Appendix A

RESUMEN EN ESPAÑOL

A.1. Introducción

En los últimos años hemos sido testigos del rápido crecimiento en el número, capacidad y diseminación de aplicaciones informáticas dedicadas a la obtención, generación, manipulación y almacenamiento de datos en diversos ámbitos de la vida humana. Esto ha propiciado una gran cantidad de colecciones de datos cuyo análisis por medios manuales se vuelve una tarea complicada. Recordemos que en muchas ocasiones los datos "crudos" necesitan ser analizados e interpretados para convertirlos en información útil y provechosa. Tal situación ha propiciado una creciente necesidad por técnicas/herramientas computacionales que nos ayuden en estas tareas. Descubrimiento de conocimiento en bases de datos (*KDD*, por sus siglas en el idioma Inglés) es definido como la extracción no trivial de información implícita, previamente desconocida y potencialmente útil a partir de datos [16]. Este es un proceso iterativo e interactivo que envuelve diferentes fases. El núcleo del proceso es la fase de minado de datos, que se conceptualiza como la aplicación de algoritmos de análisis de datos y de descubrimiento que bajo parámetros aceptables de

eficiencia computacional producen/descubren una enumeración particular de patrones sobre los datos mismos [12].

En este mismo contexto, pero enfocado al análisis y explotación de datos provenientes de fenómenos generados en, sobre, y bajo la superficie de la tierra, llamados datos espaciales, ha generado un nuevo dominio de investigación llamado Minería de datos Espaciales. De tal forma, la Minería de datos espaciales se enfoca al descubrimiento de conocimiento implícito, y previamente desconocido en datos espaciales [16]. Como resultado de esta necesidad creciente diversos enfoques para el minado de datos espaciales han sido desarrollados, entre los más representativos encontramos:

A.1.1 Métodos basados en generalización

La generalización ha demostrado ser uno de los métodos efectivos para descubrir conocimiento. Fue introducido por la comunidad de aprendizaje máquina y se basa en el aprendizaje a partir de ejemplos. El descubrimiento de conocimiento basado en generalización requiere jerarquías de conceptos (dadas explícitamente por el experto ó generadas automáticamente). En la caso de las bases de datos espaciales, pueden darse dos tipos de jerarquías de conceptos: (1) Jerarquías temáticas, por ejemplo, generalizar tomates y plátanos como frutas, las frutas y vegetales como alimentos de origen vegetal. (2) Jerarquías espaciales, por ejemplo, generalizar una serie de puntos espaciales como una región ó país.

Lu et al. [35] extienden la técnica *attribute-oriented induction* a las bases de datos espaciales. Esta técnica se basa en escalar la jerarquía de generalización e ir resumiendo las relaciones entre los datos espaciales y no espaciales a un nivel de concepto más alto. Los autores presentan dos algoritmos basados en generalización: (1) Enfoque de dominación de datos no espaciales. Este método realiza en primera instancia inducción orientada al atributo sobre los datos no-espaciales y posteriormente sobre los espaciales. (2) Enfoque de dominación de datos espaciales. Dado la jerarquía de datos espaciales, la generalización se realizado primero sobre estos datos y posteriormente sobre los datos no espaciales.

A.1.2 Agrupamiento

Agrupamiento (*clustering*) es el proceso de agrupar de manera física ó abstracta objetos en clases de objetos similares. Este enfoque de minería de datos nos ayuda a construir particiones "representativas" de un conjunto de objetos dada una medida de similitud/distancia (Ej. distancia euclidiana). Esto es, el agrupamiento de datos identifica grupos (*clusters*) ó regiones densamente pobladas de acuerdo a alguna medida de distancia en un conjunto de datos multidimensionales. Podemos clasificar a los algoritmos de agrupamiento en cuatro grupos principales: Algoritmos de particionamiento basados en los enfoques *k-means* (centro de gravedad del cluster) y *k-medoid* (objeto representativo del cluster), algoritmos jerárquicos, algoritmos basados en la ubicación de los objetos (agrupamiento por densidad), y por último los basados en *grids*.

A.1.3 Asociaciones espaciales

Una regla de asociación espacial es una regla que describe la implicación de uno o un conjunto de objetos por otro conjunto de objetos en base de datos espaciales [29]. Un ejemplo de una regla de asociación espacial podría ser "si la empresa se ubica cerca de la Ciudad de México entonces es una empresa grande". Una regla de asociación espacial es de la forma $X \rightarrow Y$, donde X y Y son conjuntos de predicados espaciales o no espaciales. Existen varios tipos de predicados espaciales que pudieran constituir una regla de asociación espacial, por ejemplo, relaciones topológicas como son intersección, traslape y orientaciones espaciales tales como Izquierda_de, y Oeste_de.

A.1.4 Aproximación y agregación

Los métodos basados en aproximación y agregación buscan analizar las características de grupos de objetos (*clusters*) en base a objetos (*features*) cercanos a ellos. Proximidad agregada es la medida de cercanía de un conjunto de puntos en un cluster a un *feature*. La idea de encontrar relaciones de proximidad no es un problema simple como podría parecer, existen tres razones para esta aseveración. Supongamos que tenemos un cluster de puntos y queremos encontrar los *k-features* más cercanos a él:

- El tamaño y forma del *cluster* y los *features* puede ser muy variado.
- Podríamos tener una gran cantidad de *features* para examinar.
- Aún en el caso de encontrar una forma conocida (Ej. polígono) que describa la forma del *cluster*, sería inadecuado reportar los *features* cuyos límites estén más

cerca a los límites de éste, porque la distribución de los puntos al interior del *cluster* puede no ser uniforme.

A.1.5 Minería de datos en imágenes

Extracción de patrones a partir de imágenes es otro enfoque de minería de datos espaciales. En la literatura existen diversos trabajos de minado de datos desarrollados bajo este enfoque. Por ejemplo, Fayyad et al. [14] presentan un sistema para la identificación y categorización de volcanes en la superficie de Venus a partir de imágenes transmitidas por la sonda espacial *Magellan*. En otro trabajo [15] (*Second Palomar Observatory Sky Survey*) se usaron árboles de decisión para la clasificación de galaxias, estrellas y otros objetos estelares. Stolorz et al. [44] y Shek et al. [41] efectuaron estudios sobre minería de datos espacio-temporal en conjuntos de datos geofísicos.

A.1.6 Clasificación de datos espaciales

La clasificación de datos espaciales tiene como objetivo encontrar reglas que dividan un conjunto de objetos en un número de grupos, donde los objetos de cada grupo pertenecen a una clase. Diversos tipos de información pueden ser usados para caracterizar los objetos espaciales. Por ejemplo, atributos no espaciales de un objeto, predicados espaciales y funciones espaciales. La idea es usar esta información para extraer ya sea atributos para la etiquetación de clases (atributos que dividen los datos en clases) y atributos predictivos (atributos cuyos valores son usados en un árbol de decisión para crear sus ramas).

A.1.7 Detección de tendencias espaciales

Detección de tendencias espaciales describe los cambios regulares de uno ó más atributos no espaciales de un objeto cuando éste se desplaza desde un punto de referencia inicial. Un ejemplo de tendencia espacial sería "alejándose del centro histórico de la ciudad de Puebla es precio de los terrenos decrece". Las trayectorias de movimiento a partir de un punto *x* son usadas para modelar dicho movimiento y análisis de regresión sobre los atributos de los objetos son usados para describir patrones de cambio. Existen dos tipos de tendencias: globales y locales.

