Explotación de bases de datos heterogéneas mediante su integración parcial

Alejandro Botello C.
Centro de Investigación en Computación, Instituto Politécnico Nacional
botello@cic.ipn.mx

Resumen

El presente trabajo de tesis se enfoca en el desarrollo de un sistema de integración de bases de datos que permita: 1) establecer conexiones con bases de datos remotas y heterogéneas para extraer sus esquemas constitutivos; 2) realizar inferencias del contenido de los datos de cada fuente de información mediante el uso de algoritmos de empatamiento para obtener una estructura común y; 3) resolver consultas globales de usuario usando una interfaz gráfica que reúna diferentes y diversas fuentes de información. Nuestra idea principal es combinar los resultados parciales obtenidos de las fuentes de bases de datos dispersas en una forma semiautomática para obtener los resultados relevantes y resolver consultas complejas.

1. Introducción

Una enorme cantidad de información se encuentra actualmente publicada en Internet, lo que permite de intercambio de datos alrededor del mundo. No obstante, esta información no tiene una estructura en común para resolver consultas globales. Por ejemplo, suponga que un usuario desea saber la edad promedio de los estudiantes de nivel superior que viven en el norte de la Ciudad de México y que sus padres trabajen en alguna oficina gubernamental. Para resolver esta consulta, tenemos que descomponer dicha consulta global en subconsultas que puedan ser resueltas en cada fuente de información correspondiente; esto es, necesitamos primeramente obtener los nombres de los estudiantes que viven al norte de la Ciudad de México, para posteriormente obtener los nombres de sus padres y finalmente la oficina del gobierno en la cual ellos trabajan. Suponemos que todas las fuentes de información están dispersas y son remotas, y no sabemos nada acerca de su estructura, ni de qué fuentes información tienen los resultados de cada subconsulta.

Para resolver el anterior problema empleamos el siguiente razonamiento: 1) inicialmente se construye una ontología con los términos más comúnmente empleados en la búsqueda de información, tales como los encontrados en oficinas del gobierno, instituciones académicas y de investigación, empresas comerciales, depositarios públicos, etc. Esta ontología permitirá al usuario guiar su búsqueda incluyendo sólo las fuentes información que ofrezcan resultados parciales. Mediante el desarrollo de una interfaz gráfica, en donde cada fuente de datos se muestra como un grafo conexo, el usuario puede dibujar líneas de conexión entre cada fuente de información para establecer las condiciones a ser aplicadas. 2) En un segundo paso, se hace la transformación del grafo en consultas que deben ser resueltas por el sistema de bases de datos que involucra a la fuente de datos respectivo, escritas en SQL. En esta fase se hará la inferencia de las relaciones entre cada fuente de información que participe en la consulta global: esto es, si el sitio que tiene la edad promedio de los estudiantes no incluye la orientación geográfica de donde viven ellos, necesitamos encontrar algún tipo de relación con otra fuente de datos no considerada inicialmente, como puede ser los mapas cartográficos de la localización de escuelas en la Ciudad de México. Una vez que hemos encontrado estas relaciones, se procederá a resolver las inconsistencias (si existen) para los datos obtenidos como resultados parciales (los nombres de estudiantes como Juan Pérez o Pérez, Juan), empleando algoritmos de inteligencia artificial para unificar los conceptos empatados. 3) finalmente tenemos que hacer la correspondencia para cada elemento de cada fuente de datos a fin de que sea consistente con el esquema común obtenido, e integrar los resultados de cada fuente de información para que sean mostrados al usuario. Si el usuario encuentra alguna inconsistencia en este proceso, puede corregirla y el sistema volverá a realizar la inferencia para aprender cómo resolver el mismo problema en otros contextos.

2.- Trabajos previos

Entre los sistemas más representativos para la integración de información se tienen Carnot e Infosleuth[5], Infomaster[4], TSIMMIS[2], Information Manifold[6], SIMS[7] y GARLIC[8]. Cada uno de ellos busca integrar fuentes de información que esta distribuida en sitios remotos, aunque algunos buscan integrarse únicamente con información obtenida del web.

Con respecto al modelo de datos, cada una propone un modelo propio, aunque [4] y [5] se basan en la interacción entre agentes, por lo que basan su modelo en una base de conocimiento, mientras que [2] busca adaptar el contenido de las fuentes al modelo orientado a objetos. No obstante, la mayoría basa su funcionamiento en el método de integración de Global As View (GAV), en donde se construye un repositorio cuyo esquema (global) está constituido de los esquemas particulares de las fuentes de datos que participen en el sistema. No obstante, otra solución es el método LAV (Local As View), en donde se define una vista para cada fuente de información participante que describe que tuplas pueden ser encontradas. Como comparación, GAV tiene como ventajas una mas eficiente reformulación de las consultas globales, mientras que LAV permite tener agregar o eliminar fuentes de información de forma simple. Lo que es importante destacar es que sólo [5] tiene una forma semiautomática (basada en agentes) para la construcción del esquema integrado, mientras que en los demás tiene que ser de forma manual.

3. Criterios de la investigación

1. Hay muchas bases de datos disponibles, y sólo algunas de ellas participan en la solución de una pregunta dada. Se contempla el modelo relacional como el más ampliamente utilizado para este propósito.

2. Cada base de datos fue diseñada de manera independiente. No fueron diseñadas para usarse en conjunto, un diseñador no tenía idea o no tomó en cuenta la existencia de otras bases.

3. Las preguntas a resolver no pueden resolverse con una base solamente, necesitan más de una.

4. Se diseñará un sistema que responda preguntas del tipo (3) con el universo U de bases de (1). Primero, las entidades de U serán agrupadas (manualmente). Si dos entidades pertenecen a un mismo grupo, significa que son semánticamente idénticas, por ejemplo "persona" de B3 y "ciudadano" de B5, de manera que Juan Pérez que aparece en B3 puede unificarse con Juan Pérez que aparece en B5. Hay otras entidades que, no siendo idénticas, obedecen cierta relación, fórmula o ley, por ejemplo, "diámetro" y "radio."

