Lecture1: Introduction to Knowledge Technology

Sarah Erfani and Karin Verspoor

July,

THE UNIVERSITY OF MELBOURNE

Who and where

Lecturers:

Dr Sarah Erfani, DMD Level 07, Room 7.14,
sarah.erfani@unimelb.edu.au
Mr Jeremy Nicholson, DMD Level 10, Room 10.03,

Oscar Fabian Correa Guerrero, oscarcg
Qingyu Chen, qingyuc1@student.uni
Daniel Ma, xingjunm@student.unimel

Lectures:

Tuesdays, 4:15pm - 5:15pm, Public Lecture Theatre in Old Arts
Fridays, 12:00pm - 1:00pm, Charles Pearson Theatre in the Eastern Resource Centre (ERC)

Who and where

Tutorials: (per your registration) Start in Week 2

Lecture Materials: Lecture slides available on LMS, lectures recorded

- Tutorials
- Discussion board
- Consultation sessions
- Assignment feedback
- Sarah/Jeremy office (by announcement or by appointment)

Prerequisites

Subjects:

- COMP20003 (433-253/293) Algorithms and Data Structures
- COMP90038 (433-521) Algorithms and Complexity

- Assignments to be completed in any p
  (Elementary C and scripts to be used in l
- Familiarity with formal mathematic
- Basic understanding of statistics an
  not essential.

This subject does not include programming language tuition.

Assessment

- Assessment: 50% final exam, 10% mid-semester test, and 40% project
-
-
-

Project 2 will be released in week 8 and due in week 11.
(Dates to be confirmed in project speci
You are expected to complete these
We will discuss the project in more deta

(Note that the non-teaching week is between weeks 9 and 10.)

COMP90049 Subject overview from handbook

"Much of the world's knowledge is stored in the form of unstructured data (e.g., text) or implicitly in structured data (e.g., databases).

"In this subject students will learn algorithms and data structures for extracting, retrieving and storing explicit knowledge from various data sources, with a focus on the web.

Learning objectives

On successful completion of the subject, st

- "To apply knowledge and skills in many fields that need extensive data analysis."
- "To describe and apply the fundamentals of knowledge systems, including data acquisition and aggregation knowledge extraction, text retrieval, machine learning and data mining"

What the subject covers

- Exposure to a range of computing techniques for:
  - Making use of uncertain, irregular, or complex data.
  -
  -

- A broader understanding of the kinds of th be accomplished computationally.
- Insight into some research activities in computing, why they are undertaken, and how.

Content

Week 1-5:

- Basic text processing
- Pattern and string matching, spelling correction
- Web and text search
- Clustering, classification
- Data mining

Along the way:

- Measurement of effectiveness (Evaluation)
- Some interesting algorithms, a little theory
- Bayesian reasoning
- Insights into current research

Beyond the scope of this subject

Far more knowledge technology topics are out there!

- Natural language processing, machin
- Image analysis, image matching.

. . . and many others.

Texts and references

There is no prescribed text. You may find these useful.

- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (2008) *Information Retrieval*, Cambridge University Press. Freely available at `informationretrieval.org`

Doing Data Science: Straight Talk from the Frontline.
(available as eBook:
`http://shop.oreilly.com/pro`

- Ian Witten, Eibe Frank, Mark Hall

Data Mining: Practical Machine Learning Tools and Techniques
`http://www.cs.waikato.ac.nz/ml/weka/book.html`

- Anand Rajaraman and Jeff Ullman

Mining of Massive Datasets
`http://infolab.stanford.edu/~ullman/mmds.html`

What's in a database?

What's in a database?

Okay, **data** obvious

What's in a database?

- ...

Data is everywhere

What's in a database?

- transactional data
- sensor data
- measurements
-

- videos
- maps
- recipes
- ...

Maybe the better question is, What's **not** in a database?

Some Data on Data

Source: Vasant Dhar “Data Science and Prediction” (2013) Communications of the ACM, Vol. 56 No. 12, Pages 64-73 doi:10.1145/2500499

What to do with all that data?

Most data generated by humans and computers today is for consumption by computers

Facilitated by:

- database schemas
-
-

Database querying, and basic computational processing of data, asks:

What data satisfies a given pattern?

- retrieval of records
- linking data across multiple data sources
- descriptive statistics
- report generation, summaries
- visualisations

Uses of computation

What is computation for?

Much of computer science concerns our attempts to coerce computers into accomplishing *tasks*, loosely definable as:

The data may be created for the task, or might b
physical world – transformed, by a device, in
events in our universe.

