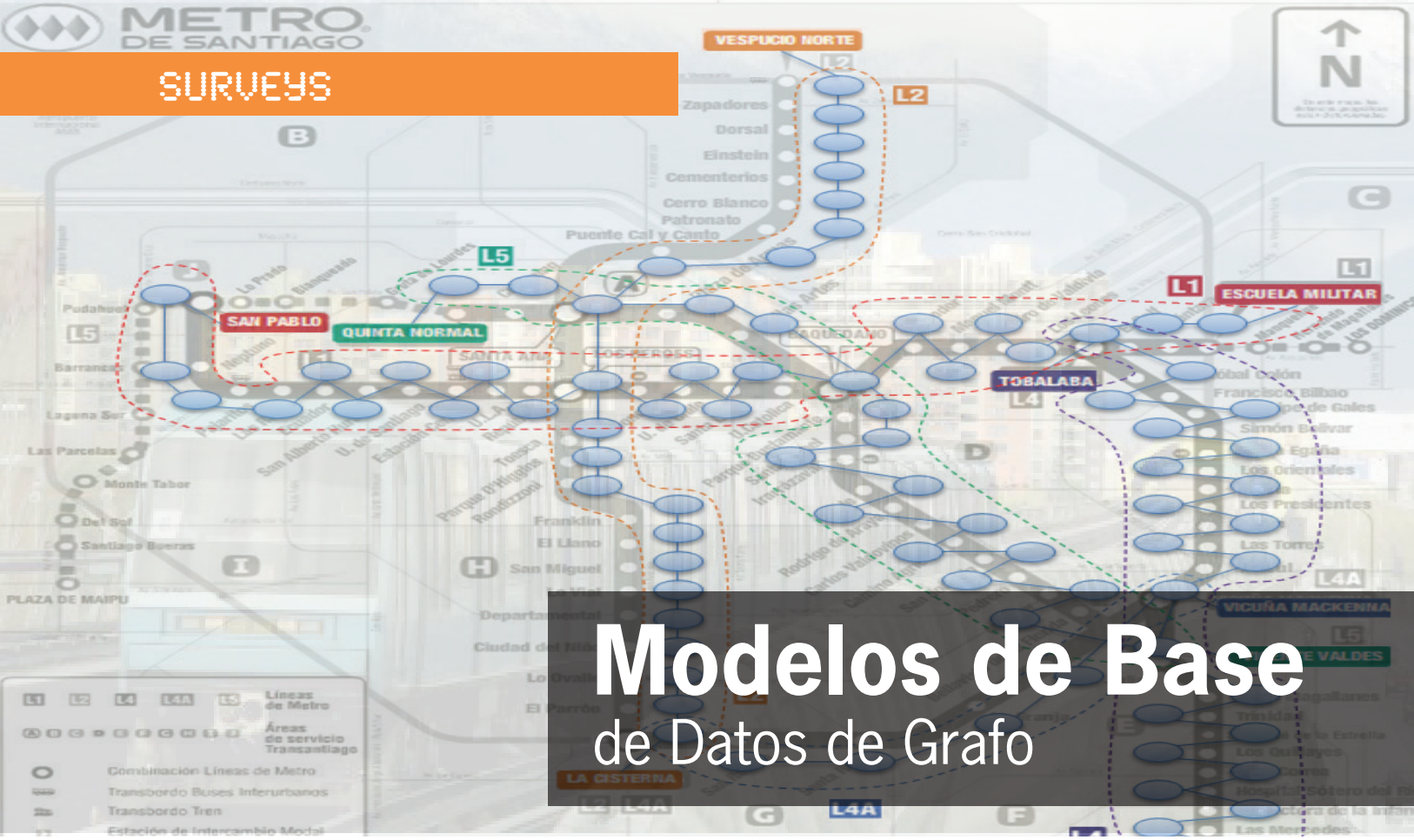
SURVEYS



Modelos de Base de Datos de Grafo

Renzo Angles

Profesor Asistente, Universidad de Talca.Doctor en Ciencias mención Computación, DCC, Universidad de Chile. Ingeniero de Sistemas, Universidad Católica de Santa María, Perú. Líneas de investigación: Bases de Datos, Web Semántica, Ingeniería de Software y Documentación Electrónica.
rangles@utalca.cl

