Digital Libraries

New Media

Week 13 Min-Yen Kan

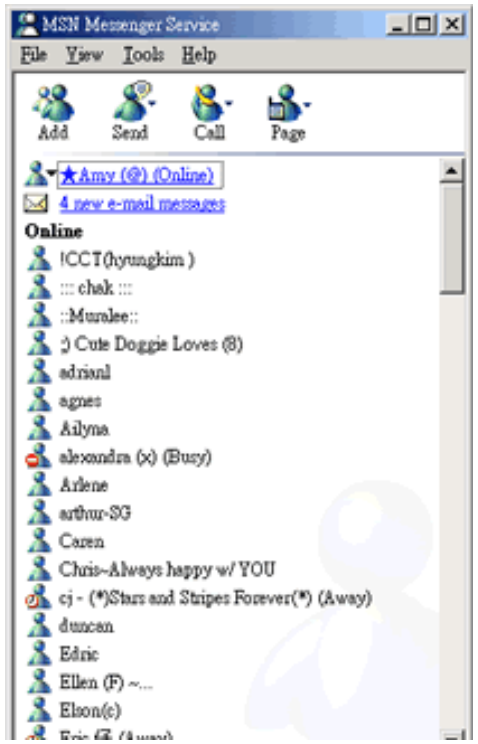
New Media

- Why important?
  - Storing knowledge in these media
  - Communicating about tasks / knowledge
  - Able to identify how information travels from place to place

- New Media to examine:
  1. Instant Messaging
  2. Email
  3. Web logs
  4. Syndication
  5. Wikis

Instant messaging

- **Synchronous**
  - Like `talk` and `IRC`, but centered around user
  - Buddy list, idle counters, emoticons
- Task-based patterns of use:
  - Mainstream users
  - Intense users (frequent, more than *x* conversations)
  - Continuously logged users (lurking)



Properties of IM

- Media switching happens frequently
  - Used to coordinate F2F meetings, telephone
  - Easily recordable
- Variable presence
  - Can be anyplace: need location and time for coordination tasks
  - Idleness hard to determine
    - Even with manually set “away” features
- Lightweight, small footprint
  - Multitasking frequently
  - Short conversations

Improving IM

- Task-related improvements
  - means that only some contacts will be active for some tasks
  - coordination with calendaring
- Turn-taking hard to thread when reviewing
  - More so in multiparty IM
  - Refactoring may be necessary
- More disruptive than email
  - But can be used as sticky note
  - Need accurate “ping”

Email – Task-centric

- Correlated in business roles
  - Not just messaging anymore
  - Has a marked interrupt effect
    - Jackson 2003 study shows people on average read email right away (within 2 minutes) and take ~ 1 minute to recover from interruption.
  - Co-opted by many functions needed in information management
    - Production, transmission and filtering of information
  - Takes the form of tasks:
    - Coordination (Time): calendar and deadlines
    - Collaboration (Other people): contacts