A context might be a specific piece of hardware or operating system, or might be assumptions such as “the numbers represent prices” or “the text is in ASCII”.

An outcome might be a number, an action, a list of results, . . .

Uses of computation

Computers and algorithms were originally developed to solve what might be called *concrete* tasks. For example (tiny selection):

- Compute a missile trajectory.
- Crack a code (decryption).
-
-
-
- Map mouse movements to cursor movements.

In common: the task is well-defined, we can a is correct.

In these tasks, the data is transformed in a mechanical way or leads to a mechanical action, but only in a very limited way do they enhance our (that is, human) knowledge.

Hence – not “knowledge technologies”.

Not

What data satisfies a given pattern?

(Actually, we want to find *interesting* y the data.)

Finding patterns

Source: Vasant Dhar “Data Science and Prediction” (2013) Communications of the ACM, Vol. 56 No. 12, Pages 64-73

doi:10.1145/2500499

Finding patterns

Source: Vasant Dhar “Data Science and Prediction” (2013) Communications of the ACM, Vol. 56 No. 12, Pages 64-73

doi:10.1145/2500499

Knowledge tasks

Consider tasks where the data is irregular or unreliable, or the outcome is not well-defined:

- Compression of an image.
- List of answers to a typical web query.
- Translation between languages.
- Distance, time, stress, fuel?)
- Deciding what movie to watch.

"What movie to watch?" (Or music to buy, or pl

This is not a computational task – but we do use computers to *mediate* between us and data, in helping to reach a decision.

Context is critical: the origin of the data, the consumer of the output.

These use, produce, or enhance human knowledge.

Knowledge task example: Translation

Why is translation between languages not well-defined?

Because translation leads to loss of meaning, and loss of nuance and "feel".

In: *Consider tasks where the outco*

Out: *Consider work where the result is*

Why would we expect a machine to do better?

What *is* a correct translation?

Paraphrasing Julian Barnes (London Review of Books, 18 Nov 2010) imagine reading a famous 1850s French novel in English

What do you want? Probably, that it provoke the same reactions in you as in a

notice?

And what English? An attempt at 1850s English, w
expression, or modernised?

And what judgements about class and educatio
equivalent.)

Are trousers held up by braces, or pants held up by suspenders?

Data serves as the raw material for creation of new knowledge.

Characteristics of knowledge technologies

Knowledge technologies tend to be either fairly general (e.g., machine learning algorithms) or fairly specific (e.g., machine translation).

components and of solutions to similar prob

- A specialized problem: parse a partic
- An approximate problem: assign a do
- A general problem: find features of the data items that discriminate between categories.

A broader perspective

- Knowledge technologies are occasionally transformational.

- The impact of social networking on email. (NB: slack is explicitly designed to reduce email)
- The impact of search on libraries and
- The impact of blogging (and tweeting newspapers.

Computational thinking

Finding solutions to tasks requires application of computational thinking:

- How should the data be represented?
- How should it be manipulated?
- Does it have properties that let it be addr
- t is possible to eliminate the need to con allowing a focus on local properties? T have to be considered holistically, or c
- How will a solution behave as the data approaches boundary conditions? (Increase or decrease in number of errors; data items unique or frequently repeated; as item size or item number grows, . . .)

Thinking tools for knowledge technologies

Consider *effectiveness* rather than *correctness*.
(Can a document ranking possibly be "correct"?)

Identify features and characteristics that can be quantified.

Identify approximations to the task.

solutions can be automatically learnt. (Which may make a solution easy, but may make it difficult to gain insight into the p

Ask: What does signal look like? What does n
What would a human do, given sufficient sta
What output would a human produce?

Is a human part of the loop in some way? How is the output to be consumed?

Example: All of these questions apply to aspects of web search.

Methods for data analysis

Supervised learning

- Classification
  predicting a discrete class
-

- Association
  detecting associations between fe
- Information organisation; Clusteri
  grouping similar instances into clust
- Reinforcement learning
- Recommender systems
- Anomaly/outlier detection

Example: Supervised Learning (Classification)

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
				no
				yes
				yes
				yes
rainy	cool	normal	TRUE	no
⋮	⋮			

Given information about current weather conditions and the forecast, can we determine whether we will go out to play?

Example: Supervised Learning (Regression)

Can we predict housing prices?

A friend has a house which is 750 square feet – h
expect to get?

(draw a straight line vs. fit a curve)

Summary

- **Next**: Document representation and str