DSCI 234: Structured and Unstructured Data

Transcript Title: Struc/Unstruc Data: DS Major

Credit Hours: 3

Pre-Requisites: DSCI 134

Description:

This course is an introduction to data models, storage, processing and analysis in connection with advanced data management and information extraction techniques for big data. The course has three components, covering structured, semi-structured, and unstructured data.

(1) *Structured data (Weeks 1-4)*. Students will first develop a basic understanding and the ability to represent, store, process and analyze structured data. Structured data include catalogs, records, tables, among others, with a fixed dimension and well-defined meaning for each data point. Suitable representation and storage mechanisms include lists and arrays. Relevant data and query processing techniques include keys, hashes, stacks, queues and trees. The course will also introduce optimization techniques to search large-scale data.

(2) *Semi-structured data (Weeks 5-8).* In the second part of the course, students will develop a basic understanding and the ability to represent, store, process and analyze semi-structured data. Semi-structured data include texts, web pages and networks, without a dimension and structure, but with well-defined meaning for each data point. Suitable representation and storage mechanisms include trees, graphs and triples. Relevant techniques include XML, RDF, JSON, parsing, annotation, and query processing strategies on semi-structured data.

(3) Un*structured data (Weeks 9-14)*. The third part of the course introduces a basic understanding and the ability to represent, store, process and analyze unstructured data. Unstructured data include images, video, and time series data, without neither a fixed dimension and structure, nor well-defined meaning for individual data points. Suitable representation and storage mechanisms include large matrices, EDF, DICOM. Relevant techniques include feature extraction, segmentation, clustering, rendering, indexing, and visualization.

The course will also cover interesting topics and applications in distributed databases and knowledge discovery of big data, with applications in social analytics, cyber security, and information networks, among others that are already in public eye.

Detailed Syllabus:

Week 1: Introduction:

- overview of data models, data type and their lifecycle
- Big Data: concepts and characterization
- From Data to Knowledge: a knowledge discovery process

Course project milestone 1
Identify and justify data models to a real-world problem

Week 2: Structured data and databases

- Relational data model
- Data storage: List, hashes, stacks, queues and trees
- Data constraints

Weak 3: Query processing

- SQL and query models
- Search and Indexing

Weak 4: Performance analysis

- Cost models and tradeoffs
- Algorithm analysis and optimization

Course project milestone 2
Develop and test search and optimization strategy over a real relational dataset

Week 5: Semi-structured data

- Basics of graph theory
- XML and RDF

Week 6: Searching Big Graphs

- Query languages: XML queries and SPARQL
- Querying techniques

Week 7: noSQL: Beyond Relational DBMS - I

- Key-value
- Column store

Week 8: noSQL: Beyond Relational DBMS - II

- Graph databases

(Midterm exams)
Course project milestone 3
Develop and test search and optimization strategy over a real network dataset

Week 9: Unstructured data

- Image and text data
- Time series data

Weak 10: Querying data streams

- Sampling, indexing and compression

Weak 11: Advanced Topic: Search Big Data

- Approximate query processing
- Parallel query processing: MapReduce

Weak 12: Advanced Topic: Data quality

- Dirty data
- Data cleaning

Week 13: Advanced Topic: Exploring and visualizing data

Week 14: Project presentation

Grading: The course consists of lectures (twice a week), 6 assignments and a course project running through the course. The course project leads to a complete implementation of an analytic tool over a specific real-world dataset in connection to a real-world problem. Students may work in teams for the course project. A final project report and oral presentation should be given by each team. A reading list and sample course project will be provided online. Final grades will be determined as follows:

o Class participation: 10%

o Homework: 40%

o Project: 40%

o Final Project report and presentation: 10%

Note: You may come up with your own project topics and proceed with the permission of the instructor. The proposed topic must be closely related with the course topics.

Textbook (Optional):

- Database Systems: The Complete Book
  https://www.amazon.com/Database-Systems-Complete-Book-2nd/dp/0131873253

A complete reading list will be posted on course website.

Policies:

Missing or late work

Except by prior arrangement, missing or late work will be counted as a zero.

Collaboration

Exploratory collaboration on assignments is allowed (and the final submissions must contain a list of all collaborators). However, students must prepare solutions individually. As an example of the ideal scenario, the following situation is permissible:

A group of students meets to develop the solution to a problem on a white board. Each student records individual notes from this problem solving meeting. All students then prepare solutions individually and without further collaboration. These solutions show the names of all members in the initial group.

Examples of collaborations that are not allowed include, but are not limited to:

 Sharing pieces of any written solution

 Sharing source code or any other computer programs

 Reviewing final written solutions

Collaboration on the projects must be discussed with the instructor.

Changes

This syllabus is subject to change. Updates will be posted on the course website.