INTRODUCCIÓN

En la comunidad de manejo de la información el término "modelo de datos" ha sido usado de manera amplia y diversa, teniendo de esta manera varios significados. En el sentido más general, un modelo de datos es una colección de herramientas conceptuales usadas para representar entidades del mundo real y las relaciones entre ellas [1]. Desde un punto de vista de base de datos, las herramientas conceptuales que definen un modelo de datos deben especificar la manera de estructurar, restringir, mantener y recuperar los datos. Acorde con estos criterios, un modelo de base de datos consiste de tres componentes: un conjunto de tipos de estructura de datos, un conjunto de operadores o reglas de inferencia, y un conjunto de restricciones de integridad [2].

Desde su introducción, a fines de los sesenta, numerosos modelos de base de datos han sido propuestos, cada cual con sus propios conceptos y principios. Los primeros se enfocaron esencialmente en especificar la estructura de los datos a nivel físico, es decir, de acuerdo al sistema de archivos. Dos modelos representativos son el modelo jerárquico y el de red. Esta perspectiva cambió rotundamente con el modelo relacional, el cual introdujo la noción de nivel de abstracción al separar el nivel físico (implementación) del lógico (modelado). Enriqueciendo el nivel de abstracción, pero desde el punto de vista del usuario, los modelos semánticos permiten representar entidades y sus relaciones de una manera clara y directa. Un ejemplo conocido es el modelo entidad-relación. Basándose en el paradigma orientado a objetos, los modelos orientados a objetos representan

los datos como una colección de objetos organizados en clases y soportando valores complejos como atributos. Posteriormente, los modelos semiestructurados fueron diseñados para modelar datos con una estructura flexible. En estrecha relación se encuentra XML (eXtensible Markup Language): un lenguaje de marcas usado para representar datos semiestructurados para el que se han desarrollado bases de datos. Mayores detalles sobre modelos de base de datos pueden encontrarse en los trabajos de Silberschatz et al. [1] y Navathe [3]. En la actualidad el paradigma relacional es el preferido por la mayoría de los sistemas de administración de base de datos.

MODELOS DE BASE DE DATOS DE GRAFO

Los Modelos de Base de Datos de Grafo (MBDG) se caracterizan porque sus estructuras para esquemas e instancias son modeladas como grafos y la manipulación de datos se basa en operaciones orientadas a grafos. Específicamente, y considerando los elementos de un modelo de base de datos, los MBDG pueden describirse de la siguiente manera:

1) Los datos y/o los esquemas son representados por grafos, o por generalizaciones de ellos (por ejemplo hipergrafos). Un aspecto a considerar es la separación entre los niveles físico y lógico; en este sentido la mayoría de los modelos distinguen el esquema de los datos (instancias).

2) La manipulación de datos es expresada por transformaciones de grafo, o por operaciones orientadas a consultar su estructura (ejemplo adyacencia, caminos, subgrafos, etc.) y propiedades (ejemplo diámetro, centralidad, etc.).

3) La consistencia de los datos se mantiene a través de restricciones de integridad aplicadas a la estructura de grafo. Por ejemplo, consistencia esquema-instancia, identidad e integridad referencial, dependencias funcionales y de inclusión.

La actividad alrededor de los MBDG fue intensa en la primera mitad de los años '90 y desde entonces el tópico casi desapareció. Las razones de esto son, entre otras, que la comunidad de bases de datos se movió hacia los modelos de datos semiestructurados; la aparición de XML capturó la atención de aquellos trabajando en Hipertexto y la Web; los investigadores del área se concentraron en aplicaciones particulares como las bases de datos espaciales o biológicas. Una revisión de los modelos de BD de grafo puede encontrarse en un artículo de ACM Surveys [4]. La Figura 1 refleja el desarrollo del área en términos de los artículos publicados, los cuales describiremos a continuación.