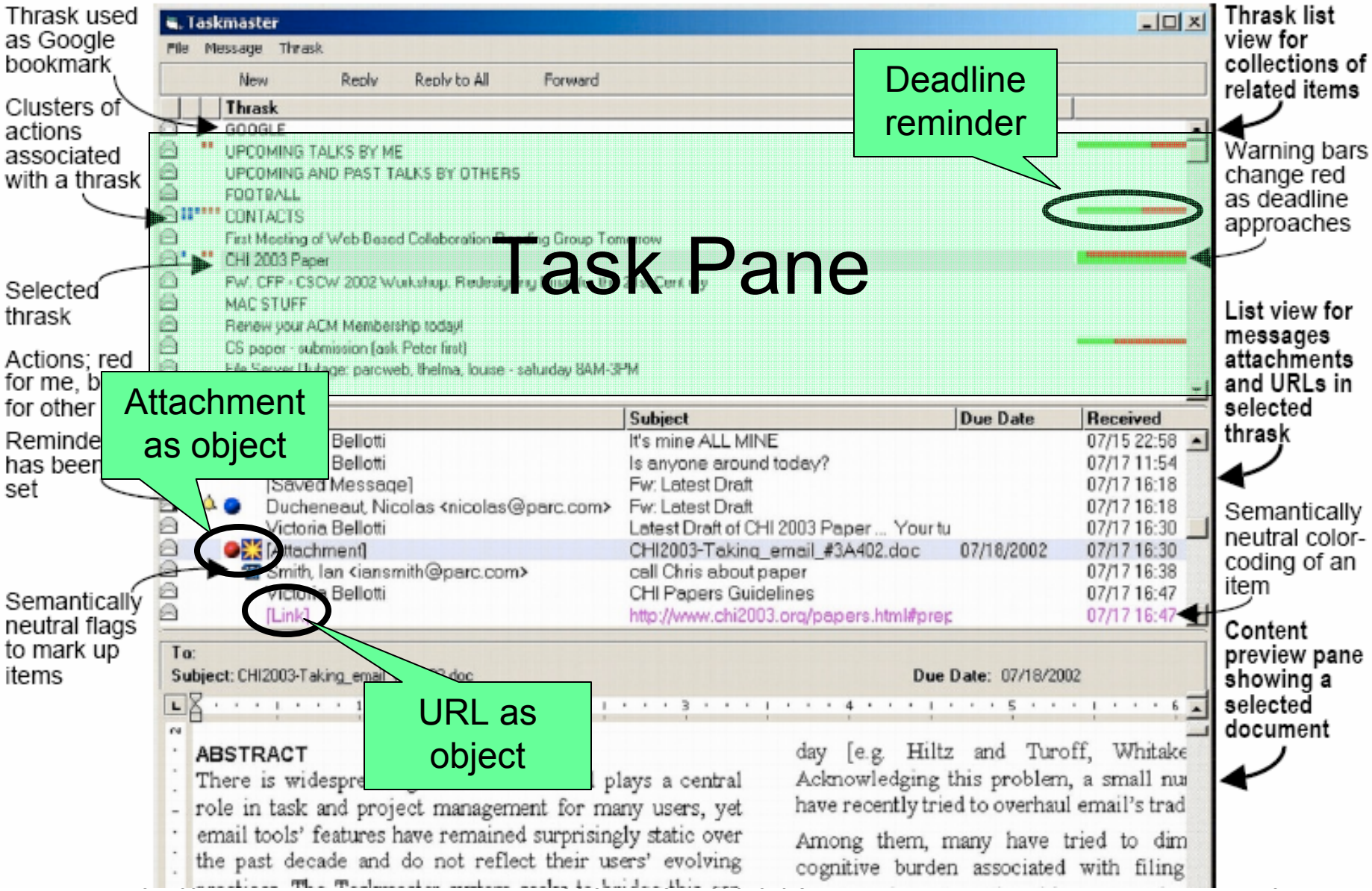
Email – Solutions

- Correlated in business roles with the Todo list
  - One's own messages as important as others'
  - Show sent-mail with incoming mail
- Tasks need support besides messaging
  - Email becomes the Personal Information Mangager (PIM)
  - Email attachments and notes need to be first-class citizens
  - Attachment synchronization (where's the most updated version?)

Email – Solutions

- Extended responses take a while to write
  - Show context of response in drafts
  - Deadlines need to shown to help prioritize
- A task involves a limited set of contacts
  - Use a separate contact list for each specific task
- Still need better solutions to identify overviews
  - Both generic and query-based summaries needed

TaskMaster email client



Finding experts using email

- One way: look in email collections for frequent keywords
- Another way: view `to:` and `from:` as citation link and analyze

