Databases And Advanced Data Techniques (CM3010)

Course Notes

Felipe Balbi

April 12, 2021

Contents

Week 1

Key Concepts

- Find, describe and evaluate sources of data
- Understand different forms in which data may come
- Evaluate data-related access and reuse rights

1.003 Introduction to working with data

Data is structured in some form, and we have to be concerned about that. There are different *levels* of structure which can be considered:

Programming Languages Data types (`float`, `int`, `double`, etc) impose a certain structure to the data.

Data Models Relations between different data. Think databases.

Data Serialization Data formats used for transmission using e.g. a network connection.

Exchange Protocols Some form of standardization for information exchange using e.g. Unix Sockets, Named Pipes, shared memory or similar methods.

User Interfaces Data is user interfaces is structured in a way that's comfortable for humans to consume.

Some of the *shapes* of data we will deal with are:

- Tables
- Trees
- Graphs
- Media (raw data)
- Documents & objects

1.005 Reading list

- Chen, P. 'The Entity-Relationship Model – Toward a Unified View of Data', ACM Transactions on Database Systems 1(1) 1976, pp.9–36.
- Codd, E. 'A relational model of data for large shared data banks', Comms of the ACM 13/6 1970, pp.377–87.
- Codd, E. 'Normalized data base structure: a brief tutorial'. In Proceedings of the 1971 ACM SIGFIDET (now SIGMOD) Workshop on Data Description, Access and Control (SIGFIDET'71). Association for Computing Machinery, New York, NY, USA (1971) pp.1–17
- Date, C.J. Database Design and Relational Theory. (Healdsburg, CA: Apress, 2019) Chapter 4. FDs and BCNF (informal)
- Härder, T and A. Reuter 'Principles of Transaction-Oriented Database Recovery', ACM Surveys, 15/4 1983
- Katie Rawson and Trevos Muñoz, 'Against Cleaning' from Matthew K. Gold and Lauren F. Klein Debates in the Digital Humanities, 5 (University of Minnesota Press, 2019).
- Lewis, D. CO2209 Database systems

1.101 Where does data come from?

Data can come from different sources:

New Data created for the sole purpose of the current application

Pre-existing Data data that already existed prior to the application being created. Perhaps it's internal *legacy* data, or it's external data that can be acquired from another supplier.

When it comes to new data, we can take different approaches:

Adding data on-demand For example, a hairdresser has bookings with clients. Either of these appointments is a new datum that gets added to the database *on-demand*, i.e. only the customer makes an appointment.

Bulk data entry Some systems can't afford to have only parts of the data available. In such cases, we can either pay for data entry services or rely on some form of crowd-sourcing.

Pre-existing data Whenever we have pre-existing data, it usually needs to be manipulated somehow in order to fit the new system. Some forms of data manipulation are:

Extraction data may already be in a spreadsheet or database and needs to be recovered, or extracted from the original source.

Conversion data may need to be converted into a new format or structure in order to fit new requirements.

Cleaning data may contain erroneous or unnecessary information. These need to be removed in order to prevent problems.

External sources of data are interesting because they amortize the cost of data entry or quality checks. When data is *purchased* from a supplier, it comes pre-cleaned and in a format that's easy to consume. Moreover, we can also have the opportunity of acquiring data produced by experts in a given field.

Conversely, when we acquire data from an external source, we relinquish control over the quality of the data and its structure. The data may also be incomplete and/or ambiguous from our point view; i.e. the level of detail to which a particular piece of information is encoded may be different from what we need. As a final concern, there may be concerns of trustworthiness with regards to the data.

1.103 Ordering some data: What's on the menu?

- Post 1: Trevor Munoz, 'What IS on the menu'
- Post 2: Trevor Munoz, 'Refining the problem'

1.105 What does your data look like?

When modelling real-life data, we must consider what sort of information is necessary for the application.

To motivate the problem, we look at the example of a book. The data required for a book may be:

Type	Book
Weight	557g
Height	172mm
Colour	Red and Green
Title	Gardener's Calendar
Authors	Thomas Mawe, John Abercrombie
Date	1803
Edition	17^{th}

Some questions arise when it comes to which form of e.g. the title to store. From the point of view of finding it in a shelf "Gardener's Calendar" is enough, from the point of view of comparison against other similar titles, a long form may be required.

1.201 Bringing data sources together

- Linked Jazz
- Pratt Institute, How Mapping Relationships Between Jazz Musicians Elevates Unsung Histories

1.203 Licenses, sharing and ethics

In academic and government circles, it's common to make data as openly available as possible. That, however, doesn't apply to all parts of government of commercial world. There are legal restrictions regarding the use of data which need to be considered.

The Linked Open Data Cloud project produces a graph of all the data openly available published in the Linked Data format. Considering the size of the graph which contains but a subset of all openly available data, the question to ask is *Why is so much data being shared for free if information is so valuable?*

To put into perspective, a furniture catalog from any given furniture company will contain many details about every item: price, sizes, materials, photos. In principle, the furniture could be copied from information that can be gathered from catalogues and manuals. However, the furniture company needs their products to be easy to find if they want to sell them. The same argument can be used for many other industries: music industry, electronics, streaming services, etc.

To summarize some of the reasons to share open data:

- To drive sales
- For the common good
- Contract requirements
- Interoperability

Conversely, here are some reasons **not** to share open data:

- Restrictions on source data
- Control of use
- Value of the data

1.204 Licensing

- Alex Ball, How to License research data