IIC3432 — Tópicos Avanzados en Bases de Datos

Programa de Curso

11 de agosto de 2015

Profesor:	Juan Reutter
	jreutter@ing.puc.cl
Clases:	martes, módulos 5,6, en el zoomverse
Atención de alumnos:	por cita concertada via correo electrónico

Descripción

Si bien los sistemas de bases de datos han logrado sobreponerse, en cuanto eficiencia, al tsunami de datos de este siglo, las formas de trabajar estos datos no han sabido actualizarse de la misma manera.
A medida que la capacidad de recopilar datos va creciendo, tareas de análisis de datos que antes eran simples hoy se vuelven engorrosas. Las aplicaciones y sus usarios demandan a los sistemas de bases de datos para que exista la capacidad de analizar datos complejos con la misma facilidad que se construye un cubo o una vista en SQL.
Durante los últimos años hemos visto una multiplicación de iniciativas que aspiran entender cuáles son los límites de lo que se puede ofrecer a los usuarios de una forma relativamente declarativa. Esto requiere saber resolver los nuevos desafíos algorítmicos que se presentan, estudiar formas de expresar nuevas tareas analíticas, y lograr que estas tareas analíticas sean procesadas (y optimizadas) por los sistemas de bases de datos de la misma forma como un sistema toma una consulta SQL.
Buscando presentar esta problemática, y algunos de los avances y preguntas claves que actualmente tiene el área, este curso tiene dos propósitos. Por un lado, proveer una perspectiva más realista de la naturaleza de las tareas de análisis de datos que son requeridas hoy en día en la práctica, y por otro, discutir sobre los desafíos que supone levantar sistemas de bases de datos que integren soporte para estas tareas de forma nativa. Por supuesto, la discusión debe ser técnica y formal, pues de otra forma nos veríamos obligados a navegar entre palabras al aire y conceptos inventados con el único propósito de vender.

Objetivo General

En cuanto a tecnología, el curso se enfoca principalmente en el paradigma de bases de datos de grafos. En cuanto a tareas de análisis, el curso se enfoca principalmente en tareas de análisis de redes y en inferencia causal.
El objetivo de este curso es, entonces, doble. Por un lado se busca que los alumnos sean capaces de usar de la tecnología de bases de datos de grafos, comprender los desafíos actuales de esta tecnología en materia de eficiencia algorítmica, y elaborar pequeños proyectos de investigación que apunten a clasificar y analizar las distintas alternativas. Por el otro lado, se busca que los alumnos estén preparados para desarrollar tareas de análisis de datos en un contexto realista. Finalmente, los alumnos conocerán como estas tareas de análisis empujan algunos de los avances teóricos más importantes en el área.

Metodología

Los contenidos del curso se aprenden mediante una mezcla de clases remotas, videos, lecturas y guías de interacción con sistemas de Bases de Datos. Existen además tres tareas de dos semanas en las que los alumnos podrán profundizar en temas que sean de su elección, más un proyecto final.

Los alumnos deberán realizar sus tareas en grupos de no más de tres personas, y cumpliendo además la condición de que los grupos de cada alumno no pueden repetir personas. La forma de entrega de las tareas es mediante un video de 15 minutos, el que será discutido, en línea, en una plataforma online.
Qué se espera de los alumnos. A lo largo del semestre, los alumnos deberán

- Trabajar las tareas en grupos, presentar sus resultados en un video de 15 minutos, pudiendo agregar además material de apoyo.
- Ver todos los videos de sus compañeros, iniciando discusiones en aquellos que les interesen más (usaremos una plataforma especial para ello)
- Contribuir a la evaluación del trabajo de sus compañeros (se entregará un formulario)
- Trabajar en un proyecto final, presentando un extended abstract y un video de sus resultados.
- Inscribirse para ver, evaluar y discutir sobre algunos de los proyectos finales (usaremos una plataforma especial para ello).

Evaluación

Se realiza en base a

- El cumplimiento de las tareas que se esperan de los alumnos.
- El desarrollo y posterior presentación de las tareas.
- El desarrollo y posterior presentación del proyecto.

Para aprobar el curso, los alumnos necesitan colaborar en la evaluación de los videos, discutir los trabajos de sus compañeros, presentar sus videos y obtener una nota igual o superior a 4 en TODAS las actividades del curso.
De aprobarse el ramo, la nota final se calcula como $= (0{,}4 \cdot \text{NP} + 0{,}2 \cdot \text{NT1} + 0{,}2 \cdot \text{NT2} + 0{,}2 \cdot \text{NT3})$, donde NP es la nota del proyecto y NT1, NT2 y NT3 son las notas de las tareas. De reprobar el ramo, la nota final es el mínimo entre la ecuación de arriba y un $3,9$.

Contenidos

Introducción
Bases de datos al servicio de la ciencia de datos

Bases de datos de grafos

1. Conceptos básicos
2. Patrones como lenguajes de consulta, algoritmos
3. Navegación en grafos, algoritmos
4. Neo4j y Cipher y otras bases de datos

Graph analytics

6. Conceptos básicos y casos de uso
7. Equilibro entre poder expresivo (lenguaje de programación) y hacer todo in-database
8. Algunas soluciones: recursión, message passing networks

Inferencia

11. Inferencia Causal
12. Ontologías
13. Desafíos: inferencia in-database

Calendario de Clases

El calendario podrá estar sujeto a cambios durante el semestre, en cuyo caso se avisará oportunamente.

Fecha	Clase	Descripción
11/08	G1:	Introducción, grafos, lenguajes, sistemas
18/08	G2:	Algoritmos, desafíos actuales
25/08	G3:	trabajo / consulta
01/09	G4:	trabajo / consulta
08/09	G5:	Presentación de proyectos de grafos
15/09	A1:	Analítica en grafos.
29/09	A2:	Message passing networks
06/10	A3:	trabajo / consulta
13/10	A4:	trabajo / consulta
20/10	A5:	Presentación de proyectos de Analytítica en grafos
27/10	I1:	Inferencia
03/11	I2:	in-database inference, in-database learning
10/11	I3:	trabajo / consulta
17/11	I4:	Presentación de proyectos inferencia
24/11	PF1:	trabajo / consulta proyectos finales
01/12	PF2:	trabajo / consulta proyectos finales
16/11	PF3:	Presentación de proyectos finales

Bibliografía

Dado el carácter del curso, existe poca bibliografía formal sobre el tema. Durante el curso se irán entregando papers de lectura obligatoria, y papers de lectura complementaria.

Otros

El Departamento de Ciencias de la Computaciónn adopta una política de tolerancia-cero frente a copias o plagios. Se sugiere revisar las política y penalidades que el departamento establece ante estas acciones.
En particular, se pide a los alumnos **citar** todo el material que usen para el desarrollo de las tareas y del proyecto. La inclusión de material de terceros sin la cita apropiada se considera como plagio, y es penalizado con una nota 1.1 en el curso.

El curso tiene dos canales de comunicación oficiales: Las clases y la página Web (en GitHub). Se asume que que toda la información que es entregada por ambos canales llega a todos los alumnos. Por lo mismo, se sugiere a los alumnos revisar la página Web constantemente.