A.2. Motivación

Aunque los enfoques antes mencionados realizan la tarea de minado de datos espaciales de manera exitosa, nuestra percepción es que éstos no consideran todos los elementos encontrados en una base de datos espaciales (datos espaciales, datos no espaciales y relaciones espaciales entre los objetos espaciales) de una manera integral. Es decir, algunos de ellos primero realizan minado de datos espaciales y posteriormente minado de datos no espaciales ó viceversa, y otros permiten combinaciones de estos elementos pero de manera restringida. Con base en lo anterior, proponemos el argumento siguiente: si somos capaces de representar los datos espaciales, no espaciales y las relaciones entre objetos espaciales como un solo conjunto de datos, y lo minamos como tal, podríamos generar/encontrar patrones de conocimiento que describan/caractericen nuestro conjunto de datos conteniendo estos tres elementos de manera conjunta. Para tal efecto en este trabajo se argumenta que

una representación basada en grafos es lo suficientemente flexible y poderosa para representar estos elementos de manera conjunta, fácilmente entendible y capaz de crear diferentes representaciones del mismo dominio. El dominio es descrito usando grafos, los grafos se convierten en datos de entrada para una herramienta de descubrimiento de conocimiento basada en grafos la cual usa heurísticas para seleccionar subgrafos que son considerados importantes (patrones).

Un grafo es definido como un par $G = (V,E)$. $V = \{v_1,\ldots,v_n\}$ denota un conjunto finito de elementos llamados vértices. E es un conjunto de arcos e satisfaciendo $E \subseteq [V]^2$. Entonces, cada arco $e \in E$ es un par (v_i,v_j). Si (v_i,v_j) es un par ordenado para cualesquiera $(v_i,v_j) \in E$, entonces se dice que $G = (V,E)$ es un grafo dirigido. Un grafo etiquetado tiene etiquetas asociadas a sus vértices y arcos.

Como comentamos anteriormente en nuestro modelo proponemos la representación de relaciones espaciales entre los objetos espaciales. En la siguiente subsección detallamos los tres tipos de relaciones espaciales que proponemos incluir.

A.2.1 Relaciones espaciales

La ubicación explícita de los objetos espaciales define relaciones implícitas de vecindad (*neighborhood*) espacial entre ellos. De tal forma, la información sobre la vecindad de los objetos espaciales constituye un elemento valioso que debe ser considerado en la tarea de minado de datos espaciales. Martin Ester et al. [9][11] introducen el concepto de grafos de

vecindad para representar explícitamente estas relaciones de vecindad implícitas. Los grafos de vecindad pueden cubrir las relaciones de vecindad siguientes:

- Topológicas. Derivadas del modelo de nueve intersecciones [6][7][8], son relaciones que permanecen invariantes bajo transformaciones lineales, es decir, si ambos objetos se rotan, se trasladan ó se escalan simultáneamente las relaciones entre ellos se preservan.
- Distancia. Compara la distancia entre dos objetos dada una constante usando operadores aritméticos tales como <, >, =. La distancia entre dos objetos se definen como la distancia mínima entre ellos (Ej. seleccionar todos los elementos dentro de una radio de 50 kilómetros desde un punto *x*).
- Dirección. La relación espacial de dirección entre 2 objetos espaciales *A* y *B* (*B R A*) se define usando un punto representativo del objeto *A* y todos los puntos del objeto destino *B*. El punto representativo del objeto fuente *A* puede ser el centro del objeto ó un punto sobre sus límites. Este punto representativo es usado como el origen de un sistema de coordenadas virtuales y su cuadrante define la dirección.

Una vez comentada la motivación de nuestro trabajo de investigación se presenta a continuación el modelo general basado en grafos para representar los datos espaciales.

A.3 Representaciones basadas en grafos

Como hemos comentado, nuestra propuesta se basa en crear un modelo basado en grafos para representar conjuntamente datos espaciales, no espaciales y relaciones espaciales entre

los objetos espaciales. La idea es usar los grafos generados como datos de entrada para un algoritmo de minado de datos, de tal forma, que el algoritmo pueda encontrar patrones involucrando estos elementos de manera conjunta y no como elementos separados. En consecuencia, proponemos el modelo de representación mostrado (en notación *UML*) en la Figura A.1.

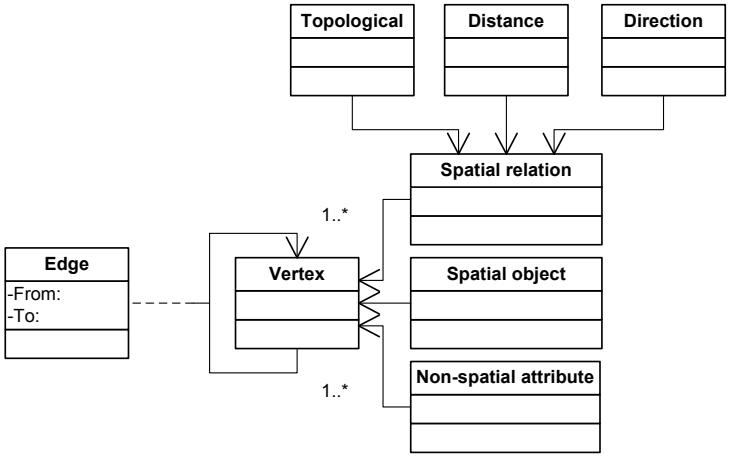

Figura A.1. Modelo basado en grafos para representar datos espaciales.

En el modelo, los datos espaciales (Ej. objetos espaciales), datos no espaciales (Ej. atributos descriptivos), y relaciones espaciales son representados como una colección de uno ó más grafos dirigidos etiquetados. Los vértices pueden representar objetos espaciales, tipos de relación espacial entre dos objetos (relación binaria), ó atributos no espaciales describiendo los objetos espaciales. Los arcos representan una liga existente entre dos vértices de cualquier tipo. Dependiendo del tipo de vértices que un arco une, éste puede representar el nombre de un atributo descriptivo ó el nombre de una relación espacial. Los nombres de atributos pueden referirse a descripciones de objetos espaciales y/ó a entidades no espaciales. Se usan arcos dirigidos para representar la información direccional de relaciones

entre elementos (Ej. objeto *x* cubre objeto *y*) y para describir atributos pertenecientes a objetos (Ej. objeto *x* tiene atributo *z*).

Actualmente se han desarrollado cinco representaciones a partir del modelo general previamente descrito. Tres tópicos definen las características de los grafos creados en cada modelo: (1) Representación de las relaciones espaciales equivalentes (Ej. tocar, traslapar). (2) Representación de las relaciones espaciales simétricas (Ej. contiene-dentro_de, Norte_de-Sur_de). (3) La manera de representar los objetos y sus relaciones en el modelo. En consecuencia, los grafos creados se diferencian de manera cuantitativa y cualitativa. En la parte cuantitativa tenemos diferencias tales como: número de vértices y arcos empleados para representar los datos espaciales, no espaciales y las relaciones, creación de grafos simples, manejo de arcos dirigidos y no dirigidos para representar relaciones espaciales y/o atributos descriptivos. En la parte cualitativa hemos observado, a través de experimentación, que ciertos modelos tienen una mayor expresividad para representar el conjunto de datos, condicionando de manera directa la calidad de los resultados generados en el proceso de minado. A continuación se presentan 3 de los 5 modelos propuestos describiendo las métricas de evaluación creadas para caracterizar a cada uno de ellos.

Modelos

Con la finalidad de describir las características de cada modelo, usaremos como conjunto de datos de ejemplo los mostrados en la Figura A.2. Como podemos observar, nuestro conjunto de datos se compone de dos objetos espaciales, objeto *A* representado una casa y objeto *B* representando un lago, y las tres relaciones espaciales siguientes: (1) Distancia,

objeto *A* cerca de objeto *B* (relación equivalente). (2) Topológica, objeto *A* toca objeto *B* (relación equivalente). (3) Dirección, objeto *A* Sur de objeto *B* (relación simétrica).