- One method to combine the two
  - use HITS algorithm

HITS-based expert finding

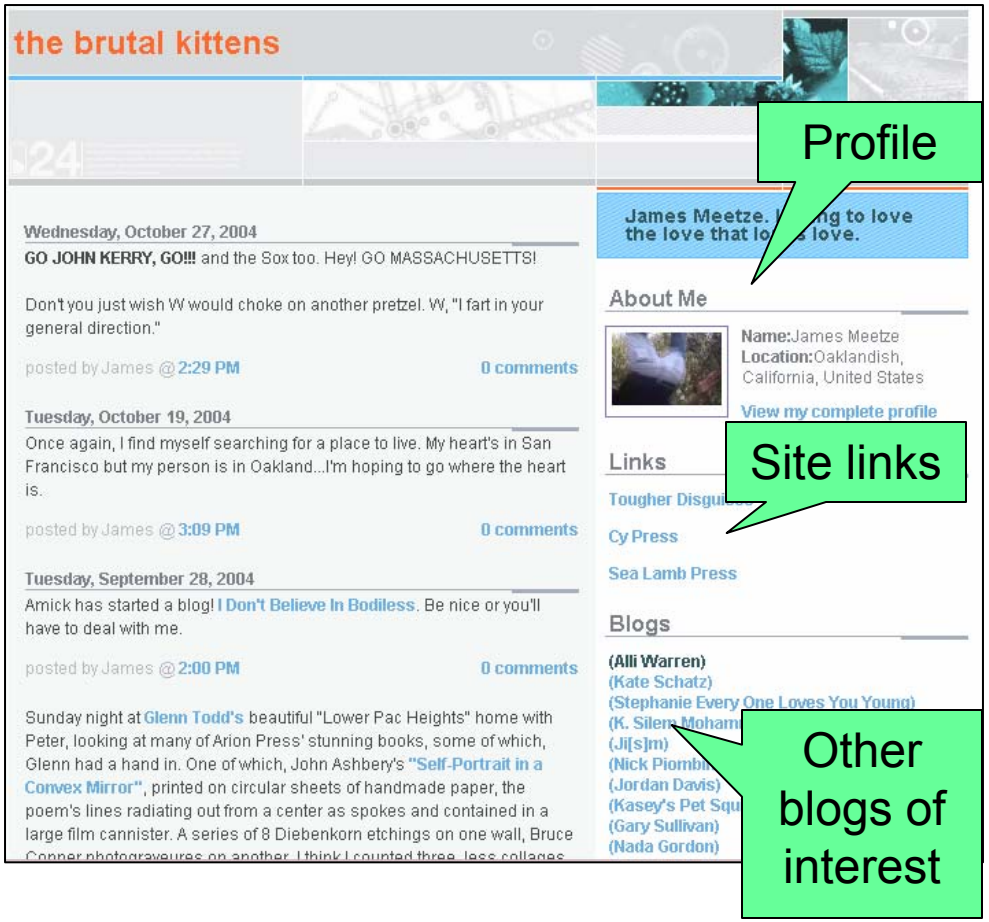
- Campbell *et al.* did exactly this (03)
  - 1. Retrieve all emails from group on subject using keywords search (e.g., “digital libraries”)
  - 2. Run HITS on this set of emails to find authorities
  - 3. Assess correlation with human judgment and compare vs. standard *tf* ranking approach
- Limitation:
  - Need access to emails
  - Email data needs to be classified to filter noise

Web logs - Blogs

History

- web log ⇒ we blog ⇒ blog
- Blogger *et al.* (1999): free web publishing

- Features
  - Chronological
  - Relatively short posts
  - Frequency
  - Vocal



Blogs – a public face of the self

- Public and private mode simultaneously
  - Implicit audience makes it more personal than typical web publishing
- Usually created for self, family and friends
  - Something to: remember, share with others, promote, comment
  - Allows tracking of thoughts in a semi-formal way
  - Hyper linking ability vital

Question: why blog to remember? Doesn't a To Do list or note do this?