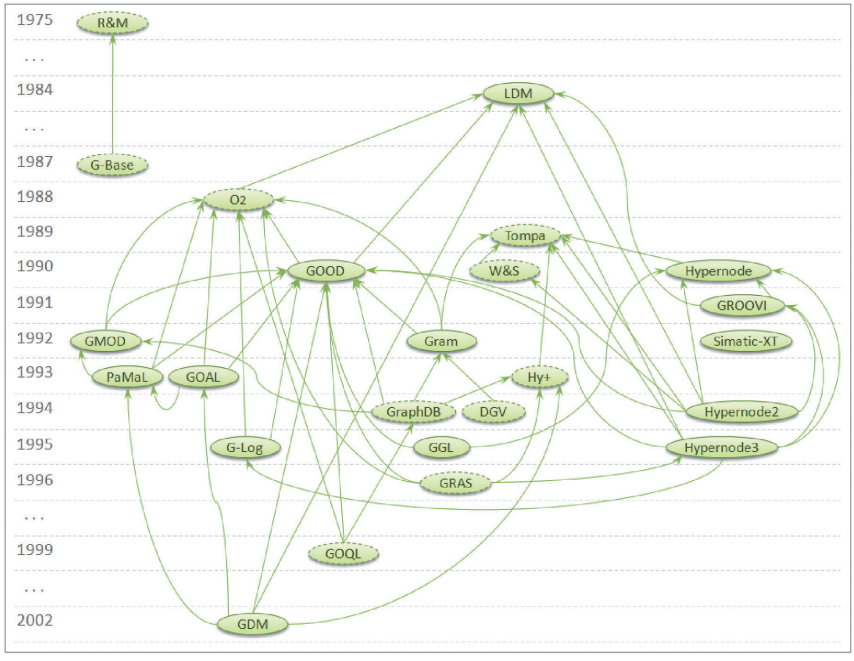


Figura 1. Desarrollo de los Modelos de Base de Datos de Grafo. Las elipses indican modelos y las flechas citaciones. Las elipses punteadas representan modelos relacionados.

En una primera aproximación, y enfocándose en la deficiencias de los sistemas (en su momento) para modelar la semántica de los datos, Roussopoulos y Mylopoulos (R&M) propusieron una red semántica para almacenar conocimiento respecto a una base de datos. Con un objetivo similar, G-Base propuso el empleo de un modelo de grafo para representar estructuras complejas de conocimiento. Un enfoque diferente fue formulado en el Logical Data Model (LDM), donde se intentó generalizar los modelos relacional, jerárquico y de red, a través de un modelo explícito basado en grafos.

A finales de los años '80 encontramos modelos orientados a objetos basados en una estructura de grafo. Por ejemplo, las relaciones entre objetos en O2 se representan a través de un grafo denominado object graph. GOOD es un modelo influyente que buscó aprovechar la orientación a objetos y la naturaleza gráfica de una estructura de grafo. La Figura 2 presenta un ejemplo de este modelo.

Entre los desarrollos subsecuentes, basados en GOOD, tenemos: GMOD, que modela la representación de interfaces de usuario en bases de datos. Gram, un modelo de grafo explícito orientado a modelar datos de Hipertexto; PaMaL, el cual extendió GOOD con una representación explícita de tuplas y conjuntos; GOAL, que introdujo la noción de nodos de asociación; G-Log, el cual propuso un lenguaje declarativo para grafos; y GDM, que incorporó la representación de relaciones simétricas n-arias. La Figura 3 muestra un ejemplo del modelo GMOD.