5. Se formarán los diagramas ERA (Entidad-Relación-Atributos) de cada Bi de U en forma gráfica, con las entidades agrupadas en (4) identificadas por color. Todas las entidades que pertenecen a un grupo de (4) tienen el mismo color.

6. El usuario expresará su pregunta P trazando una trayectoria sobre la gráfica construida en (5).

7. El sistema transformará la trayectoria (6) a sentencias SQL sobre "la base de datos que la trayectoria integra". Es sobre una base de datos "ideal" que se construiría con algunas bases de U, de tal forma que la pregunta P pueda resolverse.

8. El sistema traducirá partes de la sentencia SQL (7) a expresiones (preguntas) parciales e1, e2, .., ek, que pueden resolverse sobre las bases de datos B1, B2, ..., Bk que integran la trayectoria (6). El sistema hará las consultas a tales bases de datos.

9. El sistema integrará las respuestas parciales de (8) para obtener la respuesta total a P. Si durante la solución hubo algunas consideraciones o aproximaciones, el sistema las hará explícitas al usuario.

4. Resultados preliminares

Este trabajo de tesis se enfoca principalmente a resolver los siguientes problemas:

1. Fusión (uso efectivo) de bases de datos que no fueron creadas para trabajar en conjunto, para responder preguntas "complejas", donde las bases no se "funden" simplemente, sino que las bases son disímbolas (hablan de cosas distintas) pero varias de ellas son necesarias para resolver una pregunta.

2. Ya existen preguntas expresadas en forma gráfica. La aportación aquí es:

a) Trayectorias aproximadas (trayectorias punteadas), en los nodos donde se desea que el sistema infiera las relaciones para resolver una consulta.

b) Trayectorias donde la relación no es de "identidad", sino una transformación dada, como la relación entre "diámetro" y "radio".

3. Cálculo de la respuesta aproximada en presencia de conglomerados o valores agrupados. Por ejemplo, de "los reprobados en los años 1, 2 y 3" extraer "los reprobados en el año 3". Otro ejemplo: Deducir la venta de zanahorias en 2005, cuando se tienen los datos de 2000, 2001, 2002, 2003, 2004. (Estas extrapolaciones e interpolaciones no son nuevas, pero su uso en fusión de bases de datos es nuevo. También es nuevo que el programa escoja automáticamente cuál interpolación o aproximación emplear).

5. Referencias

[1] D. Florescu, A. Levy, A. Mendelzon. Database techniques for the World Wide Web: a survey. SIGMOD Record, September 1998.

[2] Sudarshan Chawathe, Hector Garcia-Molina, Joachim Hammer, Kelly Ireland, Yannis Papakonstantinou, Jeffrey Ullman, Jennifer Widom. The TSIMMIS Project, Integration of Heterogeneous Information Sources . Department of Computer Science, Stanford University, March 1997.

[3] William Cohen. Integration of heterogeneous databases without common domains using queries based

on textual similarity. In Proc. Of ACM SIGMOD Conf. On Management of Data, Seattle, WA, 1998.

[4] Michael R. Genesereth, Arthur M. Keller, Oliver M. Duschka. Infomaster: An Information Integration System. In Proceedings of the ACM SIGMOD Conference, May 1997.

[5] Darrell Woelk, Bill Bohrer, Nigel Jacobs, K. Ong, Christine Tomlinson, and C. Unnikrishnan. Carnot and infosleuth: Database technology and the world wide web. In Proc. of ACM SIGMOD Conf. on Management of Data, pages 443--444, San Jose, CA, 1995

[6] T. Kirk, A. Y. Levy, Y. Sagiv, and D. Srivastava. The Information Manifold. In Proc. of the AAAI Spring Symposium on Information Gathering in Distributed Heterogeneous Environments, 1995.

[7] Arens, C. A. Knoblock and C. Hsu. Query Processing in the SIMS Information Mediator. In Advanced Planning Technology, editor, Austin Tate, AAAI Press, Menlo Park, CA, 1996.

[8] Michael J. Carey, Laura M. Haas, Peter M. Schwarz, Manish Arya, William F. Cody, Ronald Fagin, Myron Flickner, Allen W. Luniewski, Wayne Niblack, Dragutin Petkovic, John Thomas, John H. Williams and Edward L. Wimmers. Towards Heterogeneous Multimedia Information Systems: The Garlic Approach. IBM Almaden Research Center, San Jose, CA 95120, 1996.

[9] Adolfo Guzmán-Arenas, Serguei Levachkine, Alexander Gelbukh, Jesús Olivares. Precision-controlled retrieval of qualitative information. Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), México, 2003.

[10] R.J. Miller, M.A. Hernandez, L.M. Haas, L.-L. Yan, C.T.H. Ho, R. Fagin, and L. Popa. The Clio Project: Managing Heterogeneity. SIGMOD Record, 30(1):78--83, 2001.

[11] R. J. Miller. Using Schematically Heterogeneous Structures. Proc. ACM SIGMOD, 27(2):189--200, Seattle, WA, June 1998.

[12] H.-H. Do and E. Rahm. COMA - A System for Flexible Combination of Schema Matching Approaches. In VLDB, 2002.

[13] E. Rahm, P.A. Bernstein. A survey of approaches to automatic schema matching. VLDB J. 10:4 (2001), pp. 334:350.

[14] Jesús M. Olivares Ceja, Adolfo Guzmán Arenas. “Measuring the understanding between two agents through concept similarity. CIC – IPN, México, 2003.