Filter blogs for knowledge aggregation

- Two types of blogs:
  - Filter: aggregator, work related
  - Journal: online diaries, personal rants
- Filter blogs
  - Earlier blogs, in which UI emphasized linking
  - Allowed community to form
  - Organized by chronology: enforces currency
  - List other blogs of interest in a *blogroll*

Blog features

Facilitate community building and awareness

- Permalinks
  - Similar to PURLs
  - Semi-transparent, with chronological info

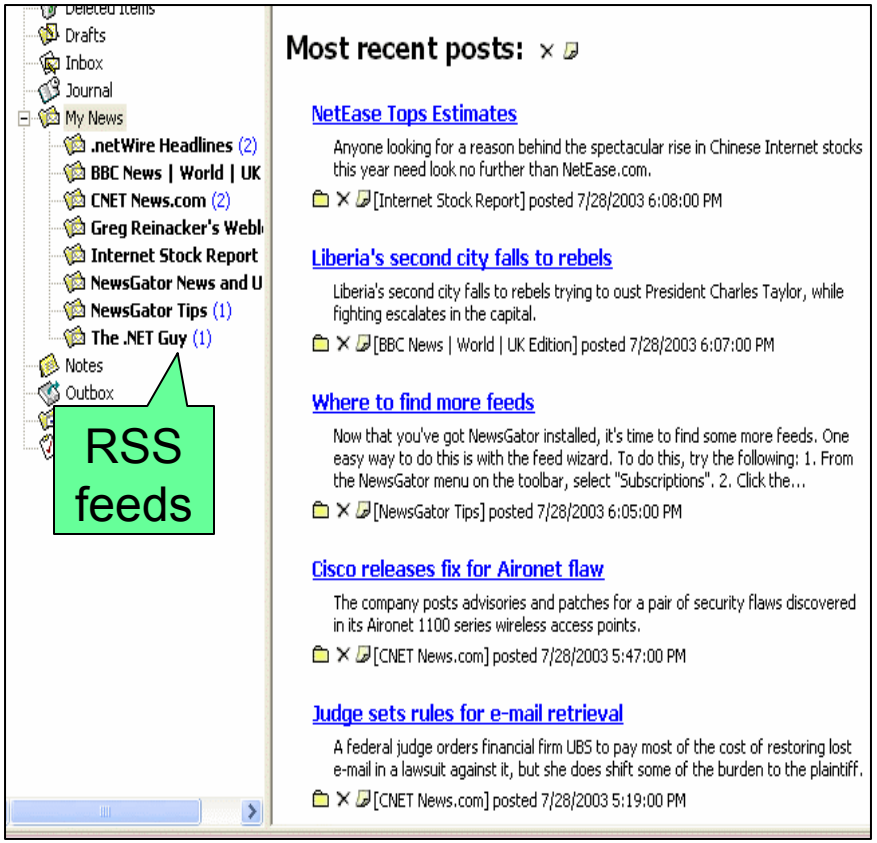
http://<username>.company/<username>/<4 digit year>/<2 digit month>/<15 character name>.html

- Trackback
  - Like SGML, automatically know which site links to yours
  - Implemented by *TrackBack ping*: a message sent back from one webserver to another.

Content Syndication

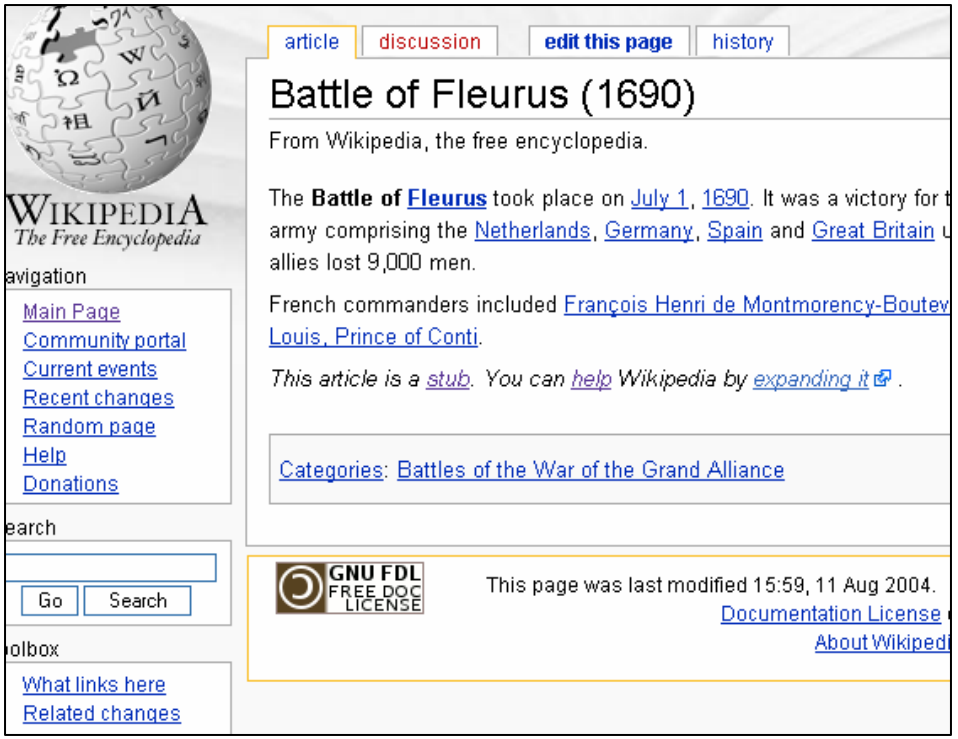
- Chronological ordering may have spurred it
  - Want the “freshest” news
  - Clipping service
- Two current standards:
  - Atom / RSS (Really Simple Syndication)
- Allows aggregation of blog items on a single reader / page