Por otra parte, encontramos modelos que usan generalizaciones de grafos. El modelo de hipernodo (Hypernode Model) presenta una estructura basada en grafos anidados

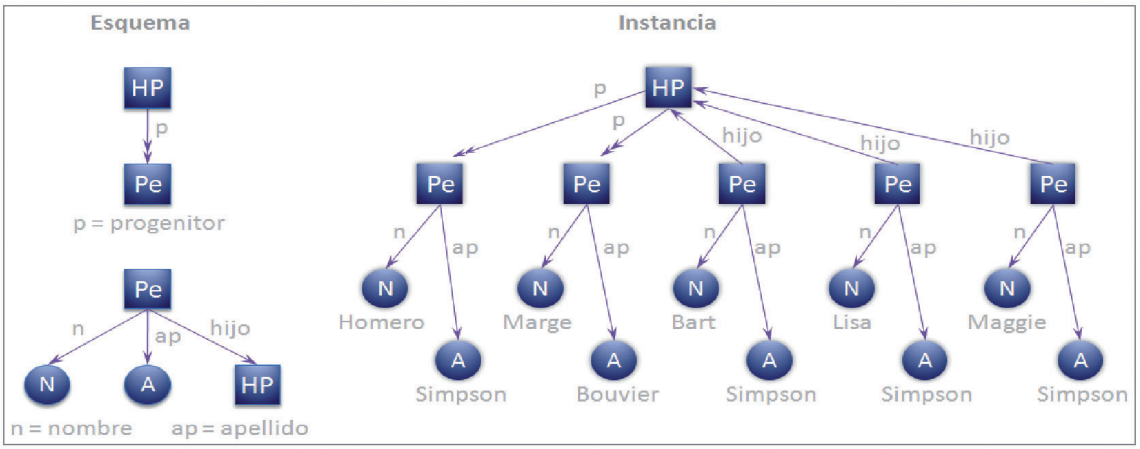


Figura 2. Ejemplo del modelo de base de datos de grafo GOOD. En el esquema: los nodos N y A, denominados imprimibles, representan nombre y apellido respectivamente; los nodos rectangulares representan entidades complejas (Persona) o relaciones (Hijo-Progenitor); una flecha simple indica una relación funcional (mono-valuada) y una doble indica una relación nofuncional (multi-valuada). La instancia consiste en asignar valores a los nodos imprimibles, e instanciar los nodos Pe y HP. En el ejemplo se modelan datos sobre Los Simpson.

(Hypernodes). La misma noción fue usada para modelar redes multiescala (Simatic-XT) y datos biológicos (GGL). GROOVY es un modelo orientado a objetos formalizado a través de hipergrafos (hypergraphs). Esta generalización fue usada en otros contextos como: consulta y visualización (Hy+); representación de esquemas, instancias y acceso a datos (W&S); representación de la estructura y contenido en bases de datos de Hypertexto (Tompa). La Figura 4 muestra un ejemplo del uso de grafos anidados.

Además encontramos otros trabajos relacionados a los MBDG. GraphDB se propuso con la finalidad de modelar y consultar grafos en una base de datos orientada a objetos. Database Graph View (DGV) propone un mecanismo de abstracción para definir y manipular grafos almacenados en bases de datos relacionales, orientadas a objetos, o en sistemas de archivos. El proyecto GRAS usó grafos para modelar información compleja desde proyectos de ingeniería de software. Otra área de desarrollo, reciente e importante, tiene que ver con los modelos de representación de datos en la Web (http://www.w3.org/). Entre ellos podemos mencionar: XML, el modelo de intercambio de datos presentando una estructura de árbol; RDF, el modelo de representación de metadatos basado en una estructura de grafo; y OWL el modelo de representación de ontologías.

