INSTITUTO POLITECNICO NACIONA

Centro de Investigación en Computación

Valoración de la similitud entre representaciones múltiples de datos geográficos

Tesis

que para obtener el grado de

Maestría en Ciencias de la Computación

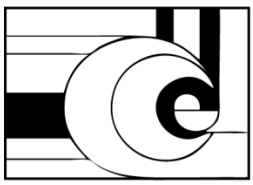
presenta

Sergio Eduardo Solano Pinedo

Asesores:

Dr. Marco Antonio Moreno Ibarra

Dr. Miguel Jesús Torres Ruiz



México DF, diciembre del 2011.

SIP-14 bis

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México, D.F. siendo las 10:00 horas del día 23 del mes de noviembre de 2011 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

Centro de Investigación en Computación

para examinar la tesis titulada:

"VALORACIÓN DE LA SIMILITUD ENTRE REPRESENTACIONES MÚLTIPLES DE DATOS GEOGRÁFICOS"

Presentada por el alumno:

SOLANO	**PINEDO**	**SERGIO EDUARDO**
Apellido paterno	Apellido materno	Nombre(s)

Con registro: **B 0 9 1 6 6 9**

aspirante de: **MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN**

Después de intercambiar opiniones los miembros de la Comisión manifestaron ***APROBAR LA TESIS***, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA

Directores de Tesis

Dr. Marco Antonio Moreno Ibarra

Dr. Miguel Jesús Torres Ruiz

Dr. Rolando Quintero Téllez

M. en C. Sandra Dinora Orantes Jiménez

Dr. Miguel Félix Mata Rivera

PRESIDENTE DEL COLEGIO DE PROFESORES

Dr. Luis Alfonso Villa Vargas

INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACION EN COMPUTACION
DIRECCION

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA CESIÓN DE DERECHOS

En la Ciudad de México el día 24 del mes noviembre del año 2011, el (la) que suscribe Sergio Eduardo Solano Pinedo alumno (a) del Programa de Maestría en Ciencias de la Computación con número de registro B091669, adscrito a Centro de Investigación en Computación, manifiesta que es autor (a) intelectual del presente trabajo de Tesis bajo la dirección de Dr. Marco Antonio Moreno Ibarra y cede los derechos del trabajo intitulado Valoración de la similitud entre representaciones múltiples de datos geográficos, al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección juancalambres@gmail.com. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

Solano

Sergio Eduardo Solano Pinedo

Nombre y firma

Resumen

Actualmente nos encontramos frente a una gran diversidad de representaciones del mismo fenómeno geográfico, en particular para el campo de la información geográfica. Dicha diversidad se conoce como representaciones múltiples ytiene su origen en aspectos como el propósito de los datos, los métodos de obtención, la digitalización, la dinámica del mundo o los errores en el proceso de elaboración.

Las representaciones múltiples poseen conocimiento abundante y significativo pero entre sus partes es común encontrar desacuerdos sobre el significado, interpretación, uso y aspecto geométrico de los datos. Dicho desacuerdo se conoce como heterogeneidad semántica y surge ya que cada parte de las representaciones múltiples se produce y administra de manera independiente.

Para hacer disponible el conocimiento dentro las representaciones múltiples esnecesario encontrar la relación semántica entre sus partes y la similitud es unmodo de hacerlo ya que es la clave para la integración semántica.

Con base en lo anterior, en este trabajo se propone un enfoque de teoría de conjuntos para valorar la similitud entre representaciones de datos geográficos. La valoración de similitud contempla también el diseño e implementación de un grado de interoperabilidad semántica que tiene la intención de servir como criterio de integración semántica entre las partes de las representaciones múltiples, y así hacer más disponible su conocimiento.

Como caso de estudio se trabaja con representaciones de localidades de la República Mexicana. Al sistema se introducen diferentes representaciones de localidades, de las cuales el usuario selecciona dos para valorar su similitud y su grado de interoperabilidad semántica. La similitud se analiza a diferentes niveles y por cada nivel, ésta se representa de forma gráfica, buscando ofrecer al usuario resultados útiles y fáciles de entender.

Abstract

Nowadays we face a variety of geographic representations of the same phenomenon, particularly in the field of geographic information. This diversity is known as multiple representations and has its origin in areas such as the purpose of data collection methods, digitization, the dynamic of the world and errors in the process itself.

Multiple representations have significant and abundant knowledge but its common to find disagreements about the meaning, interpretation, use and geometric aspect of the data in its parts. This disagreement is known as semantic heterogeneity and arises because each part of multiple representations is produced and managed independently.

To make the knowledge in multiple representations available is necessary to find the semantic relation between its parts and the similarity is one way to do it because it is the key to semantic integration.

Based on the foregoing, this paper proposes a set-theoretic approach to assess the similarity between representations of geographic data. The assessment of similarity also includes the design and implementation of a degree of semantic interoperability that is intended to serve as a criterion of semantic integration between parts of multiple representations, and thus make knowledge more available.

As case study we work with representations of locations Mexico. Different representations of locations are put into the system, then the user selects two to assess their similarities and their degree of semantic interoperability. The similarity is analyzed at different levels and at each level, similarity is represented graphically, seeking to provide the user useful and easy to understand results.

Índice

Índice de figuras

Índice de tablas

1. Introducción

Hoy en día, existe una gran diversidad de representaciones de un mismo fenomeno geografico [Scappapietra et al., 2000]. Esta observacion es particularmente cierta para el campo de la informacion geografica, ya que diferentes motivos pueden explicar lo anterior [Sheeren et al., 2009].

> Las bases de datos espaciales tienen diferentes propositos y por lo tanto diferente contenido, organizacion y granularidad;
>
> la informacion se produce con diferentes métodos y de diferentes fuentes;
>
> las fuentes, incluso siendo idénticas, pueden interpretarse y digitalizarse de diferentes maneras;
>
> el mundo geografico se encuentra en constante evolucion, así una representacion solo captura la realidad en un momento dado;
>
> finalmente, la captura de datos es un proceso complicado, por lo cual, una base de datos tiene altas probabilidades de contener errores [Sheeren et al., 2009].

De acuerdo con lo anterior, esta gran diversidad de informacion se conoce como representaciones múltiples y contiene un conocimiento abundante y significati- vo [Vangenot et al., 2002]. Las representaciones múltiples surgen en el momen- to en que diferentes grupos de usuarios producen informacion sobre un mismo fenomeno del mundo real. Como cada grupo tiene requerimientos y percepciones particulares, se genera informacion heterogénea. El problema de las representaciones múltiples está justamente en su diversidad, dado que puede ser muy heterogénea no hay un modo unificado de integrarla o compartirla.

Una de las razones que hacen particularmente difícil el soporte de representaciones múltiples es su heterogeneidad semántica. La heterogeneidad semántica es el desacuerdo sobre el significado, interpretacion, uso y aspecto geométrico de los datos [Sheth & Larson, 1990]. Por tanto, cuando entre representaciones mu ´ltiples existe heterogeneidad semántica, su soporte se complica, ya que se de- be entender el correcto significado de la informacion. En este sentido, entender la heterogeneidad semántica es la clave para establecer la similitud entre las representaciones múltiples [Mustière et al., 2009].

Un modo de abordar la heterogeneidad semántica es la similitud. En la Ciencia de la Información Geográfica (GIScience), la similitud juega un papel muy importante; es necesaria para resolver consultas vagas, consultas en lenguaje natural y es la base para la recuperacion e integracion semantica [Schwering, 2008]. En este sentido, es fundamental para el procesamiento semantico de los datos geoespaciales y se utiliza para medir el potencial de interoperabilidad semanti- ca entre representaciones múltiples [Stoimenov & Djordjevic-Kajan, 2002]. En general, la similitud entre las representaciones múltiples es la clave para hacer

realmente disponible el conocimiento de éstas.

Con base en lo anterior, en este trabajo se presenta una metodología basada en teoría de conjuntos para calcular la similitud entre representaciones múlti- ples. Actualmente, los trabajos dedicados a la similitud entre datos espaciales funcionan con un número reducido de instancias, por lo que el objetivo de este trabajo es establecer la similitud entre las representaciones; es decir, encontrar la similitud a nivel de representaciones (no solo a nivel de instancias) donde cada representacion cuente con un mayor número de instancias.

Dado que en el espacio geografico, la topología es considerada informacion de primera clase [Egenhofer & Mark, 1995], la metodología propuesta revisa las características espaciales de los datos. De acuerdo al tipo de dato se revisa su topología y/o propiedades métricas. Cuando está disponible, se da prioridad a la topología ya que las propiedades métricas (como distancia o forma) son usa- das como refinamiento que es frecuentemente capturado con menor precision [Egenhofer & Mark, 1995].

La razon de considerar las características espaciales de los datos es que a nuestro criterio es una tarea necesaria. Actualmente existen propuestas para integrar o compartir conocimiento que se apoyan en utilizar los metadatos de la informa- cion [Budak et al., 2006; Fonseca et al., 2006; Widom, 2005], pero a nuestro parecer los metadatos no bastan para una integracion, en particular una inte- gracion espacial ya que son un tanto generales y creemos que con un enfoque basado únicamente en metadatos, se pierde informacion de los detalles. Así, to- mamos en cuenta las características espaciales de los datos, ya que juegan un papel fundamental en la descripcion semantica [Schwering, 2008].

Asimismo, se propone una forma grafica para reportar la similitud. En este caso, se muestran los resultados numéricos (texto) y a partir de ellos se genera un grafico analogo que hemos denominado representación esquemática. El sistema gana expresividad con la representación esquemática ; mientras que con el texto se conserva la especificidad. La razon de presentar los resultados de esta forma es ofrecer al usuario un modo amigable de interpretar y analizar los resultados. Al ser éstos una expresion multi-modal (grafico y texto), se consigue representar la informacion con el enfoque y especificidad adecuados.

1.1. Descripción del problema

Actualmente el problema de la similitud entre datos espaciales se encuentra muy restringido a una serie de parametros o se realiza en escenas espaciales con un pequeño número de objetos geograficos. En este trabajo proponemos una me- todología que ayude a establecer la similitud entre representaciones con mayor nu ´mero de instancias y buscamos que sea aplicable a diversos casos; esto es, que los parametros para su implementacion sean mas flexibles.

El analisis de la similitud no es sencillo, y dada la complejidad espacial, es particularmente complicado para el espacio geografico. Para una adecuada identificacion de la similitud entre datos geograficos, el problema se debe abordar desde un punto de vista semantico y espacial. El enfoque semantico busca entender la informacion de los datos y así poder compararlos adecuadamente. El enfoque espacial se refiere a que dada la naturaleza de la representacion, es necesario contar con el tratamiento particular que necesita la informacion geografica.

En el analisis de la similitud se debe, como primer paso, entender la semantica de los datos para saber qué se va a comparar, esto es, alinearlos. La alineacion identifica qué atributos entre las representaciones describen a los conceptos que queremos comparar.

Una vez realizada la alineacion, se revisa la similitud para la parte tematica y espacial de los datos. El analisis tematico identifica instancias consistentes, esto es, nos ayuda a conocer qué instancias se van a comparar en su compo- nente espacial. Su funcion es similar a la alineacion, pero en lugar de identificar conceptos, identifica las instancias consistentes en cada representacion. De este modo, la similitud tematica puede pensarse como un criterio de comparacion para la similitud espacial. Una vez identificadas las instancias consistentes, solo a ellas se les revisa su similitud espacial.

El analisis espacial requiere cuidado particular debido a la complejidad del espacio geografico, en específico por el tipo de relaciones existentes entre los datos. Las relaciones espaciales se definen de acuerdo al tipo de dato. En este trabajo se consideran únicamente representaciones con tipo de dato punto y area.

Al finalizar el analisis de la similitud se deben reportar los resultados al usuario de un modo grafico, buscando que éstos sean precisos y faciles de entender.

1.2. Hipótesis

Las hipotesis de investigacion relacionadas con este trabajo de investigacion se definen a continuacion:

> El problema de la similitud se puede abordar con un enfoque de teoría de conjuntos. Si las representaciones se manejan como conjuntos, es posible procesarlas buscando tanto similitudes como diferencias y así obtener la informacion necesaria para establecer su similitud.
>
> Si se cuenta con una aplicacion que calcula la similitud entre las representaciones espaciales, entonces la similitud se convierte en un atributo mas tangible y practico que puede utilizarse para un desarrollo posterior.
>
> A partir de la similitud se puede proponer un grado de interoperabilidad entre las representaciones.

Los resultados del analisis de similitud se pueden presentar de forma grafica. Estos graficos, llamados representaciones esquemáticas, se encargan de capturar algunas características propias de cada representacion, así como la relacion de similitud entre las representaciones, lo cual hace el resultado mas descriptivo y facil de entender.

1.3. Justificación

Actualmente los sistemas para el analisis de la similitud de los datos espaciales funcionan solo para un reducido número de instancias y bajo escenarios muy controlados.

Dado que la similitud es un proceso necesario para la integracion de los datos, su intercambio, alineacion, organizacion; un sistema que establezca la similitud entre representaciones espaciales es de gran utilidad para el desarrollo subsecuente en el area de la Geomatica.

1.4. Ob jetivo general

Diseñar e implementar una metodología para valorar la similitud entre representaciones múltiples de datos geograficos, considerando el número de instancias, aspectos topologicos y propiedades tematicas. Con lo cual se establezca una m étrica denominada grado de interoperabilidad semántica para describir la simi- litud entre las representaciones de datos geograficos, favoreciendo la integracion de diferentes fuentes.

1.5. Ob jetivos particulares

Valorar la similitud entre representaciones espaciales con un gran número de instancias.

Diseñar e implementar un grado de interoperabilidad basado en la simili- tud entre las representaciones.

Diseñar e implementar un grafico que capture la similitud y despliegue los resultados de un modo grafico, claro y preciso.

Valorar la similitud entre representaciones de localidades de la República Mexicana.

1.6. Aportaciones

Se encuentra la similitud y se genera un grado de interoperabilidad entre representaciones espaciales de localidades de la República Mexicana.

Se identifica la similitud semantica entre representaciones. Esto tanto a nivel de instancias como de representaciones.

Se propone un grado de interoperabilidad semántica.

La relacion de similitud se representa por medio de un grafico que captura tanto la asimetría de la relacion como otras características propias de las representaciones.

1.6.1. Publicaciones

Artículo: Evaluacion de la integracion semantica en representaciones multiescala.

Presentado en el Congreso Internacional sobre Innovacion y Desarrollo Tecnologico 2010.

Artículo: Evaluacion de la integracion semantica en representaciones multiescala.

Presentado en la Conferencia Iberoamericana en Sistemas de Informacion Geografica 2011.

1.7. Limitaciones

El criterio de comparacion del sistema es la parte que decide qué instancias se van a comparar y funciona de la siguiente manera: revisa las propiedades tematicas de las instancias y halla la correspondencia entre representaciones; una vez establecido el criterio de comparacion, se revisa la consistencia espacial y se determina la similitud (encontrando si las representaciones hablan sobre los mismos objetos geograficos). La limitacion radica en que no se puede establecer primero un criterio de comparacion espacial y después revisar las propiedades tematicas.

El sistema trabaja solo con datos tipo punto o area y entre ellos establece las relaciones topologicas y métricas que considera. La limitacion del sistema es que no puede trabajar con datos de tipo línea y que tiene definidas u´nicamente cuatro relaciones espaciales entre datos de tipo area (descritas en la seccion Consistencia espacial de la Metodología).

Usuario experto. Los datos se introducen al sistema de forma manual y se puede seleccionar una ponderacion sobre las relaciones espaciales; esto requiere la operacion de un usuario que tenga conocimiento en el area. Este requerimiento es mas marcado dado que la importacion implica una alineacion entre conceptos.

Archivos de entrada. Actualmente el sistema solo contempla representaciones en formato shapefile cuando cualquier representacion que pueda importarse a PostgreSQL debe de funcionar.

1.8. Organización de la tesis

El resto de la tesis se encuentra organizada de la siguiente manera, el capítulo 2 trata sobre el estado del arte, en él se presentan los trabajos que sirven como base para esta investigacion. En el capítulo 3 se presenta la metodología que proponemos para valorar la similitud entre representaciones y se divide principalmente en adecuacion de datos, analisis de la similitud y la visualizacion de resultados. El capítulo 4 trata sobre las pruebas que se realizaron, los resultados que se obtuvieron y su implicacion en modificaciones al sistema. El capítulo final trata sobre las conclusiones del trabajo así como la propuesta de investigacion futura.

2. Estado del arte

En este capítulo se presentan trabajos relacionados con la representacion, integracion y comparacion de datos. Los trabajos se dividen en cuatro secciones. Primero se habla de representaciones múltiples, una forma de representar el conocimiento de un mismo fenomeno geografico con diferentes puntos de vista. Después se habla de ontologías, se describe como sirven para capturar la semantica de los datos e integrarlos. A continuacion se habla sobre integracion de datos, se presenta la teoría existente, algunos procedimientos y ejemplos. Por último se habla sobre similitud semantica. Se explica que es la característica fundamental para la integracion de informacion y se describen algunos modelos para su analisis. Al finalizar el capítulo se presenta un ejemplo de la integracion de datos utilizando similitud.

2.1. Representaciones múltiples

Mientras el mundo real se asume único, el modo de representarlo depende del uso que se le dará a la informacion. Así, diferentes aplicaciones que tienen interés en el mismo fenomeno del mundo real podran tener diferentes percepciones y requerir diferentes representaciones, a estas representaciones se les conoce como representaciones múltiples. Por ejemplo, esto ocurre cuando los productores de mapas producen mapas sobre la misma region geografica a diferente nivel de detalle.

Cada representacion es una descripcion de la realidad que se materializa en un nivel de detalle que contiene las características espaciales y tematicas del fenomeno [Vangenot et al., 2002]. Estas características son definidas por las necesidades de la aplicacion y difieren por diversas causas [Sheeren et al., 2009]:

> Las bases de datos tienen diferentes propositos y por lo tanto diferente contenido, organizacion y granularidad;
>
> la informacion se produce de diferentes métodos y a partir de diferentes fuentes: las fuentes, incluso siento idénticas, pudieron interpretarse y digitalizarse de manera diferente;
>
> el mundo geografico se encuentra en constante evolucion, así una representacion solo captura la realidad en un momento dado;
>
> la captura de datos es un proceso complejo, así, una base de datos puede contener errores

Dicho de forma general, los usuarios no comparten el mismo punto de vista. Como punto de vista nos referimos a necesidades e intereses específicos sobre un fenomeno del mundo real.

Una base de datos de representaciones múltiples o MRDB es una base de datos espacial que se usa para almacenar el mismo fenomeno de la realidad a diferentes niveles de precision, exactitud y resolucion [Weibel & Dutton, 1999]. En una MRDB, pueden almacenarse y relacionarse entre sí las diferentes representaciones de un mismo fenomeno [Anders & Bobrich, 2004]. La diversidad de representaciones puede surgir de las diversos modos de ver el mundo, de diversas aplicaciones, o de diversas resoluciones. Esto lleva a diferencias en las representaciones tanto en la semantica como en la geometría [Bédard & Bernier, 2002].

El nivel de detalle espacial o resolucion espacial determina el aspecto geométri- co del fenomeno, esto es, de acuerdo al punto de vista, el aspecto geométrico de un mismo fenomeno puede cambiar en representaciones múltiples [Vangenot et al., 2002]. Considerando que los datos geograficos tienen dos componentes: atributos espaciales y atributos no-espaciales [Fonseca & Egenhofer, 1999], la informacion en representaciones múltiples variará tanto en la componente espa- cial como en la no-espacial (componente tematica), incluso cuando se habla del mismo fenomeno.

Actualmente, es muy difícil tener un GIS que soporte representaciones múlti- ples. Por lo general se tienen varias bases de datos, una por escala, y entre ellas no existe propagacion de actualizaciones. Sin embargo la coexistencia de representaciones múltiples del mismo fenomeno se ha vuelto usual. Esta observacion general es particularmente cierta para el campo de informacion geografica. Diferentes motivos puede explicar esto. por ejemplo, la obtencion de datos geograficos se ha vuelto mas sencilla gracias a la evolucion de técnicas y herramientas: GPS, imagenes digitales de alta resolucion, correlacion automatica en fotometría, analisis de imagenes automatizado en percepcion remota, entre otras. Esta diversidad también origina la necesidad de manipular los diferen- tes puntos de vista (tanto tematicos como topologicos) [Sheeren et al., 2009]. En general, la administracion de representaciones múltiples es actualmente uno de los requerimientos de los modelos de datos espaciales que tiene poco apoyo [Vangenot et al., 2002].

Una de las razones que hacen particularmente difícil el soporte de representaciones múltiples es que en ellas existe heterogeneidad semántica. La heterogeneidad semántica es el desacuerdo sobre significado, interpretacion, uso [Sheth & Larson, 1990] y aspecto geométrico de los datos. Cuando existe heterogeneidad semántica en las representaciones múltiples, el proceso de actualizacion se hace mas complicado dado que se debe revisar el correcto significado de la actualizacion.