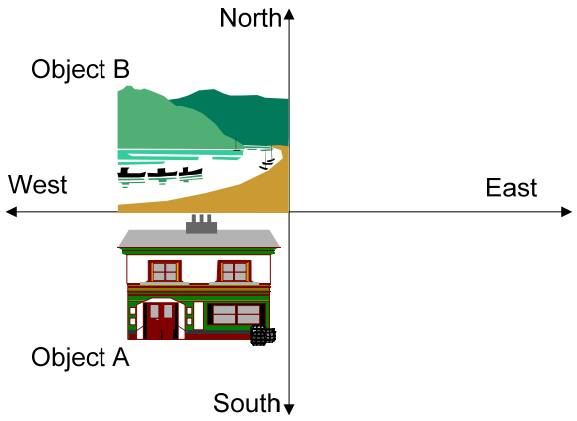

Figura A.2. Base de datos de ejemplo para caracterizar los 3 modelos propuestos.

Modelo #1 - modelo base

La Figura A.3 muestra el primer modelo creado para representar datos espaciales bajo el enfoque propuesto. Las características del modelo de acuerdo a las métricas creadas para su caracterización son:

- **Num. vértices:** 2 vértices, cada uno representando un objeto espacial (objeto *A* y objeto *B*).
- **Num. arcos:** 4 arcos, 3 arcos para representar las relaciones espaciales originales existentes en nuestro conjunto de datos de ejemplo (“cerca”, “toca” y “Sur_de”) y un arco para representar la relación “Norte_de” creada de la relación simétrica original “Sur_de”. La relación “Norte_de” es en si misma una relación simétrica.
- **Tamaño (vértices + arcos):** 6
- **% incremento:** 0%, este es el *modelo base*.
- **Grafo simple.** No, Es un grafo complejo con 4 arcos uniendo 2 vértices.

- **Arco dirigido.** Si, son usados para representar las relaciones simétricas “Sur_de” y “Norte_de”. La dirección de los arcos va en concordancia con la lectura de las relaciones entre los objetos.
- **Arco no dirigido.** Si, son usados para representar las relaciones equivalentes “cerca” y “toca”.
- **Información completa.** Si, en el grafo es representada la relación simétrica “Norte_de” creada a partir de la relación simétrica original “Sur_de”.
- **Arco “Relation” redundante.** No, en el modelo no se usan arcos “Relation”.

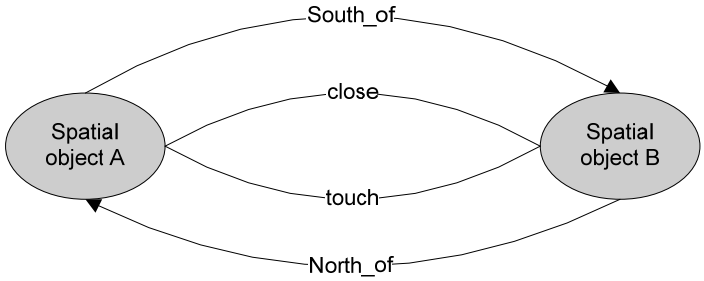

Figura A.3. Modelo #1 - modelo base.

Modelo #2 - replicación simple de tipos de relación, información completa

En la Figura A.4 presentamos el segundo modelo creado para representar datos espaciales. Las características del modelo de acuerdo a las métricas son:

- **Num. vértices:** 5 vértices, 2 vértices para representar los objetos espaciales y 3 vértices para representar los tipos de relaciones espaciales “topológica”, “distancia” y “dirección”. Por cada tipo de relación especial existente entre dos objetos, se añade un vértice etiquetado con el nombre del tipo de la relación espacial. En el ejemplo existen una relación “topológica”, una relación de “distancia” y una relación de “dirección”.

- **Num. arcos:** 6 arcos, 3 arcos para representar las relaciones originales, 2 arcos para representar las relaciones equivalentes (“cerca” y “toca”) creadas a partir de las relaciones originales y un arco para representar la relación simétrica (“Norte_de”) creada también de las relaciones originales.
- **Tamaño (vértices + arcos):** 11
- **% incremento:** +83.33%
- **Grafo simple.** Si, existe a lo más un arco entre cualesquiera 2 vértices dados.
- **Arco dirigido.** Si, son usados para representar todas las relaciones. La dirección de los arcos va de los vértices representando los objetos espaciales a los vértices representado los tipos de relaciones espaciales.
- **Arco no dirigido.** No, en el modelo no se usan arcos no dirigidos.
- **Información completa.** Si, representamos las relaciones simétricas creadas a partir de las relaciones originales.
- **Arco “Relation” redundante.** No, en el modelo no usamos arcos “Relation”.

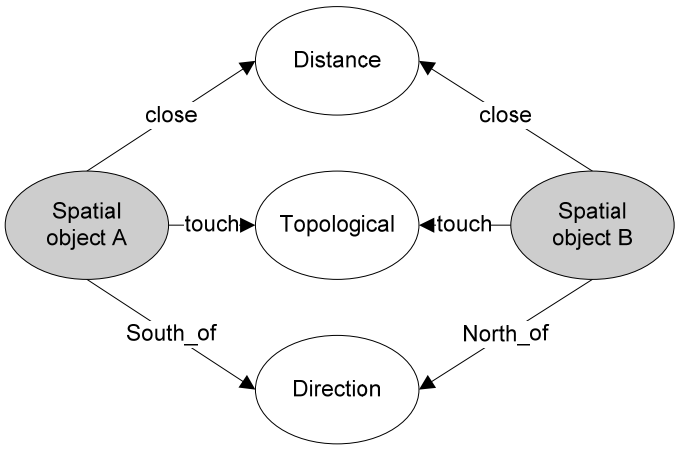

Figura A.4. Modelo #2 - replicación simple de tipos de relación, información completa.

Modelo #3 - doble replicación de tipos de relación, información no completa

En la Figura A.5 se presenta el tercer modelo creado para representar datos espaciales usando un enfoque de grafos. Las características del modelo de acuerdo a las métricas son:

- **Num. vértices:** 8 vértices, 2 vértices para representar los objetos espaciales y 6 vértices para representar los tipos de relaciones espaciales ("distancia", "topológica" y "dirección"). Por cada tipo de relación espacial entre dos objetos espaciales, se añaden 2 vértices etiquetados con el nombre del tipo de la relación espacial. Por ejemplo, en nuestros datos de prueba existen tres tipos de relaciones: 1 relación "topológica", 1 relación "distancia" y 1 relación "dirección", de tal forma, se añaden 6 vértices, 2 por cada tipo de relación espacial.
- **Num. arcos:** 9 arcos, 6 arcos "Relation" para unir los vértices representando los objetos espaciales con los vértices representando los tipos de relaciones espaciales (desde cada vértice representando un objeto espacial nacen 3 arcos ya que existen 3 tipos de relación), y 3 arcos para presentar las relaciones espaciales originales. Estos 3 arcos son usados para unir los vértices representando los tipos de relaciones espaciales.
- **Tamaño (vértices + arcos):** 17
- **% incremento:** +183.33%
- **Grafo simple.** Si, existe a lo más un arco entre cualesquiera 2 vértices dados.
- **Arco dirigido.** Si, son usados para representar las relaciones simétricas y los arcos "Relation". La dirección de los arcos "Relation" va desde los vértices representando los objetos espaciales a los vértices representando los tipos de relaciones espaciales.

La dirección de los restantes arcos va en concordancia a la lectura de las relaciones espaciales entre los objetos.

- **Arco no dirigido.** Si, son usados para representar las relaciones equivalentes.
- **Información completa.** No, en el modelo no se representan las relaciones simétricas que son creadas a partir de las relaciones espaciales originales.
- **Arco "Relation" redundante.** Si, en el modelo se usan arcos "Relation" para representar explícitamente la existencia y tipo de una relación espacial entre 2 objetos espaciales. Adicionalmente, se emplean estos arcos para evitar la creación de grafos complejos.