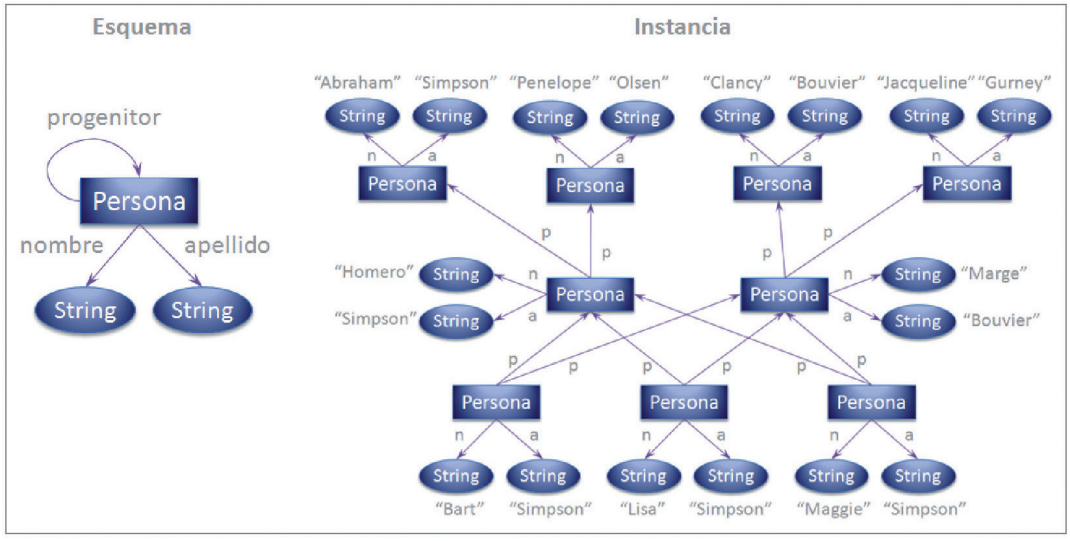


Figura 3. Ejemplo del modelo GMOD. En el esquema: los nodos rectangulares representan objetos complejos (Persona); los nodos elípticos representan tipos de datos primitivos; las aristas representan propiedades (nombre, apellido) y relaciones (progenitor) de los objetos. La instancia se construye creando subgrafos en base al esquema y asignando valores a los nodos elípticos.

MOTIVACIONES

Debido a la importancia filosófica y práctica del modelado conceptual, los modelos de base de datos han llegado a ser herramientas de abstracción esenciales para el desarrollo de sistemas de administración de base de datos (DBMS). La evolución y diversidad de los modelos muestra que hay muchos factores que influencian el desarrollo de los mismos. Algunos de los más importantes son: las características o estructura del dominio a ser modelado; el tipo de herramientas teóricas que interesan a los usuarios esperados; y, por supuesto, las restricciones de hardware y software.

Además cada propuesta de modelo se basa en ciertos principios teóricos fundamentales como es el caso de la teoría de grafos en los MBDG.

Los MBDG son aplicados en áreas donde la información sobre las relaciones entre los datos es tan importante, o más importante que los datos en sí. En este sentido, estos modelos intentan cubrir las limitaciones de los modelos tradicionales con respecto a capturar la estructura de grafo intrínseca en aplicaciones donde la interconectividad o topología de los datos es relevante. Un ejemplo claro son las redes sociales (por ejemplo Facebook), donde la importancia de una persona (dato) se debe principalmente a su red de contactos (relaciones). De hecho, el uso de grafos como una herramienta de modelado trae muchas ventajas para este tipo de datos:

- Modelado de datos más natural. La estructura de grafo es visible para el usuario y permite una manera natural de manejar los datos de las aplicaciones (por ejemplo al modelar una red de contactos). El uso de grafos ofrece la ventaja de mantener toda la información de una entidad en un único nodo, el cual muestra su información relacionada a través de los arcos conectados a éste.

- Consultas propias de grafos. Se tienen operaciones que consultan directamente la estructura y propiedades de un grafo. Por ejemplo, retornar los nodos adyacentes a un nodo, encontrar los caminos entre dos nodos (observe el ejemplo de la Figura 5), encontrar los subgrafos que satisfacen un patrón, etc. Estas estructuras y operaciones de grafos permiten al usuario expresar consultas con un alto nivel de abstracción. En algunos casos no es necesario tener un conocimiento completo de la estructura de los datos para poder expresar consultas. Esto último es particularmente útil al momento de explorar los datos.

- Implementación física ad hoc. Las bases de datos de grafo pueden entregar estructuras y algoritmos especiales para el almacenamiento y consulta de grafos. Considerando la complejidad intrínseca en varios problemas de grafos, se pueden seleccionar estrategias de implementación apropiadas dependiendo de las características de los grafos. Por ejemplo, considere la evaluación de la consulta transitiva presentada en la Figura 6.

APLICACIONES

Los MBDG están orientados a aplicaciones de la vida real donde la interconectividad es una característica clave. Nosotros dividiremos estas áreas de aplicación en aplicaciones clásicas y redes complejas.