2.2. Ontologías

Cada base de datos espacial refleja una conceptualizacion particular del mundo. Entender la heterogeneidad semántica de diferentes representaciones geografi- cas es la clave para establecer su similitud. Hay un acuerdo en este aspecto: la

semantica de la informacion sería mucho mas útil si ésta se representa de forma explícita. Y las ontologías son una forma de hacerlo [Mustière et al., 2009].

Un rapido acceso y una interpretacion inteligente de diferentes tipos de datos espaciales requiere de integrar y compartir informacion eficientemente entre sistemas y diseños heterogéneos. Con las ontologías se aborda un punto importante en la investigacion sobre metadatos y semántica que busca proveer una correspondencia entre esquemas e instancias, y proveer un acceso unificado a la informacion [Budak et al., 2006]. Sin embargo como nuevo reto se requiere ir mas allá de un enfoque tematico e incluir ahora dimensiones de tiempo y espacio.

Gruber define que "una ontología es una especificacion explícita de una conceptualizacion" [Gruber, 1992]. Guarino dice que "una ontología es una teoría logica que considera la intencion de un vocabulario formal, esto es, la teoría considera un compromiso con una conceptualizacion particular del mundo. Una ontología refleja indirectamente su compromiso y su conceptualizacion subyacente" [Guarino, 1998]. Smith, mas enfocado en sistemas de informacion, dice que una ontología puede verse como un diccionario de términos o un vocabulario común compartido por diferentes comunidades de sistemas de informacion [Smith, 2003].

Fonseca las define informalmente como "acuerdos sobre conceptualizaciones compartidas" [Fonseca et al., 2006]. Wiederhold dice que los acuerdos son representados como ontologías, una por area de estudio [Wiederhold, 1994]. Las ontologías contienen objetos, propiedades de los objetos y posibles relaciones entre los objetos de un dominio específico del conocimiento [Chandrasekaran et al. 1999].

La definicion que mas se acopla a este trabajo es la que define Fonseca como: las ontologías capturan la semantica de la informacion, pueden ser representadas en un lenguaje formal y pueden usarse para almacenar metadatos relacionados que permitan un enfoque semantico para la integracion de informacion [Fonseca et al., 2006].

2.2.1. Geo-ontologías

Una geo-ontología u ontología geoespacial, ademas de las características usuales de una ontología común, provee informacion sobre las entidades geográficas. Sin embargo, estas entidades no estan únicamente ubicadas en el espacio, estan atadas intrínsecamente al espacio geografico. Toman del espacio algunas de sus características estructurales como propiedades mereologicas, topologicas o geométricas [Smith & Mark, 1998]. Una geoontología es diferente al resto de las ontologías principalmente porque las relaciones topologicas juegan un rol principal en el dominio geografico. La tecnología semantica juega un papel principal en la interpretacion espacial de las fuentes. Con ella, los métodos para el

analisis espacial tratan de incorporar consideraciones del contexto para examinar asociaciones y relaciones espaciales [Budak et al., 2006].

Los esfuerzos actuales para integrar informacion geografica adoptan la idea de usar metadatos estandarizados como una parte clave del analisis de la informacion. El analisis tradicional de informacion usa frecuentemente un enfoque u ´nicamente cuantitativo para presentar e inferir relaciones tematicas entre entidades. Este enfoque tiene serios defectos al trabajar con informacion cualita- tiva, espacial o temporal ya que es frecuentemente imprecisa. El trabajar con la semantica de los datos ya sea para integrar, compartir o analizar informacion geografica permite una interpretacion mas significativa. La semantica espacial puede manejarse mediante el uso de geo-ontologías ya que permiten un sofisticado analisis de informacion de diferentes dominios [Budak et al., 2006].

A las entidades geograficas, Couclelis y Goodchild las conceptualizan en dos dimensiones: campo y objeto [Couclelis, 1992; Goodchild, 1992]. La dimension de campo considera que los datos espaciales son un conjunto de distribuciones continuas. La dimension de objeto ve el mundo como formado por objetos discretos [Fonseca et al., 2006].

[Fonseca propone dividen los conceptos dentro de una ontología en dos grupos: (a) conceptos que corresponden a un fenomeno físico en el mundo real (variacio- nes sobre la superficie de la tierra) y (b) conceptos que corresponden a atributos del mundo creados por el hombre para representar conceptos sociales o institucionales. Al primer grupo de conceptos los llaman conceptos físicos y al segundo grupo conceptos sociales [Fonseca et al., 2006].

En [Frank, 2001] se describe a las geoontologías y se dividen sus conceptos en capas. La capa 0 es ocupada para campos reales continuos en el espacio-tiempo. La capa 1 está dedicada a mediciones de los campos reales. La capa 2 es para objetos creados por humanos que fueron derivados de las mediciones. La capa 3 es para objetos sociales, construidos por acuerdos de las personas. Y por último, la capa 4, es para conceptos subjetivos sobre el espacio.

2.3. Integración de datos

La integracion de datos es el problema de combinar datos residentes en diferentes fuentes y proveer al usuario de un modo unificado de ver los datos [Lenzerini, 2002]. Ademas, es un proceso fundamental para los sistemas que manejan múltiples fuentes autonomas y heterogéneas [Halevy et al., 2006].

La forma de la organizacion de datos se denomina esquema. En la integracion de datos lo mas común es que los datos a integrar tengan esquemas diferentes. Para que la informacion de las fuentes sea integrada, se debe resolver que la informa- cion relacionada está almacenada en diferentes esquemas, esto se conoce como heterogeneidad de esquemas semanticos. Para la integracion entre esquemas es

necesario generar un esquema global que incluya el a los diferentes esquemas de las fuentes, para identificar las correspondencias entre los esquemas locales y el global con el fin de permitir que el sistema traduzca las consultas para cada esquema [Pottinger, 2008].

Los componentes principales de un sistema de integracion de datos son: el esquema global, las fuentes y sus correspondencias. Las fuentes contienen los datos reales mientras que el esquema global provee una vista integrada y virtual de las fuentes incluidas. Antes de que los esquemas puedan integrarse, las similitudes entre atributos deben ser identificadas, se deben conocer las correspondencias que describan las relaciones semanticas entre el esquema global y los esquemas de las fuentes [Havely, 2006]. Estas correspondencias son el componente principal de la descripcion de las fuentes. Esto es, modelar las relaciones entre las fuentes y el esquema global.

Para el modelado de estas relaciones existen dos enfoques. El primero, llamado Global-as-View (GAV), requiere que el esquema global sea expresado en términos de las fuentes de datos. El segundo enfoque, llamado Local-as-View (LAV), requiere que el esquema global sea especificado de manera independiente a las fuentes, y que las relaciones entre el esquema global y las fuentes se establezca definiendo cada fuente como una vista del esquema global [Lenzerini, 2002].

Lenzerini dice que independientemente del método empleado para especificar las relaciones entre el esquema global y las fuentes, un servicio fundamental que debe contemplar un sistema de integracion de informacion, es la solucion a las consultas en términos del enfoque global. Esto es, la consulta al esquema global debe ser replanteada para cada una de las fuentes, considerando la integracion de datos en la respuesta y la solucion de la consulta con informacion incompleta.

Dado que las fuentes son generalmente autonomas, surgen problemas de fuentes con datos inconsistentes entre sí y a pesar de lo complicado de la tarea, lo que se espera de una computadora es que tome decisiones lo mas acertadas posibles. Este problema se trata con procedimientos de transformaciones y filtro de la informacion obtenida de las fuentes [Lenzerini, 2002]. Los datos que se deben considerar para la integracion de esquemas son: el nombre del esquema, el tipo de dato que maneja, su estructura y el formato de las instancias [Pottinger, 2008].

De forma general, las consultas en LAV son un procedimiento complicado. En LAV el esquema global se comunica con las fuentes por medio de vistas, o sea, el enfoque no cuenta con la informacion suficiente para saber como adaptar la consulta, por lo que se debe especificar de forma manual cada una de las fuentes de datos. Por otro lado, en el enfoque GAV, dado que el esquema global está especificado con base en cada uno de las fuentes es mas sencillo adaptar las consultas [Lenzerini, 2002], pero como desventaja, si se quiere agregar una nueva fuente de datos, el esquema global se tiene que replantear.

Dado que estos enfoques de integracion de la informacion estan basados en logica de primer orden, no es posible manejar inconsistencias entre las fuentes de datos. La consistencia se refiere a la falta de contradicciones logicas dentro del modelo o representacion [Egenhofer et al, 1994]. Esto es, si en el sistema de integracion de datos se presentan datos que no cumplen con las mismas reglas del sistema, algunos datos perderan sentido. Para manejar las inconsistencias entre las fuen- tes, en la practica generalmente se realizan procedimientos de transformacion o filtrado de datos para alcanzar sentido en los datos [Lenzerini, 2002]. Para conseguir sistemas interoperables es necesario que los sistemas sean capaces de valorar la certidumbre y consistencia de los datos en las diferentes fuentes. Si esto no es posible, los sistemas deberan preguntar al usuario cual de las dos fuentes es mas confiable [Halevy et al., 2006].

En [Taylor & Ives, 2006] se presenta un enfoque que propone un sistema en donde sus componentes se actualizan independiente uno de otro. Se presenta una sobreposicion de informacion (overlapping) o redundancia, para esto se im- plementa un algoritmo de reconciliacion para saber qué cambios hubo desde la u ´ltima actualizacion. Con esto se verifica si los nuevos cambios pueden llevar- se a cabo en cada una de las partes del sistema sin producir inconsistencias. De esta manera se permite cierta independencia entre las fuentes, actualizando los campos consistentes entre ellas y permitiendo la existencia de campos locales.

[Widom, 2005] desarrolló el proyecto Trio que contempla la precision y el linaje tanto en los datos como en consultas. Debido a diferentes métodos de obtencion de datos, algunos elementos de una base de datos pueden tener un cierto nivel de certidumbre o confiabilidad. Saber como se obtuvo un dato, si fue derivado de otro (probablemente inexacto), puede ser un factor importante, a veces tan importante como el dato mismo [Widom, 2005]. El proyecto Trio utiliza tres componentes: datos, precisión y linaje. Esto permite que los datos puedan ser inexactos y manejarse con un rango. También pueden contar con un registro que indique su precision. Estas características se relacionan con areas de tra- bajo llamadas informacion sobrepuesta y manipulacion de las anotaciones que estan enfocadas en las de los metadatos. Estos metadatos ayudan a organizar, acceder, conectar y reutilizar la informacion proveniente de diferentes fuentes. Dentro de los metadatos debe haber, por ejemplo, informacion sobre precision y linaje.

De forma general, la mayoría de los algoritmos de integracion de esquemas tratan al esquema como un grafo y para la asociacion de conceptos consideran prin- cipalmente la estructura esquematica e instancias. Tres ejemplos de algoritmos para integracion esquematica son:

LSD: es un algoritmo que aprende de asociaciones iniciales que el usuario establece y a partir de ellas trata de encontrar el resto [Mitchell, 1997].

Similarity Flooding: donde la primera asociacion se establece entre los

conceptos A y B al comparar nombres, tipos y datos. Después revisan los conceptos relacionados con A y B, dentro de sus esquemas respectivos, y se buscan nuevas asociaciones hasta agotar las posibilidades [Melnik et al., 2002].

Clio: Recibe una serie de asociaciones ya establecidas entre los esquemas y, usando la informacion del esquema y datos, se busca el modo de traducir instancias entre sí. Su resultado es una consulta en SQL que puede traducir datos de una fuente a otra [Pottinger, 2008].

La integracion de datos es un campo tan rico que aún existen problemas, como la necesidad de saber qué valores en diferentes fuentes se refieren a los mismos objetos en el mundo real. Según Lenzerini, en el tema de la integracion de informacion quedan aún abiertos muchos temas para investigar. Queda por investigar a mayor detalle las relaciones entre los enfoques LAV y GAV, investigar algorit- mos para la solucion de consultas, el tratamiento de fuentes inconsistentes y el razonamiento de sus consultas. Ademas, falta por investigar mejores formas de crear un esquema global, cómo manejar las posibles limitaciones para acceder a la informacion, como incorporar la nocion de calidad y filtros de datos en un esquema de integracion, como establecer reglas para encontrar las relaciones entre las fuentes de forma automatica, entre muchos otras tareas.

Finalmente, vale la pena recordar que la integracion de datos es hoy una necesidad, debido a la gran cantidad de datos disponibles. La informacion debe ser compartida de manera apropiada entre diferentes fuentes y los individuos deben poder encontrar la informacion correcta en el momento correcto sin importar donde se encuentre.

2.3.1. Interoperabilidad

El Open Geospatial Consortium (OGC) define interoperabilidad como “la capacidad para comunicar, ejecutar programas, o transferir datos entre varias unidades funcionales de una manera que requiere poco conocimiento del usuario sobre características particulares de dichas unidades” [OpenGIS, 1996].

La interoperabilidad apoya la comunicacion y acceso entre repositorios de datos, y da al usuario una oportunidad de encontrar informacion complementaria sobre el mismo fenomeno (o fenomenos relacionados) desde diferentes fuentes que se han creado de manera independiente [Vangenot et al., 2002]. La interoperabilidad es una propiedad parecida y relacionada a la integracion de datos.

Para lograr interoperabilidad en un sistema es necesario encontrar la correspondencia entre los conceptos de las fuentes de datos [Fonseca et al., 2006]. O como Vangenot propone, una solucion para alcanzar la interoperabilidad es identificar el conocimiento relacionado y proveer un mecanismo que establezca la relacion entre las diferentes representaciones [Vangenot et al., 2002].

La interoperabilidad en la informacion geografica ha ganado importancia debido a las nuevas posibilidades que surgen del mundo interconectado y de las disponibilidad de la informacion geografica [Goodchild et al., 1999]. Es necesario que se encuentren métodos innovadores para encontrar sentido y relacion entre toda la informacion disponible hoy en día [Fonseca et al., 2006].

En el area de Sistemas de Informacion Geografica, diversos autores han propuesto modelos basados en ontologías para alcanzar la interoperabilidad. El uso de ontologías mejora la interoperabilidad entre diferentes sistemas de informacion en general [Mena et al., 1996] y de manera específica en sistemas de informa- cion geografica [Fonseca & Egenhofer, 1999]. El uso de ontologías para modelar entidades geograficas busca capturar las conceptualizaciones compartidas por comunidades específicas de usuarios y así mejorar la interoperabilidad entre diferentes bases de datos geograficas [Smith & Mark, 1998].

Por otro lado, existe mucho desarrollo en el area de ontologías y hay un gran esfuerzo enfocado en usarlas como herramientas de interoperabilidad. Se cree que el soporte y uso de múltiples ontologías debe ser una característica basica de los sistemas de informacion modernos, si quieren poder manejar adecuadamente la semantica para integrar la informacion [Fonseca et al. 2006].

Como herramientas de integracion, las ontologías se usan para representar explícitamente la semantica de la informacion. Una vez que se cuenta con una ontología por cada representacion, estas deben alinearse y compararse. Se debe encontrar una alineacion entre las dos ontologías que exprese la correspondencia entre sus entidades. De acuerdo a su tipo de procesamiento se distinguen algunas técnicas de alineacion [Mustière et al., 2009]:

Terminologicas, ocupan las cadenas en las entidades ontologicas;

Estructurales, trabajando con la estructura ontologica; Extensionales,

su procesamiento está basado en las instancias;

Semanticas, ocupan el modelo relacionado a cada representacion.

En [Kashyap & Sheth, 1996] consideran que compartir una misma ontología es un requisito previo para compartir e integrar la informacion. Opinan que debe existir una ontología que capture el compromiso en común de las fuentes de datos y en caso de que no exista una ontología compartida, una solucion común es derivar una ontología global a partir de la informacion que se tenga sobre los esquemas a integrar. [Bergamaschi et al., 1998] implementaron esta solucion, mientras que Rodríguez et al. [Rodríguez et al., 1999] resuelven el problema de interoperabilidad utilizando un proceso que revisa la consistencia de atributos y la distancia semantica.

Un ejemplo de interoperabilidad con ontologías es el mapeo de ontologías. El mapeo de ontologías (ontology mapping) es el nombre que se le da a relacionar el vocabulario de dos ontologías que comparten el dominio en un modo en que la estructura matematica de la ontología, sus axiomas e interpretaciones sean respetados [Kalfoglou & Schorlemmer, 2003]. Un proceso conocido como intermapeo de ontologías (interontology mapping) es donde el usuario define las reglas para generar mapeos específicos entre dos ontologías [Wache et al, 2001], pero a pesar de que es un proceso muy flexible pocas veces preserva la semantica.

Fonseca, Cmara y Monteiro desarrollaron un marco de trabajo para medir la interoperabilidad de geoontologías donde asumen que los conceptos en una ontología estan separados de las instancias en un base de datos [Fonseca et al. 2006]. Esto porque consideran su problema de interoperabilidad como un caso de la administracion de modelos de Bernstein [Bernstein, 2003]. Así, solo analizan conceptos en la ontología. Sin embargo afirman que es importante extender la investigacion a los datos mismos, a las instancias.

De forma general, para crear sistemas interoperables con ontologías, se realiza primero la alineacion de ontologías y se revisan las características tematicas de cada representacion, sus diferentes enfoques, su nivel de detalle, etc. Para esto existen algunos métodos como la particion de ontologías, evaluacion de ontologías, aptitud de uso, visualizacion de ontologías o se usan analogías entre mundos conceptuales y físicos [Mustière et al., 2009].

Actualmente, se ha hecho gran avance en el estudio de la interoperabilidad, principalmente en cuestiones de interoperabilidad sintactica (tipos de datos y formatos) y estructural (integracion de esquemas, interfaces). Sin embargo, la interoperabilidad es aún un gran problema para la siguiente generacion de sistemas de informacion [Sheth, 1998].

2.4. Similitud

Los humanos usamos la similitud para almacenar y recuperar informacion, para comparar nuevas situaciones con experiencias anteriores. En la Ciencia de Informacion Geografica (GIScience), la similitud juega un papel importante en muchas aplicaciones como sistemas de toma de decisiones, minería de datos o reconocimiento de patrones.

La similitud es necesaria para resolver consultas vagas, conceptos vagos o consultas en lenguaje natural y es la base para la recuperacion e integracion semantica. Mientras que las computadoras pueden procesar una decision de equivalencia o no-equivalencia binaria de modo muy rapido, el procesamiento de la similitud es un problema complejo y no trivial [Schwering, 2008]. Es fundamental para el funcionamiento de procesamiento semantico de datos geoespaciales y se usa para medir el potencial de la interoperabilidad semantica entre datos de diferentes sistemas de informacion geografica.

En los sistemas de informacion, la semantica se refiere al contenido y representacion de entidades o conceptos del mundo real [Meersman, 1997]. El problema de la interoperabilidad semantica es la identificacion de objetos semanticamente similares en diferentes fuentes de datos [Kashyap & Sheth, 1996]. Estudios han sugerido el uso de ontologías como un marco de trabajo para la deteccion de la similitud semantica [Bishr, 1997]. Un posible enfoque es crear un base de conocimiento en términos de una ontología común, sobre la cual sea posible detectar la similitud semantica y definir las relaciones entre los conceptos [Lenat & Guham, 1990; Kahng & McLeod, 1998].

Existen dos conceptos principales en las medidas de similitud semantica: similitudes - diferencias y la distancia semantica (Figura 2.1).

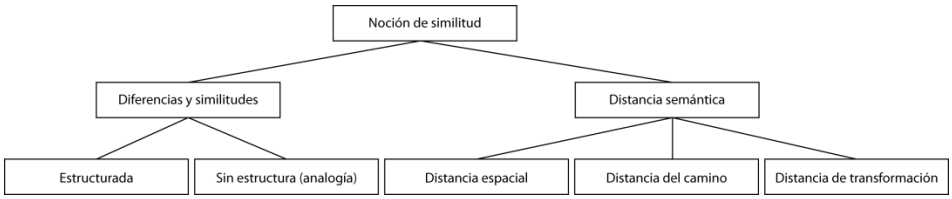


Figura 2.1: Diferentes nociones de similitud.

Mientras algunas definiciones de similitud se basan en una comparacion sin estructura, otras lo hacen de forma estructurada, de modo que los elementos en común debe tener un rol analogo en la representacion para incrementar las similitudes entre dos conceptos [Schwering, 2008].

Para aplicar la distancia semantica es necesario un marco de trabajo común con una métrica específica. Algunas medidas de similitud usan un espacio multidimensional como marco de trabajo y métricas euclidianas o “city-bloc” como m étricas para medicion de distancias. La distancia semantica en una estructura arbol o de red está definida por el camino mas corto entre nodos.

En el area de GIS la similitud semantica se realiza trabajando con objetos y conceptos geoespaciales. Un concepto es una idea que caracteriza un conjunto de objetos. [Sloman et al., 1998]. Un objeto geoespacial se refiere a una instan- cia ubicada en el espacio geografico. Las medidas de similitud deben trabajar con objetos y conceptos para medir la interoperabilidad entre fuentes de datos geoespaciales.

La semantica de objetos geoespaciales es compleja y tiene algunas características particulares: es descrita por propiedades como forma, tamaño y ubicacion. Ademas que las relaciones, y en particular las relaciones espaciales juegan un papel principal en la descripcion semantica [Schwering, 2008].

