Carleton
UNIVERSITY

Investigación
en
Ciencia de Computación
y
Manejo de Datos

(visión y trayectoria personales)

Leopoldo Bertossi

Contenido:

A. Algunos aspectos de mi trayectoria académica

B. Un área de investigación de los últimos años

C. Proyección de mi investigación

A. Un Poco de Historia Personal (y algo más)

Mi formación básica es en matemática

Hice la licenciatura, magíster y doctorado en matemática, este último en lógica matemática

Lógica matemática "clásica", como usualmente desarrollada por matemáticos

Estudiando con los métodos matemáticos los mismos métodos de la matemática

¿Métodos? Entre ellos ...

- Razonamiento matemático: las leyes/reglas que gobiernan el razonamiento matemático y la obtención de conclusiones de premisas dadas

- Construcciones usuales en matemática

  Por ejemplo, construcciones algebraicas/conjuntistas de estructuras matemáticas a partir de otras

  La abstracción y estudio matemático general de estas construcciones constituyen la "teoría de modelos''

Desde muy temprano en mi formación matemática me interesé en los aspectos lógico-matemáticos de la computación

No en los computadores *per se* ni en la tecnología particular ni en la programación tradicional

Sí en la ciencia de computación:

- Computabilidad:

  Modelos matemáticos de la noción intuitiva de algoritmo

  Modelos matemáticos de computación

  De lo computable y de lo *no* computable

- Complejidad computacional:

  ¿Cuál es la dificultad intrínseca de un problema computacional?

  En tiempo, espacio, aleatoridad, etc.

- ...

¿Y qué tiene esto que ver con la lógica matemática?

Lógica y Computación

- La ciencia de computación parte en los años 30 dentro de la comunidad de lógica matemática, y como parte de ésta

K. Goedel, A. Turing, A. Church, S. Kleene, E. Post, etc.

En gran parte por problemas propuestos explícitamente a principios del siglo XX por D. Hilbert

- La lógica matemática está en la base y raíz mismas de la (ciencia de) computación

La computación en la forma propuesta y desarrollada por J. von Neumann debe mucho a Goedel, Church y Turing

La idea de almacenar programas en computadores generales ya está implícita en Goedel, y explícita en Church y Turing

• Desde hace algunas décadas y especialmente hoy en día la lógica matemática es desarrollada principalmente en ambientes de ciencia de computación

• La ciencia y práctica de la computación han sido una fuente permanente de problemas y desafíos para la lógica matemática

En un círculo virtuoso de retroalimentación positiva

Similar a la relación estrecha entre física y matemática a principios del siglo XX (cuando era más difícil distinguirlas)

- Lógica matemática tiene aplicación directa en computación
  - Modelos de computación y complejidad computacional
  - Lenguajes de programación, incluyendo semántica y lenguajes propiamente tales (e.g. LISP, Prolog)
  - Ingeniería de software (e.g. especificaciones formales, verificación automatizada)
  - ...
  - Representación del conocimiento:

    ¿Cómo representar conocimiento en el computador y usarlo en sistemas computacionales?

    Inteligencia Artificial, Ontologías, Web Semántica, etc.

    Las "bases de conocimiento"son esencialmente teorías expresadas en lenguajes de la lógica matemática

    La extracción de conocimiento implícito se hace mediante razonamiento lógico (muchas veces, no clásico)

- Manejo de datos (data management)

  Como una extensión del área de "bases de datos"

  Las bases de datos relacionales nacen directa y explícitamente de una aplicación de la lógica matemática

  (E. Codd, $\sim$ 1970)

  ¡Un éxito científico, tecnológico, y comercial!

Volviendo a lo personal ...