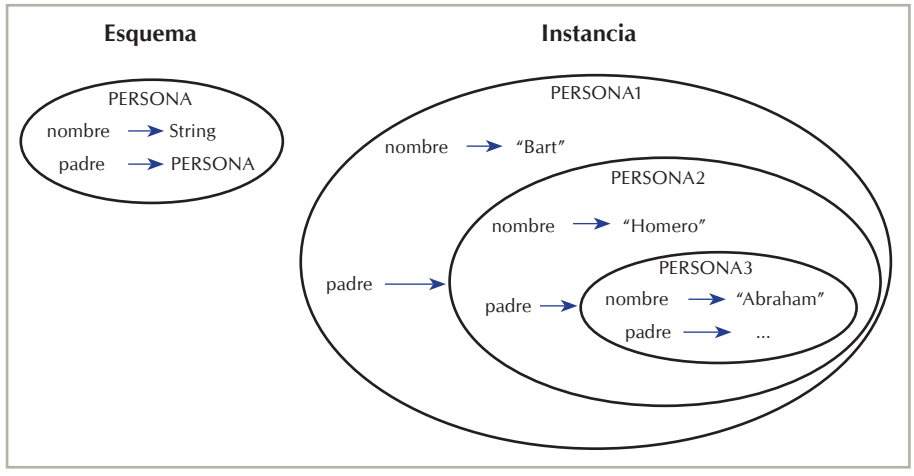


Figura 4. El modelo de Hipernodo (Hypernode model) extiende la definición básica de grafo al permitir que los nodos sean a su vez grafos (hypernodes). En el ejemplo usamos esta característica para representar una estructura anidada de objetos de tipo PERSONA. Cabe resaltar que, aunque no se observa en el ejemplo, el esquema podría definir una relación recursiva (por ejemplo amigo) que podría resultar en una representación cíclica en la instancia.

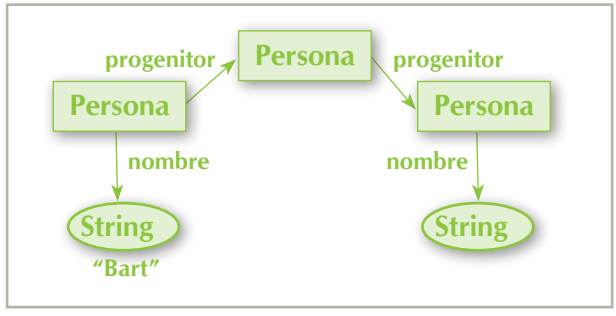


Figura 5. Ejemplo de una consulta de grafo (específicamente de caminos). Considerando el grafo de la Figura 3, la consulta retorna los nombres de los abuelos de "Bart", estos son "Abraham", "Penelope", "Clarcy" y "Jacqueline".

