UNIVERSIDAD DE CHILE
ESCUELA DE INGENIERIA
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACION

Tópicos Avanzados en Bases de Datos - CC60W
Integración de Datos
Programa de Curso - Segundo semestre 2007

Horario cátedra	:	Miércoles y Viernes, 10:15 a 11:45
Profesor	:	Pablo Barceló (pbarcelo@dcc.uchile.cl)
Atención de alumnos	:	Acordar cita con el profesor

Objetivo

La integración de datos es el problema de combinar información presente en diferentes bases de datos, y de proveer una vista unificada de esta información para el usuario. Este importante problema emerge en una variedad de situaciones tanto comerciales (e.g. cuando dos compañías necesitan unir sus bases de datos) como científicas (e.g. al combinar resultados de investigación de repositorios bioinformáticos diferentes). El problema de integración de datos aparece cada vez más frecuentemente, debido a que el volumen y la necesidad de combinar fuentes de datos es cada vez más apremiante. Este problema ha sido el foco de un intenso trabajo teórico, y muchos problemas interesantes en el área aún esperan ser resueltos.

El curso intentará esclarecer los principales conceptos, problemas, y técnicas utilizadas en el área de la integración de datos, con énfasis en los aspectos de modelos de datos, procesamiento de consultas, y consistencia. En particular, trataremos los dos paradigmas más utilizados para integrar datos: La integración material de los datos, y la integración virtual de ellos.

Metodología

La parte inicial del curso se basa en clases expositivas de 80 minutos cada una. El resto del curso estará basado en presentaciones de los profesores y los alumnos sobre artículos de investigación. En cada clase habrá una presentación de 40 minutos, seguida de un recreo de 5 minutos y una discusión de 35 minutos sobre el artículo presentado.

Evaluación

Cada alumno deberá presentar a lo menos cuatro artículos. Estas presentaciones son individuales, y la nota del curso será el promedio de las notas de estas presentaciones (60 %) y de la participación en clases (40 %).

Contenido

1. Conceptos básicos del modelo relacional.
   a) Lenguajes de consulta: Algebra relacional y lógica de primer orden.
   b) Inclusión de consultas. Relación con la noción de homomorfismo.
   c) Restricciones de integridad. El problema de implicación y la utilización de chase.
2. Integración de datos.
   a) Introducción.
      - Laura Haas. *Beauty and the beast: The theory and practice of information integration.* ICDT 2007: 28-43.
      - Alon Y. Halevy, Anand Rajaraman, Joann J. Ordille. *Data integration: The teenage years.* VLDB 2006: 9-16.

   b) Dos enfoques para la integración de datos: GAV y LAV.
      - Jeffrey D. Ullman. *Information integration using logical views.* ICDT 1997: 19-40.
      - Maurizio Lenzerini. *Data integration: A theoretical perspective.* PODS 2002: 233-246.

   c) Reescritura de consultas.
      - Alon Y. Levy, Alberto O. Mendelzon, Yehoshua Sagiv, Divesh Srivastava. *Answering queries using views.* PODS 1995: 95-104.
      - Rachel Pottinger, Alon Y. Halevy. *MiniCon: A scalable algorithm for answering queries using views.* VLDB Journal 10(2-3): 182-198 (2001).
      - Serge Abiteboul, Oliver M. Duschka. *Complexity of Answering Queries Using Materialized Views.* PODS 1998: 254-263.
3. Intercambio de datos.
   a) Introducción.
      - Renee J. Miller, Mauricio A. Hernandez, Laura M. Haas, Ling-Ling Yan, C. T. Howard Ho, Ronald Fagin, Lucian Popa. *The Clio Project: Managing Heterogeneity.* SIGMOD Record 30(1): 78-83 (2001).
      - Laura M. Haas, Mauricio A. Hernandez, Howard Ho, Lucian Popa, Mary Roth. *Clio grows up: from research prototype to industrial tool.* SIGMOD Conference 2005: 805-810.
      - Phokion G. Kolaitis. *Schema mappings, data exchange, and metadata management.* PODS 2005: 61-75.

   b) Un enfoque formal para el problema de intercambio de información.
      - Ronald Fagin, Phokion G. Kolaitis, Renee J. Miller, Lucian Popa. *Data exchange: semantics and query answering.* Theoretical Computer Science 336(1): 89-124 (2005).
      - Ronald Fagin, Phokion G. Kolaitis, Lucian Popa. *Data exchange: getting to the core.* ACM Transactions on Database Systems 30(1): 174-210 (2005).

c) Complejidad del problema de calcular soluciones.

- Phokion G. Kolaitis, Jonathan Panttaja, Wang Chiew Tan. *The complexity of data exchange.* PODS 2006: 30-39.
- Georg Gottlob. *Computing cores for data exchange: new algorithms and practical solutions.* PODS 2005: 148-159.
- Georg Gottlob, Alan Nash. *Data exchange: computing cores in polynomial time.* PODS 2006: 40-49.

d) Reescritura de consultas.

- Marcelo Arenas, Pablo Barceló, Ronald Fagin, Leonid Libkin. *Locally consistent transformations and query answering in data exchange.* PODS 2004: 229-240.

e) El problema de manejo de meta-datos.

- Philip A. Bernstein, Alon Y. Halevy, Rachel Pottinger. *A vision of management of complex models.* SIGMOD Record 29(4): 55-63 (2000).
- Philip A. Bernstein. *Applying model management to classical meta data problems.* CIDR 2003.
- Ronald Fagin, Phokion G. Kolaitis, Lucian Popa, Wang Chiew Tan. *Composing schema mappings: Second-order dependencies to the rescue.* ACM Transactions on Database Systems 30(4): 994-1055 (2005).
- Ronald Fagin. *Inverting schema mappings.* PODS 2006: 50-59.

f) El problema de intercambio de datos XML.

- Marcelo Arenas, Leonid Libkin. *XML data exchange: consistency and query answering.* PODS 2005: 13-24.

4. Consistencia de los datos.

- Marcelo Arenas, Leopoldo Bertossi, Jan Chomicki. *Consistent Query Answers in Inconsistent Databases.* PODS 1999: 68-79.
- Leopoldo Bertossi. *Consistent query answering in databases.* SIGMOD Record 35(2): 68-76 (2006).
- Jan Chomicki. *Consistent query answering: Five easy pieces.* ICDT 2007: 1-17.
- Leopoldo Bertossi, Loreto Bravo. *Consistent Query Answers in Virtual Data Integration Systems.* Inconsistency Tolerance 2005: 42-83.