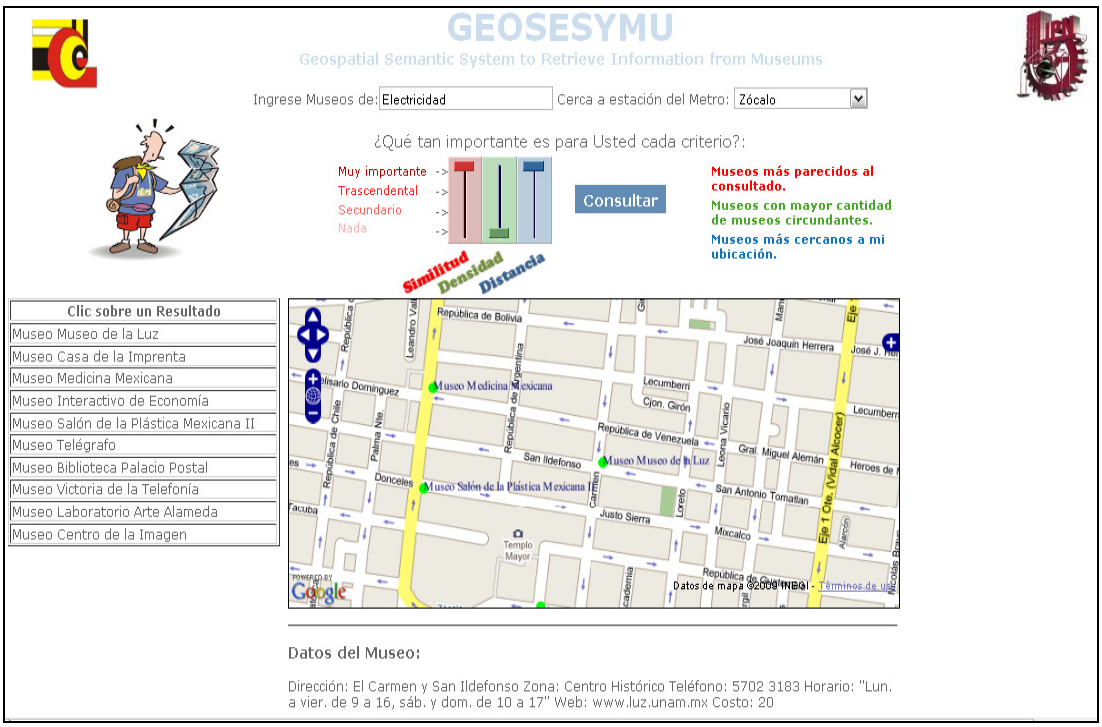

Figura. 5.22 Resultados cuando no hay museos exactamente iguales pero hay otros similares y muy próximos, ejemplo 2.

Se presenta ahora, un aspecto trascendental a la hora de hablar de sistemas que van más allá de la coincidencia sintáctica, y toman partida de otros aspectos de la realidad de un dominio para resolver este problema, en este caso se toma ventaja del dominio geoespacial. Por ejemplo, piense en lo siguiente: se puede aceptar que un usuario ingrese una consulta no prevista o incluso una consulta sin sentido, es decir, alguien quizá por error, o de manera involuntaria introdujo los siguientes caracteres buscando el museo llamado: "%7!$@¿ójó(:-)W", por supuesto que no hay un museo con ese nombre, pero lo que no puede suceder, es que un usuario haga una consulta desde un lugar que no existe, pues todos y en todo momento tienen una posición sobre el globo terráqueo cualquiera que sea. Lo anterior lleva al caso en que un usuario desea poner a prueba el sistema, por tanto introduce

como consulta una combinación extraña de caracteres como "%7!$@¿ójó(:-)W", entonces el sistema, descarta el componente semántico, ya que no existirá correspondencia de este término con ninguno otro en el repositorio, y regresará los museos más próximos a la posición suministrada por el usuario (i.e, la estación del metro o en versiones futuras, las coordenadas tomadas directamente desde el sistema de posicionamiento global de su dispositivo móvil), observe por ejemplo la Figura 5.23,

Figura. 5.23 Resultados cuando se realiza una consulta lexicalmente inconsistente, se regresan los museos más próximos.

Se han introducido los caracteres mencionados, pretendiendo buscar un museo con este nombre o de este tipo, ubicados por ejemplo en la estación del metro Bellas Artes, los resultados en este caso son los museos: Museo Palacio de Bellas Artes, Museo Manuel Tolsá, Museo Nacional de Arte, entre otros, porque son los museos menos distantes (solo 250 metros) de la ubicación actual del usuario.