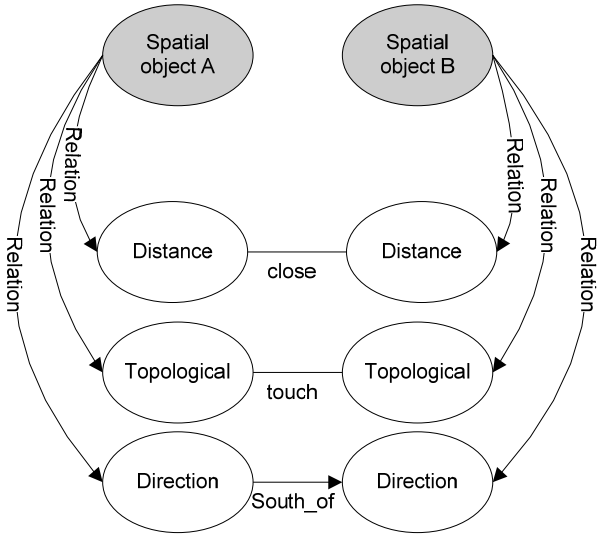

Figura A.5. Modelo #3 - doble replicación de tipos de relación, información no completa.

La Tabla A.1 presenta los resultados de las nueve métricas desarrolladas para evaluar las características de cada modelo propuesto (actualmente 5 modelos). El modelo #1 es llamado el modelo base.

Modelo	Num. Vértices (1)	Num. Arcos (2)	Tamaño (v + e) (3)	% Incremento (4)	Grafo Simple (5)	Arco Dirigido (6)	Arco no Dirigido (7)	Información Completa (8)	Arco "Relation" Redundante (9)
#1	2	4	6	-	No	Yes	Yes	Yes	No
#2	5	6	11	+83.33	Yes	Yes	No	Yes	No
#3	8	9	17	+183.33	Yes	Yes	Yes	No	Yes
#4	5	6	11	+83.33	Yes	Yes	Yes	No	Yes
#5	8	12	20	+233.33	Yes	Yes	No	Yes	Yes

Tabla A.1. Características de los modelos de representación basados en grafos.

Las métricas fueron propuestas basaron en el causas/efectos que cada uno de éstas tiene tanto en el grafo creado como en el algoritmo de minado. Visualizamos cuatro tópicos significativos relacionados directamente a estas métricas:

1. Espacio de búsqueda

El espacio de búsqueda en un algoritmo de minado de datos basado en grafos consiste de todos los subgrafos que pueden ser derivados de su grafo de entrada, de tal forma, el número de vértices (1) y arcos (2) del grafo creado (3) definen el tamaño del espacio de búsqueda para el sistema del descubrimiento. Por lo tanto, el objetivo debe ser minimizar el número de vértices y arcos usados para crear los grafos pero al mismo tiempo maximizar la representatividad de los mismos. Como podemos ver en la Tabla A.1, el modelo usando el número mínimo de vértices y arcos para representar el conjunto de datos de ejemplo es el modelo #1 (2 vértices y 4 arcos) mientras que el modelo #5 es el caso opuesto (8 vértices y 12 arcos).

2. Tiempo de procesamiento

El tamaño del espacio de búsqueda juega un papel relevante con respecto al tiempo de procesamiento usado para descubrir patrones. Si tenemos un espacio búsqueda "grande" se requeriría más tiempo para evaluar todos los subgrafos posibles. Por lo tanto, una comparativa de la métrica "porcentaje de incremento" (4) entre los modelos propuestos se presenta en la Tabla A.1. Recordemos que esta métrica compara el tamaño de un modelo dado con respecto al modelo #1 (modelo base). Por ejemplo, el modelo #5 tiene un incremento de tamaño del grafo de 233.33% respecto al modelo #1. Es decir, el algoritmo de minado requerirá la evaluación de 233.33% más vértices y/o arcos usando el modelo #5 en vez del modelo #1 para el mismo conjunto de datos.

3. Complejidad del grafo

En el capítulo 4 de la disertación se describe el sistema Subdue, nuestra herramienta de minado de datos basada en grafos. Cómo ahí describimos, existe una mayor complejidad para el algoritmo de minado trabajar con grafos complejos en vez de grafos simples (Ej. a lo máximo un arco uniendo cualesquiera dos vértices dado y no ciclos). Por ejemplo, en el proceso de macheo de grafos, la fase de expansión (Subdue emplea un enfoque de "expansión" para descubrir patrones), y la etapa de compresión del grafo. Por lo tanto, el objetivo fue proponer modelos basados en grafos que nos permitieran crear grafos simples. Como podemos ver en la Tabla A.1, únicamente el modelo #1 no permite crear grafos simples.

En consecuencia, como estrategia para romper la multiplicidad de arcos entre dos vértices (por ejemplo, vértices representando dos objetos espaciales con dos ó más relaciones espaciales entre ellos) se emplean los enfoques siguientes:

- Añadir un nuevo vértice etiquetado con el nombre del tipo de la relación (Ej. topológica, distancia y dirección) por cada relación espacial entre los objetos espaciales. Este enfoque es usado en el modelo #2.
- Agregar un nuevo vértice (modelo #4) o dos nuevos vértices (modelo #3 y modelo #5) etiquetado como el nombre de tipo de la relación espacial (Ej. topológica, distancia y dirección) por cada relación espacial entre los objetos espaciales y unir este nuevo vértice (modelo #4) o nuevos vértices (modelo #3 y modelo #5) con los vértices representando los objetos espaciales por medio de arcos etiquetados como "Relation". El enfoque empleado para unir los vértices difiere para cada modelo tal y como se ha comentado en la definición de cada uno de ellos. Esta nomenclatura se utiliza para representar el hecho de que existe una relación espacial entre los objetos espaciales. Estos arcos son conocidos como arcos “Relation” redundantes (9).

4. Representatividad de los datos

Las métricas arcos dirigidos (6), arcos no dirigidos (7), y información completa (8) son usadas para maximizar la representatividad de los datos pero minimizando, tanto como sea posible, el tamaño del grafo y su complejidad. Los arcos dirigidos son usados para representar las relaciones espaciales simétricas (objeto *A* Norte_de objeto *B*, implica, *B* Sur_de *A*), los arcos “Relation” redundantes, los atributos no-espaciales que describen a los objetos espaciales. Arcos no dirigidos son usados para representar las relaciones espaciales

equivalentes (la relación es representada por un no dirigido en vez de dos arcos dirigidos). Finalmente, información completa significa que las relaciones espaciales simétricas entre objetos espaciales también son representadas en el modelo.

A.4 Minando el grafo

La característica de overlap (traslape entre vértices pertenecientes a diferentes instancias de una subestructura) desempeña un rol importante en el sistema de descubrimiento de patrones (subestructuras) en nuestra herramienta de minado de datos basada en grafos, el sistema Subdue. En consecuencia, los resultados generados están condicionados, en un alto porcentaje, al funcionamiento de esta característica. Sin embargo, su implementación actual es ortodoxa: se permite el overlap entre todas las instancias de una subestructura (sin ninguna regla) ó no se permite el overlap entre ninguna instancia perteneciente a una subestructura. Es decir, todo ó nada. En este contexto, proponemos un nuevo enfoque llamado overlap limitado. Una de las ventajas principales de este nuevo enfoque es la capacidad que se le da al usuario para especificar el conjunto de vértices donde el overlap será permitido. Estos vértices podrían representar elementos significativos en el contexto de trabajo. Visualizamos directamente tres motivaciones para proponer el nuevo algoritmo, los cuales serán explicados en las subsecciones siguientes:

1. Reducción del espacio de búsqueda

En sistemas de descubrimiento de conocimiento basados en grafos, el algoritmo de minado de datos usa grafos como su representación de conocimiento. El espacio de búsqueda de un

algoritmo de este tipo consiste de todos los subgrafos que puedan ser derivados del grafo de entrada. El proceso de descubrimiento de subestructuras en Subdue comienza con la creación de subestructuras de un solo vértice a partir del grafo de entrada (una subestructura por cada etiqueta de vértice que exista al menos 2 veces en el grafo). En cada iteración del proceso de descubrimiento, el algoritmo selecciona las mejores subestructuras y expande las instancias de esas subestructuras añadiendo un arco vecino (ó un arco y un nuevo vértice) en todas las direcciones posibles.