Terminado mi doctorado en matemática me cambié a ciencia de computación, tanto en ambiente académico como en área de investigación

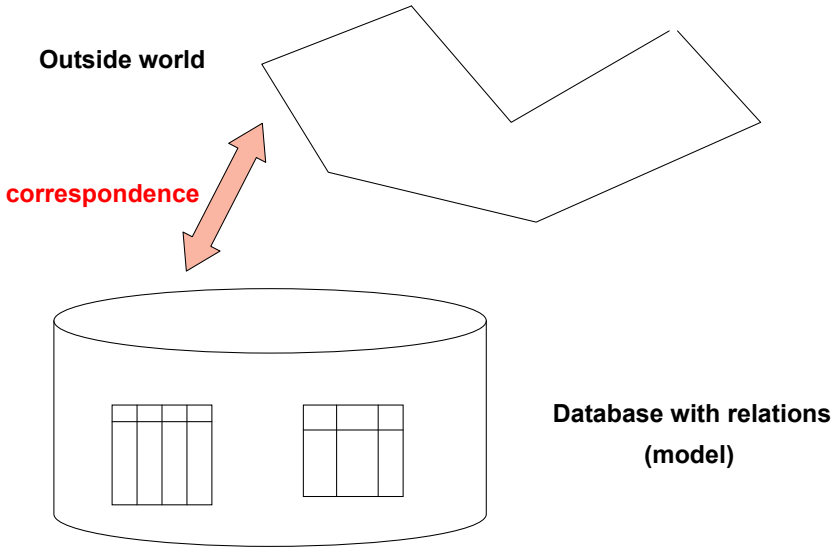
Un cambio natural que estuvo madurando por un cierto tiempo

Conocimientos y destrezas adquiridas en lógica matemática me fueron fundamentales en ciencia de computación

Mi principal área de investigación actual es "data management''

Mi investigación sigue siendo de naturaleza matemática, en términos de problemas de investigación, métodos, y tipo de mis publicaciones

B. Manejo de Inconsistencia en BDs

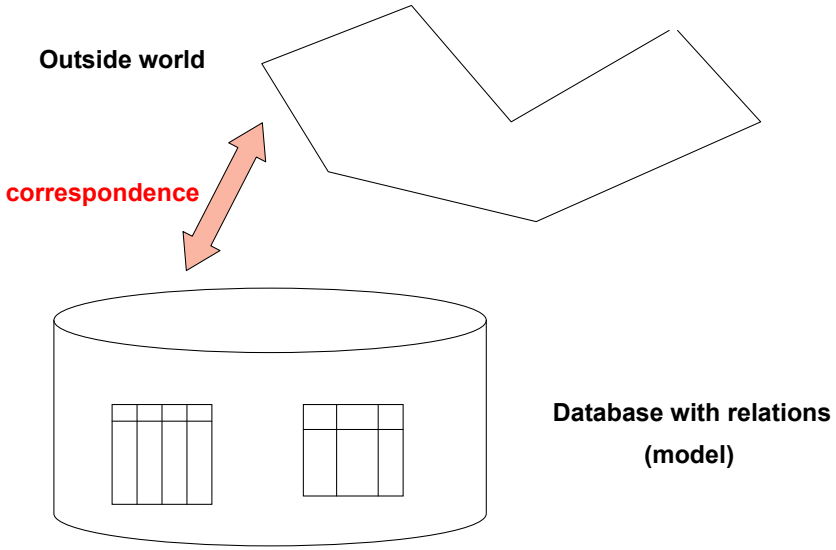


Empleado	*Nombre*	*Salario*
	page	5K
	smith	3K
	stowe	7K

Una base de datos (BD) relacional D es un modelo de una realidad externa

El modelo está expresado en términos de relaciones finitas sobre un cierto dominio de valores

Una restricción de integridad (RI) sobre D es una condición que D debería satisfacer para capturar la semántica del dominio de aplicación



Empleado	*Nombre*	*Salario*
	page	5K
	smith	3K
	stowe	7K

Un conjunto de RI ayuda a mantener la correspondencia entre D y la realidad externa

Por ejemplo, la siguiente es una típica RI, llamada *dependencia funcional* (DF):

"El salario depende funcionalmente del nombre" o

"Cada nombre está asociado a lo más a un salario"

Empleado	*Nombre*	*Salario*
	page	5K
	page	8K
	smith	3K
	stowe	7K

DF expresada como fórmula de lenguaje simbólico de la lógica matemática:

$$\forall x \forall y \forall z (Empleado(x, y) \wedge Empleado(x, z) \rightarrow y = z) \quad (*)$$

Por otro lado, una BD D es una estructura matemática: un conjunto (dominio o universo) con ciertas relaciones sobre él

Como estructura matemática D puede satisfacer o no la RI (*)

En el caso de arriba, no ...