Después de observar que el sistema regresa resultados bastante razonables, se presenta un resumen de los *valores* que están por detrás de estos resultados, es decir, invisibles al usuario. En la Tabla 5.6 se ejemplifican los resultados de una consulta del tipo Q2, es decir, un usuario que pregunta por un tipo de museo, en esta tabla se puede observar la dificultad de tener ordenados los resultados mediante un solo criterio, ya que al ordenarlos por ejemplo, según la mayor similitud, se incurre en el problema de restar relevancia a un museo que por su distancia debería aparecer primero, si así lo hubiese preferido el usuario.

Tabla 5.6 Ejemplo de resultados sin integrar, ordenados en función de la similitud

Q2= Museo de Pintura cerca al Metro Bellas Artes

Museo	ε	σ	x
Museo Pinacoteca de la Casa Profesa	0.0	0.14	0.695
Museo Antiguo Colegio de San Ildefonso	0.0	0.14	+1.00
Museo Manuel Tolsá	0.0	0.06	0.246
Museo Pinacoteca Virreinal de San Diego	0.0	0.06	0.465
Museo Mural Diego Rivera	0.0	0.09	0.570
Museo Salón de la Plástica Mexicana II	0.0	0.06	0.931
Museo del Palacio de Bellas Artes	0.0	0.06	0.144
Museo Palacio Cultural Banamex	0.0	0.14	0.414
Museo Nacional de San Carlos	0.14	0.06	+1.00
Museo José Luis Cuevas	0.14	0.06	+1.00

Ahora bien, si se enfatiza en dos resultados, particularmente la segunda fila y la séptima fila, se puede observar que en algunas situaciones no son suficientes los valores de la medida de similitud semántica, pues a pesar de que dos museos sean del misma tipo, uno puede estar muy distante, mientras que otro puede estar a solo unos pasos, por ejemplo, el valor de similitud de Museo Antiguo Colegio de San Ildefonso es ε=0, o sea, es un museo de pintura, pero entre los museos de pintura, este es el museo más lejano, mientras que el Museo del Palacio de Bellas Artes está a tan solo una calle (i.e. la distancia desde el usuario hasta dicho museo es x=144 metros), y su valor de similitud semántica igualmente es ε=0. Lo anterior hace pensar que si un usuario usa un sistema para buscar un museo y después de usarlo se da cuenta que, el sistema lo envió a un museo a más de diez o quince calles, en lugar de indicarle un museo del mismo tipo que estaba a solo una o dos

calles, es muy posible que el usuario pierda completamente el interés por un sistema con estas deficiencias y busque una mejor alternativa, una un poco más "inteligente" o al menos, más cercana a la "intuición humana".

Ahora se examina y compara, empleando el modelo de recuperación e integración propuesto, pueden obtenerse resultados más intuitivos, más cercanos a los gustos particulares de un usuario, ya que, dada la consulta "Museos *de Pintura*" y las preferencias del usuario para cada criterio, se establecen los correspondientes pesos para similitud, densidad y distancia en $W\varepsilon=0.9$, $W\sigma=0.2$ and $W_x=0.7$ respectivamente, se observa que en primer lugar aparece un museo que satisface en mayor grado todos los criterios simultáneamente, pues es un *museo de pintura*, y a la vez es el más cercano según la ubicación de usuario; dichos resultados se ilustran en la Tabla 5.7

Tabla 5.7 Resultados de la consultas usando el modelo propuesto
Q2= Museo de Pintura cerca al Metro Bellas Artes

Museo	R(ε,Wε, σ,Wσ, x,Wx)
Museo del Palacio de Bellas Artes	0.58
Museo Manuel Tolsá	0.56
Museo Palacio Cultural Banamex	0.53
Museo Pinacoteca Virreinal de San Diego	0.51
Museo Pinacoteca de la Casa Profesa	0.50
Museo Mural Diego Rivera	0.49
Museo Salón de la Plástica Mexicana II	0.47
Museo Antiguo Colegio de San Ildefonso	0.42
Museo Nacional de San Carlos	0.41
Museo José Luis Cuevas	0.39

Claramente los más relevantes son: *Museo del Palacio de Bellas Artes*, *Museo Manuel Tolsá* y *Museo Palacio Cultural Banamex*, porque son los más similares al tipo de museo que se busca, es decir, son Museo de Pintura, y simultáneamente, éstos son los museos más cercanos a la ubicación especificada por el usuario "están entre 100 y 200 metros" y tienen alta densidad.

Otro experimento muestra de igual manera cómo los usuarios ponderan los criterios a través de sus preferencias, en este ejemplo, el usuario quiere visitar museos algo similar al de la consulta, pero sobre todo, muy cercanos a la ubicación especificada, restando importancia a la densidad, es decir, $W\varepsilon = 0{,}5$, $W\sigma = 0{,}1$ y $Wx = 0.9.$, los resultados sin ordenar son los mostrados en la Tabla 5.8.