Pero como parte del proceso para seleccionar las mejores subestructuras y entonces expandirlas existe un proceso de filtrado. En este proceso, de acuerdo al valor del parámetro overlap las instancias de una subestructura son evaluadas: si el overlap es permito entonces las instancias compartiendo vértices son mantenidas, de lo contrario si el overlap no es permito las instancias que comparten vértices son descartadas.

La mejor subestructura descubierta por Subdue (de acuerdo a sus métricas de evaluación), en cada iteración, puede ser empleada para comprimir el grafo de entrada el cual entonces puede convertirse en el nuevo grafo de entrada para una siguiente iteración. Después de varias iteraciones, Subdue crea una descripción jerárquica de los datos, donde subestructuras descubiertas en cierta iteración pueden estar definidas en base a subestructuras descubiertas en iteraciones previas. De tal forma, el número de instancias de una subestructura define el espacio de búsqueda (en cada iteración) en el proceso de descubrimiento de subestructuras. Como podemos observar, a través de uso del overlap limitado, se obtiene una reducción del espacio de búsqueda dado que el número de

instancias candidatas a ser expandidas se condiciona a los valores (vértices) donde el overlap es permitido, siendo éstos establecidos por el usuario.

2. Reducción del tiempo de procesamiento

La reducción en el número de instancias candidatas a ser expandidas trae consigo una reducción del espacio de búsqueda, y esto beneficia en una reducción del tiempo de procesamiento para la búsqueda de subestructuras (patrones).

Permitiendo el overlap en Subdue, hace que éste sea considerablemente más lento (en cuanto a tiempo de procesamiento) dado que el número de instancias candidatas a ser expandidas, evaluadas, comparadas, y descubiertas se incrementa como hemos descrito. Sin embargo, con la implementación del overlap limitado, el número de instancias a ser procesadas en estas fases decrece resultando en una reducción del tiempo de procesamiento en todo el proceso de descubrimiento de subestructuras.

3. Búsqueda orientada de patrones con overlap (traslape) selectivo

El overlap limitado da a usuario la capacidad para definir el conjunto de elementos donde el overlap será permitido y que a su juicio considera relevantes para su contexto de trabajo (Ej. un objeto espacial, un atributo descriptivo). Estos elementos son representados usando vértices de acuerdo al modelo propuesto. En contra parte, el algoritmo descartará aquellos elementos traslapados que el usuario no consideró significativos.

Por lo tanto, el overlap limitado proporciona al usuario un medio para implementar una búsqueda orientada de patrones con overlap. Es decir, el usuario delimita el conjunto de

elementos que tendrán un papel preponderante en el proceso de descubrimiento de subestructuras. Adicionalmente, esta característica ofrece la ventaja de que el proceso de evaluación de patrones se simplifica dado que el conjunto de resultados generados es menor porque éstos se centran sobre los requerimientos del usuario.

A.5 Resultados

Como parte de las pruebas para evaluar nuestra propuesta de modelado y minado de datos espaciales usando un enfoque basado en grafos se desarrollaron tres casos de uso ilustrativos. Los dos primeros casos fueron implementados usando un censo de población del centro histórico de la ciudad de Puebla en el año de 1777. El tercer caso de uso fue desarrollado usando una base de datos espaciales de la región del volcán Popocatépetl.

En esta sección presentamos ejemplos de resultados obtenidos con el caso de uso del volcán Popocatépetl. Supongamos que deseamos conocer características comunes entre las poblaciones (asentamientos poblacionales), carreteras y ríos en la zona que nos ayuden a evaluar/implementar planes de evacuación en caso de una contingencia volcánica. Por ejemplo, características de carreteras empezando en ó cruzando una población, material usado para construir esas carreteras y su estado actual (Ej. pavimentadas, terracería), características de carreteras y ríos que tengan alguna relación entre ellos (Ej. se crucen, se tocan), ríos cerca de poblaciones que en caso de alta precipitación pluvial pudieran representar peligros potenciales.

Los experimentos fueron desarrollados usando los 5 modelos de representación actualmente propuestos. En esta sección se presentan patrones encontrados con el modelo #1 entre carreteras y ríos, carreteras y poblaciones, y por último ríos y poblaciones. La idea de organizar la presentación de resultados de esta manera es mostrar los diversos patrones que se pueden encontrar entre estos elementos. Al final de la sección se presentan tablas comparando los resultados obtenidos con cada uno de los 5 modelos.

La Figura A.6 muestra el patrón más significativo descubierto entre carreteras y ríos usando el modelo #1. El patrón describe una relación entre "carretera categoría terracería traslapando un río categoría escurrimiento" en la zona. Este patrón puede considerarse como un indicador del número de carreteras que necesitan ser supervisadas en caso de una contingencia volcánica dado el tipo de material con el que están construidas, y porque ellos atraviesan ríos (la lectura puede ser hecha en orden inverso) que en caso de altas concentraciones pluviales pueden desbordarse e inutilizarlas. Subdue encontró con no overlap 46 instancias del patrón en la segunda iteración; vía overlap estándar encontró 85 instancias en la primera iteración; y a través de overlap limitado también encontró 85 instancias en la segunda iteración. Como podemos observar overlap estándar y overlap limitado encontraron el mismo número de instancias, pero overlap limitado necesitó dos iteraciones para encontrar el mismo patrón. Sin embargo, esto no quiere decir que overlap estándar es mejor que overlap limitado (respecto a tiempo de procesamiento) porque analizando el tiempo global de procesamiento requerido por overlap limitado para finalizar la fase de descubrimiento de subestructuras notamos que es menor que el requerido por overlap estándar.