La BD es inconsistente, lo que es indeseable ...

Nuestra gran pregunta: (en algún momento)

¿Si una BD es inconsistente, cuál es la información consistente en ella?

¿Cuáles son los datos consistentes en una BD inconsistente?

En particular:

¿Cuáles son las respuestas consistentes a una consulta hecha a una BD inconsistente?

Queremos obtener información consistente, pero: ¿Cuál es ésa?

Consistencia es global, y buscamos consistencia local ...

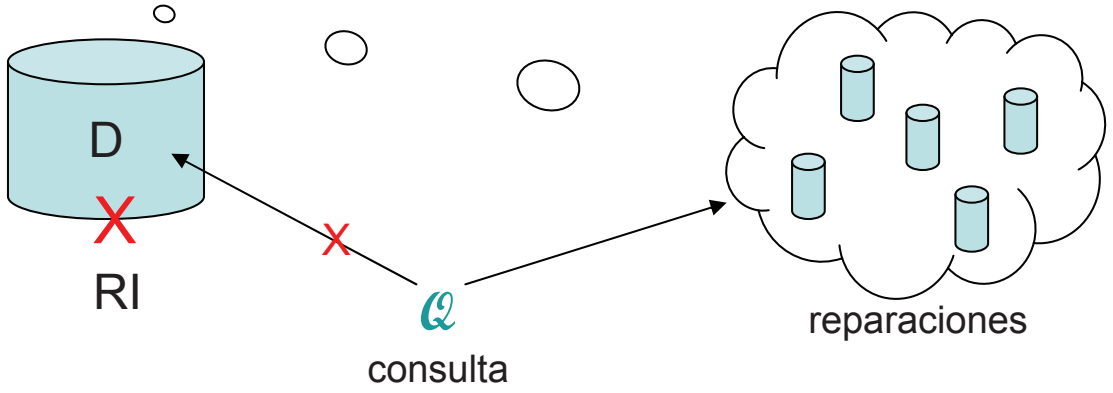
¿Caracterización precisa (definición matemática)?

Propusimos una ... (Arenas,Bertossi,Chomicki; PODS99)

Intuitivamente: Los datos consistentes en una BD inconsistente D son invariantes bajo todas las formas minimales de restitución de la consistencia de D

Es decir, la información consistente persiste a través de todas las versiones mínimamente reparadas de D

Estas últimas son las reparaciones de D



Ejemplo: La instancia D viola la DF

$$\forall x \forall y \forall z (Empleado(x, y) \wedge Empleado(x, z) \rightarrow y = z) \quad (*)$$

Employee	*Name*	*Salary*
	page	5K
	page	8K
	smith	3K
	stowe	7K

Hay dos reparaciones (minimales) si se acepta sólo eliminaciones e inserciones de tuplas (filas) completas:

D_1

Employee	*Name*	*Salary*
	page	5K
	smith	3K
	stowe	7K

D_2

Employee	*Name*	*Salary*
	page	8K
	smith	3K
	stowe	7K

Cada una de estas reparaciones es consistente, satisface (*)

Lo consistente es lo que tienen en común ...

D_1

Employee	*Name*	*Salary*
	page	5K
	smith	3K
	stowe	7K

D_2

Employee	*Name*	*Salary*
	page	8K
	smith	3K
	stowe	7K

$(stowe, 7K)$ persiste en todas las reparaciones: es información consistente

$(page, 8K)$ no lo es (no es verdadera en D_1)

Una respuesta consistente a una consulta $\mathcal{Q}$ desde una DB D es una respuesta a $\mathcal{Q}$ que puede ser obtenida de manera usual a partir de toda posible reparación de D (con respecto a las RI)

- $\mathcal{Q}_1 : Empleado(x, y)?$

  Respuestas consistentes: $(smith, 3K), (stowe, 7K)$

- $\mathcal{Q}_2 : \exists y Empleado(x, y)?$

  Respuestas consistentes: $(page), (smith), (stowe)$

Nuestra definición de información consistente en una BD inconsistente es matemáticamente precisa

Es un modelo matemático, una abstracción, a partir de un problema práctico real

El modelo matemático se puede investigar como tal ...