Dada la consulta: "Museos de Arqueología" cerca al "metro Zócalo", los resultados ordenados por menor distancia, es decir, sin integración serían:

Tabla 5.8. Resultados de la consulta = "Museos de Arqueología" cerca al "metro Zócalo"

Museo	ε	σ	x
Museo Introductor a la Catedral Metropolitana	0.43	0.58	**0.095**
Museo del Templo Mayor	0.0	0.64	**0.140**
Museo Puro Corazón	0.14	0.52	**0.155**
Museo Mexicano de Diseño	0.43	0.5	**0.160**
Museo Galería Cocina Duque de Herdez	0.14	0.64	**0.165**
Museo SHCP	0.43	0.65	**0.211**
Museo Cocina Mexicana	0.14	0.65	**0.266**
Museo Casa de la Imprenta	0.43	0.65	**0.244**
Museo Nacional de las Culturas	0.14	0.66	**0.324**

En este resultado, se puede observar que si el sistema sólo tomara en cuenta un criterio para ordenar los resultados (e.g, criterio de distancia), los resultados serán menos satisfactorios, debido a que, a pesar de que "*Museo Introductor a la Catedral Metropolitana*" es el más cercano (95 metros), no es un museo similar a la consulta, es decir, no es un *museo de Arqueología*. Sin embargo, en segundo lugar aparece "*Museo del Templo Mayor*", que es efectivamente un museo de Arqueología, y está a sólo 140 metros; incluso, se observa en último lugar, un museo de arqueología a solo 324 metros.

Ahora bien, si se compara con los resultados obtenidos mediante la integración de criterios en la Tabla 5.9, claramente el museo más relevante es: "*Museo del Templo Mayor*" que es un *Museo de Arqueología* y, simultáneamente, entre todos los *Museos de Arqueología* es el más cercano (i.e. menos de 150 metros). Además, primero se muestran *Museos de Arqueología*, y luego los más similares, es decir,

Museos de Etnología; ya que los *museos de Arqueología y Etnología* son subclase de los *museos de Antropología*, y luego se muestra el resto, de acuerdo a la relevancia de los museos. En conclusión, aparecen primeros museos más similares y más cercanos simultáneamente.

Tabla 5.9. Resultados usando la expresión de integración propuesta.

Museo	$R(\varepsilon, W\varepsilon, \sigma, W\sigma, x, Wx)$
Museo del Templo Mayor	0.57
Museo Nacional de las Culturas	0.52
Museo SEP	0.50
Museo Puro Corazón	0.50
Museo Galería Cocina Duque de Herdez	0.49
Museo Cocina Mexicana	0.47
Museo Recuerdo de la Academia Mexicana de la Lengua	0.46
Museo Introductor a la Catedral Metropolitana	0.45
Museo Mexicano de Diseño	0.45
Museo Numismático	0.44
Museo Serfín de Indumentaria Indígena	0.43
Museo SHCP	0.43

Es importante señalar aquí que, las preferencias del usuario "perfil" (es decir, la configuración de los pesos w_i) permiten dar resultados precisos, de manera que los resultados (instancias de la ontología u objetos geográficos de la base de datos geográfica) se ajustan a lo que prefiere el usuario. De esta manera el ordenamiento de los resultados obtenido, está en función del usuario, es decir, de sus preferencias. En el último caso presentado, el usuario prefirió museos muy similares y muy cercanos de la ubicación indicada en la consulta, asignando *menos importancia* a aquellos museos que tienen en su vecindad mayor cantidad de Puntos de Interés Cultural.

Los experimentos muestran resultados de un largo camino de investigación, que está dirigido cada vez más a satisfacer las necesidades específicas de un usuario, es decir, personalizar los servicios, de tal manera que la "inteligencia" de las máquinas y más bien, de los programadores e investigadores se ve retada, para ofrecer servicios en una forma similar a como un humano ofrece alternativas a otro

humano en función de sus preferencias. Como resultados de esta investigación se sugiere que, el modelo aquí presentado es una aproximación a la forma como el ser humano recomienda opciones en función de varios criterios y preferencias del usuario.

5.3. Comentarios finales

El modelo expuesto no está atado a la *teoría de confusión*, por tanto su funcionalidad, precisión y futura expansión a otros dominios no dependen de la misma, sino de la estructuración conceptual de la información. Por el contrario, si se usa otra teoría para medir la similitud semántica, o en trabajos posteriores se desarrolla otra versión de *confusión* más precisa o más cercana a la semántica del dominio, entonces el modelo también mejorará y retornará resultados más ajustados a la realidad del dominio donde se esté aplicando el modelo.

