Digital Libraries

Revision
Min-Yen Kan
Information Retrieval

- Text
- Audio, Image, Video
- Synchronized

Access

- Persistent identifiers
- Content: TF × IDF
- Metadata: Indexing, Bibliometrics
Information Retrieval and Multimedia

• Traditional Information Retrieval
  – Lexicon and posting file construction and compression
  – Euclidean and cosine similarity

• Multimedia
  – Textual Images: CCITT, OCR sensitivities
  – Image: vector vs. raster graphics
  – Audio: perceptual coding for human limitations

  – Markup Languages
    • SGML to:
      • HTML and XML
      • XML variants: TEI, SMIL, SVG
Indexing and Metadata

• Dublin Core addresses all aspects of metadata
  – Administrative, structural, use, IP and descriptive
• Indexing as one part of descriptive metadata

• Tradeoff in specificity and exhaustiveness in indexing
• Controlled vocabulary
  – Objectives: distinctive terms, help bridge ASK
• Classification
  – Exhaustive, 1 to 1 mapping of possible subjects
  – Faceted indexing for faceted metadata
Identifiers

- Identifiers
  - Properties: persistent, unique, fast resolution, decentralized
  - Two systems: PURL, DOI
- OpenURL – solve appropriate copy problem
Bibliometrics

• Originated in social networks
  – Find power laws exponential distributions
  – Decay in citation rates, impact of time
  – Co-citation and bibliographic coupling
  – Centrality (undirected) and prestige (directed)

• Applying it to the web:
  – Pagerank: iterative prestige, rank only
  – HITS: hubs and authorities on a expanded base set
DL Policy

• Economics of the DL
  – Volume of knowledge vs. publishers’ cost
  – Search engines acting as marketing;
    Websites act as publishing house

• Social Aspects
  – Self-archiving
  – Preservation: Digital Deposit, Internet Archives

• Digital Divide
  – Rich have access, get richer ... poor get poorer
  – Bridge divide through access to resources and education
Information Seeking

• Types of Questions in RI
  – In contrast to the DL and Web

• Seeking as berry-picking
  – Finding and evaluating sources
  – Using others: collaborative filtering
    • Ask-A services and user-user recommender systems

• Aspects of seeking
  – Affective, accessibility and quality factors

• Information Chain
  – And its relationship to citations
  – Evaluating sources
User Interfaces

• HCI goals
  – Feedback, reduce memory load, scaffolding

• Different interfaces for different parts of the seeking process
  – Query specification, Results display, Relevance feedback

• Systems and their properties
  – VQuery, Filter/Flow, QBIC, Flamenco, Tilebars, Infocrystal, Superbook, Tablelens, Startree, Magic Lens
Patterns of Use

• DL, articles have distinct uses
  – Browsing, searching modes
  – Particular to user’s role
• Web users have limited actions, too
  – Case study: the “back” button

In both cases, optimize UI to account for these specifics
Applications

Both applications can be structured as a machine learning problem

• Recommender Systems
  – Memory vs. Model
  – Shilling

• Authorship attribution
  – Non-content word patterns

• Duplicate detection
  – R-measure
Evaluation

• IR based metrics
  – P / R / Sn / Sp and compound metrics

• Library metrics
  – Use centered vs. materials centered
  – Micro vs. macro evaluation
Cyberinfrastructure

• Jim Gray’s 4\textsuperscript{th} paradigm
• How scholarship is changing
  – Elsevier’s applications of linking data
  – Information Velocity
  – What’s the endpoint look like? To you? To others?
Final Exam

• 1 ½ hours, 20% of final grade
  – Definitions
  – Calculation
  – Critical essays
Digital Libraries

Presentation Guidelines

Min-Yen Kan
Presentation format & timing

• 10 minutes of presentation (max 10 slides)
  – 2 minutes (1 slide) to introduce the problem
  – 2 minutes to define the problem
  – 2 minutes evaluation
  – 2 minutes conclusions
  – The rest is up to you.

• 5 minutes for questions
• Only one group member has to be present
• You should be prepared to ask questions of other projects
  – Not graded, but encouraged
Other details

• Will be the same grade for all students unless your team tells me otherwise

• Practice at least once
  – Otherwise, you’ll probably run over time
  – Anticipate questions

• Send me your slides (.PDF or .PPT) to post to IVLE after your presentation
  – Think about publishing your slides, survey paper on the web to help others
Some presentation guidelines

• **Introduction:**
  – Involve your audience immediately and throughout the presentation
  – (1) Tell them what you're going to say, (2) say it, & (3) tell them what you said

• **Questions:**
  – Carefully listen to questions before answering
  – Acknowledge the validity of an appropriate question
  – Don't answer a question that you don't know

• **Visual aids:**
  – Use 1 figure per minute at most, & 1 figure per 2 minutes at best
  – Make every figure interesting
  – Simplify your figures, and then make them simpler.
  – Explain your figures in detail (including defining axes)
  – Use figures as a memory (numbers & words) crutch
  – Don't read from text figures (face audience & paraphrase).
  – Use a CONCLUSION or SUMMARY figure to show you're done

- From Russ Flegal’s class notes
Overall grading metrics

- **Oral Presentation Skills:**
  - Correct use of English.
  - Logical presentation.
  - Conclusions demonstrate critical thinking.
  - Emphasize important points.
  - Good eye contact, do not read presentation.
  - Appropriate non-verbal communication

- **Slides:**
  - Make sure your slides are readable.
  - Use short phrases on slides, say full sentences.
  - Chose a high contrast color scheme and font (generally sans-serif).
  - Don’t put too much text on a slide.
  - Make use of graphics but make sure the graphics do not distract.
Grading metrics

• **Organization**
  – State what his topic is?
  – Main point presented clearly?
  – Speech clearly organized into a few sections?

• **Scientific Presentation**
  – Cite scientific facts, statistics, statements from authorities?
  – Use scientific terms and define these terms for the class?

• **Analysis and Synthesis**
  – Synthesize and compare different articles?

• **Use of Visual Aids**
  – Visual aids add quality to the presentation?

• **Sources**
  – Give proper credit to people whose ideas he borrowed?
  – Figures properly attributed?

• **Questions**
  – Show respect for those who asked questions?
  – Understood question?
  – Answered question well?

• **Overall Quality**
  – Speaker prepared?
  – Present adequate information?
  – Interesting?
  – Understand the material?
That’s all folks!

• Thanks very much!
• Hope it has been a fun and worthwhile course for you…