Theme: Making Sense of Complex Data

J. McNaught
A. Galata
What kind of complex data and why?

- Intelligent life revolves around processing of symbolic information
  - Recognise, interpret (make sense of), create
- Layers of symbolic information involving
  - Perception, learning, communication
  - Creativity, technology, culture
- Two major types of symbolic information
  - Images and language (latter: text)
Images and text

• Both are
  • Highly complex
  • Demand much of our visual/language processing apparatus
  • Infinite in variety, laden with ambiguity
  • Involve complex interactions of events that have to be recognised
• Represent bulk of humanity’s knowledge
  • In vast stores
• Challenging to interact with such stores
The theme

- Theme explores theory and practice
- Via Computer Vision techniques: how we can program computers to recognise and extract information from images
  - In the face of uncertain and noisy data
  - Inferencing becomes critical
- Video, medical scanner images, 3D images, …
The theme

• Via Text Mining techniques: how we can program computers to extract information from very large amounts of text
  • Automatically discover previously unknown (by anyone) knowledge
  • Enable sophisticated semantic search beyond the capabilities of the conventional search engine.
Text Mining

• Discover new knowledge from old
• Battery of techniques from
  – IR: indexing, clustering, classification
  – Information extraction: NLP
  – Data mining
• Enables semantic search: goes far beyond conventional search engine capabilities
• 1st (in world) National Centre for Text Mining located in UoM

www.nactem.ac.uk for demos/services
Text mining in a nutshell

Unstructured Text (implicit knowledge)

Information Retrieval

Knowledge Discovery

Information Extraction

Semantic Metadata Annotations

Structured Content (explicit knowledge)

Applications
Semantic search
Data mining

Other data

Text mining in a nutshell
Text Mining: (some) application areas and strategic importance

- Drug discovery
- Competitive intelligence
- Hypothesis generation
- Sentiment analysis, opinion mining
- Security
- Disease surveillance
- Knowledge management
- Semantic search
Market aspects

• Value of global markets
  • NLP in general: $16.07bn by 2021 (MarketsandMarkets)
  • NLP in Healthcare and Life Sciences industry: $2.65bn by 2021 (MarketsandMarkets)
  • Text mining: $6.5bn by 2020 (Allied Market Research)

• Value of European text mining market:
  • $1.23bn by 2019 (MicroMarketMonitor)
Simple? Or not so simple?

A term cloud generated by TM from a BBSRC CEO's blog
TM tool: NaCTeM's TerMine
(note quality of terms and fact they are **multiword** terms)
Visualisation: www.wordle.net (IBM)
Trying to find documents on a protein zinc finger domain called RING

Classic keyword search for “ring” returns 143368 hits
“Power user” tries Boolean search

Classic Boolean keyword search for “ring AND finger” does not help much.
We can use semantic search to greater effect: Named entity search for “PROTEIN:ring AND finger”

Choosing entity facets narrows search further…
A couple of steps thus leads to the 1 document on RING related to rice

Choosing a facet value automatically expands the query for the user

**KLEIO**

```
AND finger AND SPECIES_OFFICIAL_NAME:"oryza sativa"
```

Search Results: 1

Facets:

- **PUBLICATION_TYPE(1+)**
- **MESHHEADING(0)**
- **PROTEIN(6+)**
- **GENE(5+)**
- **METABOLITE(0)**
- **DRUG(0)**
- **BACTERIA(0)**
- **DISEASE(0)**
- **SYMPTOM(0)**
- **ORGAN(0)**
- **DIAG_PROC(0)**
- **THERAPEUTIC_PROC(0)**
- **GENERAL_PHENOM(0)**
- **HUMAN_PHENOM(0)**
- **NATURAL_PHENOM(0)**
- **INDICATOR(0)**
- **SPECIES_OFFICIAL_NAME(2+)**
- **SPECIES_COMMON_NAME(1+)**

Articles: 1 -- 1 of 1

1. Molecular characterization and concerted evolution of two genes encoding RING-C2 type proteins in rice.

Journal: Gene 15/08/2012;505(1):9-18

RING (Really Interesting New Gene) finger proteins are believed to play a critical role in canonical types. RING-H2 and RING-HC, have been well-characterized, the molecular functions of the modified types, particularly the RING-C2 types, remain elusive. We isolated two rice genes ...
Generating questions *with known answers* to cut down search space

<table>
<thead>
<tr>
<th>Results: All citations (120859)</th>
<th>Full text articles (40480)</th>
<th>Sort by: Relevance</th>
<th>Data</th>
</tr>
</thead>
</table>

1. **Improved candidate drug mining for Alzheimer’s disease.**
   
   (PMCID:PMC3955684)  Free resource
   
   Cheng YH, Chuang LY, Cheng HW, Yang CH
   
   
   Cited: 0 times

2. **7.0T nuclear magnetic resonance evaluation of the amyloid beta (1–40) animal model of Alzheimer’s disease: comparison of cytology verification.**
   