A través de cada uno de los resultados se puede observar la capacidad del sistema para ajustarse a las preferencias que cada usuario establece a través del ecualizador, logrando así que su búsqueda arroje resultados más satisfactorios, más ajustados a lo que el propio usuario desea; ésta es claramente la tendencia de los sistemas actuales "la personalización de servicios".

Ahora bien, al presentar resultados que debido a su similitud semántica satisfacen en un menor grado las preferencias del usuario, hace que se evite retornar resultados vacíos, pero sin incurrir en arrojar ruido, es decir, resultados que son completamente alejados de lo que busca el usuario.

Se pudo ver como las potencialidades de la semántica aplicada a este dominio, extienden las posibilidades de reducir enormemente el "silencio" o resultados nulos en los sistemas de recuperación, y en segundo lugar, se mejoran los sistemas GIR

al dotarlos de esa cercanía con la "intuición humana", al ofrecer alternativas que, aunque aparentemente no tengan conexión geográfica y topológica, los humanos son capaces de ofrecerlas como alternativas, aún cuando no se haya preguntado por ellas.

Ahora, es posible culminar con algunos puntos que ratifican algunas de las bondades ya expuestas, por ejemplo:

- Se mostró que al usar el análisis semántico es posible aproximarse a una forma de procesamiento inteligente, integrando criterios de dominios comúnmente aislados, y ponderando dichos criterios según las preferencias del usuario "denominado aquí como perfil del usuario".
- Dado que el sistema está accesible vía internet, ya lo han usado aún personas sin conocimiento del dominio, logrando así una retroalimentación que ha permitido afinar el sistema.
- Al comparar con formas tradicionales de búsqueda como los sistema mostrados en el estado del arte, se mostró que los resultados regresados por el modelo propuesto no obedecen exclusivamente a una coincidencia sintáctica, sino que aprovechan las bondades del análisis semántico para regresar resultados conceptualmente similares, esto incita a formular la siguiente pregunta, ¿Se debe olvidar el enfoque geoespacial clásico?, por supuesto que no, se mostró que éste enfoque además de refinar los resultados del análisis semántico, sus componentes espaciales permiten obtener resultados que auxilian al análisis semántico cuando hay algún error sintáctico del usuario, ofreciendo resultados que por su proximidad espacial, podrían interesarle al usuario dado su corta distancia.
- En definitiva, la manera como se han integrado los criterios, permitió evitar los resultados vacíos, presentando siempre al usuario alternativas, ya sea, desde el punto de vista geoespacial o desde la perspectiva de la similitud semántica.

CAPÍTULO 6.

CONCLUSIONES Y TRABAJO A FUTURO

No os engañéis; Dios no puede ser burlado: pues todo lo que el hombre sembrare, eso también segará.
Gálatas 6:7
Las revoluciones se producen en los callejones sin salida.
Bertolt Brecht

6.1 Conclusiones

A continuación se describen las principales conclusiones en torno a los resultados obtenidos con el presente trabajo de investigación, enmarcado en el ámbito de la recuperación de información geográfica (GIR), con especial enfoque en la recuperación semántica geoespacial.

Se destaca que para cada consulta se recupera un conjunto de resultados ordenados en función de su grado de relevancia. Los resultados son Puntos de

Interés Cultural (CPI), en particular Museos que representan una tripleta de atributos que, según sus valores se ajustan en mayor o menor medida al museo que busca un usuario.

Se mostró que a pesar de no existir información que satisfaga sintácticamente una consulta, puede haber información similar que, con un pequeño margen de error satisface la búsqueda del usuario, esto lleva a afirmar que la metodología propuesta permite obtener resultados similares, en adición a los que tienen correspondencia exacta con una consulta sobre Puntos de Interés Cultural, evitando así resultados vacíos.

A través de cada uno de los resultados, pudo observarse la tendencia del sistema para ajustarse a las preferencias que cada usuario establece a través del ecualizador, logrando que su búsqueda arroje resultados más relevantes, pero al mismo tiempo, con mayor precisión y exactitud mediante el control de la similitud semántica. Ésta es claramente la tendencia de los sistemas actuales "*la personalización de servicios*".

Al ofrecer alternativas al usuario, es decir, resultados que debido a su similitud semántica satisfacen en un menor grado la consulta y las preferencias del usuario, se evitan los resultados vacíos, pero sin incurrir en retornar ruido, es decir, resultados que son completamente alejados de lo que busca el usuario.

La tarea de integración semántica geoespacial no es trivial, ya que algunos datos proceden de un contexto cualitativo y otros de un contexto cuantitativo. Se presentó un enfoque para aproximarnos a este problema mediante la utilización de jerarquías como caso particular de las ontologías. Además, se usaron medidas de similitud semántica para permitir la integración de conceptos con el análisis

geoespacial. Los resultados obtenidos muestran que es posible la integración de criterios en un solo criterio utilizando el enfoque semántico geoespacial.