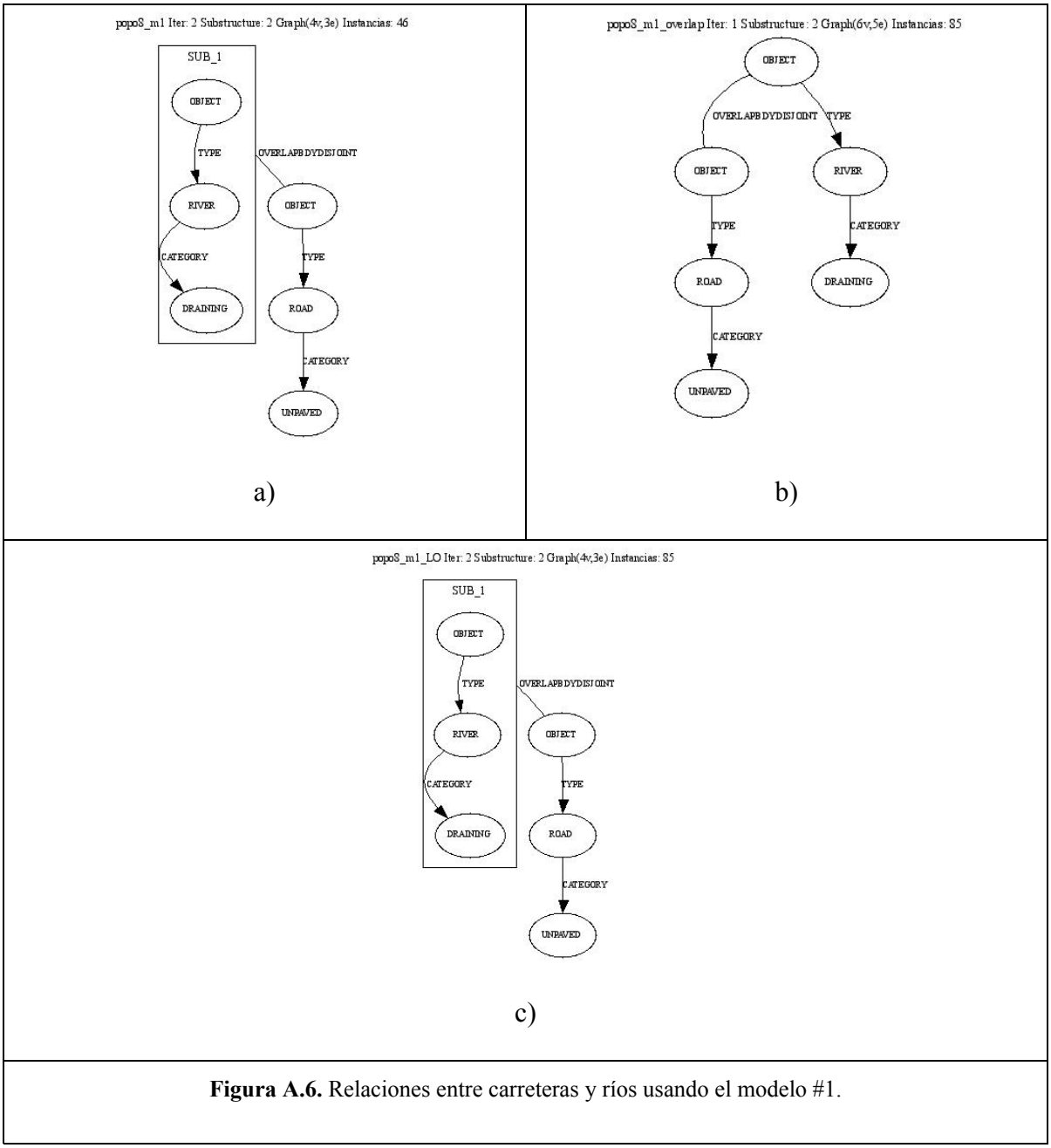

Figura A.6. Relaciones entre carreteras y ríos usando el modelo #1.

El patrón más significativo, usando el modelo #1, encontrado entre carreteras y poblaciones es presentado en la Figura A.7. Este describe una relación entre "carretera categoría terracería tocando un asentamiento poblacional categoría construcción". "Asentamiento poblacional categoría construcción" representa en la capa de datos espaciales "asentamientos" de la base de datos del volcán áreas habitacionales con alta población,

edificios y una gran cantidad de construcciones usadas para ofrecer servicios a los habitantes. Si asumimos que la gente podría necesitar ser evacuada en caso de una erupción y que las carreteras que serían usadas para ese propósito son de terracería, entonces, esta situación podría convertirse en un problema (Ej. cuello de botella). En este experimento Subdue encontró vía no overlap 6 instancias del patrón en la novena iteración; a través de overlap estándar 9 instancias en la cuarta iteración; y usando overlap limitado 8 instancias en la décima iteración.

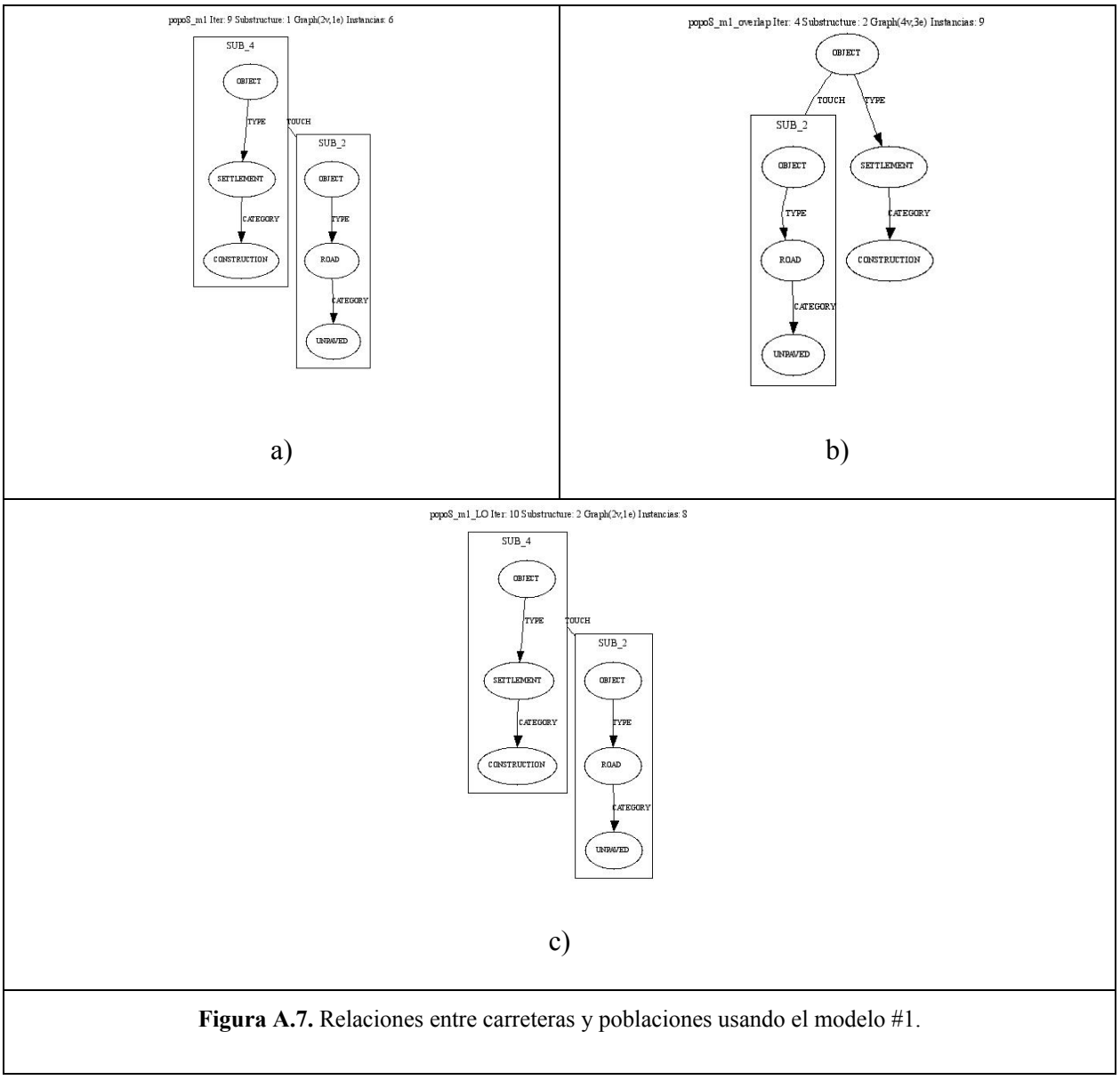

Figura A.7. Relaciones entre carreteras y poblaciones usando el modelo #1.

La Figura A.8 muestra el patrón más significativo encontrado entre ríos y poblaciones usando el modelo #1. El patrón describe una relación entre “río categoría escurrimiento cruzando un asentamiento poblacional categoría manzana o construcción" en la zona. “asentamientos poblacionales categoría manzana" representa en la capa de datos espaciales “asentamientos” de la base de datos del volcán áreas habitacionales con poca población, de hecho con muchas áreas deshabitadas, escasos edificios y construcciones. El patrón puede ser utilizado para identificar zonas potenciales de inundación, habitadas por personas, dada la cercanía de ríos. A través de no overlap Subdue encontró 5 instancias del patrón en la décima segunda iteración; usando overlap estándar encontró 5 instancia en la octava iteración; y vía overlap limitado también encontró 5 instancias en octava iteración. Subdue encontró el mismo patrón en los tres casos, sin embargo, usando overlap estándar y overlap limitado la lectura del patrón es más simple.