A pesar que se ha realizado investigacion en la sicología para entender el proceso cognitivo, no existe una teoría común en la medicion de la similitud semantica. A continuacion, con base en [Schwering, 2008], se presenta un analisis de diferentes modelos para mediciones de similitud semantica en el contexto de la GIScience, en cada uno se analiza su forma de representar el conocimiento, su medicion de la similitud, sus propiedades métricas, sus requerimientos y suposiciones, se muestran algunos ejemplos y finalmente se presenta su evaluacion para la medicion de similitud.

2.4.1. Modelo geométrico (geometric model)

Explota la analogía con el espacio para medir la similitud, los conceptos son modelados como puntos dentro de un espacio multidimensional y su distancia espacial indica la similitud semantica.

Representacion del conocimiento

Está basado en la nocion de espacios vectoriales multidimensionales. Cada dimension se usa para describir propiedades de objetos y conceptos. La mayoría de los modelos geométricos se enfocan en modelar únicamente objetos.

Medicion de la similitud

La distancia semantica es analoga a la distancia espacial. La similitud es medi- da como una funcion de la distancia espacial. Las mediciones mas comúnmente usadas son las de Minkowski [Schwering, 2008].

Propiedades métricas

Los modelos geométricos de similitud semantica deben cumplir con tres axiomas:

Minimalidad: Si la distancia espacial entre dos conceptos es cero, entonces

los conceptos son iguales.

Simetría: La distancia, y por lo tanto la similitud semantica, es la misma entre un concepto y otro.

Inequidad triangular: La distancia entre dos conceptos es siempre menor o igual que la distancia entre dos conceptos por medio de un tercer concepto. Este axioma no se cumple cuando las comparaciones se realizan en diferentes dimensiones (por ejemplo, balon-luna-lampara). Por lo mismo la similitud en dimensiones diferentes no es transitiva.

Requerimientos y suposiciones

Los modelos geométricos de similitud semantica deben asumir las siguientes suposiciones:

Independencia del elemento de representacion: se asume que las propieda- des son independientes entre sí.

Solubilidad: el conjunto de propiedades usadas para describir un concepto debe ser lo suficientemente rico y representativo para la conceptualizacion. Un conjunto de propiedades que no refleje la conceptualizacion humana no podrá dar buenos resultados de similitud.

Constante de los elementos representativos: la equivalencia de intervalos en las dimensiones debe preservarse a través de las dimensiones, esto es, las dimensiones debe estar normalizadas.

Complejidad de la representacion: al agregar mas informacion a la descrip- cion de la distancia semantica, ésta solo puede incrementarse y disminuir la similitud. Así, los modelos geométricos son adecuados solo para la com- paracion de conceptos con idéntico número de dimensiones.

Ejemplos del modelo geométrico

Las representaciones mas comunes son los espacios conceptuales [Gardenfors, 2000]. Los espacios conceptuales representan la informacion a un nivel concep- tual y estan formados por un conjunto de dimensiones de cualidad. Las dimensio- nes son conectadas a cualidades perceptibles por el sistema sensorial humano. Los objetos son representados como puntos y los conceptos como regiones n- dimensionales.

Evaluacion de la medición de similitud entre datos geoespaciales

Los objetos y conceptos son representados del mismo modo. La medicion de similitud solo se puede realizar entre puntos en el espacio conceptual. Por este motivo es difícil realizar mediciones entre conceptos (ya que involucran varios objetos), y se proponen diferentes soluciones, como reducir un concepto a un solo punto, o tomar el promedio de diferentes distancias.

Las propiedades son representadas como dimensiones cualitativas pero no hay una representacion de las relaciones entre conceptos.

2.4.2. Modelo de propiedades (feature model)

Como en el modelo geométrico, el modelo de propiedades también utiliza las propiedades para describir los conceptos. Mientras en el modelo geométrico las propiedades son dimensiones, en el modelo de propiedades son valores boolea- nos, (las propiedades o atributos) estan o no asociados a un concepto. Se basan en las suposicion de que dos conceptos con el mismo atributo son similares en algún aspecto. Las medidas de similitud de modelos de características hacen la consideracion que la similitud de conceptos se incrementa mientras mas atribu- tos tienen en común.

Representacion del conocimiento

El modelo de propiedades tiene una representacion del conocimiento con base en la teoría de conjuntos: los objetos y propiedades que tiene un determinado concepto son representados en conjuntos de atributos sin estructura. Al igual que las dimensiones en el modelo geométrico, las características pueden repre- sentar variables nominales, ordinales o intervalos.

Existen atributos de adicion o sustitucion. Para los atributos de adicion no es necesario revisar el resto de atributos, mientras que los de sustitucion tienen una permanencia en el conjunto restringida, esto es, un atributo sustituirá a otro y viceversa, pero no podran estar presentes los dos atributos en un mismo conjunto. Los atributos de sustitucion pueden verse como colecciones de atribu- tos, donde un objetos podrá tener solo un atributo de la coleccion.

Medicion de la similitud

Se realiza mediante un modelo de correspondencia de características. Este mo- delo establece que la similitud no es necesariamente métrica. Los conceptos son representados como colecciones de características. Al representar cada concep- to con diferentes conjuntos de características, se les puede aplicar operaciones elementales para estimar similitudes y diferencias.

[Tversky, 1977] propuso, con base en teoría de conjuntos, una medicion de la similitud entre conceptos a y b como una funcion de sus atributos en similares y discordes :

$$s(a, b) = F(A \cap B, A - B, B - A) \qquad (2.1)$$

Al basarse en logica de conjuntos, este modelo no soporta correspondencias par-

ciales.

Propiedades métricas

Este modelo es no-métrico. La funcion tiene como componentes la diferencia de A a B, la diferencia de B a A y su similitud, siendo A y B dos objetos a com- pararse. Y dependiendo del peso de estos componentes, la funcion de similitud F (Ecuacion 2.1) es no-métrica.

Tversky probó empíricamente que los tres axiomas métricos no son consistentes en el establecimiento de la similitud humana, dijo que: el axioma de minima- lidad es problematico, la simetría es aparentemente falsa y que la inequidad triangular apenas es convincente.

Requerimientos y suposiciones

Independencia entre los elementos: el grado en el que una característica compartida por dos conceptos afecta su similitud no debe depender de otra característica.

Solubilidad: el conjunto de características debe ser los suficiente rico y representativo. Esto es, si no se tiene un adecuado conjunto de carac- ter ísticas, los resultados seran también inadecuados.

Constante de los elementos representativos: los intervalos entre carac- terísticas se asumen equivalentes.

Ejemplos del modelo de propiedades

El modelo de propiedades mas conocido es el modelo de contraste y proporcion de Tversky, que se basa en la funcion de similitud como una version normalizada de la diferencia entre atributos similares y distintos.

La funcion de similitud puede determinarse por la cardinalidad de los conjuntos o con base en la importancia de algunos atributos.

El Matching distance similarity measurement (MDSM) de [Rodríguez & Egen- hofer, 2004] es un método para la medicion de la similitud entre conceptos geoespaciales que combina dos estrategias: la similitud de características (featu- re matching) y el calculo de la distancia semantica. Al tomar en consideracion las propiedades cognitivas, este modelo busca representar un método para la medicion del grado de interoperabilidad. En su propuesta, un proceso de simili- tud de propiedades y el calculo de la distancia semantica, provee una estrategia para crear un modelo que capture la similitud. En la practica confirman, como dice Tversky, que la similitud no siempre es simétrica [Tverski, 1977]. Aunque MDSM utiliza datos espaciales considerando la distancia semantica y la simili- tud de propiedades, no considera la componente espacial de los datos.

Evaluacion de la medición de similitud entre datos geoespaciales

Objetos y conceptos: Tversky aplicó este modelo solo para objetos.

Propiedades y relaciones espaciales: En un modelo de propiedades es impo- sible relacionar dos objetos de un modo estructurado. En estos casos la re- lacion es expresada en términos de atributos compuestos (por ejemplo, jun- to a). Pero hay que recordar que no se detectan correspondencias parciales, por lo que no habrá similitud entre junto a Río y junto a CuerpoDeAgua.

2.4.3. Modelo de red

Este modelo está basado en la teoría de grafos para evaluar la similitud y se basa en redes semanticas para la representacion del conocimiento.

Representacion del conocimiento

Las redes semanticas estan compuestas por nodos y aristas etiquetadas. Los no- dos representan unidades de conocimiento: objetos, conceptos o propiedades; las aristas conectan nodos y representan relaciones entre ellos de forma explícita. Una red semantica necesita una terminología estandar.

A pesar que los modelos de red siempre tienen la misma estructura, se diferencian en que dependiendo de su implementacion, algunos modelos permiten solo relaciones taxonomicas, otros de hiponimia y paronima, y otras permiten todo tipo de relaciones. Algunas redes semanticas restringen la direccion de la relacion y asignan pesos para modelar su importancia.

Medicion de la similitud

Como el modelo geométrico, los modelos de red basan su medicion de similitud en la nocion de distancia. Y es aquí donde se aplican algoritmos para grafos como el camino mas corto o algoritmos que miden la distancia de acuerdo a pesos.

Propiedades métricas

Las medidas de similitud en teorías basadas en grafos son métricas, si la distancia entre conceptos se mide sin importar la direccion de las aristas. Cuando se considera la direccion de las aristas el computo resulta en una similitud no- m étrica: la similitud es asimétrica, pero se cumplen los axiomas de inequidad triangular y de la similitud con uno mismo es igual a cero.

Requerimientos y suposiciones

Solubilidad: La similitud entre dos conceptos solo puede medirse si existe un camino entre ellos.

Constante de los elementos representativos: se asume que cada relacion es relevante y que tienen la misma influencia en el valor de la similitud.

Ejemplo del modelo de red

En [Rada et al. 1989] se propone una medida denominada Distance, la cual utiliza de relaciones taxonomicas y relaciones de asociacion. Mide la distancia entre dos nodos o conjuntos de nodos. Supone que el juicio humano de similitud es métrico y que la asimetría de la similitud entre conceptos no se deriva de la asimetría propia de la similitud sino de una asimetría entre conceptos, como por ejemplo una categorizacion difusa.

Por otra parte, [Resnik, 1995] propone una medicion de similitud basada en la nocion de contexto de informacion. En enfoques como Distance las relaciones representan distancias uniformes, lo que no es cierto en taxonomías reales. La medicion de Resnik con base en el contenido supera estas deficiencias. La probabilidad de un concepto se incrementa mientras mas alto está en la jerarquía. Su enfoque es simétrico y transitivo, pero no cumple que la similitud entre un concepto y sí mismo es cero, esto se cumple solo para el concepto en la cima de la jerarquía.

Evaluacion de la medición de similitud entre datos geoespaciales

Objetos y conceptos: este modelo fue pensado en modelar solo conceptos pero funcionan también con objetos. En cuanto a las propiedades y relaciones espaciales, la fortaleza del modelo está en la representacion de sus relaciones.

2.4.4. Modelo de alineacion

Como en el modelo de propiedades, el modelo de alineacion usa las similitudes y diferencias como nociones de similitud, pero contempla también la estructura relacional en donde las propiedades y relaciones se encuentran.

Representacion del conocimiento

Se realiza en una forma estructurada adoptando el marco de trabajo de alineacion estructural de Gentner [Gentner & Markman, 1995]. Los objetos son representados por sus propiedades incorporadas en un sistema de relaciones. Como parte central de este modelo hay relaciones de alineacion que indican la analogía estructural de dos elementos, propiedades o relaciones, pertenecientes a dos objetos diferentes.

Las relaciones de alineacion deben ser estructuralmente consistentes y sistematicas. Para ser estructuralmente consistentes deben cumplir que cada elemento corresponda cuando mas a un elemento (correspondencia uno a uno) y que los argumentos correspondientes a cada par de correspondencias tengan también correspondencia. La alineacion se llama sistematica si existe una estructura mas profunda de atributos y relaciones correspondientes.

Medicion de la similitud

Consiste en revisar la similitud entre las estructuras relacionales, mientras el modelo geométrico y de atributos revisa únicamente las similitudes de atribu- tos, el modelo de alineacion también considera si estos elementos se alinean o no. Así, las relaciones alineables incrementan la similitud mas que las relaciones no alineables.

Requerimientos y suposiciones

Independencia de los elementos de representacion.

Estructura homogénea: las ventajas de este modelo solo son evidentes cuando se trabaja en una estructura homogénea. Las reglas de alineacion necesitan una estructura uniforme para trabajar de manera automatica.

Solubilidad

Constante de los elementos representativos: este modelo asume que cada elemento tiene el mismo peso en la similitud.

Ejemplos del modelo de alineacion

El modelo Similitud como activacion y mapeo interactivo (SIAM) se utiliza pa- ra medir la similitud entre escenas espaciales, fue propuesto por Goldstone y Medin [Goldstone & Medin, 1994]. La similitud entre escenas espaciales se mide en un proceso de aprendizaje iterativo basado en redes neuronales. SIAM calcu- la todas las posibles alineaciones de relaciones y atributos de manera iterativa hasta que las alineaciones son consistentes y puede determinarse la similitud.

Evaluacion de la medición de similitud entre datos geoespaciales

Objetos y conceptos: SIAM fue definido para medir la similitud entre objetos en escenarios espaciales. La regla de correspondencia uno a uno es difícil de lograr en la similitud entre conceptos si son descritos con diferente nivel de granularidad. Se podría interpretar que un concepto corresponde a un elemento a mas detalle de otro concepto.

Propiedades y relaciones espaciales: SIAM soporta relaciones jerarquicas así como relaciones relaciones espaciales. Las propiedades son también consideradas en la medicion de similitud.

2.4.5. Modelo de transformacion

El computo de la similitud se realiza de manera diferente que en el resto de los modelos, aquí se definen las transformaciones necesarias para distorsionar un concepto en otro y la similitud está definida en términos del número de trans- formaciones necesarias para hacer los conceptos iguales.

Representacion del conocimiento

Para este modelo, se plasma en la representacion un conjunto de transformacio- nes que pueden ser aplicadas para transformar un concepto. Este conjunto de transformaciones depende de la naturaleza de los conceptos. A pesar de que las transformaciones no tienen que ser perceptivas, la identificacion de éstas no es

una tarea facil. Y para aplicar el modelo de transformacion se necesita que se defina antes un conjunto finito de transformaciones.

Medicion de la similitud

La base de la medicion de similitud es el número de transformaciones necesarias para convertir un concepto en otro. Se asume que la similitud decrece de forma monotonica cuando el número de transformaciones aumenta [Imai, 1977; Wiener-Ehrlich et al., 1980]. En [Hahn et al., 2003] se propone el uso de la complejidad de Kolmogorov para calcular la similitud.

Propiedades métricas

El modelo de transformacion es asimétrico pero los axiomas de minimalidad e inequidad triangular se conservan [Hahn & Chater, 1997]. La similitud entre un concepto y él mismo es maxima, ya que no se necesita de ninguna transforma- cion. Hanh y Chater dicen que la similitud es asimétrica ya que las transforma- ciones inversas pueden no ser de la misma complejidad.

Requerimientos y suposiciones

Solubilidad: El modelo debe contener al menos todas las transformaciones para llegar de un concepto a otro.

Comparacion de elementos de representacion: para considerar las transformaciones como un método para medir la similitud, cada transformacion debe afectar en mismo grado la similitud, esto es, cada transformacion debe ser de la misma complejidad. La complejidad de Kolmogorov hace posible la comparacion entre dos transformaciones de diferente compleji- dad.

Complejidad de la representacion: la transformacion de un concepto en otro de la misma granularidad requerirá una transformacion simple, mien- tras que transformaciones entre conceptos de diferente granularidad reque- riran transformaciones mas complejas.

Ejemplos del modelo de transformacion

Este modelo se ha aplicado principalmente en estímulos perceptivos, como cadenas alfabéticas [Wiener-Ehrlich et al. 1980], cadenas de alveolos [Imai 1977] o complejos geométricos [Hahn et al., 2003]. Las transformaciones se enfocan has- ta la fecha en atributos percetibles únicamente. Algunas de las transformaciones usadas para modificar las cadenas son espejo, reversa o agregar símbolos. Transformaciones para modificar el arreglo geométrico son rotacion, reflejo, traslacion y dilatacion [Goldstone & Son, 2005].

Evaluacion de la medición de similitud entre datos geoespaciales

Objetos y conceptos: Hasta ahora se ha aplicado a objetos simples y con una similitud perceptual, pero en teoría se puede aplicar a conceptos geoespaciales y con una similitud conceptual.

Propiedades y relaciones espaciales: Las transformaciones se han aplicado a conceptos y objetos únicamente, pero se puede aplicar también a relaciones. Para aplicar el modelo a relaciones es necesario definir el tipo de transformaciones posibles dentro del marco de trabajo.

En la Figura 2.2 se presenta una descripcion grafica de las propiedades de cada modelo. En la Tabla 2.1 se resumen las características de cada modelo, se señala su modelo matematico, sus elementos y estructura de representacion, su nocion de similitud y se indica si el modelo es aplicable a objetos o conceptos.

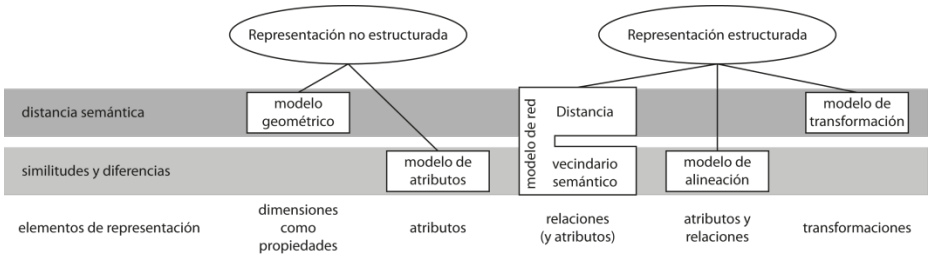


Figura 2.2: Diferentes nociones de similitud.

Tabla 2.1: Caracteristicas de los modelos de similitud

	Modelo geométrico	Modelo de atributos	Modelo de red	Modelo de alineación	Modelo de transformación
Modelo matemático - Espacio multidimensional - Teoría de conjuntos - Teoría de grafos - Redes neuronales - Transformaciones	X	X	X	X	X
Elementos de representación - Dimensiones - Atributos - Relaciones - Transformaciones	X	X	X	X X	X
Estructura de representación - Con estructura - Sin estructura	X	X	X	X	X
Aplicable a - Objetos - Conceptos	X (X)	X (X)	X X	X	X X
Noción de similitud - Distancia semántica como disimilitud - Similitud contra diferencias - Analogía estructural	X	X	X X	X X	X

2.4.6. Similitud en el contexto geográfico

[Schwering, 2008] dice que la seleccion de un modelo para medir la similitud siempre será subjetiva, pero una aproximacion es el analizar, por cada dominio y tarea, qué enfoque satisface las restricciones. Otra aproximacion es conside- rar el contexto, que muchos autores consideran fundamental para la similitud semantica [Tversky, 1977; Krumhansl, 1978] y para la interoperabilidad [Kash- yap & Sheth 1996; Bishr, 1997]. Fonseca y Li proponen un modelo llamado Topología-Distancia-Dirección (TDD) para evaluar la similitud [Li & Fonseca, 2006]. TDD es un modelo desarrollado en específico para datos geoespaciales, ya que toma en cuenta la topología, la direccion y la distancia. Adopta un enfoque de teoría de conjuntos (en lugar de una métrica) y toma en cuenta diferencias y semejanzas para establecer la similitud.

TDD modelo mide la similitud entre escenas espaciales. Antes de comenzar a revisar la similitud, su sistema realiza una alineacion espacial. La alineacion es- pacial identifica qué conceptos deben compararse, esto se hace definiendo una serie de transformaciones de objetos. Debido a la complejidad de alineacion en- tre escenas, su modelo funciona bien para escenas con hasta tres objetos .

Una vez realizada la alineacion, se mide la similitud topologica como funcion de una distancia dentro de una vecindad conceptual. La similitud direccional se evalúa por una funcion de distancia en un modelo direccional de 5 nodos. Para la similitud de la distancia métrica se emplea una distancia de cuatro granularidades donde establecen si los objetos estan lejos, a distancia media,

cercanos, o son iguales.

2.5. Fusión de objetos geográficos

Es una aplicacion de integracion de datos usando similitud. En GIS, la fusion de datos geograficos (geographic data conflation) se refiere a combinar informacion geografica de diferentes fuentes y obtener datos precisos, minimizar la redundancia y conciliar conflictos entre datos [Longley et al., 2001]. La necesidad de fusionar datos surge de actualizarlos para compensar la precision o compensar la carencia de atributos respecto nuevas fuentes con la misma cobertura. Ademas, se relaciona con la adecuacion de características y atributos de las fuentes de GIS adyacentes, eliminando discrepancias de posicion y atributos [Sharad & As- hok, 2008]. Los datos usados en la fusion son punto, línea, area, así como sus atributos.

En la Figura 2.3 se presenta un prototipo de un sistema de fusion de atributos. No es necesario que todos los sistemas de fusion presenten todos estos pasos, ya que pueden contar con menos o incluir algunos mas.