Con este trabajo se contribuye a disminuir la carencia de sistemas que integren aspectos semánticos conjuntamente con aspectos espaciales, de allí que el modelo de integración presentado en este trabajo, expone el análisis semántico como un excelente complemento mutuo para las técnicas tradicionales de análisis geoespacial.

La metodología aquí propuesta, es expandible en cuanto al número de criterios, sumado a esto, es fácil su migración a otros dominios donde es necesaria la recuperación controlada de información desde repositorios de datos.

Las bondades del espacio de integración propuesto están dirigidas a ofrecer como resultados, alternativas que bajo un solo criterio no tienen conexión o relación alguna.

Las potencialidades de la semántica aplicada a este dominio extiende las posibilidades de reducir enormemente los resultados nulos en los sistemas de recuperación; en particular, se mejoran los GIR, al dotarlos de esa cercanía con la "intuición humana", ya que se ofrecen alternativas que, aunque aparentemente no tienen conexión geoespacial o topológica, los humanos son capaces de ofrecerlas como alternativas, aún cuando no se les haya preguntado por ellas.

6.2 Aportaciones

Finalmente, se puntualizan algunas de las principales contribuciones de este trabajo:

- Uso de las preferencias del usuario para mejorar la *precisión* de los resultados e incrementar la *interacción con el usuario,* para *personalizar los servicios*.

- Una API desarrollada en Java con métodos específicos para:
  - Calcular la *confusión* entre conceptos de una jerarquía de clases en un archivo OWL
  - Generar de manera automática la tabla de *confusión* de una jerarquía de clases.

- Enriquecimiento de la *teoría de confusión*.
  - Mecanismos geoespaciales que posibilitan el reordenamiento y refinación intra-clase "amplitud", para entidades con igual valor de similitud, de igual modo, se supera la limitación de instancias únicas en las jerarquías propuestas por dicha teoría.

- Un procedimiento que posibilita la integración de criterios para flexibilizar y controlar la recuperación de información.

- Modelo en el cual se complementan mutuamente el análisis Geoespacial con el análisis Semántico, ofreciendo resultados más satisfactorios.

- Una aproximación de tipo formal mediante expresiones matemáticas (Expresión $\boldsymbol{R_w}$), para abordar el problema de integración de criterios del dominio semántico y geoespacial.

6.3 Trabajo a futuro

- Desarrollar experimentos aplicando la propuesta metodológica de este trabajo en otros dominios o campos, que requieren mecanismos de integración de criterios y recuperación de información.

- Refinar la forma de obtener los valores en cada uno de los criterios espaciales, de tal manera que, por ejemplo, el cálculo de distancia geométrica no sea simplemente una distancia lineal y simétrica entre dos puntos, sino que se consideren más aspectos de la realidad geoespacial de un entorno socio-urbano.

- Contemplar el uso del tiempo como una dimensión que podría enriquecer el trabajo, dado que el tiempo representa una dimensión no homogénea. El uso de esta componente temporal es factible ya que, el modelo también usa una dimensión no homogénea como es la similitud semántica.

- Complementar el modelo de recuperación con un módulo de Procesamiento de Lenguaje Natural, para procesar consultas lexicalmente más complejas, por ejemplo, si en lugar de buscar solo un museo de pintura, la búsqueda fuera: Museo de Pintura Rupestre.

- Indagar y proponer más componentes de análisis geográfico y topológico, para refinar y complementar las medidas de similitud semántica.

- Estudiar más medidas de similitud en estructuras conceptuales, con miras a retornar resultados que por su similitud con la entidad buscada, se acerquen más a la "intuición humana" o a la realidad del dominio.

- Ampliar el uso de perfiles de usuario, de tal manera que el perfil sea un conjunto de criterios dinámicos respecto al comportamiento del usuario a través del tiempo. Además contemplar la posibilidad de incluir historicidad y mejores formas de interacción con el usuario y la captura de sus preferencias.

REFERENCIAS

[BCJ05] Bucher, B. P., Clough, H. and Joho, R. 2005. Geographic IR Systems: Requirements and Evaluation. En Proceedings of the 22nd International Cartographic Conference. 2005.

[BG05] Borg, I. and Groenen, P.: "Modern Multidimensional Scaling: theory and applications" (2nd ed.), Springer-Verlag New York, 2005.

[BLH01] Berners-Lee, T. and Hendler, J. The Semantic Web: A new form of Web. content that is meaningful to computers will unleash a revolution of new possibilities, Scientific American, 2001.