   (PMCID:PMC41446193)  Free resource
   
   Zhang L, Dong S, Zhao G, Ma Y
   
   
   Cited: 0 times

3. **Lipid rafts participate in aberrant degradative autophagic-lysosomal pathway of amyloid-beta peptide in Alzheimer’s disease.**
   
   (PMCID:PMC41446310)  Free resource
   
   Zhou X, Yang C, Liu Y, Li P, Yang H, Dai J, Qu R, Yuan L
   
   Neural Regen Res [2014, 9(1):92-100]
   
   Cited: 0 times

4. **Dab2 attenuates elementary in ABCR2 via dictating transforming growth...**

---

**Ask a question**

- What is associated with Alzheimer's disease? (231)
- What is implicated in Alzheimer's disease? (210)
- What treats Alzheimer's disease? (180)
Known answers to “what is linked to AD”
Finding associations

- Known associations (but not to you)
  - Unknown knowns
- Unknown associations (to anyone)
  - Statistics of surprise
  - A related to B, B to C, so infer A to C
Reproducing a finding - reported (11/11) in *Nature Medicine* - with FACTA+, using MEDLINE 2009 data
A complex problem

- TM involves
  - Many **components** (converters, analysers, miners, visualisers, ...)
  - Many **resources** (grammars, ontologies, lexicons, terminologies, thesauri, CVs)
  - Many **combinations** of components and resources for different applications
  - Many different **user requirements** and scenarios, training needs
TM course unit

• Principles of linguistics and natural language processing
  – Not all of linguistics, just what is necessary to support NLP of the type for TM

• The text mining pipeline
  – Information retrieval, information extraction, data mining
TM course unit

• Approaches to TM
  – ML based
  – Rule based
  – Hybrid

• Domain adaptation
  – Most TM systems need to be adapted to handle different domains, different text types, ...
Dealing with real text

- Formats, encodings, markup, low-level processing, tokenising, chunking, ...
- Term extraction, named entity recognition, relation extraction, fact/event extraction
- Partial vs full analysis
- Discourse, sentiment, opinion, ...
TM course unit

• Architectures for TM
  – Complex mix of components
  – Interoperability issues
• Evaluation
  – Developer oriented
  – User oriented
• Development of resources for TM
• Issues in large scale processing for TM
Why is NL hard?
Ambiguity, ambiguity, ambiguity

• Picking up your litter puts road-workers at risk (Motorway sign)
• Women bitten by rabid bat found in crates outside pub
• “Last night I shot an elephant in my pyjamas. How he got into my pyjamas I’ll never know.” (Groucho Marx)
• FOOT HEADS ARMS BODY (headline)
• Donald Trump, Mr Trump, Trump, the President of the United States, POTUS, he, that man
Q/A

• 50% exam, 50% coursework
• Not much programming
  • We work mainly in Java
• More use of applications and TM tools
• Some knowledge of machine learning is assumed
• First lectures are intro to linguistics and NLP, we assume no knowledge of these
• Mix of lectures and labs to back up lectures. No auditing: register for unit
COMP61342
Advanced Computer Vision

Dr Aphrodite Galata & Dr Carole Twining
Can computers understand the visual world?
Why study computer vision?

- Vision is useful
  - Images and video are everywhere

- Vision is interesting

- Vision is difficult
Challenges

Viewpoint variation

Illumination

Ambiguous matches

Self Occlusion

Scale
Challenges

Geometrically under-constrained
Current state of the art
Earth viewers (3D modelling)

**Bing maps**, Google Streetview

Source: S. Seitz
Face detection and Mobile Visual search

Google Goggles in Action
Click the icons below to see the different ways Google Goggles can be used.

Landmark  Book  Contact info  Artwork  Places  Wine  Logo
Toys and robots

http://www.robocup.org/

http://marsrover.nasa.gov/
Vision-based interaction and games

Microsoft’s Kinect

Sony EyeToy

Assistive technologies

Source: S. Seitz
Example Applications covered in the course

Articulated body tracking

Vision for HCI

Motion Segmentation
http://syllabus.cs.manchester.ac.uk/pgt/COMP61342/

- Very basic knowledge of linear algebra & calculus
  - Probability and statistics (desirable but not compulsory)

- Lectures
  - Demonstrations / tutorials and MATLAB practicals
  - Introductory MATLAB exercises
  - Group exercise / presentations

- Assessment:
  - Coursework: 50%
    - MATLAB exercises, essay, group presentation
  - Final Exam: 50%
Summary

- Computer Vision Goals
  - be exposed to many areas of current Computer Vision research.
  - Acquire knowledge of core machine learning techniques used in this area.
  - implement a number of practical assignments to get hands-on experience. No artistic talent required!
  - find out that all that linear algebra and calculus you learned is actually useful for something real

- Who should take Visual Computing
  - Interest in: Computer Vision, AI (e.g. robotics, NLP, learning), Machine Learning, Graphics.
  - basic tools & techniques will be really useful in many other areas.
Contact

a.galata@manchester.ac.uk

Kilburn Building
Room 2.101