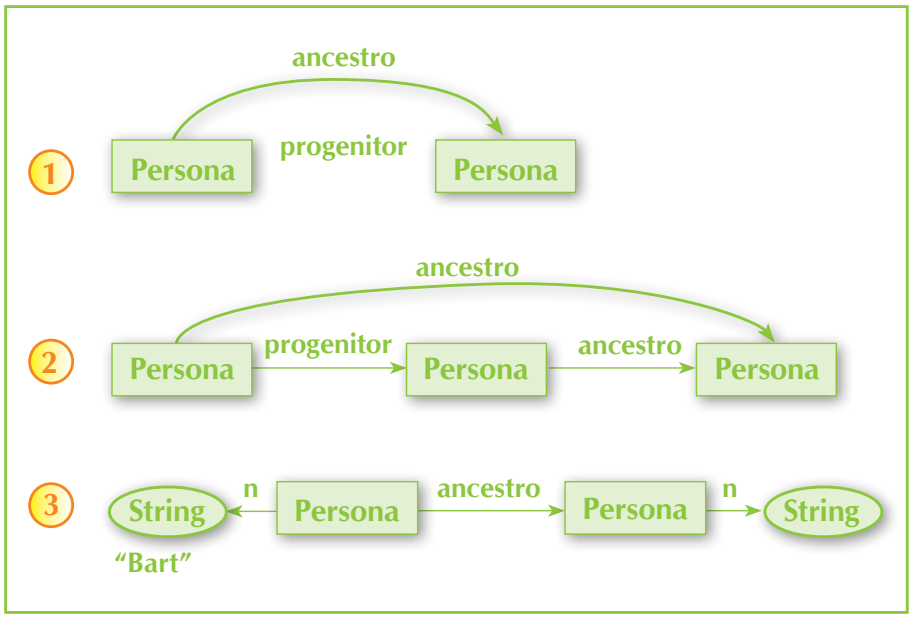


Figura 6. Consulta recursiva en grafos. La consulta retorna los nombres de los ancestros de "Bart" en el grafo de la Figura 3. La consulta define la relación "ancestro" como la clausura transitiva de la relación "progenitor". Esto se expresa a través de tres patrones de grafo: el primer patrón define el caso base para la relación "ancestro"; el segundo patrón define el paso recursivo; y el tercer patrón usa las definiciones anteriores para definir la salida de la consulta.

Aplicaciones Clásicas

El desarrollo de bases de datos de grafo fue motivado por varias aplicaciones clásicas. A continuación describimos algunas.

La generalización de los modelos de base de datos clásicos. Estos modelos fueron criticados por su falta de semántica, la monotonía de las estructuras de datos permitidas, las dificultades del usuario para "observar" las conexiones entre los datos y la dificultad para modelar objetos complejos.

Las aplicaciones donde la complejidad de los datos excede las capacidades del modelo relacional fueron también fuente de motivación. Por ejemplo, los sistemas de gestión de redes de transporte (ejemplo, rutas de trenes y aviones) o las redes en datos espaciales (ejemplo, redes de autopistas y transporte público). Muchas de estas aplicaciones se encuentran en sistemas de información geográfica y bases de datos espaciales.

Las limitaciones en el poder expresivo de los lenguajes de consulta motivaron la búsqueda y definición de modelos que permitieran una mejor representación y consulta de aplicaciones complejas. Otras limitaciones se encuentran en los sistemas de representación del conocimiento, donde se observó la necesidad de técnicas intrínsecas pero flexibles para la representación de éste.

El uso de grafos en el diseño de modelos semánticos y orientados a objetos motivó la idea de definir modelos de datos basados "netamente" en una estructura de grafo. Esto fue acompañado de la necesidad de mejorar la funcionalidad entregada por los modelos orientados a objetos. Por ejemplo, considere aplicaciones como CASE, procesamiento de imágenes y análisis de datos científicos. La aparición de Hipertexto en línea evidenció la necesidad de otros modelos, principalmente semiestructurados. Además la Web creó la necesidad de un modelo de datos más apropiado para la representación e intercambio de información.

Redes Complejas

Diversas áreas han corroborado la aparición de grandes redes de datos con propiedades matemáticas interesantes llamadas Redes Complejas [5].

En las redes sociales [6], los nodos representan personas o grupos y las aristas relaciones o flujos entre los nodos. Algunos ejemplos son las redes de amistad (ejemplo, Facebook), contactos de negocio, patrones de contacto sexual, redes de investigación (ejemplo, colaboración, coautoría), registros de comunicación (ejemplo correo postal, llamadas telefónicas, correo electrónico), redes de computadoras, etc. Hay una actividad creciente en el área de análisis de redes sociales [7], así como en la visualización y técnicas de procesamiento de datos de estas redes.

Las redes de información modelan relaciones que representan flujo de información, por ejemplo, las citaciones entre artículos de investigación, información en la Web [8], redes peer-to-peer, redes de preferencia, etc.

En las redes tecnológicas los aspectos espaciales y geográficos de los datos son dominantes. Algunos ejemplos son Internet (como una red de medios físicos), redes de distribución de servicios básicos (ejemplo, luz, agua, gas, teléfono), rutas y espacio aéreo, redes de entrega postal, etc. Hoy en día, el área de Sistemas de Información Geográfica (GIS) [9] cubre en gran parte esta área.