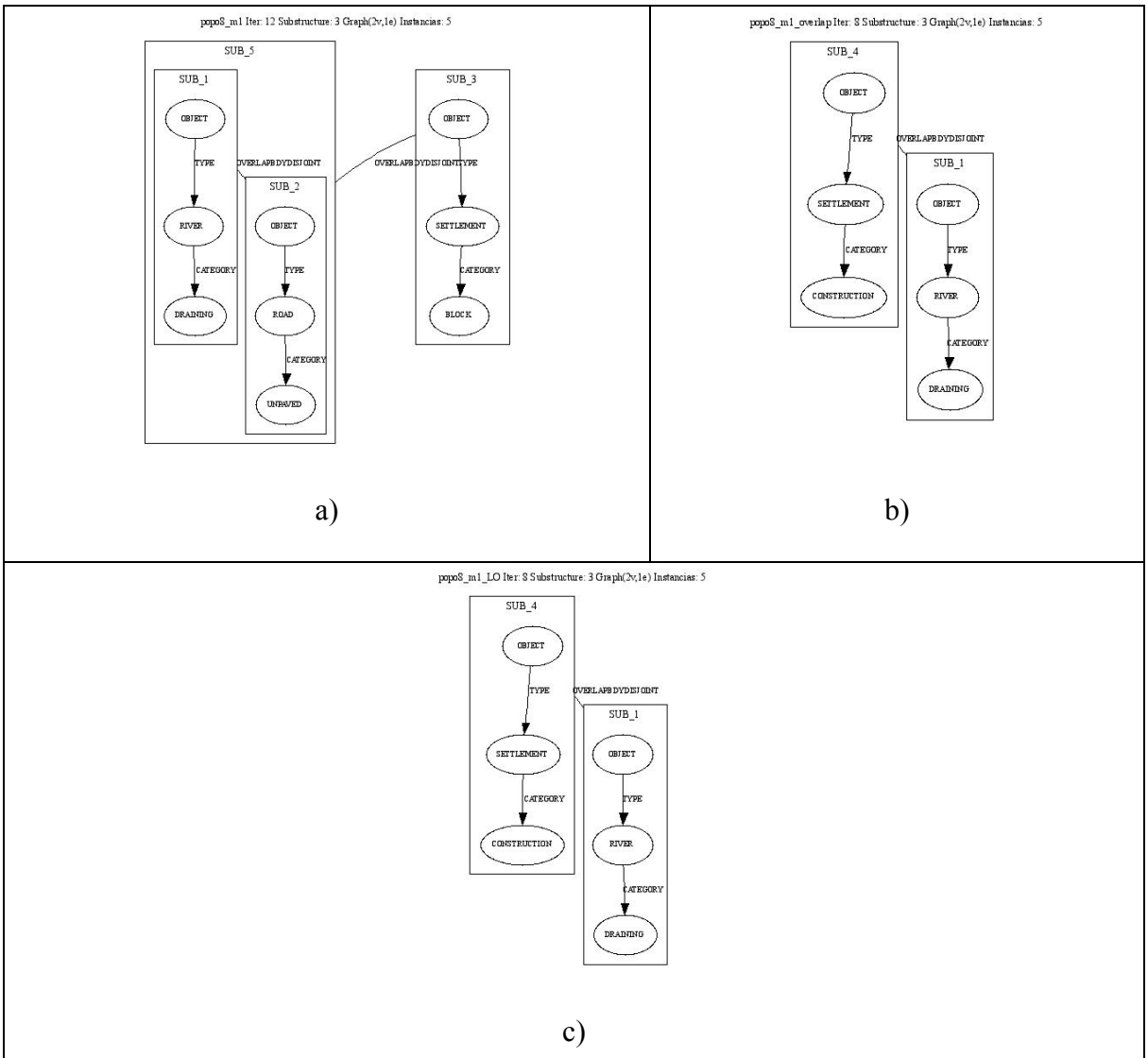

Figura A.8. Relaciones entre ríos y poblaciones usando el modelo #1.

La Tabla A.2 presenta una comparación, por modelo, entre el número de instancias descubiertas/iteraciones necesitadas para descubrirlas y las tres implementaciones de overlap. Por ejemplo, usando el modelo #1, Subdue encontró 46 instancias (en la segunda iteración) de un patrón "completo" (nuestra definición para reportar un patrón "completo" es que éste contengan al menos dos objetos espaciales y la relación espacial entre ellos)

conteniendo los objetos espaciales carretera-río vía no overlap. Un valor más alto significa un modelo que permite descubrir más instancias de una subestructura (patrones). Recordemos que Subdue reporta como el mejor patrón (por iteración) la subestructura con el número más alto de instancias descubiertas de esa subestructura. Esta comparación es reportada por cada estructura "objeto-objeto" (Ej. carretera-río).

Nota: NO (no overlap), SO (overlap estándar), LO (overlap limitado).

	Modelo #1			**Modelo #2**			**Modelo #3**			**Modelo #4**			**Modelo #5**		
	NO	**SO**	**LO**	**NO**	**SO**	**LO**	**NO**	**SO**	**LO**	**NO**	**SO**	**LO**	**NO**	**SO**	**LO**
Carretera-Río															
Instancias	46	85	85	41	85	64	39	85	34	39	85	60	45	85	45
Iteraciones	2	1	2	2	3	2	2	2	9	2	1	2	5	1	5
Carretera-Población															
Instancias	6	9	8	5	8	7	4	8	5	6	8	8	6	0	7
Iteraciones	9	4	10	14	6	10	15	6	13	12	10	7	7	0	7
Río-Población															
Instancias	5	5	5	5	10	5	5	19	5	5	10	5	5	5	5
Iteraciones	12	8	8	16	7	14	12	4	10	6	6	11	13	6	10

Tabla A.2. Instancias/iteraciones por cada modelo basado en grafos: caso de uso Popocatépetl.

La Tabla A.3 presenta una comparación de máximo/mínimo instancias descubiertas por cada implementación de overlap. Un modelo con el valor más alto es mejor porque permite descubrir más instancias de una subestructura. La comparación es presentada por cada estructura "objeto-objeto". Por ejemplo en la estructura carretera-río el modelo #1 reportó 46 instancias descubiertas en la segunda iteración (el valor más alto).

	Máximo	Mínimo
Carretera-Río		
No overlap	modelo #1 (segunda iteración)	modelos #3 y #4 (segunda iter.)
Overlap estándar	modelos #1, #4 y #5 (primera iter.)	modelo #2 (tercera iteración)
Overlap limitado	modelo #1 (segunda iteración)	modelo #3 (novena iteración)
Carretera-Población		
No overlap	modelo #5 (séptima iteración)	modelo #3 (quinceava iteración)
Overlap estándar	modelo #1 (cuarta iteración)	modelo #5 (patrón no completo)
Overlap limitado	modelo #4 (séptima iteración)	modelo #3 (treceava iteración)
Río-Población		
No overlap	modelo #4 (sexta iteración)	modelo #2 (dieciseisava iteración).
Overlap estándar	modelo #3 (cuarta iteración)	modelo #1 (octava iteración).
Overlap limitado	modelo #1 (octava iteración)	modelo #2 (catorceava iteración)

Tabla A.3. Max/Min de instancias descubiertas por "objeto-objeto"/característica overlap.