[BT06] Blázquez del Toro, J. M., et al. 2006. Ordenación de resultados en consultas semánticas sobre base de datos. Madrid : s.n., 2006

[Chu06] Chaudhuri, S., et al. 2006. Probabilistic information retrieval approach for ranking of database query results. ACM Press. pp. 1134-1168, 2006

[DFMP92] L. De Floriani and P. Marzano and E. Puppo. Spatial queries and data models, Information Theory: a Theoretical Basis for {GIS},Springer-Verlag, Lecture Notes in Computer Science, N.716;ed, I. Campari A. U. Frank and U. Formentini",pages:113-138,1992.

[Ege02] Max J. Egenhofer, Toward the semantic geospatial web, Proceedings of the 10th ACM international symposium on Advances in geographic information systems, 2002.

[FK82] Francis, W. and Kucera, H. 1982. Frequency Analysis of English Usage, Houghton Mifflin Co. 1982.

[Fon00] Fonseca, F. 2000. Users, Ontologies and Information Sharing inUrbanGIS in ASPRS. 2000.

[Fra91] Frank, A. U. Properties of Geographic Data: Requirements for Spatial Access Methods. In Guenther, O & Schek, H.-J. (eds) Advances in Spatial Databases, (pp. 225-234). Berlin : Springer-Verlag, 1991.

[GAL04] Guzmán-Arenas A. & Levachkine S., "Graduated Errors in Approximate Queries using Hierarchies and Ordered Sets", in Monroy R., Arroyo-Figueroa G., Sucar L.E. & Sossa-Azuela J.H. (Eds.), Advances in Artificial Intelligence, Third Mexican International Conference on Artificial Intelligence (MICAI' 2004), Mexico City, Mexico, Lecture Notes in Computer Science 2972, Springer-Verlag, 119-128, 2004.

[GP04] Gómez-Pérez et. Al. 2004. "Ontological Engineering". Springer (2004)

[GGW03] Ganesan, P., Garcia-Molina, H. and Widom, J. Exploiting hierarchical domain structure to compute similarity. ACM Trans. Inf. Syst. 21(1): 64-93, 2003.

[Gru95] T. Gruber, Toward principles for the design of ontologies used for knowledge sharing, International Journal of Human-Computer Studies archive, Volume 43, Issue 5-6 Nov./Dec. 1995 table of contents,
Special issue: the role of formal ontology in the information technology, pp. 907-928, ISSN:1071-5819, 1995

[Gua95] Guarino, N.: Formal ontology, conceptual analysis and knowledge representation. International Journal of Human-Computer Studies. Vol. 43 Num:5-6, pp. 625-640, 1995.

[Her03] Herman, I. 2003. W3C Head of Offices Helsinki, Introduction to the Semantic Web. 2003.

[HF05] G. Hobona, P. James, and D. Fairbairn. An evaluation of a multidimensional visual interface for geographic information retrieval. In the CIKM Workshop on Geographic Information Retrieval, pages 5–8. ACM Press, 2005.

[HSO98] Hirts, G. and St-Onge, D. 1998. Lexical chains as representation of context for the detection and correction of malapropisms, en Fellbaum, C., WordNet: An Electronic Lexical Database, MIT Press. pp. 265–283, 1998.

[HSW96] Van Heijst, G., Schereiber, A. T. and Wielinga, B. J. 1996. "Using Explicit Ontologies in KBS Development" en International Journal of Human and Computer Studies. 1996.

[ICO01] ICOM. 2001. International Council of Museums, http://icom.museum/hist_def_eng.html, último acceso Febrero 2009.

[INE05] INEGI. 2005. Conjunto de datos vectoriales, Diccionario de Datos Turísticos, Escala 1: 1 000 000.

[JAT01] Jones, C.B., H. Alani, and D. Tudhope Geographical information retrieval with ontologies of place, in Spatial Information Theory, LNCS 2205, 322-35, 2001.
[Jen07] Jena 2. "A Semantic Web Framework". http//www.hp1.hp.com/semweb/jena.html. 2007, último acceso: mayo 2009.

[KK90] Kim, Y. W. and J.H. Kim. "A Model of Knowledge Based Information Retrieval with Hierarchical Concept Graph". Journal of Documentation, 46(2): 113-136, 1990.

[Lar07] Ray R. Larson, Geographic Information Retrieval and Spatial Browsing, University of California, Berkeley. https://sherlock.ischool.berkeley.edu/geo_ir/PART1.html, último acceso: Septiembre 2009.

[LP04] Ray R. Larson, Patricia Frontiera, Geographic Information Retrieval (GIR) Ranking Methods for Digital Libraries, Pages: 415 – 415, Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, Tuscon, AZ, USA June 07 - 11, 2004.

[LC98] Leacock, C. and Chodorow, M. 1998. Combining local context and WordNet similarity for word sense identification, en Fellbaum, C., WordNet: An Electronic Lexical Database, MIT Press. pp. 265–283, 1998.