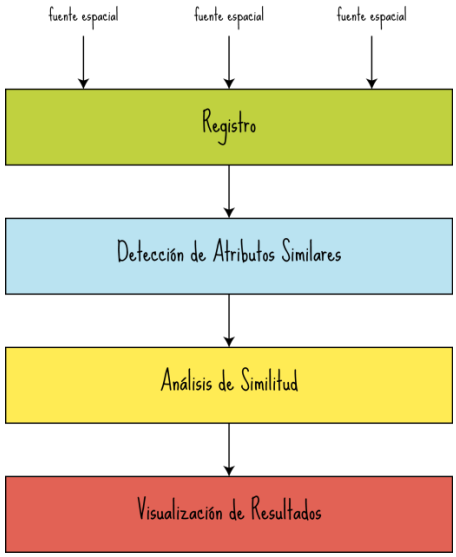


Figura 2.3: Prototipo de un sistema de fusion de atributos.

Al sistema de fusion se introducen representaciones que se quieren comparar, después se identifican los atributos similares. Una vez que los atributos similares se han identificado se revisa su similitud. Un modo de capturar su similitud y normalizar la similitud es con la siguiente formula:

$$s(A, B) = 1 - \frac{d(A, B)}{U} \tag{2.2}$$

Donde; d(A, B), es la disimilitud entre los conceptos A y B; U es el factor de normalizacion. Dado que la similitud y normalizacion son subjetivas se debe llegar antes a un acuerdo para evitar conflictos.

Por lo regular, los atributos de los datos se evalúan a nivel sintactico y espacial. A nivel sintactico se busca identificar la semantica y que ésta sea la misma en ambas representaciones, independientemente del modo de representarla. A nivel espacial, la evaluacion depende del tipo de dato. Los datos de tipo punto se evalu´an con una distancia euclidiana. Para líneas existen mediciones como la distancia Hausdorff o se revisa el porcentaje de consistencia entre líneas [Goodchild & Hunter, 1997]. Para objetos de tipo area es importante analizar su forma, por ejemplo, revisar el porcentaje de area en común.

En muchos casos, no basta considerar la similitud, independientemente del contexto, éste debe ser considerado. [Samal et al., 2004] dicen que el contexto geografico define la relacion espacial (topología, distancia y direccion) entre objetos.

Una vez que se han analizado las similitudes particulares entre objetos se establece un criterio de similitud general de las representaciones. Dicho de otra forma, la similitud general debe tomar en cuenta la similitud particular de cada objeto y con ella capturar la similitud a nivel de representaciones.

Algunas aplicaciones de la fusion de atributos son: consolidacion de cobertura, actualizacion de datos espaciales, registro de cobertura y deteccion de errores entre otras.

2.6. Discusión estado del arte

En este capítulo se presentaron trabajos relacionados con la similitud de datos. Se mostró su importancia y relevancia para otras areas de investigacion como integracion, alineacion, almacenamiento, consultas vagas, etc; y se ha mostrado que la similitud no es un problema sencillo particularmente para datos espacia- les.

Se presentaron modelos para evaluar la similitud entre datos, se señalaron sus propiedades y alcances. Como punto importante se observa que el modelo métri- co es poco convincente para utilizar como herramienta de similitud; el modelo de transformacion requiere tener establecidas previamente las transformaciones posibles entre estados; para el modelo de red, agregar una nueva representacion es un proceso complicado y crucial pues de él depende la correcta medicion de la similitud; y en el modelo de alineacion el desarrollo de un sistema que identifique conceptos equivalentes y logre la alineacion es bastante complicado.

Con estas observaciones encontramos el modelo de propiedades el mas adecuado para evaluar la similitud entre representaciones espaciales. Si los campos a

comparar pueden representarse como conjuntos determinados por sus elementos y atributos, el modelo de propiedades es el adecuado.

Se observa también que a pesar del desarrollo existente, aún hace falta un siste- ma que pueda establecer la similitud para representaciones espaciales de forma practica, en especial para aquellas que contengan un gran número de instancias.

3. Metodología

La metodología establece la similitud entre dos representaciones espaciales. Su analisis se divide principalmente en tres aspectos: estructural, tematico y espacial. El aspecto estructural se refiere al modo en que cada representacion organiza su informacion. El aspecto tematico verifica que ambas representaciones cuenten con las mismas instancias y sirve como criterio de alineacion para el analisis espacial. El aspecto espacial toma las instancias alineadas tematicamente y revisa sus propiedades espaciales, de acuerdo a ellas actualiza el valor de similitud y determina si las representaciones hablan del mismo objeto geografico.

En el aspecto estructural (o similitud estructural) se utiliza una ontología para alinear los conceptos entre representaciones. De forma manual el usuario establece la alineacion de los conceptos de cada representacion con los conceptos de la ontología. Dicha alineacion es necesaria para el establecimiento de la correspondencia entre representaciones, ya que relaciona los conceptos que describen a las mismas entidades del mundo real. Así, la ontología es necesaria para realizar una alineacion semantica entre representaciones.

La metodología tiene un enfoque de teoría de conjuntos, esto es, las instancias de cada representacion se manejan como elementos de un conjunto. Los elementos, a su vez, determinan el conjunto correspondiente a la representacion. Una vez formados los conjuntos se revisa su interseccion. El conjunto interseccion está formado por las instancias que se encuentren en ambos conjuntos y se asume que mientras mayor sea el conjunto interseccion (respecto al tamaño de cada conjunto), mayor será la similitud entre representaciones.

La metodología tiene como salida dos indicadores principales: las representaciones esquemáticas y un grado de interoperabilidad semántica. Los diagramas esquematicos son graficos que describen la similitud entre las representaciones. Se genera una representación esquemática por cada nivel de analisis.

El grado de interoperabilidad semántica es un valor entre 0 y 1 que establece qué tan interoperables son dos representaciones. A diferencia de la interoperabilidad, la interoperabilidad semantica va mas allá de encontrar y homogeneizar (por medio de estandares) las diferencias; trata de aceptar la diversidad geografica y hallar una forma de consolidar la diferencia en los significados [Harvey et al., 1999]. El grado de interoperabilidad semántica se calcula como una funcion de la similitud ya que se asume que mientras mas similares sean dos representaciones, seran mas interoperables. Mientras mayor sea el grado de interoperabilidad semántica, mayor será la interoperabilidad de las representaciones; así un grado de interoperabilidad semántica igual a 1 significa que las representaciones son totalmente interoperables entre sí, mientras que 0 significa que no son nada interoperables.

3.1. Marco de trabajo

La metodología se divide en tres etapas: Adecuacion de Datos, Analisis de la Similitud y Visualizacion de Resultados. Entre las características a destacar es que el procesamiento es semiautomatico; asimismo, la componente de Datos de Entrada se realiza de forma manual mientras que el resto del procesamiento se lleva a cabo en forma automatica.

El funcionamiento general del sistema consiste en introducir representaciones de un objeto geografico al sistema, seleccionar dos de ellas para compararlas, posteriormente realizar el analisis de la similitud y finalmente desplegar los resultados. En la Figura 3.1 se muestra el marco de trabajo general implementado en esta metodología.

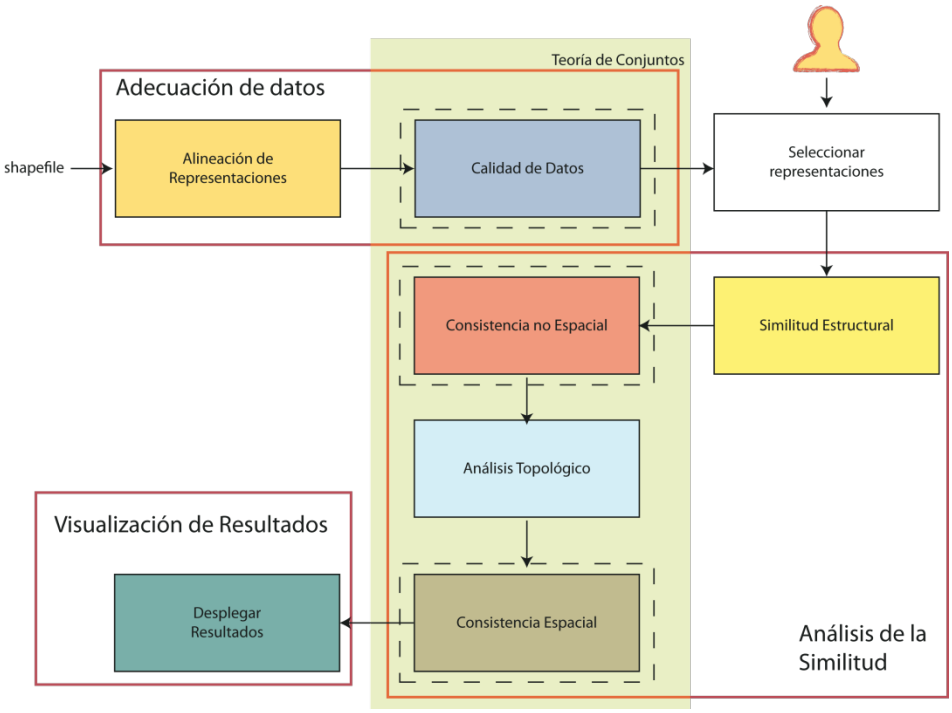


Figura 3.1: Metodología.

El sistema se diseñó basado en un enfoque de teoría de conjuntos, así la mayor parte del sistema maneja los datos como conjuntos definidos por sus elementos y propiedades. En la Figura 3.1 sobre un area verde al centro, se muestran los bloques o tareas principales, encargados basicamente de manejar explícitamente los datos espaciales como conjuntos.

Una vez que los datos geoespaciales han sido introducidos al sistema, éstos pueden ser comparados. Para ello, se despliega una lista de las representaciones de

donde el usuario selecciona dos. Por tanto, el sistema recupera los metadatos de cada representacion, donde se especifica qué tabla y atributos estan relacionados con las representaciones.

El Analisis de la Similitud revisa las propiedades estructurales y de instancias de los datos para determinar qué tan parecidas son las representaciones entre sí.

En la Figura 3.1, los recuadros de línea discontinua son aquellos bloques o tareas que discriminan a los datos. En otras palabras, al descartar instancias o elementos que no cumplen con ciertas restricciones, se actualiza el conjunto de datos con el que se trabaja. Cabe señalar, que estos bloques son los únicos que modifican los conjuntos de datos espaciales, el resto de los bloques solo consulta las representaciones sin alterarlas.

Finalmente, en la etapa de Visualizacion de Resultados se generan las representaciones esquemáticas que reflejan la similitud semantica entre los conjuntos. Las representaciones esquemáticas se pensaron como una solucion basada en los diagramas de Venn. Por tanto, se generan cuatro representaciones esquemáticas que capturan: la consistencia de las entidades, la consistencia del municipio, la consistencia de las localidades y la consistencia espacial.

En este trabajo, la consistencia se refiere específicamente a las instancias presen- tes en ambas representaciones. Si una instancia está en ambas representaciones se considera consistente, esto es, las representaciones no se contradicen. También se asume que mientras mas instancias consistentes tengan entre sí dos representaciones, mas similares son. Así, al capturar las representaciones esquemáticas la consistencia, capturan también la similitud entre representaciones.

3.2. Adecuación de los datos

Esta etapa se encarga de importar shapefiles1 al sistema. En esta seccion se introducen aquellas representaciones que se quieren comparar. La importacion se realiza por medio de dos herramientas: Protégé y PostgreSQL (Anexo A). En Protég é se almacenan metadatos de las representacion, con base en la construc- cion de una ontología de tarea. Este almacenamiento implica una alineacion de conceptos entre las representaciones, el cual se lleva a cabo en forma manual. Por su parte, PostgreSQL administra las representaciones en una base de datos espacial. Una vez que los datos se encuentran en la base de datos, nos referimos a ellas indistintamente como representaciones o conjuntos.

El bloque relacionado con la Calidad de los Datos revisa la correcta importacion de los datos en PostgreSQL. En particular, se examina la componente espacial cuando el tipo de dato es un area ya que una correcta geometría sustenta el adecuado funcionamiento del sistema.

3.2.1. Alineacion de las representaciones

La alineacion se lleva a cabo al importar los datos de entrada al sistema. Los datos se introducen en formato shapefile y son agregados de forma manual a PostgreSQL y Protégé. El ingresar los datos a PostgreSQL permite realizar con ellos las operaciones que definiran su similitud. Cuando los datos se importan en Protégé, se introducen sus metadatos, identificando principalmente qué atri- butos de la representacion corresponden a las propiedades de la ontología. Por tanto, al poblar la ontología se alinean las representaciones; es decir, al definir las propiedades de los individuos (instancias de la ontología) se establece la co- rrespondencia con la estructura ontologica y a su vez, la correspondencia con otras representaciones.

Por otra parte, las fuentes de informacion consideradas para el caso de estudio fueron el Instituto Nacional de Estadística, Geografía e Informatica (INEGI), así como la Comision Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO). Cabe señalar que ambas instituciones manejan y procesan sus datos en formato shapefile, lo cual origina que la entrada de datos al sistema se lleve a cabo en forma natural y sin la necesidad de realizar algún tipo de conversion entre formatos.

PostgreSQL

Dado que el shapefile no contiene explícitamente la informacion sobre la topo- logía de los datos, ésta se analiza en forma separada. Este analisis se realiza en PostgreSQL, utilizando su extension para manipular objetos geograficos, deno- minada PostGIS. Antes de poder trabajar los datos en PostgreSQL, éstos son importados de forma manual a una base de datos. La importacion se realiza con PostGIS utilizando las instrucciones shp2pgsql y psql2 (Figura 3.2).

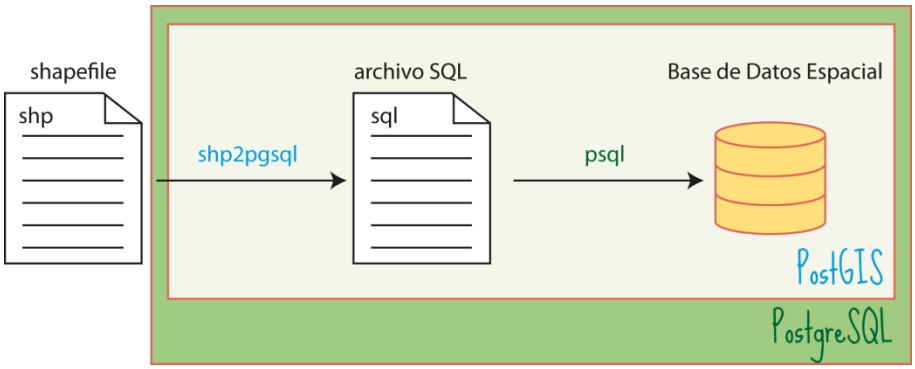


Figura 3.2: Importacion de shapefile a base de datos.

La instruccion shp2pgsql es parte de PostGIS y se encarga de convertir la infor- macion del shapefile en instrucciones adecuadas para su insercion en una base de

datos PostgreSQL (PostGIS - Data Management and Queries). La instrucción se ocupa de la siguiente manera:

```
shp2pgsql -W latin1 [shapefile] [tabla] >[nombreArchivo].sql
```

Donde -W especifica la codificacion de las instrucciones, para el caso de estudio se ocupa una codificacion latin1; el campo [shapefile] es el nombre del archivo shapefile; [tabla] indica el nombre de la tabla que se creará y que incluirá la informacion del shapefile; [nombreArchivo] indica el nombre que recibirá el archivo sql.

Una vez que se ejecuta la instruccion shp2pgsql, se crea un archivo sql que contiene la informacion para crear la tabla en PostgreSQL e introducirle los datos del shapefile.

Para crear la tabla a partir del archivo sql se ocupa la instruccion psql. Psql es una instruccion propia de PostgreSQL y, entre otras operaciones, permite la ejecucion de consultas y actualizaciones desde archivos (PostgreSQL Documentation - psql). Cuenta con un gran número de parametros pero para este trabajo solo cinco son ocupados. La instruccion se ocupa de la siguiente manera:

```
psql -d tesis -h localhost -U postgres -p 5433 -f [nombreArchivo].sql
```

Donde el parametro -d tesis indica la base de datos en donde se creará la tabla (en este caso tesis); -h localhost y -p 5433 indican el nombre del servidor y puerto respectivamente donde se encuentra la la base de datos; -U postgres es el nombre del usuario de la BD; finalmente el parametro -f indica el archivo fuente de las instrucciones a ejecutar. Este nombre debe coincidir con el establecido en la instruccion shp2pgsql.

Después de la ejecucion psql se cuenta con la nueva tabla en la base de datos. Se puede observar que la tabla contiene la informacion del archivo shapefile mas una columna llamada the geom (Figura 3.3 y Figura 3.4). the geom es el atri- buto que ocupa PostGIS para describir los datos espaciales.

CLAVE	COUNT	CVE_ENT	NOM_ENT	CVE_MUN	NOM_MUN	CVE_LOC	NOM_LOC
010010001	1	01	Aguascalientes	001	Aguascalientes	0001	Aguascalientes
010010479	1	01	Aguascalientes	001	Aguascalientes	0479	Villa Lic. Jess Tern (Ca...
010011025	1	01	Aguascalientes	001	Aguascalientes	1025	Pocitos
010020001	1	01	Aguascalientes	002	Asientos	0001	Asientos
010020059	1	01	Aguascalientes	002	Asientos	0059	Villa Jurez
010030001	1	01	Aguascalientes	003	Calvillo	0001	Calvillo
010030055	1	01	Aguascalientes	003	Calvillo	0055	Ojocaliente
010040001	1	01	Aguascalientes	004	Coso	0001	Coso
010050001	1	01	Aguascalientes	005	Jess Mara	0001	Jess Mara
010050041	1	01	Aguascalientes	005	Jess Mara	0041	Jess Gmez Portugal (M...
010060001	1	01	Aguascalientes	006	Pabelln de Arte...	0001	Pabelln de Arteaga
010070001	1	01	Aguascalientes	007	Rincn de Romos	0001	Rincn de Romos
010070030	1	01	Aguascalientes	007	Rincn de Romos	0030	Pabelln de Hidalgo
010080001	1	01	Aguascalientes	008	San Jos de Gracia	0001	San Jos de Gracia
010090001	1	01	Aguascalientes	009	Tepezal	0001	Tepezal
010100001	1	01	Aguascalientes	010	Llano, El	0001	Palo Alto
010110001	1	01	Aguascalientes	011	San Francisco d...	0001	San Francisco de los R...
020010001	1	02	Baja California	001	Ensenada	0001	Ensenada

Figura 3.3: Representacion en shapefile (detalle).

clave character var	cour num	cve_ent characte	nom_ent character varying(45)	cve_mun character	nom_mun character varying(45)	cve_loc character varying(6)	nom_loc character varying(60)	the_geom geometry
010010001	1	1	Aguascalientes	1	Aguascalientes	1	Aguascalientes	01010000
010010479	1	1	Aguascalientes	1	Aguascalientes	479	Villa Lic. Jesús Terán (Cc	01010000
010011025	1	1	Aguascalientes	1	Aguascalientes	1025	Pocitos	01010000
010020001	1	1	Aguascalientes	2	Asientos	1	Asientos	01010000
010020059	1	1	Aguascalientes	2	Asientos	59	Villa Juárez	01010000
010030001	1	1	Aguascalientes	3	Calvillo	1	Calvillo	01010000
010030055	1	1	Aguascalientes	3	Calvillo	55	Ojocaliente	01010000
010040001	1	1	Aguascalientes	4	Cosío	1	Cosío	01010000
010050001	1	1	Aguascalientes	5	Jesús María	1	Jesús María	01010000
010050041	1	1	Aguascalientes	5	Jesús María	41	Jesús Gómez Portugal (Marg	01010000
010060001	1	1	Aguascalientes	6	Pabellón de Artea	1	Pabellón de Arteaga	01010000
010070001	1	1	Aguascalientes	7	Rincón de Romos	1	Rincón de Romos	01010000
010070030	1	1	Aguascalientes	7	Rincón de Romos	30	Pabellón de Hidalgo	01010000
010080001	1	1	Aguascalientes	8	San José de Graci	1	San José de Gracia	01010000
010090001	1	1	Aguascalientes	9	Tepezalá	1	Tepezalá	01010000
010100001	1	1	Aguascalientes	10	Llano, El	1	Palo Alto	01010000
010110001	1	1	Aguascalientes	11	San Francisco de	1	San Francisco de los Romo	01010000
020010001	1	2	Baja California	1	Ensenada	1	Ensenada	01010000

Figura 3.4: Representacion importada a tabla (detalle).

Protégé

Cada shapefile que se importa a PostgreSQL es una representacion de algún fenomeno de la realidad. En Protégé se introducen los metadatos de las representaciones. Los metadatos describen el nombre de las tablas de la representacion, sus atributos (entidad, municipio, localidad), su tipo de dato (punto, area) y su escala.