Las redes biológicas representan información de esta área cuyo volumen ha llegado a ser un interesante problema debido a la necesidad de automatizar el proceso de recolección y análisis de datos. Un buen ejemplo son los datos del genoma [10], con sus problemas de identificación y búsqueda de genes, secuencias metabólicas, estructuras químicas, etc.

Cabe resaltar que los lenguajes de consulta clásicos ofrecen poca ayuda al tratar con los tipos de consulta necesitadas en las áreas mencionadas anteriormente. Por ejemplo, en los GIS se tienen operaciones geométricas

El uso de grafos en el diseño de modelos semánticos y orientados a objetos motivó la idea de definir modelos de datos basados "netamente" en una estructura de grafo. Esto fue acompañado de la necesidad de mejorar la funcionalidad entregada por los modelos orientados a objetos.

(ejemplo, área, intersección, inclusión, etc.), operaciones topológicas (ejemplo, adyacencia, caminos, vecindad, etc.) y operaciones métricas (ejemplo, distancia entre entidades, diámetro de una red, etc.). En las redes genéticas se buscan componentes conectados (ejemplo, interacción entre proteínas), grado de cercanía y vecindad (ejemplo, correlaciones fuertes entre pares). En las redes sociales se mide la distancia entre nodos, el vecindario y coeficiente de agrupamiento de un vértice, tamaño de componentes conectados. En una red de información, como la Web, es natural la necesidad de consultar las conexiones entre los recursos.

RESUMEN

Un modelo de base de datos es una herramienta conceptual que permite definir la estructura de los datos, sus restricciones y la manera de manipularlos (actualización y consulta). Los modelos de base de datos de grafo permiten: modelar naturalmente datos con estructura de grafo, manipular directamente esta estructura, definir restricciones de integridad acorde con ésta e implementar estructuras y algoritmos especiales para almacenar y consultar grafos. Aunque estos modelos perdieron fuerza después de su mayor desarrollo a mediados de los años '80, el incremento de aplicaciones que usan datos con estructura de grafo y las limitaciones de los modelos de base de datos tradicionales (como el modelo relacional) para soportarlas, ha generado el resurgimiento del área. Ejemplos claros de su aplicación son las redes sociales (Facebook), la Web (RDF), información biológica (genoma), etc.BITS

REFERENCIAS

[1] Avi Silberschatz, Henry F. Korth, and S. Sudarshan. Data Models. ACM Computing Surveys, 28(1):105–108, 1996.

[2] E. F. Codd. Data Models in Database Management. In Proceedings of the 1980 Workshop on Data abstraction, Databases and Conceptual Modeling, pages 112–114. ACM Press, 1980.

[3] Shamkant B. Navathe. Evolution of Data Modeling for Databases. Communications of the ACM, 35(9):112–123, 1992.

[4] Renzo Angles and Claudio Gutiérrez. Survey of graph database models. ACM Computing Surveys (CSUR), 40(1):1–39, 2008.

[5] M. E. J. Newman. The Structure and Function of Complex Networks. SIAM Review, 45(2):167–256, 2003.

[6] Robert A. Hanneman. Introduction to Social Network Methods. Technical report, Department of Sociology, University of California, Riverside, 2001.

[7] Ulrik Brandes. Network Analysis. Number 3418 in LNCS. Springer-Verlag, 2005.

[8] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D. Sivakumar, Andrew Tomkins, and Eli Upfal. The Web as a Graph. In Proceedings of the 19th Symposium on Principles of Database Systems (PODS), pages 1–10. ACM Press, May 2000.

[9] Shashi Shekhar, Mark Coyle, Brajesh Goyal, Duen-Ren Liu, and Shyamsundar Sarkar. Data Models in Geographic Information Systems. Communications of the ACM, 40(4):103–111, 1997.

[10] Mark Graves, Ellen R. Bergeman, and Charles B. Lawrence. Graph Database Systems for Genomics. IEEE Engineering in Medicine and Biology. Special issue on Managing Data for the Human Genome Project, 11(6), 1995.