[Len95] D.B. Lenat: CYC: a large-scale investment in knowledge infrastructure, Communications of the ACM 38 (11) (1995)33.

[LG04] Levachkine S. & Guzmán-Arenas A., "Hierarchies Measuring Qualitative Variables", in Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing (CICLing' 2004), Lecture Notes in Computer Science, Vol. 2945, Springer-Verlag, 262-274, 2004.

[LG07] Levachkine, S. and Guzmán-Arenas, A. 2007. Hierarchy as a new data type for qualitative variables. Expert Systems with Applications: An International Journal, 32(3): pp. 899-910, 2007.

[Lin98] Lin, D. An information-theoretic definition of similarity, Proc 15th International Conference on Machine Learning. pp. 296–304, 1998.

[Lor03] Lord, P. W., et al. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation, Bioinformatics, 19(10). pp. 1275– 1283, 2003.

[Mat09] Mata Rivera, Miguel Félix. Recuperación y Ponderación de Información Geográfica desde Repositorios No Estructurados Conducidas por Ontologías, México D.F., Presentada en el Centro de Investigación en Computación CIC-IPN para obtención del grado de Doctor en ciencias de la Computación. 2009.

[McG01] McGuinness, D.L.: Ontologies come age, in D. Fensel, J. Hendler, H. Lieberman and W. Wahlster (eds), Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential, MIT Press, Cambridge, MA, 2001.

[MC91] Miller, G. A. and Chrales, W. G. 1991. Contextual Correlates of Semantic Similarity. Language and Cognitive Processes. 1991. pp. 6(1):1–28.

[Mee99] R. Meersman: Semantic Ontology Tools in IS Design. ISMIS'99 Conference in Warsaw, June, 1999.

[Mil90] Miller, G. A., et al. 1990. WordNet: an on-line lexical database. International Journal of Lexicography, 3(4). 1990.

[McG83] M. J. McGill. Introduction to Modern Information Retrieval. McGrawHill, 1983.

[Nig85] Nigam, M.L., Fundamentals of Museology, Hyderabad, p.p. 127-139, 1985.

[OGC] Open Geospatial Consortium, Inc., http://www.opengeospatial.org/

[PGGU08] Perea, J.M, García-Cumbreras, M.A., García-Vega, M., Ureña, L.A. Sistemas de Recuperación de Información Geográfica multilingües en CLEF. Procesamiento de Lenguaje Natural, 40. 129- 136, Abril 2008.

[Pro08] Protégé. 2008. The Protégé Ontology Editor and Knowledge Acquisition System. URL: http://protege.stanford.edu/, último acceso: Marzo 2009.

[REI09] Tumasch Reichenbacher. Geographic Relevance in Mobile Services. In Proceedings of the Second International Workshop on Location and the Web (LocWeb 2009), Boston, Massachusetts, April 2009.

[RER99] Rodriguez, M. A., M. J. Egenhofer and R. D. Rugg. "Assessing semantic similarities among geospatial feature class definitions". Interop'99. A. Vckovski, K. Brassel and H.-J. Schek (eds). Berlin, Springer. Lecture Notes in Computer Science 1580: 189-202, 1999.

[Res93] P. Resnik. Selection and Information: A Class-Based Approach to Lexical Relationships, Ph.D Dissertation. 1993.

[RG65] Rubenstein, H. and Goodenough, J. B. Contextual Correlates of Synonymy. Communications of the ACM, 8(10). pp. 627-663, 1965.

[RIS+94] P. Resnick, N. Iacovou, M. Sushak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In Proc. Computer Supported Collaborative Work Conference, 1994.

[RMBB89] Rada, R., H. Mili, E. Bicknell and M. Blettner. "Development and application of a metric on semantic nets." IEEE Transactions on Systems, Man and Cybernetics 19(1): 17-30, 1989.

[Rus79] Rusbergen, C. J. Information Retrieval. 2nd ed., Butterworths. London: s.n., 1979.

[Sal89] Salton, G. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Reading, Mass.: Addison-Wesley, 1989.

[SCG09] Secretaria de Cultura del Gobierno Distrito Federal 2009 http://www.df.gob.mx, último acceso: Febrero 2009.

[SJAB04] Mark Sanderson, Kalervo Järvelin, James Allan, Peter Bruza (Eds.): SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, ACM 2004, ISBN 1-58113-881-4, 2004.

[Sar08] Sarabia López, Gerardo. Búsqueda y Ponderación de información Contenida en Bases de Espaciales, Utilizando Jerarquías. México D.F., Presentada en el Centro de Investigación en Computación CIC-IPN para obtención del grado de Maestro en ciencias de la Computación.