En Protégé se cuenta con una ontología que contiene como conceptos: representacion, entidad, municipio y localidad (Figura 3.5). Tiene solo una relacion "es-un" (is-a) que representa la division de una representacion (de la República Mexicana) en entidades y a su vez en municipios y localidades. En la ontología, cada representacion es procesada como un individuo con las propiedades: nom-

bre de la representacion, nombre de la entidad, nombre del municipio, nombre de la localidad, tipo de dato y escala.

Figura 3.5: Ontología de tarea en Protégé.

El nombre de la representacion indica el nombre de la tabla en la base de datos correspondiente a la representacion.

Los nombres de la entidad, municipio y localidad especifican los atributos dentro de la tabla que corresponden con los datos de entidad, municipio y localidad de la representacion. Si el valor de estos campos en la ontología es "vacío", significa que la representacion no cuenta con esa informacion, dicho de otra forma, estos campos capturan la granularidad de la representacion. Así, una representacion que contenga datos de entidad y municipio únicamente, tendran una granula- ridad diferente a otra representacion que tenga datos de entidad, municipio y localidad.

Finalmente, los metadatos de escala y tipo de dato contienen la informacion de escala de la representacion y el tipo de dato que maneja (area o punto). El tipo de dato sirve para saber qué operaciones se realizaran en el analisis topologico. La escala sirve para determinar la consistencia espacial. Con ella se establece un parametro que permite decidir si cada instancia de las representaciones es o no consistente.

3.2.2. Calidad de los datos

En nuestro caso de estudio, las localidades se representan con datos de tipo punto o area. Cuando las localidades se representan con puntos los datos no presentan problemas geométricos. Sin embargo, cuando el tipo de dato es un area, pueden existir diferentes problemas. Por ejemplo, que los vértices no formen un area cerrada o que haya intersecciones entre la misma area. Por este motivo, una vez que los shapefiles son importados a PostgreSQL, las representaciones de tipo area son revisadas.

Esta revision es un control sobre la calidad de los datos a nivel de instancia [Widom, 2005]. Se seleccionan y descartan aquellas instancias que presentan errores (Figura 3.6). El motivo es que PostgreSQL necesita representaciones geométricas validas para realizar un análisis espacial. Si la representacion geométrica contiene errores, PostgreSQL no puede realizar consultas espaciales con estos datos.

Para verificar la validez de la componente espacial de los datos, se ocupa la instruccion ST_IsValid de PostGIS. Con ST_IsValid se actualiza cada representacion y se conservan únicamente las instancias validas. Es importante re-

cordar que únicamente las representaciones con datos de tipo area pasan por este proceso. Durante la importacion de datos, PostgreSQL no realiza el analisis de validez de los datos porque requiere mucho tiempo de procesamiento, en especial para geometrías complejas [PostGIS - Ensuring OpenGIS compliancy of geometries]. No obstante, para este trabajo es un proceso necesario pues una correcta geometría sustenta el adecuado funcionamiento del sistema.

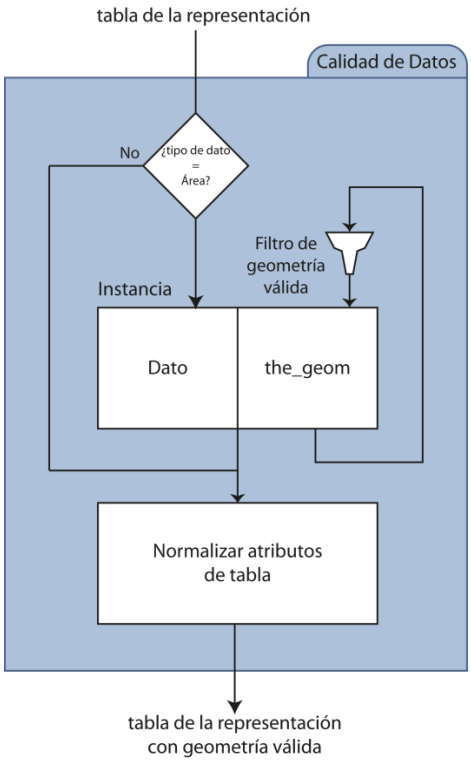


Figura 3.6: Revision de la calidad de los datos.

Normalizar los atributos de tabla

La normalizacion se refiere a establecer, por cada representacion, los atributos que se van a comparar como cadenas de caracteres.

INEGI y CONABIO, las fuentes del caso de estudio, manejan claves para entidades, municipios y localidades. Estas claves son números pero no siempre se encuentran en un formato estandarizado, es común encontrarlas como enteros o cadenas de caracteres. La normalizacion de atributos hace posible una comparacion a gran escala. Esto es, el estandarizar los atributos conlleva el realizar un analisis sobre un gran número de datos. Con esta conversion la semantica de los datos no es alterada, solo se cambia el tipo de dato para homogeneizar las

claves de cada representacion.

La razon de usar caracteres como el formato de la clave es que el sistema trabaje con texto, así, en trabajo a futuro se pueden analizar directamente los nombres de las localidades, y mas aún, agregar un analisis léxico para mejorar el desarrollo del sistema.

3.3. Análisis de la similitud

En esta etapa se evalúa la similitud de las representaciones que selecciona el usuario. Cuando los datos han sido introducidos al sistema el usuario selecciona dos representaciones para comparar. Se revisan cuatro propiedades de las representaciones: su similitud estructural, la consistencia no espacial, su similitud topologica y la consistencia espacial (Figura 3.7). Estas cuatro propiedades se ocupan para establecer la similitud.

La etapa de Similitud Estructural analiza la granularidad de las representacio- nes, por ejemplo, una representacion con informacion sobre entidad, municipio y localidad tiene diferente granularidad que otra con informacion solo de loca- lidad. La consistencia no espacial revisa la consistencia de las instancias en su componente no espaciales, en particular, revisa las instancias a nivel de entidades, municipios y localidades; una instancia se considera consistente solo si se encuentra presente en ambas representaciones. El analisis topologico revisa la relacion espacial que existe entre las instancias de cada representacion. Finalmente, la consistencia espacial determina, de acuerdo con la distancia entre localidades, si éstas son consideradas consistentes; se revisa la distancia entre instancias consistentes y con base en la escala de las representaciones se propo- ne un umbral. Si la distancia entre instancias no supera el umbral entonces las instancias se consideran espacialmente consistentes y se asume que ambas representaciones, en esa instancia en particular, se refieren al mismo objeto geografico.

El Analisis de la Similitud tiene como datos de entrada las representaciones que el usuario desea comparar. A su salida entrega cuantro conjuntos que reflejan la similitud entre las representaciones. Los tres primeros conjuntos representan aquellas entidades, municipio y localidades que se consideran consistentes, esto es, las entidades, municipios y localidades que se manejan en ambas representaciones. El cuarto conjunto representa las instancias totalmente consistentes, esto es, aquellas que además se su consistencia tematica se consideran ser el mismo objeto greografico. En conjunto, (asumimos que) los cuatro conjuntos generados capturan la similitud entre representaciones. Los datos de salida son ocupado por la etapa de Visualizacion de Resultados para mostrar el resultado de modo grafico.

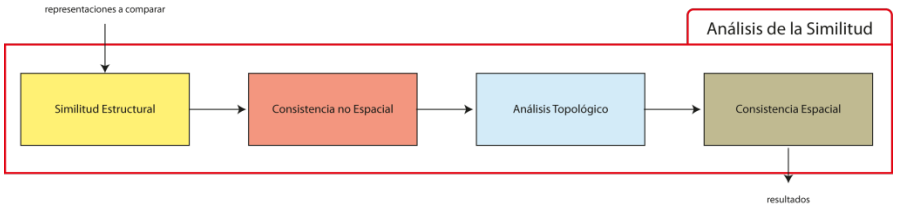


Figura 3.7: Analisis de la similitud.

3.3.1. Similitud estructural

En este proceso se revisa si la representacion contiene informacion de entidades, municipios y localidades, esto es, se revisa la granularidad de la representacion. En este paso, la revision es una verificacion booleana, no se revisa el contenido de la informacion, solo se revisa que la informacion se encuentre presente.

En la ontología, los individuos tienen las siguientes propiedades: tieneNombre-DeLocalidad, tieneNombreDeMunicipio, tieneNombreDeEntidad. La granulari- dad de la representacion queda capturada por dichas propiedades. Cuando una representacion no incluye alguno de estos niveles, la propiedad correspondiente tendrá "vacío" como su valor (ver Figura 3.8). Así, si existen dos representacio- nes, una sin ninguna propiedad vacía y la otra con al menos una, su granularidad será diferente. Es importante señalar que la Figura 3.8 representa propiedades de un individuo en la ontología, no conceptos.

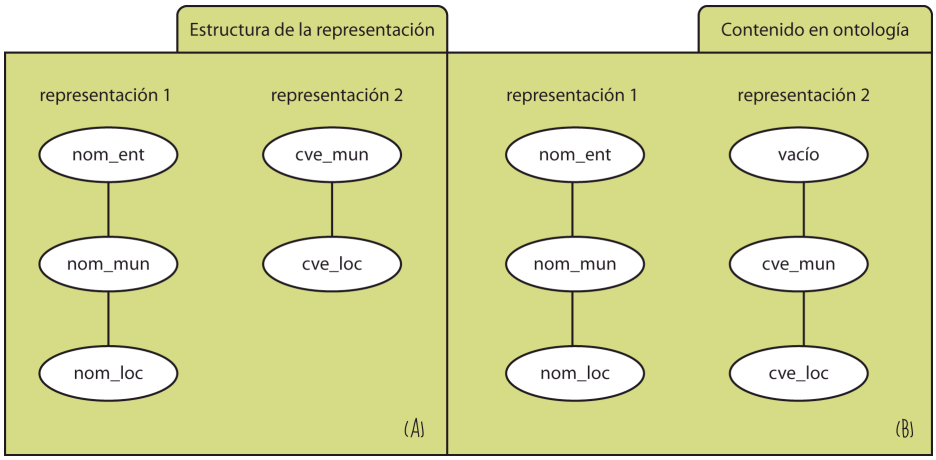


Figura 3.8: Ejemplo de representaciones con granularidad diferente(A) y su in-corporacion a la ontología (B).

Para conocer la similitud estructural se realiza una consulta a la ontología que

regresa las propiedades de las representaciones a analizar. A partir de las propiedades se crea un arreglo unidimensional que captura la granularidad. Por ejemplo, un individuo con propiedades:

tieneNombreDeEntidad = nom ent
tieneNombreDeMunicipio = nom mun
tieneNombreDeLocalidad = nom loc

En este caso se genera el arreglo 111, que indica que existe informacion para cada nivel de la ontología, en cambio un individuo con propiedades:

tieneNombreDeEntidad = ''vacio''
tieneNombreDeMunicipio = cve mun
tieneNombreDeLocalidad = cve loc

Por tanto, éstas generan el arreglo 011.

El arreglo para la similitud estructural tiene la forma: representacion1.entidad, representacion2.entidad, representacion1.municipio, representacion2.municipio, representacion1.localidad, representacion2.localidad (Figura 3.9). Así, considerando el ejemplo anterior, se tiene el arreglo similitudEstructural = 101111.

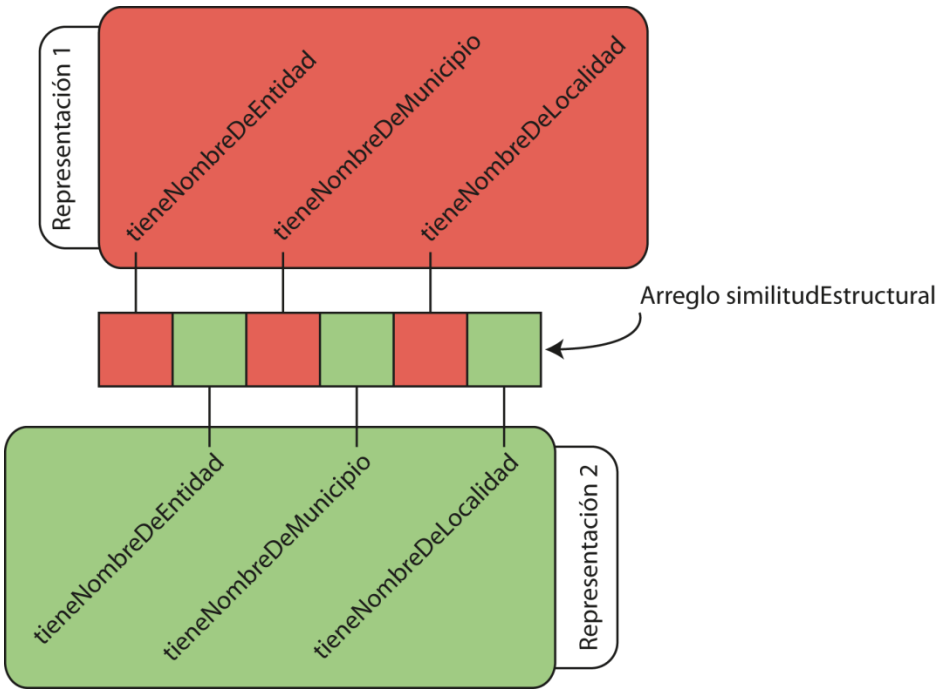


Figura 3.9: Configuracion del arreglo similitudEstructural.

La similitud estructural es el primer parametro que modifica el grado de interoperabilidad semántica de las representaciones.

Una vez que las representaciones se han introducido al sistema, éste genera una lista de ellas de donde el usuario selecciona dos para compararlas. La primer representacion será la representacion a revisar con respecto a la segunda. El orden en que sean seleccionadas importa debido a que la similitud se considera no simétrica. La primer representacion será A y la segunda B, así, la similitud entre representaciones se escribe S(A, B).

A partir de este punto, las representaciones se procesan como conjuntos, de los cuales contaremos: el número de elementos de cada conjunto, el número de elementos en la interseccion de conjuntos y el número de elementos en la diferencia.

El analisis de consistencia se divide en dos secciones, la primera se refiere a los componentes no espaciales de los datos (consistencia no espacial) y la segunda a la componente espacial (consistencia espacial). La consistencia no espacial, a su vez se divide en tres secciones: entidad, municipio y localidad.

3.3.2. Consistencia temática

Dentro de esta seccion se establece la similitud no espacial entre dos conjuntos que se trabaja como un modelo de atributos donde la funcion de similitud depende de los elementos en común y los elementos discordes, como se plantea en [Tversky, 1997]. La similitud se establece conforme al número de instancias consistentes; se asume que la similitud es proporcional a las consistencias e inversamente proporcional a las inconsistencias.

La funcion ocupada para la similitud es:

$$s(a, b) = F(A \cap B, A - B, B - A) \qquad (3.1)$$

donde a y b son los conjuntos; A es el número de elementos en a; B el número de elementos en b; $A \cap B$ es el número de elementos consistentes; $A - B$ es el número de elementos de a y que b no tiene; $B - A$ es el número de elementos que tiene b pero que a no tiene.

El analisis no espacial de los datos se refiere a la revision de las representaciones por niveles (entidad, municipio y localidad). El primer paso se realiza sobre el nivel entidad.

Dado que la similitud se considera asimétrica, es importante definir como referirse a la similitud. A la funcion s(a, b) se le denomina similitud de A respecto B. Del mismo modo, a s(B, A) se le llama similitud de B respecto A. No hay que olvidar que s(A, B) = s(B, A).

De cada representacion, se seleccionan sus entidades evitando datos duplicados. A nivel de entidad, esta consulta sirve para conocer el número de elementos en los conjuntos A y B. Estos valores se almacenan para posteriormente realizar el calculo de la similitud y para reportar los resultados. La similitud a nivel entidad se considera una similitud parcial y se almacena en un archivo con formato Json3.

Después, se realiza una consulta que da como resultado el número de elementos dentro del conjunto $A \cap B$; es decir, se obtiene el número de entidades consistentes entre las dos representaciones. Esto se resuelve ejecutando una consulta similar pero ahora se cuenta el número de instancias consistentes entre A y B.

Por otra parte, ya que se tienen las entidades consistentes, se puede realizar el siguiente nivel de analisis. De las entidades consistentes se revisara, por cada entidad, qué municipios son consistentes. Cada analisis de los niveles va discriminando datos inconsistentes, (Figura 3.10). Esto es, que no tiene sentido revisar un municipio perteneciente a una entidad inconsistente, al ser inconsistente su entidad no podrá encontrarse una correspondencia con la otra representacion. Por tanto, no existirá ninguna correspondencia de una entidad ni de sus datos derivados si ésta es inconsistente. Haciendo la discriminacion de datos inconsis- tentes el conjunto con el que se trabaja se actualiza, y no se revisan las instancias inconsistentes, reduciendo el tiempo de computo.
Para el analisis de las localidades se revisa que los municipios sean consistentes, que las localidades se encuentren en ambas representaciones. De la misma manera que para los municipios, se seleccionan subconjuntos de los creados anteriormente, para encontrar el conjunto de localidades con municipios y entidades consistentes. Este procesamiento se puede definir como un filtrado recursivo de instancias sobre el mismo conjunto (Figura 3.10). Al término de este procesamiento, se cuenta con el conjunto de instancias consistentes. En el siguiente analisis de consistencia se utilizará este conjunto para identificar las instancias espacialmente consistentes.

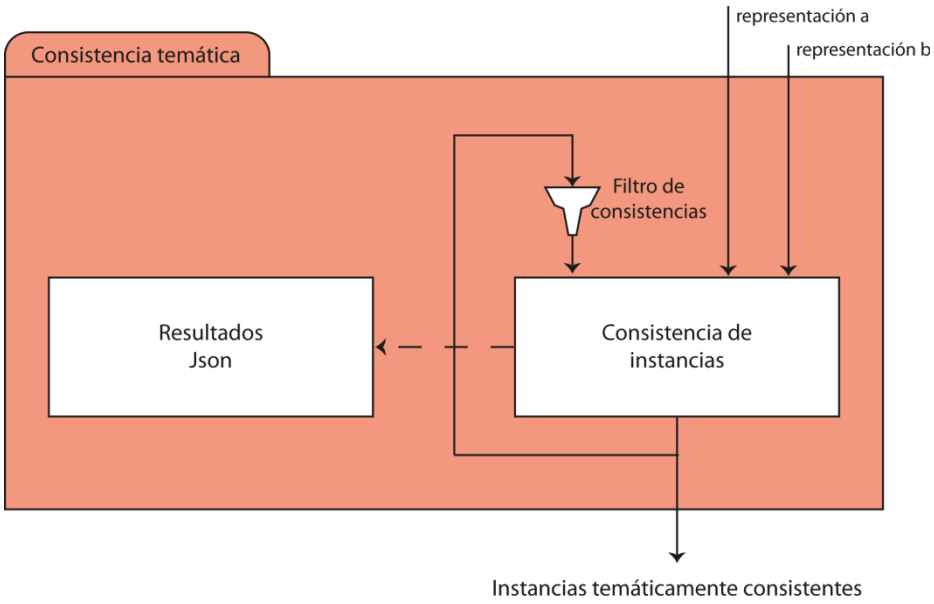


Figura 3.10: Identificacion del conjunto de instancias consistentes.

3.3.3. Análisis topologico

Esta etapa tiene como datos de entrada las instancias consistentes y trabaja con el tipo de dato de las representaciones. El tipo de dato es parte de los meta- datos de cada representacion y se obtiene de la consulta realizada en Protégé. Cada representacion puede tener tipo de dato punto o area. El tipo de dato es necesario para el analisis topologico, ya que las operaciones espaciales entre las representaciones dependen de él.

Si ambas representaciones tienen datos tipo punto, ninguna relacion topologica se verifica. Si ambas representaciones tienen datos tipo area, el usuario puede seleccionar la relaciones Son-disjuntos (Disjoint); Se-tocan (Touches); y/o Están-sobrepuestos (Overlaps). Si el tipo de dato en las representaciones es punto y area, las relaciones espaciales disponibles son Disjoint, Touches y/o Contains. En este caso la definicion de estas relaciones se presenta a continuacion:

Disjoint : Regresa verdadero si las geometrías no se intersectan espacialmente, es decir, si no comparten ningún espacio [PostGIS - ST Disjoint].

Touches : Regresa verdadero si las geometrías tienen al menos un punto en común, pero sus interiores no se intersectan [PostGIS - ST Touches].

Overlaps : Regresa verdadero si las geometrías se sobreponen espacialmente, pero una no contiene completa a la otra [PostGIS - ST Overlaps].

Contains: Regresa verdadero si y solo si todos los puntos de una geometría estan contenidos en su totalidad por la otra. Esto es, ningún punto de B está en el exterior de A [PostGIS - ST Contains].

En caso de que el usuario tenga conocimiento sobre los datos espaciales, puede seleccionar la ponderación de las relaciones topologicas. La ponderacion se refiere al aporte que tendrá cada relacion sobre el grado de interoperabilidad semántica. La ponderacion sobre las relaciones topologicas es útil en caso que se quiera dar prioridad a alguna relacion espacial, por ejemplo, si el usuario está interesado únicamente en instancias disjuntas, debe establecer una ponderacion 100-0-0 para el Analisis Topologico (Figura 3.11), de este modo, el sistema el resto de las relaciones no afectará el grado de interoperabilidad semántica.