Ojalá dando aplicaciones de vuelta al problema que lo motivó

Fue sólo el comienzo: Se abrió toda un área de investigación matemática y computacional!

Algunos problemas ...

- ¿Propiedades del modelo?

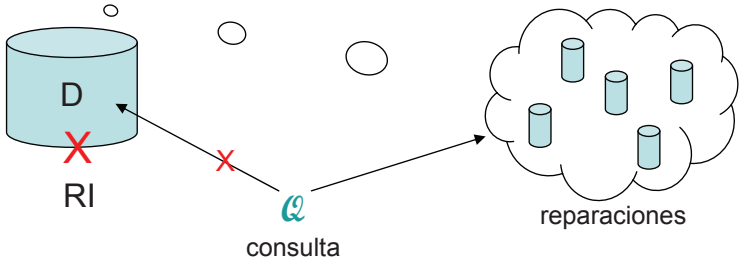
  ¿Propiedades de las respuestas consistentes?

- ¿Si se actualiza la BD D, cómo cambian las respuestas consistentes?

  ¿Computación incremental, sin partir de cero?

- ¿Cómo podemos obtenerlas/computarlas?

  Idea ingenua: Producir todas las reparaciones de la BD inconsistente y ver qué tienen en común ...



Mala idea: ¡Demasiadas reparaciones!

¿Dónde se crean y mantienen para consultarlas?

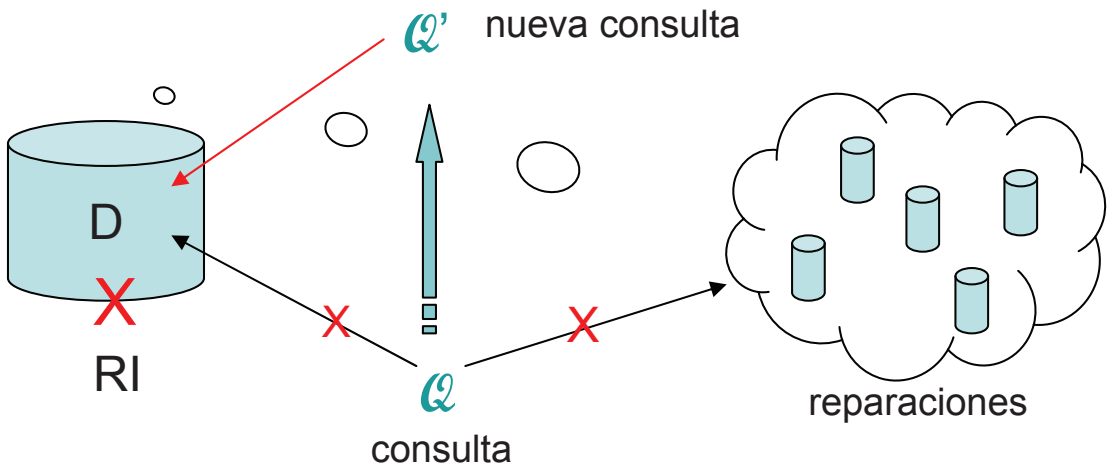
- ¿Cuál es la complejidad intrínseca del problema de computar respuestas consistentes?

- Si es (demostradamente) alta ....

¿Hay clases de RI y consultas para las cuales el problema puede ser resuelto eficientemente?

- ¿Se puede computar las respuestas consistentes a Q en D consultado sólo D? (sin producir las reparaciones)

En algunos casos, reescritura ...



Respuestas consistentes a $\mathcal{Q}_1 : Empleado(x, y)?$

Hágale a D la siguiente consulta y respóndala de manera usual:

$\mathcal{Q}_1' : Empleado(x, y) \wedge \neg\exists z(Empleado(x, z) \wedge y \neq z)?$

Obtenida por "interacción lógica" entre la consulta original y la DF (*)

$\mathcal{Q}_1'$ puede ser respondida eficientemente, como siempre, en un SABD (no hay necesidad de computar reparaciones)

- ¿Hay límites, y cuáles, pare este método de reescritura?