[SB98] Salton, G. and Buckley, C. Term-weighting approaches in automatic text retrieval. Inf. Process. Manage, 24(5): 513-523, 1998.

[STC09] Secretaria de Turismo de la Ciudad de México, http://www.mexicocity.gob.mx/museos.php, último acceso: Febrero 2009.

[SGM00] A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In Proc. AAAI Workshop on AI for Web Search, 2000.

[She62] Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function, I. Psychometrika, 27(2): 125-140, 1962.

[Sig] SIG Metrópoli 2025. http://www.sigmetropoli2025.com/, último acceso: Abril 2009.

[Spa07] SPARQL W3C. URL: http://www.w3.org/TR/rdf-sparql-query/. 2007. último acceso: mayo 2009.

[SR92] Shapiro, S.C., and Rapaport, W.J., "The SNePS Family" Computers and Mathematics with Applications, 1992, 23: 243-275.

[Sow00] Sowa, J. "Knowledge Representation: Logical, Philosophical, and Computational Foundations", Pacific Grove, CA., Brook/Cole, a division of Thomson Learning, 2000.

[Tve77] Amos Tversky. Features of similarity. Psychological Review, 84(4): 327-352, 1977.

[W3C01] W3C. 2001. URL: http://www.w3.org/2001/sw/. 2001. último acceso: mayo 2009.

[Wor] WordNet, An electronic lexical database, http://WordNet.princenton.edu, último acceso: Septiembre de 2009.

[Xaq] xaqui. http://www.xaqui.com.mx/index.htm, último acceso: Abril 2009.

ANEXO A. Jerarquía del dominio

Este anexo está comprendido por la *jerarquía diseñada* para Museos, tomando como fuentes de información las descritas en el apartado de conceptualización del dominio del capítulo 4. Dicha jeraquía se presenta en formato OWL, dada su amplia difusión y sumado a que, de esta forma se facilita la interoperabilidad con otros sistemas y futuros trabajos de investigación.

A manera de ejemplo se presenta un fragmento de dicha jerarquía:

```
<owl:Class rdf:ID="Pintura">
   <rdfs:subClassOf rdf:resource="#Artes_Plasticas"/>
   <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
   >Plasticas</rdfs:label>
 </owl:Class>
<owl:Class rdf:ID="Litografia">
   <rdfs:subClassOf>
     <owl:Class rdf:ID="Artes_Graficas"/>
   </rdfs:subClassOf>
<rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
   >Artes_Graficas</rdfs:label>
```

La información detallada puede consultarse en el archivo digital adjunto a esta tesis.

ANEXO B. Cálculo del valor de *confusión*

Este anexo está comprendido por el código fuente desarrollado para calcular automáticamente la tabla de *confusión*, a partir de la *jerarquía de museos* contenida en el archivo OWL.

```
//---------------------- Cálculo de los valores de confusión conf(fila "r", columna "s") -----------------------------------
for(c=1; c<=w;c++)
for(f=1; f<=w;f++){
  vec[f][c]=new Vector();
   vec[f][c].add(Confusion( (String)vec[f][0].get(0) , (String)vec[0][c].get(0) ) );
  System.out.println("Conf("+(String)vec[f][0].get(0)+","+
     (String)vec[0][c].get(0)+")="+vec[f][c].get(0)+" en f,c: "+f+","+c);
}
//----Guardar Tabla de confusion: TablaConfusion.txt, confusion conf(fila "r", columna "s")------
try{
     BufferedWriter bw = new BufferedWriter(new FileWriter(sFichero));
    for(f=0; f<=w;f++){
      for(c=0; c<=w;c++){
            if(vec[f][c].get(0).getClass()==vec[0][0].get(0).getClass())
                  bw.write( Integer.toString(((Double)vec[f][c].get(0)).intValue()) +"\t");
           else
                  bw.write( (String)vec[f][c].get(0) +"\t");
      }
         bw.write("\n");
         bw.flush();
    }
    bw.close();
} catch (IOException ioe){
        ioe.printStackTrace();
}
```

El código fuente completo puede consultarse en el material adjunto a este trabajo.

ANEXO C. Implementación del sistema

Este anexo versa sobre el código fuente de la aplicación web desarrollada en *php* versión 5 en conjunto con el visor OpenLayer, el cual permite la visualización de la cartografía y la superposición de capas dentro de la página Web, inclyendo servicios como WMS y WFS para visualizar información cartográfica de Google Maps, Yahoo!, Microsoft entre otros.

La aplicación Web accede a la base de datos implementada en *PostgreSQL/PostGIS* versión 8.3 y al servidor Web Apache que trabaja en conjunto con el servidor de mapas GeoServer.

En el material adjunto se anexan tanto el código fuente en php, la base de datos y los archios de instalación para los servidores.