Para modificar la ponderación sobre la relaciones topologicas se deben asignar valores a cada funcion revisando que el total sume 100. De esta forma, si le interesa alguna relacion en particular puede asignarle un valor mas alto, así el grado de interoperabilidad capturará dicha prioridad (Figura 3.11). En el caso de usuarios inexpertos, el sistema tiene predeterminada una configuracion 50-0-50 para las relaciones topológicas. Se decidió por esta configuracion dado que la mayoría de instancias tienen una relacion Disjoint u Overlaps. Así, se ignora la componente menos significativa para que no afecte al grado de interoperabilidad semántica.

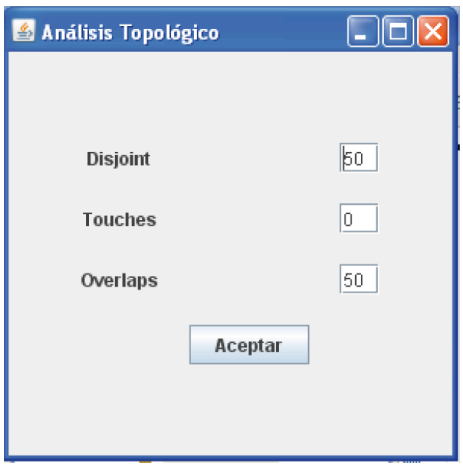


Figura 3.11: Ponderacion de las relaciones topologicas.

Las funciones en PostGIS revisan la relacion espacial entre dos objetos. Dado que los datos se manejan como conjuntos, las funciones se aplican por cada instancia consistente. Los parámetros son cada instancia de la representacion A y la instancia correspondiente en la representacion B. Así, la funcion Touches revisa la relacion entre la localidad de la representacion A y la misma locali- dad en la representacion B. Dado que las relaciones topologicas son excluyentes

podrían trabajarse como filtros y aplicarse en orden únicamente a los datos que no cumplan con alguna de ellas. Por ejemplo, si algunas instancias resultaron disjuntas (Disjoint), hay que evitar su revision para Touches y Contains. No obstante, para evitar separar instancias en tablas o con consultas elaboradas, si el usuario decide revisar las relaciones topologicas, éstas se revisan para todos los datos (Figura 3.12). Los resultados de cada relacion topologica actualizan el grado de interoperabilidad pero los datos siguen aún considerandose consisten- tes.

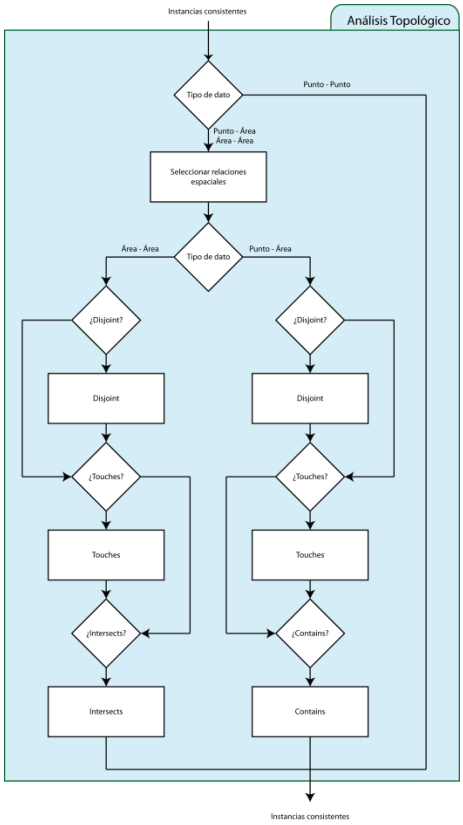


Figura 3.12: Diagrama de flujo para el analisis topologico.

3.3.4. Consistencia espacial

El factor determinante en la consistencia espacial es la distancia. Las relaciones topologicas sirven para actualizar el grado de interoperabilidad, ellas no determinan la consistencia espacial entre las representaciones.

De acuerdo con la escala de los datos, se propone un umbral o filtro espacial (Fi-

gura 3.13). El umbral representa la distancia espacial maxima que puede existir entre las instancias para ser consideradas espacialmente consistentes. Del mis- mo modo que con las relaciones espaciales, la distancia tiene como parametros las componentes espaciales de la instancia en la representacion A y la instancia correspondiente en la representacion B.

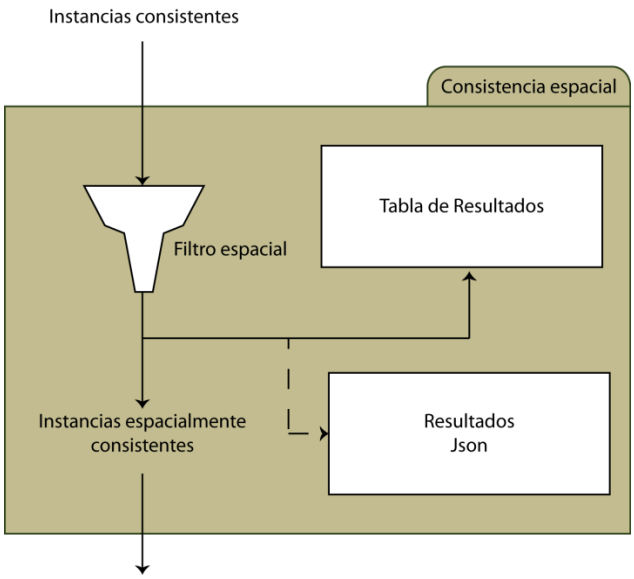


Figura 3.13: Diagrama de flujo para el analisis espacial.

En la Tabla 3.1 se observa que no se realiza ninguna revision topologica entre los datos de tipo punto. Por el contrario, la distancia se aplica a todo par de instancias consistentes y se utiliza como parametro para determinar su consistencia espacial. A la salida de esta etapa se obtienen todas las instancias consistentes, tanto en la componente no espacial como en la espacial.

Tabla 3.1: Relaciones entre los tipos de datos.

	Punto		Área	
	Relación Topológica	Relación Métrica	Relación Topológica	Relación Métrica
Punto		Distancia	Son disjuntos Se tocan Está contenido	Distancia
Área	Son disjuntos Se tocan Está contenido	Distancia	Son disjuntos Se tocan Están sobrepuestos	Distancia

3.4. Visualización de resultados

Durante cada nivel del analisis de consistencia (entidad, municipio, localidad y nivel espacial) se almacena la informacion que describe a los conjuntos. La informacion que nos interesa es el tamaño de cada conjunto y el tamaño de su interseccion. La informacion captura la consistencia en cada nivel de analisis.

De esta manera, por cada nivel se almacenan las variables N, n,y nc; donde N representa el número de instancias del conjunto A; n el número de instancias del conjunto B y nc representa las instancias consistentes $nc = N \cap n : N \in A, n \in B, nc \in A \cap B$.

Durante el analisis de consistencias, se genera un archivo php que contiene la informacion referente a cada nivel de analisis. El archivo php es un arreglo en formato Json, llamado vector, con la estructura mostrada en la Tabla 3.2.

Tabla 3.2: Estructura del arreglo de resultados.

	Nivel	A	B	$A \cap B$
vector[0]	entidad	N	n	nc
vector[1]	entidad	N	n	nc
vector[2]	entidad	N	n	nc
vector[3]	entidad	N	n	nc

Los resultados se muestran en un navegador que despliega una pagina web. La pagina consulta el arreglo de resultados y a partir de él se crean cuatro representaciones esquemáticas. Las representaciones esquemáticas reflejan la similitud entre los conjuntos, donde cada conjunto se representa como un rectangulo (Fi- gura 3.14).

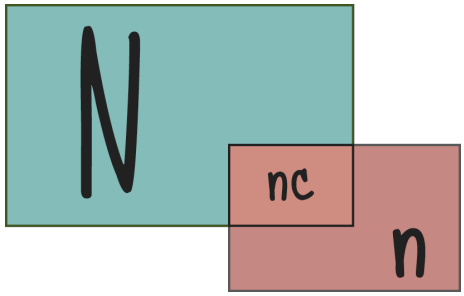


Figura 3.14: Representacion esquematica.

La representación esquemática se despliega en un elemento canvas o lienzo de

HTML5. El primer paso para desplegar la informacion es leer el arreglo de resultados y con él determinar las dimensiones (altura a y base b) de los conjuntos en las representaciones esquemáticas (Figura 3.15).

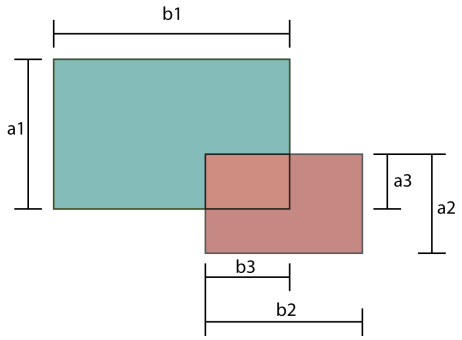


Figura 3.15: Dimensiones de los conjuntos de una representación esquemática. A continuacion se calculan los lados b1 y a1 del conjunto A con las siguientes expresiones matematicas:

$$b_1 = 3k/2 \quad (3.2)$$

$$a_1 = 2b_1/3 \quad (3.3)$$

Donde k es una contante que representa el area del rectangulo mayor que cabe dentro del lienzo.

Posteriormente, se calculan los lados del segundo rectangulo:

$$b_2 = 3 \cdot n \cdot k/2N \quad (3.4)$$

$$a_2 = 2b_2/3 \quad (3.5)$$

Esta vez, los lados estan en funcion del tamaño del conjunto B, el conjunto A y de la constante. Con esta funcion, si el conjunto B tiene la mitad de los elementos que el conjunto A, el tamaño de su rectangulo será también la mitad del tamaño del rectangulo de A.

Finalmente, se traza el rectangulo de consistencia $A \cap B$. Sus lados b3 y a3 se calculan con las siguientes ecuaciones:

$$b_3 = 3 \cdot nc \cdot k/2N \quad (3.6)$$

$$a_3 = 2b_2/3 \quad (3.7)$$

Una vez que se tiene el valor de cada conjunto, se traza el primer rectangulo con origen en el inicio del lienzo. La ubicacion del segundo rectangulo se calcula como sigue:

$$origen_X = b_1 - b_3 \tag{3.8}$$

$$origen_Y = a_1 - a_3 \tag{3.9}$$

Al establecer el origen del segundo rectangulo de esta forma, se consigue trazar el segundo conjunto de forma tal que se sobreponga al primero, formando a su vez el conjunto consistente $A \cap B$.

Ahora se tiene un grafico que representa la similitud entre los dos conjuntos, denominado representación esquemática (Figura 3.14).

3.5. Grado de interoperabilidad semántica

Una vez que se tiene los resultados de cada nivel de analisis, con ellos se genera el grado de interoperabilidad semántica. El grado de interoperabilidad semántica es un valor entre 0 y 1 que establece qué tan interoperables son las representaciones y se calcula a partir de las tres componentes de la similitud: similitud estructural, similitud tematica y similitud espacial.

La similitud estructural contiene los resultados de comparar la estructura de ca- da representacion; la parte tematica contiene la similitud de todos los atributos no espaciales; y la parte espacial contiene los resultados del analisis espacial.

Cada componente tiene el mismo aporte al grado de interoperabilidad y a su vez está normalizada a ser un valor entre 0 y 1, de modo que el grado de interoperabilidad se calcula con la funcion:

$$i = \frac{S_e(A, B) + S_T(A, B) + S_E(A, B)}{3} \tag{3.10}$$

donde i es el grado de interoperabilidad semántica, S_e la similitud estructural, S_T la similitud tematica y S_E la similitud espacial.

De acuerdo al tipo de dato de las representaciones la parte espacial de los resultados puede variar. Para similitud entre representaciones punto-punto el resultado espacial se genera con las instancias consideradas espacialmente consistentes (de acuerdo a la distancia entre instancias); para representaciones punto-area el resultado se genera revisando la relacion espacial Contains; mientras que para representaciones area-area el resultado se genera a partir de las relaciones espaciales Disjoint, Touches e Intersects (Figura 3.16).

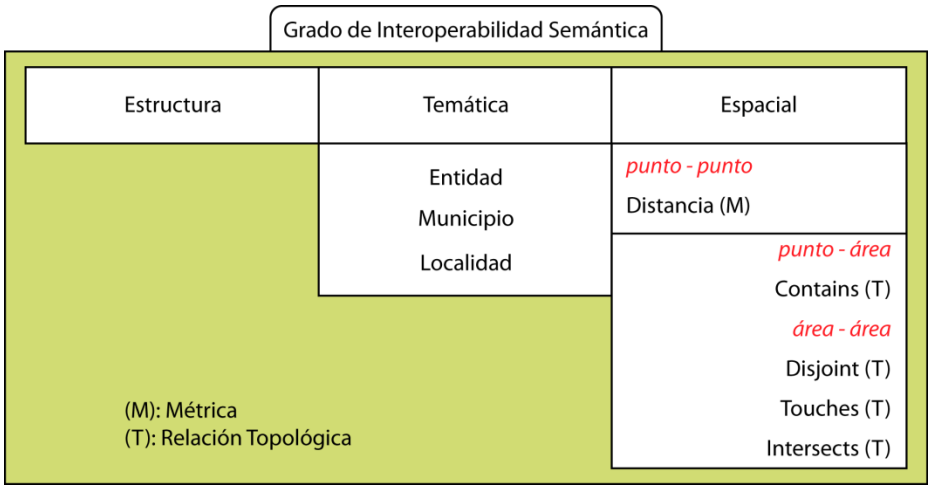


Figura 3.16: Componentes del grado de interoperabilidad semántica.

Al aplicar la funcion del grado de interoperabilidad semántica (Ecuacion 3.10) se genera un valor entre 0 y 1 que representa, con base en la similitud, qué tan interoperables son dos representaciones. El grado de interoperabilidad semántica se interpreta de la siguiente manera: 0 significa que las representaciones no son nada parecidas y por lo tanto tiene interoperabilidad nula; 1 significa que las representaciones son iguales y por lo tanto tienen alta interoperabilidad (Figura 3.17).

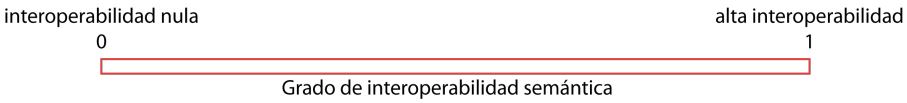


Figura 3.17: Rango del grado de interoperabilidad semántica.

La similitud entre representaciones se establece al identificar los objetos geografi- cos presentes en ambas representaciones y es representa graficamente por medio de representaciones esquemáticas. Las representaciones esquemáticas reflejan la similitud por cada uno de los niveles de analisis, hacen mas descriptivos los resultados y, al ser acompañadas por el texto analogo, conservan su precision.

El problema de la similitud se trabaja con un enfoque de teoría de conjuntos, donde cada representacion se ve como un conjunto definido por sus elementos; éstos a su vez se definen por sus propiedades. Una ventaja de esto es la posibilidad de aplicar esta metodología para cualquier dominio que pueda trabajarse con un enfoque de teoría de conjuntos. A pesar que el caso de estudio son localidades de la República Mexicana, una correcta adecuacion de parametros,

permitiría también utilizar la metodología para diferentes dominios.

Dicha adecuacion de parametros debe prestar especial atencion a la importacion de datos al sistema, en particular a su importacion a Protégé, ya que es en esta herramienta que se establecen los conceptos y propiedades a comparar. Por otro lado, en nuestro caso de estudio los datos de entrada son shapefiles, sin embargo, la metodología funciona para cualquier representacion espacial que pueda importarse a PostgreSQL, y en general a cualquier manejador de base de datos con soporte para datos espaciales.

4. Pruebas y resultados

En este capítulo se presentan algunas pruebas realizadas y su implicacion; algunos resultados, ejemplos de analisis y se comenta su interpretacion.

El capítulo se divide en Normalizacion de atributos y Análisis de similitud. La seccion de Normalizacion de atributos se refiere a las pruebas realizadas, los resultados observados y a las decisiones que se tomaron para mejorar el desempeño del sistema, en particular sobre los atributos que se iban a comparar.

En Analisis de similitud se seleccionaron y analizaron algunas representaciones, con ellas se muestra el funcionamiento del sistema. Se despliegan las representaciones, sus resultados y se explica como deben interpretarse. Así mismo se seleccionaron ejemplos de instancias que muestran la funcionalidad del sistema, se seleccionaron cuatro ejemplos, dos de ellos son comparaciones punto-punto, uno es punto-area y el cuarto es una comparacion area-area. La segunda comparacion punto-punto se escogió principalmente para ejemplificar como los resultados (en particular, las representaciones esquemáticas) reflejan la similitud entre representaciones.

4.1. Normalización de atributos

En la metodología existe una fase llamada Normalizar atributos de tabla. Esta fase se agregó al realizar pruebas con representaciones y observar que el número de instancias consistentes se podía mejorar.

Inicialmente el sistema trabajaba directamente con los campos que contienen el nombre textual de entidades, municipios y localidades. Se observó que cada campo manejaba también una clave y que esta clave era consistente entre representaciones, por ejemplo, en todas las representaciones a nivel de entidades, a Aguascalientes le corresponde la clave 1, a Baja California la clave 2, etc. Así mismo, la correspondencia entre claves se conserva para municipios y localidades.

La ventaja de trabajar con la clave y no con el nombre textual de cada instancia es que se mejora la similitud, por ejemplo, el sistema identifica como diferentes los municipios “Dr. Arroyo” y “Doctor Arroyo”, cuando sabemos que son equivalentes (municipio de Nuevo Leon). Afortunadamente en las claves sí existe consistencia en ambas representaciones: Nuevo Leon es la entidad 19, Doctor Arroyo y Dr. Arroyo son el municipio 14 de Nuevo Leon. Por este motivo se agregó la parte de normalizacion que se refiere a establecer como cadenas de caracteres a todas las claves, de esta forma se mejora significativamente el analisis de similitud.

La siguientes graficas reflejan la mejora que se logró al implementar la fase normalizacion de atributos. En la Figura 4.1 se observa la grafica de la comparacion de las representaciones loc95cw y loc2000cw. En rectangulos rellenos se muestra la correspondencia que existía al trabajar con los nombres textuales de las instancias. Los rectangulos sin relleno (y mas altos) reflejan la correspondencia que se alcanzó al trabajar con las claves de las instancias. La consistencia se presenta como porcentaje de cada fase de analisis: entidad, municipio, localidad y topología.

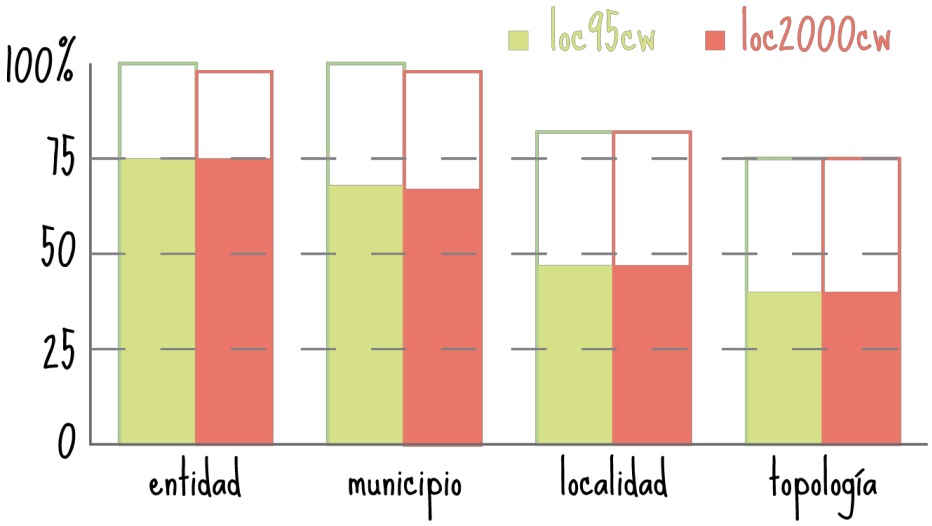


Figura 4.1: Porcentaje de instancias consistentes entre las representaciones loc95cw y loc2000cw.

Para cada fase de analisis se contó el total de instancias por representacion y con el número de instancias consistentes se obtuvo el porcentaje de instancias consistentes por representacion. En la Figura 4.2 se observa una grafica que indica que aún cuando se trabaja con las claves, el porcentaje de instancias consistentes puede ser bajo y particularmente bajo para una sola representacion, esto se explica recordando que la similitud es asimétrica; $s(A, B) = s(B, A)$.