La Tabla A.4 presenta una comparación entre el promedio de instancias descubiertas por modelo. Un valor más alto significa un modelo permitiendo descubrir más instancias de una subestructura. Cada valor representa el promedio de subestructuras descubiertas usando no overlap, overlap estándar y overlap limitado. La comparación es reportada por cada estructura "objeto-objeto".

	Modelo #1	Modelo #2	Modelo #3	Modelo #4	Modelo #5
Carretera-Río	72.0	63.3	52.7	61.3	58.3
Carretera-Población	7.7	6.7	5.7	7.3	4.3
Río-Población	5.0	6.7	9.7	6.7	5.0

Tabla A.4. Promedio de instancias descubiertas por modelo/"objeto-objeto".

La Tabla A.5 presenta una comparación entre el promedio de instancias descubiertas por modelo. Un valor más alto significa un modelo permitiendo descubrir más instancias de una subestructura. La comparación es reportada por cada implementación de overlap.

Modelo #1			Modelo #2			Modelo #3			Modelo #4			Modelo #5		
NO	**SO**	**LO**	**NO**	**SO**	**LO**	**NO**	**SO**	**LO**	**NO**	**SO**	**LO**	**NO**	**SO**	**LO**
19.0	33.0	32.7	17.0	34.3	25.3	16.0	37.3	14.7	16.7	34.3	24.3	18.7	30.0	19.0

Tabla A.5. Promedio de instancias descubiertas por modelo/característica overlap.

La Tabla A.6 presenta un comparativo final entre el promedio de instancias descubiertas por modelo. Podemos ver en la tabla que el modelo #1 reporta el valor más alto de instancias descubiertas (de acuerdo a nuestros parámetros para reportar instancias completas) en este caso de uso ilustrativo. Los siguientes modelos son el modelo #2 y el modelo #4 respectivamente.

Modelo #1	Modelo #2	Modelo #3	Modelo #4	Modelo #5
28.2	25.6	22.7	25.1	22.6

Tabla A.6. Promedio de instancias descubiertas por modelo.

A.6 Conclusiones

La constante interacción entre los seres humanos y su hábitat natural, el planeta de la tierra, genera día a día, nuevos requerimientos asociados al manejo y explotación de datos espaciales. Por ejemplo, el análisis urbano, la prevención los riesgos naturales, la exploración del espacio estelar, la contaminación en los océanos, y la reforestación de los

suelos, por nombrar algunos de ellos. La minería de datos espaciales involucra la integración de métodos y técnicas provenientes de diversos campos científicos los cuales nos ayudan, por medio de algoritmos de análisis y de descubrimiento, a producir una enumeración particular de patrones sobre los datos espaciales.

Nuestra argumentación en esta disertación doctoral se basa en la idea de que los enfoque de minaría de datos espaciales descritos no consideran todos los elementos encontrados en una base de datos espaciales (datos espaciales, datos no espaciales y relaciones espaciales entre los objetos espaciales) de una manera integral. En consecuencia, se propuso emplear un enfoque basado en grafos para representar estos elementos como un solo conjunto de datos, minarlos como un todo, para de estar forma poder encontrar patrones conteniendo ambos tipos de datos y relaciones espaciales (patrones más descriptivos).

En nuestro modelo las relaciones espaciales entre los objetos espaciales son incluidas porque una característica significativa de los datos espaciales es la influencia que los vecinos de un objeto pueden tener en el objeto mismo. En el modelo incluimos tres tipos de relaciones espaciales. Derivado del modelo general se propusieron cinco modelos operativos. Tres aspectos definen las características de un grafo creado con estos modelos: (1) Representación de las relaciones espaciales equivalentes. (2) Representación de relaciones espaciales simétricas. (3) La manera de representar los objetos y sus relaciones.

Como parte integrante de nuestra metodología para el minado de datos espaciales usando un enfoque basado en grafos, usamos el sistema Subdue como nuestra herramienta de minado. Se propuso un nuevo algoritmo llamado overlap limitado el cual le da al usuario la

capacidad de especificar el conjunto de vértices sobre los cuales el overlap es permitido. Visualizamos tres motivaciones para proponer este nuevo enfoque: (1) Reducción del espacio de búsqueda. (2) Reducción del tiempo de procesamiento. (3) Búsqueda orientada de patrones con overlap (traslape) selectivo.

Para demostrar la viabilidad, capacidad de minado y descubrimiento de patrones usando el enfoque propuesto, se desarrolló un prototipo implementado nuestro modelo para crear los conjuntos de datos basados en grafos, minar esos grafos (a través del llamado al sistema Subdue) y visualización de patrones encontrados. Los resultados generados de los casos de uso desarrollados nos dan un panorama respecto a cómo y qué podríamos obtener usando este enfoque. Es importante comentar el hecho de que podemos utilizar esta metodología de modelado y representación en cualquier dominio que pueda ser representado como grafo.

Una vez demostrada la viabilidad de nuestra propuesta, las perspectivas relacionadas a mejorar nuestro trabajo (modelo de representación basado en grafos, algoritmo de minado de datos y sistema prototipo) incluyen los puntos siguientes:

- **Visualización de conocimiento descubierto.** Por ejemplo, visualización de resultados sobre las capas espaciales, a través del uso de gráficas, y navegación en la jerarquía de patrones descubiertos usando un enfoque de hypergrafo.
- **Mejoramiento de los algoritmos empleados para crear los conjuntos de datos basados en grafos de acuerdo a los modelos propuestos.** La validación de relaciones espaciales entre objeto espaciales es una fase que en la mayoría de los casos requiere gran cantidad de recursos de computacionales.

- **Minado de los grafos.** Se empleó el sistema Subdue como herramienta de minado de datos. Así mismo, se propuso un nuevo algoritmo llamado overlap limitado. El isomorfismo de grafos es un problema NP-completo, de tal forma, debemos ser capaces de que nuestros tiempos de procesamiento para la búsqueda de patrones cumpla parámetros aceptables de eficiencia.
- **Manejo y representación de relaciones entre datos no espaciales.** Las relaciones implícitas y explicitas entre los atributos describiendo los objetos espaciales pueden ser incluidos en el modelo con la finalidad de mejorar la representación de los datos.

A.7 Contribución

La contribución al descubrimiento de conocimiento en el dominio de los datos espaciales, descrito en esta disertación, es el desarrollo de un nuevo enfoque para el modelado y minado de datos espaciales usando una representación basada en grafos. Este enfoque incluye los aspectos siguientes:

- Se propuso una nueva representación de datos basada en grafos para datos espaciales. Se visualizaron dos objetivos para crear un modelo de datos con estas características. El primero de ellos es crear un único conjunto de datos, basado en grafos, representando estos elementos relacionados. El segundo es emplear este conjunto de datos para alimentar a un sistema de minado de datos basado en grafos, de tal forma que pudiéramos descubrir patrones conteniendo datos espaciales, no espaciales y relaciones espaciales los cuales nos ayuden a describir/entender los

datos, basado en la premisa, de que estos son elementos relacionados en el mundo real.

- Se propuso un nuevo algoritmo para descubrir subestructuras (patrones) usando un enfoque de overlap limitado en el sistema Subdue, nuestra herramienta de minado de datos. Visualizamos directamente tres motivaciones para proponer la implementación del nuevo algoritmo: reducción del espacio de búsqueda, reducción del tiempo de procesamiento y búsqueda orientada de patrones con overlap selectivo (*specialized overlapping pattern oriented search*).
- Se diseñó e implementó un sistema prototipo implementado el modelo propuesto. El prototipo ofrece una interfase de usuario amigable para el manejo de las capas espaciales con las que se trabajará, para la creación de grafos espaciales y no espaciales, para el minado de estos grafos (a través del llamado al sistema Subdue) y para el despliegue de los resultados generados.