  Sí los hay: para ciertas consultas y RI, se puede demostrar que la computación es "intratable"

- Para los casos "duros":

  ¿Alguna lógica con lenguajes más expresivos para reescritura de consultas?

Todos estos son problemas matemáticos difíciles

Aún hay muchos problemas abiertos

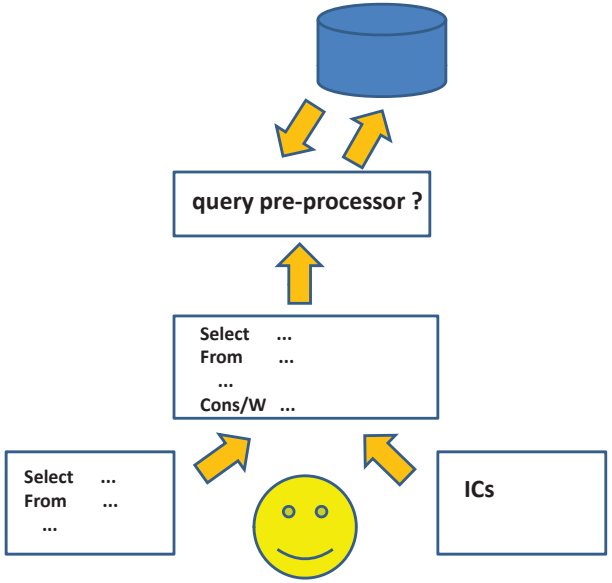
Desde matemáticos, pasando por problemas de implementación, hasta problemas de aplicación ...

Con impacto en aplicaciones reales ...

¿Cómo podrían ser las cosas en la práctica?

Los SABD deberían proveer a los usuarios interfaces y mecanismos más flexibles

Que les permitan especificar sus RI de usuario junto con las consultas



¿Un SQL extendido?

SELECT	Name, Salary
FROM	Employee
CONS/W	FD: Name -> Salary;

(DF no mantenida por el SABDS)

Aunque las RI no sean satisfechas por la BD, un SABD debería entregar las respuestas consistentes

Notar en cambio de paradigma: ¡RI son restricciones sobre respuestas a consultas, y no sobre estados de la BD!

C. Un Par de Direcciones de Investigación

Hay varios temas de investigación en manejo de datos que he estado desarrollando

Sin embargo, hay un par de temas nuevos en mi agenda de investigación

(a) Complejidad Estructural y Contenido de Información

Bases de datos (y sistemas de información) representan y almacenan información

No han sido investigados en profundidad desde la perspectiva de "la teoría de la información"

Como se la encuentra en sistemas termodinámicos y mecánica estadística

En general, la entropía juega un papel crucial, como medida de (des)orden y complejidad estructural

¿Cómo aplicar la noción de entropía en BDs?

Una BD es un objeto estructurado, con una complejidad intrínseca que refleja su organización interna

Las restricciones de integridad (RI) imponen condiciones adicionales sobre esta estructura y organización

¿Cómo cuantificar (modelar y estudiar matemáticamente) el "grado de satisfacción de RI''?

Aplicaciones? Calidad de datos ...

(b) La Emergencia del Contexto y el Sentido

La noción de "contexto" aparece en muchas áreas de computación (context-aware devices ...)

Otra noción de aparición repetida es la de "sentido" (sense)

Ambas aparecen a menudo en manejo de datos:

- La calidad de los datos depende del contexto

  (He hecho investigación inicial en este problema)

- La calidad de un dato depende de su sentido

  Sobre esta noción se define otros predicados de calidad, de más alto nivel

  Sentido parece depender del contexto ...

Se requiere un tratamiento preciso, formal, matemático de contexto y sentido, al menos desde el punto de vista de la ciencia de computación (y manejo de datos, en particular)

¿Cómo emergen los contextos a partir de interacciones sociales entre agentes computacionales? (o humanos)

¿Cómo se usan en la interacción?

¿Cómo modelar, representar, actualizar contextos, y usarlos computacionalmente?

¿Cómo emerge la noción de sentido desde- y es determinada por contextos?

Creo que la "teoría de sistemas" como se ha investigado y estudiado en biología y sociología tiene mucho que aportar a esta discusión ...