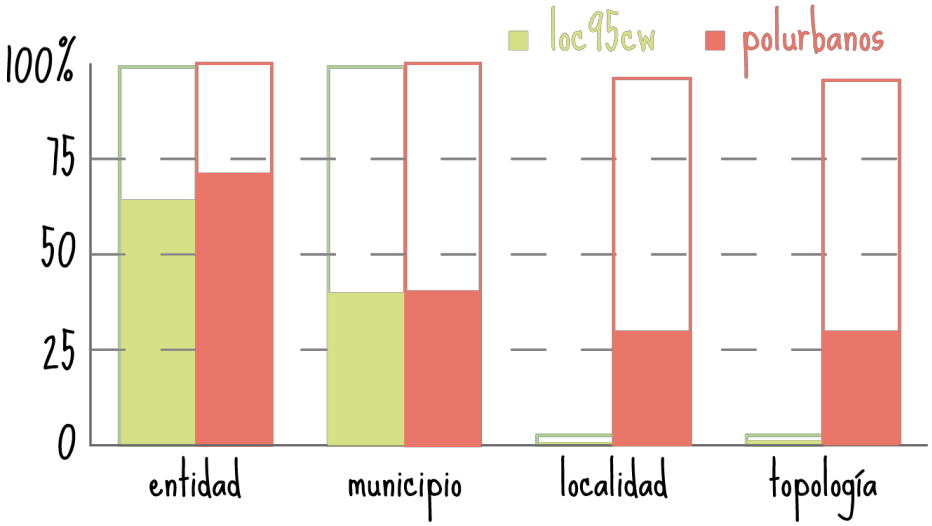


Figura 4.2: Porcentaje de instancias consistentes entre las representaciones loc95cw y polurbanos.

De los resultados en la Figura 4.2 se concluye que la representacion polurbanos es muy parecida a la representacion loc95cw. Sin embargo la representacion loc95cw no lo es tanto a la representacion ttpolurbanos.

4.2. Análisis de similitud

En esta seccion se presentan algunas pruebas realizadas para encontrar la similitud entre representaciones, primero se muestran las representaciones posteriormente los resultados del analisis.

Por cada prueba se señala el nombre de las representaciones, su tipo de dato, se despliegan sus shapefiles y se explican los resultados.

Prueba 1
Representaciones: loc2000cw - loc95cw. Tipo
de dato: punto - punto.
Grado de interoperabilidad: 0.89587325.

En esta prueba se comparan dos representaciones muy similares. De la Figura 4.3 podemos intuir que tienen casi el mismo número de instancias y podemos decir que las representaciones son muy parecidas. Los resultados concuerdan con esto; en la parte de entidades se observa que ambas representaciones tienen 32 entidades y éstas son consistentes entre sí.

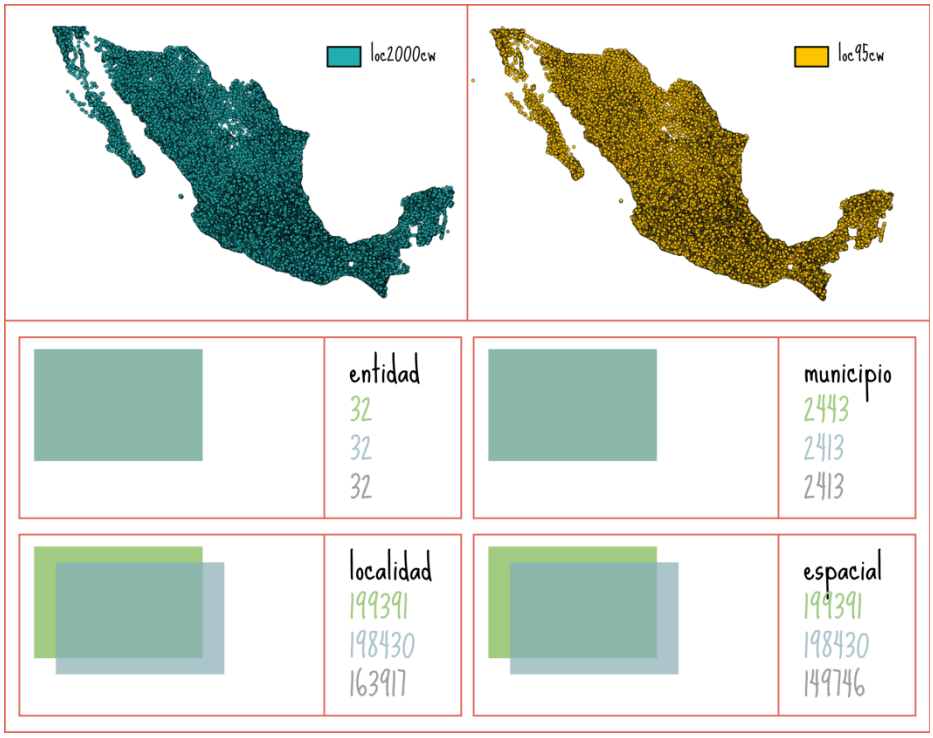


Figura 4.3: Prueba 1: loc2000cw - loc95cw.

Para el caso de los municipios, se observa casi el mismo fenomeno. La representacion loc95cw tiene el 100 % de sus municipios consistentes, esto es, sus 2413 municipios se encuentran también en la representacion loc2000cw. Mientras que para loc2000cw, el 98.77 % de sus representaciones se consideran consistentes.

Es en las localidades donde se comienza a observar la diferencia entre representaciones. La razon de que los conjuntos se observen del mismo tamaño es que loc2000cw tiene 199,391 localidades y loc95cw tiene 198,391. Sin embargo los conjuntos ya no estan totalmente acoplados, esto es porque solo 163,917 localidades se consideran consistentes. Esto es 82.2 % de las localidades de loc2000cw y 82.82 % de las de loc95cw son consistentes y así, solo ese porcentaje de cada conjunto se encuentra traslapado.

Para el analisis espacial el caso es similar. 149,746 instancias se consideran consistentes, lo cual produce una representación esquemática similar a la de localidades.

El grado de interoperabilidad se genera, como en todos los casos, con las componentes de estructura, tematica y espacial del analisis. Pero en esta prueba, por ser ambas representaciones de tipo punto, la componente espacial del grado

de interoperabilidad se genera únicamente a partir de las instancias espacial- mente consistentes. Esto es, que no se revisa relacion alguna entre instancias, simplemente se mide la distancia entre ellas y el total de instancias consistentes se toma como componente espacial para el grado de interoperabilidad semántica.

Prueba 2
Representaciones: locurbanas 1995 - loc95cw. Tipo de dato: punto - punto.
Grado de interoperabilidad: 0.90499778.

Nuevamente se comparan dos representaciones de tipo punto, pero en este caso, la diferencia entre representaciones es clara. En la Figura 4.4 se observa que la representacion locurbanas 1995 cuenta con menos instancias que loc95cw. El analisis a nivel entidad resulta igual, las 32 entidades son consistentes entre representaciones. El resultado para el caso de los municipios el resultado es muy parecido: el 100 % de los municipios de loc95cw resultan consistentes y para locurbanas 1995 el 99 %.

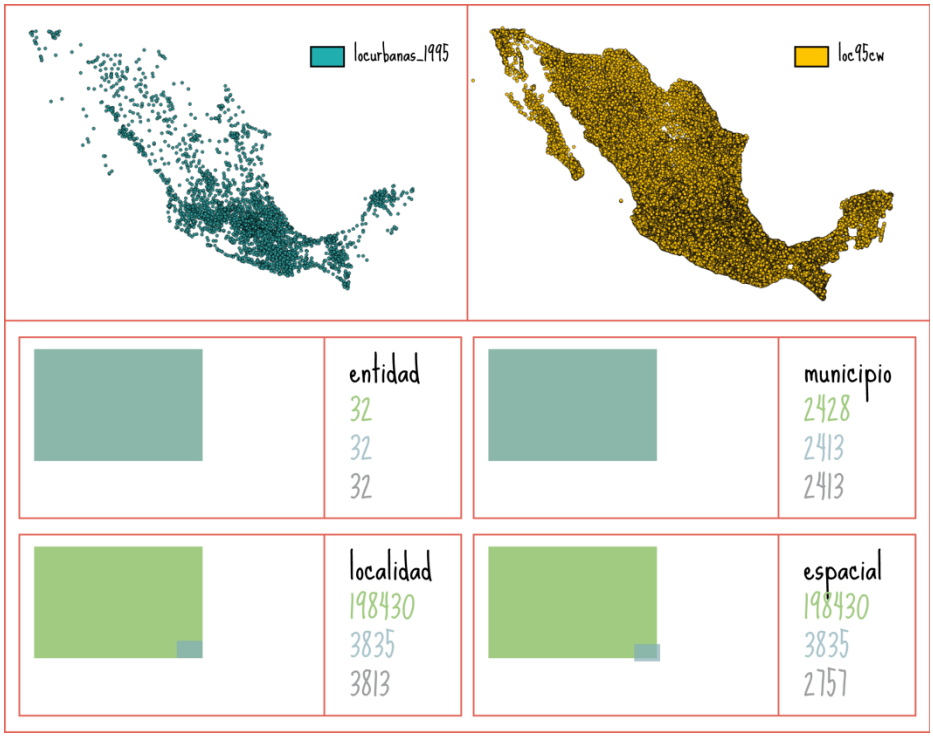


Figura 4.4: Prueba 2: locurbanas 1995 - loc95cw.

El cambio mas interesante se observa en el analisis de localidades. La representa-

ción esquemática muestra dos conjuntos de tamaño muy diferente, esto se debe a la cantidad de localidades en cada representacion (y por ende al tamaño de sus conjuntos) y concuerda con lo observado al inicio de la Figura 4.4. La representacion loc95cw tiene 198,430 localidades, mientras que locurbanas 1995 tiene solo 3,835, de las cuales 3,813 son consistentes, es decir mas del 99 %. Por este motivo, su conjunto se observa contenido totalmente en el conjunto de loc95cw.

Para el caso del analisis espacial, solo el 71.89 % de las localidades de locurbanas 1995 son consistentes, es por esto que se alcanza a observar el conjunto ligeramente fuera de loc95cw.

En esta prueba, al igual que en la Prueba 1, la componente espacial del grado de interoperabilidad semántica se toma como aquellas instancias espacialmente consistentes.

Prueba 3
Representaciones: locurbanas 1995 - polurbanos. Tipo de dato: punto - área.
Grado de interoperabilidad: 0.9956495.

En esta prueba se revisa la similitud entre representaciones con diferente tipo de dato. Comparamos nuevamente la representacion locurbanas 1995 que tie- ne punto como tipo de dato y la representacion polurbanos con tipo de dato area.

En la Figura 4.5 se muestra un detalle de la sobreposicion de ambas representaciones. Los resultados son similares a pruebas anteriores pero hay una característica particular, para esta prueba se revisa que los puntos se encuen- tren dentro de las areas; esto es, se revisa la relacion espacial existente entre instancias. Esto se realiza ya que, como se menciona en la Metodología, para datos tipo area se puede revisar algunas relaciones espaciales.

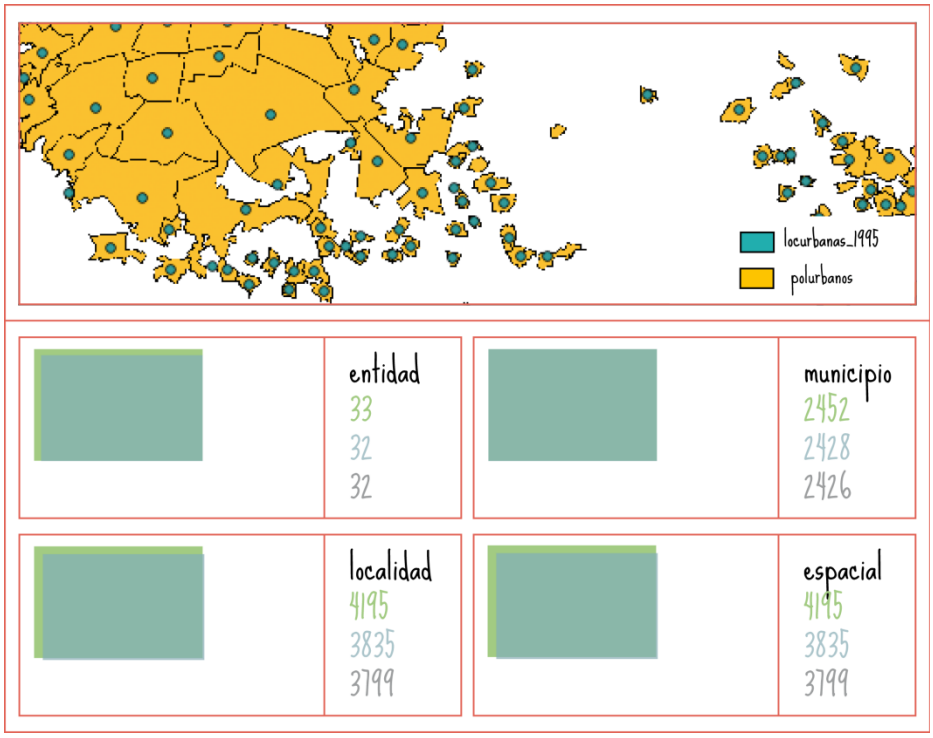


Figura 4.5: Prueba 3: locurbanas 1995 - polurbanos.

El resultado del analisis de relaciones espaciales modifica el grado de interoperabilidad, sin embargo no se toma en cuenta para la decision de consistencia entre localidades. La consistencia se determina de acuerdo a la distancia entre localidades y como se observa en los resultados, la mayor parte de las localidades pasaron la prueba. También se observa que las representaciones son muy similares en cada nivel de analisis.

En esta prueba, dado que existe un dato de tipo area, se revisa la relacion espacial Contains y se toma como la componente de similitud espacial para el grado de interoperabilidad semántica. En la prueba 3799 instancias se consideran espacialmente consistentes y de ellas 3798 cumplen la relacion espacial Contains, esto es, 3798 las localidades area contienen a sus localidades punto correspondientes.

Prueba 4
Representaciones: localidades urbanas - polurbanos. Tipo de dato: área - área.
Grado de interoperabilidad: 0.998857.

En esta prueba se analiza la similitud entre representaciones ambas con tipo

de dato area. El procedimiento es el mismo y los resultados similares, pero dado que se trabaja con datos de tipo area se revisan algunas relaciones espaciales.

En la Figura 4.6 se muestran las dos instancias correspondientes a la localidad de Paseo Nuevo, en Nogales, Veracruz. Las localidades se encuentran con un patron de rayas y se observa que son areas disjuntas. El sistema reporta otras dos localidades disjuntas, tres en total; 4,150 localidades que se intersectan; ninguna que únicamente se toquen (Touches). Estas son las relaciones espaciales que se revisan para analisis area-area.

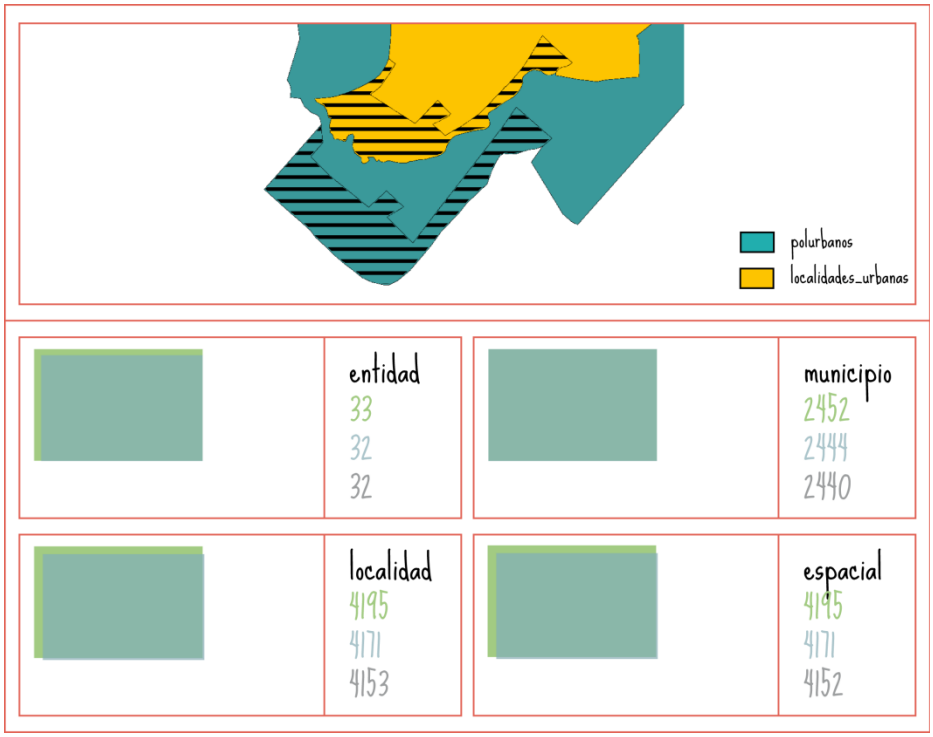


Figura 4.6: Prueba 4: polurbanos - localidades urbanas.

4.3. Grado de interoperabilidad semántica

En esta seccion se presentan los grados de interoperabilidad semantica que resultaron en cada prueba y, de acuerdo al tipo de dato en cada prueba, se explica como se generaron.

El grado de interoperabilidad es un valor que indica, con base en la similitud, qué tan interoperables son dos representaciones. Tiene un rango entre 1 y 0, y mientras mayor sea, mayor será la interoperabilidad entre las representaciones.

En este caso, dado que se tiene representaciones con tipo de dato area-area, la componente espacial del grado de interoperabilidad semántica se genera revisan- do tres relaciones espaciales: Disjoint - 3 instancias, Touches - 0 instancias e Intersects - 4150 instancias. Donde Intersects y Touches se consideran favora- bles al grado de interoperabilidad, no así Disjoint. De este modo se consideran 4147 instancias como consistentes para la componente espacial del grado de interoperabilidad.

5. Conclusiones

En este capítulo se presentan las conclusiones que se alcanzaron con el trabajo y se presenta también la propuesta de investigacion futura.

5.1. Conclusiones

Con este trabajo se encuentra la similitud y se genera un grado de interoperabili- dad entre representaciones espaciales de localidades de la República Mexicana. La similitud se describe en tres aspectos: estructural, tematico y espacial. A partir de cada aspecto de la similitud se propone un grado de interoperabilidad semántica que captura qué tan interoperables son dos representaciones.

Ademas de identificar la similitud semantica entre representaciones, el sistema describe la similitud espacial entre instancias. Específicamente genera una ta- bla en la base de datos con aquellas instancias consistentes donde almacena sus características tematicas así como su distancia. Esta distancia es la que se midió entre las instancias correspondientes (entre las representaciones) y que finalmente determin ó a las instancias como espacialmente consistentes.

La similitud espacial se describe también a nivel de relacion espacial. Para ins- tancias de tipo area, el sistema revisa su relacion espacial y como resultado da una lista de instancias espacialmente disjuntas, instancias que se intersectan, se tocan o que se contienen (en el caso de representaciones con tipo de dato punto y area). El revisar las relaciones espaciales hace al sistema mas descriptivo y ayuda al usuario a conocer la similitud espacial a mayor detalle.

Las representaciones y su similitud se trabajan con un enfoque de teoría de conjuntos, donde una representacion es vista como un conjunto definido por sus elementos (instancias de la representacion) y éstos a su vez se definen por sus propiedades o atributos. Se demuestra, con base en las pruebas y resultados (capítulo Pruebas), que con un enfoque de teoría de conjuntos puede trabajarse el problema de la similitud entre representaciones espaciales.

Algunas de las ventajas de trabajar las representaciones como conjuntos son: (1) que puede aplicarse un modelo de similitud de atributos (feature model), aceptado y probado como modelo para el establecimiento de la similitud; (2) al tratar la informacion como conjuntos, ésta se puede manejar dentro de una base de datos, que es un entorno de trabajo relativamente estable y robusto gracias a la tecnología madura existente; y (3) mientras las representaciones y sus propiedades puedan representarse como conjuntos formados por elementos, esta metodología es una opcion para el analisis de la similitud.

Para realizar un analisis de similitud sobre representaciones espaciales se es- tablece primero un criterio de comparacion para saber qué es lo que se va a

comparar. El sistema establece un criterio tematico de comparacion para despues analizar de forma espacial los datos. El criterio de comparacion revisa las características no-espaciales de los datos e identifica instancias tematicamente correspondientes entre representaciones. Una vez establecida la correspondencia tematica se revisa su consistencia espacial para determinar si las representaciones hablan del mismo objeto geografico.

De acuerdo al número de instancias consistentes se identifica la similitud entre representaciones. La similitud se revisa en tres aspectos principales: aspecto estructural, aspecto tematico y aspecto espacial. Con base en la similitud y sus tres aspectos, se propone un grado de interoperabilidad semántica que sirve como criterio para integracion o intercambio de informacion entre las representaciones.

La relacion de similitud se representa por medio de un grafico (llamado representación esquemática) y éste captura la asimetría de la relacion junto con otras características propias de los conjuntos (por ejemplo, su cardinalidad).

Se crean representaciones esquemáticas para capturar la similitud y se acompañan del texto analogo para mejorar su interpretacion. La razon de ocupar representaciones esquemáticas y texto es dar un resultado tanto descriptivo como preciso. La descripcion la da la representación esquemática (el grafico) mientras que la precision se encuentra en los resultados (el texto). De este modo se busca que el usuario reciba informacion clara, precisa y facil de entender.

5.2. Trabajos futuros

Encontrar el modo de probar el sistema, en particular el grado de interoperabilidad. Hallar la forma de conocer si el grado de interoperabilidad es correcto y útil para algunos sistemas o usuarios. Esto serviría como retroalimentacion para revisar, justificar y mejorar el desempeño del sistema.

Cambiar el criterio de comparacion. Actualmente el sistema tiene un criterio tematico de comparacion para evaluar espacialmente la informacion; esto es, se encuentran localidades con las mismas características tematicas y después se evalúa si son el mismo objeto geografico. Como trabajo a futuro se desea ahora crear un criterio espacial de comparacion y a partir de él revisar las características tematicas de los datos. Esto sería equivalente a encontrar localidades geograficamente cercanas y revisar posteriormente sus propiedades tematicas para determinar si corresponden al mismo objeto geografico.

Considerar un número mayor de relaciones espaciales. Actualmente, el sistema revisa, dependiendo del tipo de dato, cuatro relaciones espaciales. Al

agregar relaciones el sistema haría un analisis mas fino, sería mas descriptivo y se conseguiría un grado de interoperabilidad mas preciso.

Extender el analisis espacial a líneas. Establecer un vecindario conceptual de relaciones topologicas como [Li & Fonseca, 2006] donde se establezcan las relaciones (y sus distancias) entre datos de tipo punto, línea y area; y así ampliar el rango de aplicacion del sistema.

Actualmente el sistema compara cadenas de caracteres, como trabajo a futuro se proponer agregar una fase de analisis léxico para mejorar el analisis de consistencia tematica. Al agregar un analisis léxico se lograría una mejor comparacion, por ejemplo, con ella se podrían establecer como equivalentes las instancias: "Veracruz" y "Veracruz de Ignacio de la Llave". El estado actual del sistema no considera equivalentes estas instancias. Con el analisis léxico el sistema extendería su aplicacion a cualquier representacion espacial con conceptos expresados en forma de cadenas de caracteres.

Automatizar la importacion de representaciones al sistema. Para esto sería necesario dos aspectos; primero, se necesita de una forma automatizada de importar shapefiles; y segundo, un modelo de alineacion automatico que se encargue de identificar el conocimiento relacionado entre representaciones.

6. Referencias

[Anders & Bobrich, 2004]
Anders, K. H. & Bobrich, J. (2004). MRDB approach for automatic incremental update. ICA WORKSHOP on Generalisation and Multiple Representation. Leicester, Inglaterra.

[Bédard & Bernier, 2002]
Bédard, Y. & Bernier, E. (2002). Supporting multiple representation with spa- tial databases views management and the concept of "VUEL". ISPRS / ICA Joint Workshop on Multi-Scale Representations of Spatial Data. Ottawa, Ca- nada.

[Bergamaschi et al., 1998]
Bergamaschi, S., Castano, S., Vermercati, S., Montanari, S., & Vincini, M. (1998). An Intelligent Approach to Information Integration. En N. Guarino, Formal Ontology in Information Systems. Amsterdan, Países Bajos.

[Bernstein, 2003]
Bernstein, P. (2003). Applying Model Management to Classical Meta Data Problems. En Conference on Innovative Database Research. Asilomar, California. EE. UU.

[Bishr, 1997]
Bishr, Y. (1997). Semantic Aspects of Interoperable GIS. Thesis, Wageningen Agricultural University and International Institute for Aerospace Survey and Earth Science (ITC). Enshede, Países Bajos.

[Budak et al., 2006]
Budak Arpinar, I., Sheth, A., Ramakrishnan, C., Lynn Usery, E., Azami, M. & Kwan, M. (2006). Geospatial Ontology Development and Semantic Analytics. Handbook of Geographic Information Science. Ed. Blackwell.

[Chandrasekaran et al. 1999]
Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. (1999). What Are Ontologies? Why Do We Need Them? IEEE Intelligent Systems, 14(1), pp. 20-26.

[Couclelis, 1992]
Couclelis, H. (1992). People Manipulate Objects (but Cultivate Fields): Beyond the Raster-Vector Debate in GIS. En A. U. Frank, I. Campari & U. Formentini, Theories and Methods of Spatio-Temporal Reasoning in Geographic Space (Vol. 639, pp. 65-77). Springer-Verlag. Nueva York, EE. UU.

[Egenhofer et al., 1994]
Egenhofer, M., Clementini, E. & Di Felice, P. (1994). Evaluating inconsistencies

among multiple representations. En Proceedings of the 6th International Symposium on Spatial Data Handling (SDH94), pp. 901920.

[Egenhofer & Mark, 1995]
Egenhofer, M. & Mark, D. (1995). Naive Geography. En A. Frank and W. Kuhn (Eds.). Spatial Information Theory A Theoretical Basis for GIS (COSIT 95), Semmering, Austria (Lecture Notes in Computer Science, 988). Berlin: Springer, 1-15.

[Frank, 2001]
Frank, A. (2001). Tiers of Ontology and Consistency Constraints in Geographical Information Systems. International Journal of Geographical Information Science, 15(7), 667-678.

[Fonseca et al., 2006]
Fonseca, F., Cmara, G. & Monteiro, A. M, (2006). A Framework for Measuring the Interoperability of Geo-Ontologies. Spatial Cognition & Computation, Vol. 6, No. 4

[Fonseca & Egenhofer, 1999]
Fonseca F., & Egenhofer M. (1999). Ontology-Driven Geographic Information Systems. En Medeiros C B, (Ed.) 7th ACM Symposium on Advances in Geographic Information Systems, Kansas City, Misuri, EE. UU.

[Gardenfors, 2000]
Gardenfors P. (2000). Conceptual Spaces: The Geometry of Thought. , MIT Press. Cambridge, Massachusetts, EE. UU.

[Gentner & Markman, 1995]
Gentner, D. & Markman, A. B. (1995). Similarity is like analogy: Structural alignment in comparison. En Cacciari, C. (ed) Similarity in Language, Thought and Perception. Ed. Brepols: 11147. Bruselas, Bélgica.

[Goldstone & Medin, 1994]
Goldstone, R. L. & Medin, D. L. (1994). Similarity, interactive activation and mapping: An overview. En Barnden, J. and Holyoak, K. J. (eds) Advances in Connectionist and Neural Computation Theory: Volume 2, Analogical Connections. Ed. Ablex. Norwood, Nueva Jersey, EE. UU.

[Goldstone & Son, 2005]
Goldstone, R. L. & Son, J. (2005). Similarity. En Holyoak K and Morrison R (ed) Cambridge Hand- book of Thinking and Reasoning. Cambridge University Press: 1336. Cambridge.

[Goodchild, 1992]
Goodchild, M. (1992). Geographical Data Modeling. Computers and Geoscien-

ces, 18(4), 401-408.

[Goodchild et al., 1999]
Goodchild, M., Egenhofer, M., Fegeas, R., & Kottman, C. (1999). Interoperating Geographic Information Systems. Norwell, MA: Kluwer Academic.

[Goodchild & Hunter, 1997]
Goodchild, M. F., Hunter, G. J. (1997). A Simple Positional Accuracy Measure for Linear Features. Int. J. Geogr. Inf. Sci. 11(3), 299306.

[Gruber, 1992]
Gruber, T. (1992). A Translation Approach to Portable Ontology Specifications (Technical Report No. KSL 92-71). Stanford, CA: Knowledge Systems Labora- tory, Stanford University.

[Guarino, 1998]
Guarino, N. (1998). Formal Ontology and Information Systems. In N. Guarino (Ed.), Formal Ontology in Information Systems (pp. 3-15). Amsterdam, Netherlands: IOS Press.

[Kahng & McLeod, 1998]
Kahng, J. & McLeod, D. (1998). Dynamic Classificational Ontologies: Mediation of Information Sharing in Cooperative federated Database Systems. en: Papazoglou, M. & Schlageter, G. Cooperative Information Systems: Trends and Directions. Academic Press, London, UK.

[Krumhansl, 1978]
Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. Psychological Review 85: 445 63.

[Hahn et al., 2003]
Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. Cognition 87: 132.

[Hahn & Chater, 1997]
Hahn, U. & Chater, N. (1997). Concepts and similarity. In Lamberts, K. and Shanks, D. (eds) Knowledge, Concepts and Categories. Hove, UK, Psychology Press: 4392.

[Halevy et al., 2006]
Halevy, A., Rajaraman, A., & Ordille, J. (2006). Data integration: The teenage years. In VLDB, pages 916.

[Harvey et al., 1999]
Harvey, F., Kuhn, W., Pundt, H., Bishr, Y., & Riedemann, C. (1999). Semantic

Interoperability: A Central Issue for Sharing Geographic Information. Annals of Regional Science, 1999. 33 (2)(Geospatial data sharing and standardization): pp. 213-232.

[Imai, 1977]
Imai, S. (1977). Pattern similarity and cognitive transformations. Acta Psychologica 41: 43347.

[Kalfoglou & Schorlemmer, 2003]
Kalfoglou, Y. & Schorlemmer, M. (2003). Ontology mapping: the state of the art. The Knowledge Engineering Review, 18(1), 1-31.

[Kashyap & Seth, 1996]
Kashyap, V. & Sheth, A. (1996). Semantic Heterogeneity in Global Informa- tion System: The Role of Metadata, Context and Ontologies. In M. Papazoglou & G. Schlageter (Eds.), Cooperative Information Systems: Current Trends and Directions (pp. 139-178). London: Academic Press.

[Lenat & Guham, 1990]
Lenat, D. & Guham, R. (1990). Building Large Knowledge Based Systems: Representation and Inference in the Cyc Project. Addison-Wesley Publishing Company, Reading, MA.

[Lenzerini, 2002]
Lenzerini, M. (2002). Data integration: A theoretical perspective. In Proceedings of the Symposium on Principles of Database Systems (PODS), pages 233246.

[Li & Fonseca, 2006]
Li, B. & Fonseca, F. T. (2006). TDD - A Comprehensive Model for Qualitative Spatial Similarity Assessment. Spatial Cognition and Computation.

[Longley et al., 2001]
Longley, P. A., Goodchild, M. F., Maguire, D. J. & Rhind, D. W. (2001). Geographic Information Systems and Science. John Wiley & Sons, Chichester, West Sussex, England.

[Meersman, 1997]
Meersman, R. (1997). An Essay on The Role and Evolution of Data(base) Semantics. en: Meersman, R. & Mark, L. (Eds.) DataBase Application Semantics. Chapman Hall, Longon, UK.

[Melnik et al., 2002]
Melnik, S., Garcia-Molina, H. & Rahm, E. (2002). Similarity flooding: a versitile graph matching algorithm and its application to schema matching. In: Proceedings of the International Conference on Data Engineering 2002, June 36, Madison, WI.

[Mena et al., 1996]
Mena, E., Kashyap, V., Sheth, A., & Illarramendi, A. (1996). OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies. Paper presented at the First IF- CIS International Conference on Cooperative Information Systems (CoopIS'96), Brussels, Belgium.

[Mitchell, 1997]
Mitchell, T. M. (1997). Machine Learning, 1st edn. McGraw-Hill, New York.

[Mustière et al., 2009]
Mustière, S., Reynaud, C., Safar, B. & Abadie, N. (2009). Same words? Same worlds? Comparing ontologies underlying geographic data.

[OpenGIS, 1996]
OpenGIS. (1996). The OpenGIS Guide-Introduction to Interoperable Geoprocessing and the OpenGIS Specification. Wayland, MA: Open GIS Consortium, Inc.

[Pottinger, 2008]
Pottinger, R. (2008). Database schema integration. En S. Shekhar and H. Xiong. Encyclopedia of GIS.

[Rada et al. 1989]
Rada, R., Mili, H., Bicknell, E., & Bletner, M. (1989). Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics 19: 1730.

[Resnik, 1995]
Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. En Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI), Montreal, Quebec.

[Rodríguez & Egenhofer, 2004]
Rodríguez, M. A. & Egenhofer, M. J. (2004). Comparing Geospatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure. International Journal of Geographical Information Science.

[Rodríguez et al., 1999]
Rodríguez, M. A., Egenhofer, M. J. & Rugg, R. (1999). Assessing Semantic Similarity Among Geospatial Feature Class Definitions. In A. Vckovski, K. Brassel & H.-J. Schek (Eds.), Interoperating Geographic Information SystemsSecond International Conference, INTEROP'99 (Vol. 1580, pp. 1-16). Berlin: Springer-Verlag.

[Samal et al., 2004]
Samal, A., Seth, S. & Cueto, K. (2004). A feature-based approach to conation of geospatial sources. Int. J. GIS 18(5), 459589.

[Scappapietra et al., 2000]
Spaccapietra, S., Parent, C. & Vangenot, C. (2000). From Multiscale to Multi-representation. In Choueiry, B., Walsh, T., eds.: Proceedings 4th International Symposium, SARA-2000, Horseshoes Bay, Texas, USA, Springer-Verlag, LNAI 1864.

[Schwering, 2008]
Schwering, A. (2008). Approaches to semantic similarity measurement for geo-spatial data - a survey. Transactions in GIS, 12(1):529.

[Sharad & Ashok, 2008]
Sharad, S. & Ashok, S. (2008). Conflation of Features. En S. Shekhar and H. Xiong (Eds). Encyclopedia of GIS.

[Sheeren et al., 2009]
Sheeren, D., Mustière, S., Zucker, J. D. (2009). A datamining approach for assessing consistency between multiple representations in spatial databases. Int. J. of Geographical Information Science. Vol. 23.

[Sheth, 1998]
Sheth, A., (1998). Changing Focus on Information Systems: From System, Syn- tax, Structure to Semantics. en: M. Goodchild, M. Egenhofer, R. Fegeas, y C. Kottman, Interoperating Geographic Information Systems. Kluwer Academic Press.

[Sheth & Larson, 1990]
Sheth, A., & Larson, J. (1990). Federated Databases Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. ACM Computing Sur- veys, 22(3), 183-236.

[Sloman et al., 1998]
Sloman, S. A., Love, B. C. & Woo-Kyoung, A. (1998). Feature centrality and conceptual coherence. Cognitive Science 22: 189228.

[Smith, 2003]
Smith, B. (2003). Ontology. In L. Floridi (Ed.), The Blackwell Guide to the Philosophy of Computing and Information (pp. 155-166). Malden, MA: Blackwell.

[Smith & Mark, 1998]
Smith, B., & Mark, D. (1998). Ontology and Geographic Kinds. Paper presented at the International Symposium on Spatial Data Handling, Vancouver, BC, Canada.

[Stoimenov & Djordjevic-Kajan, 2002]
Stoimenov, L. & Djordjevic-Kajan, S. (2002). Framework for Semantic GIS Interoperability, FACTA Universitatis, Series Mathematics and Informatics, Vol.17 (2002), pp.107-125.

[Taylor & Ives, 2006]
Taylor, N. & Ives, Z. (2006). Reconciling while tolerating disagreement in collaborative data sharing. In Proc. of SIGMOD.

[Tversky, 1997]
Tversky, A. (1977). Features of similarity. Psychological Review, 84, 327-352.

[Vangenot et al., 2002]
Vangenot, C., Parent, C., & Spaccapietra, S. (2002). Modeling and Manipulating Multiple Representations of Spatial Data. En Proceedings of Spatial Data Handling Conference SDH02 (Ottawa).

[Wache et al., 2001]
Wache, H., Voegele, T., Visser, U., Stuckenschmidt, H., Schuster, G. & Neumann, H. (2001). Ontology-based integration of information - a survey of existing approaches. Paper presented at the IJCAI-01 Workshop on Ontologies and Information Sharing, Seattle, WA.

[Weibel & Dutton, 1999]
Weibel, R. & Dutton, G. (1999). Generalising spatial data and dealing with multiple representations. In: P.A. Longley, M.F. Goodchild, D.J., Maguire & D.W. Rhind, editors, Geographic Information Systems Principles and Techni- cal Issues, volume 1. John Wiley & Sons, 2 edition, pp. 125155.

[Weiderhold, 1994]
Wiederhold, G. (1994). Interoperation, Mediation and Ontologies. Paper presented at the International Symposium on Fifth Generation Computer Systems (FGCS94), Tokyo, Japan.

[Widom, 2005]
Widom, J. (2005). Trio: A System for Integrated Management of Data, Accuracy, and Lineage. In Proc. of CIDR.

[Wiener-Ehrlich et al., 1980]
Wiener-Ehrlich, W. K., Bart, W. M. & Millward, R. (1980). An analysis of generative representation systems. Journal of Mathematical Psychology 21: 21946.

[HTML5, 2011] http://
www.w3.org/TR/html5/
Consultado 14-10-2011

[Protégé, 2011] http://
protege.stanford.edu/overview/
Consultado 14-10-2011

[PostgreSQL, 2011] http://
www.postgresql.org/about/
Consultado 14-10-2011

[PostGIS, 2011] http://
postgis.refractions.net/
Consultado 14-10-2011

[Jena, 2011] http://
jena.sourceforge.net/
Consultado 14-10-2011

[Protégé-Jena, 2011] http://protege.stanford.edu/plugins/owl/jena-integration.html Consultado 14-10-2011

[Resource Description Framework, 2011]
http://www.w3.org/RDF/
Consultado 14-10-2011

[W3C, 2011] http://www.w3.org/TR/owl-features/ Consultado 14-10-2011

[Web-Ontology Working Group, 2011]
http://www.w3.org/2001/sw/WebOnt/
Consultado 14-10-2011

7. Anexo A - Marco Teórico

En este capítulo se describen las herramientas utilizadas para el desarrollo de la tesis. Las herramientas descritas son: Protégé, Jena, PostgreSQL, PostGIS y HTML5.

En la tesis, Protegé se utiliza para crear la ontología (en OWL) y alinear las representaciones. Al momento de importar representaciones, éstas se alinean de forma manual cuando se introducen como individuos a la ontología. Una vez que las representaciones se encuentra dentro del sistema y como la ontología se creó en OWL, se utiliza Jena para realizar consultas. Desde Java se utiliza el API de Jena para realizar consultas y conocer los metadatos de las representa- ciones a analizar.

PostgreSQL es la base de datos que contiene la informacion de las representaciones. Se usa su extension espacial PostGIS para almacenar y revisar la componente espacial de los datos.

Finalmente, cuando se conoce la similitud entre las representaciones, se ocupa HTML5 para desplegar los resultados de forma grafica.

7.1. Protégé

Protégé es una plataforma de codigo libre, basada en Java, que provee herramientas para construir modelos de dominio y aplicaciones basadas en conocimiento, usando ontologías. Implementa un conjunto de estructuras para modelado de conocimiento y acciones que soportan la creacion, visualizacion y manipulacion de ontologías en varios formatos de representacion. Se puede personalizar para proveer un soporte amigable del dominio para crear modelos de conocimiento e introducir datos. También se puede expandir su funcionalidad por medio de otros componentes basados en Java [Protégé, 2011].

7.2. Jena

Jena es un marco de trabajo de Java para construir aplicaciones semanticas en la web. Provee un entorno pragmatico para RDF, OWL, SPARQL e incluye un mecanismo de inferencia basado en reglas [Jena, 2011].

Jena es una de las APIs mas usadas para RDF y OWL de codigo libre. Provee herramientas para servicios sobre modelos de representacion como analisis, persistencia en bases de datos, consultas y visualizacion [Protégé-Jena, 2011].

RDF (Resource Description Framework) es un modelo estandar para el intercambio de datos en la web. RDF tiene características que facilitan la fusion de datos incluso para diferentes esquemas, y específicamente soporta la evolucion

de esquemas en el tiempo sin requerir modificar a todos consumidores [Resource Description Framework, 2011].

OWL (Ontology Web Language) es un lenguaje de ontologías y un estandar de la red semantica que se diseñó para ser usado por aplicaciones que necesitaran procesar el contenido de la informacion en lugar de solo presentar datos a los usuarios [Web-Ontology Working Group, 2011]. Provee un marco de trabajo pa- ra establecer la administracion, integracion, distribucion y reutilizacion de datos en la web. También facilita la interoperabilidad del contenido web mas que los esquemas XML y RDF al proveer vocabulario adicional junto con semantica formal [W3C, 2011].

7.3. PostgreSQL

PostgreSQL es una poderosa base de datos objeto-relacional de codigo libre. Tiene interfaces de programacion nativa para C/C++, Java, .Net, Perl, Pyt- hon, Ruby y otras. Ademas, con PostGIS, puede manejar objetos geograficos [PostgreSQL, 2011].

7.4. PostGIS

PostGIS agrega el soporte para objetos geograficos a PostgreSQL. Habilita es- pacialmente el servidor de PostgreSQL para ser el soporte de sistemas de in- formacion geografica, de forma similar como el SDE de ESRI o la extension espacial de Oracle. También sigue las especificaciones de atributos simples para SQL establecidos por el OpenGIS [PostGIS, 2011].

7.5. HTML5

HTML5 es la quinte revision del HTML. En esta version, nuevas características se han agregado para ayudar a autores de aplicaciones web. Las nuevas carac- ter ísticas se decidieron con base en la investigacion de las practicas de autores de la web y se puso especial atencion en definir los criterios de conformidad para usuarios, con el fin de mejorar la interoperabilidad [HTML5, 2011].

La razon de usar HTML5 es que cuenta con un elemento llamado lienzo o can- vas. El elemento lienzo es una herramienta que puede dibujar graficas, mostrar graficas de juegos, y desplegar otro tipo de imagenes sobre la marcha.