Question: How is it different from mailing lists? From news groups?



Wikis – open to the world

- Wiki wiki = Hawai'ian for "very quick"
- First used in Portland Pattern Repository in 1995
- Allows anyone to post or modify pages
  - Adds edit and create new page buttons to a page
  - Blurs author and reader



- wikipedia.org

Wiki Properties

- Extremely easy to add a link
  - Use CamelCase
  - If page with title “CamelCase” doesn’t exist, it will be created as a stub
- A collaboration tool for webpages
  - Currently hampered by non-WYSIWYG editing (need to know HTML)
- Navigation and linking difficult
  - Anarchic link policy too loose
    - Most sites impose guidelines (although most not enforced)
  - Recency difficult to see
  - Refactoring (page restructuring) necessary

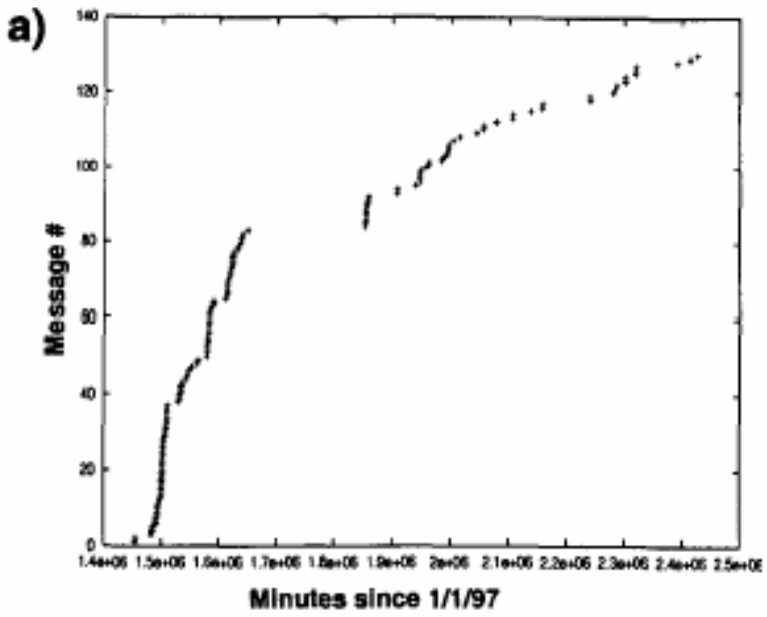
Wiki uses and other hazards

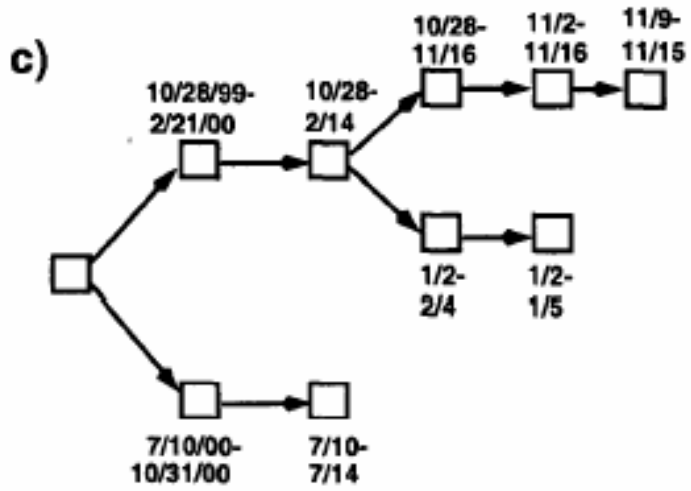
- Structured knowledge base
  - Customer support
  - Reference sites
  - Digital Libraries?
- Skirts issue of trust
  - Shilling possible
  - Link spam

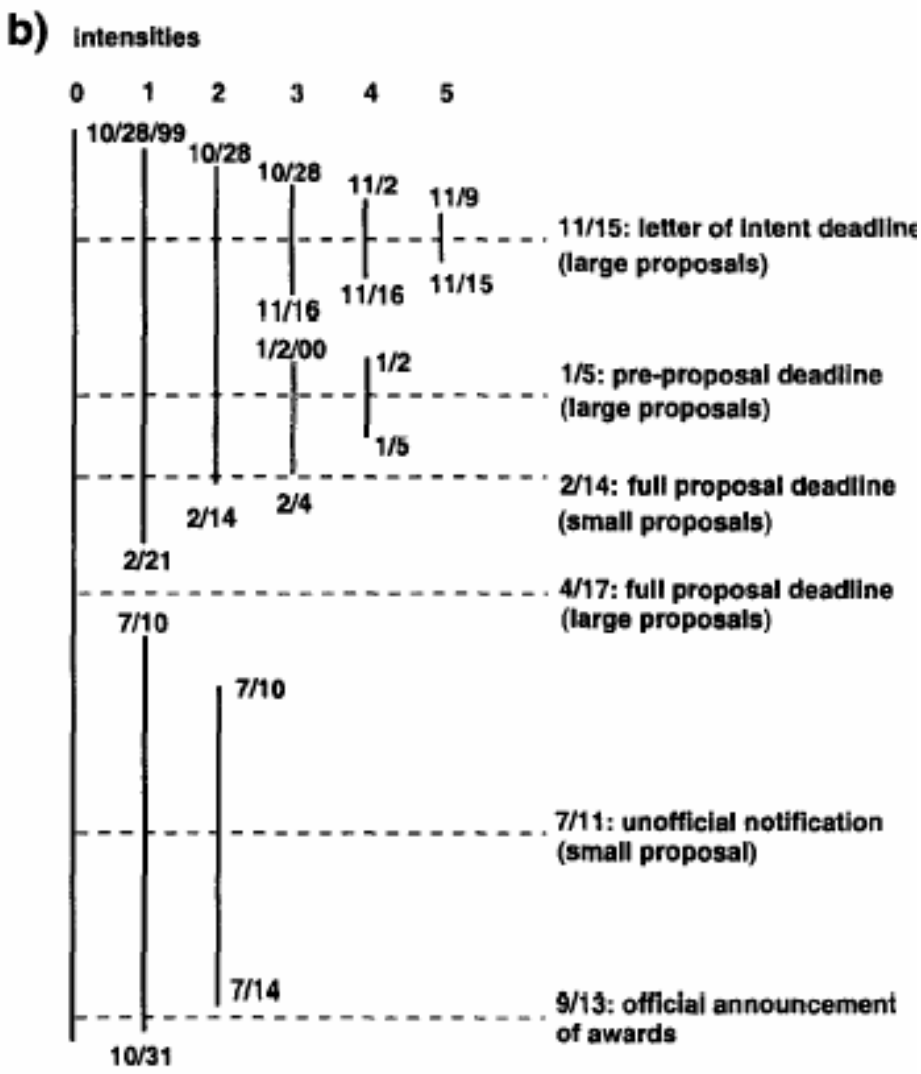
Digital Libraries

Analyzing new media

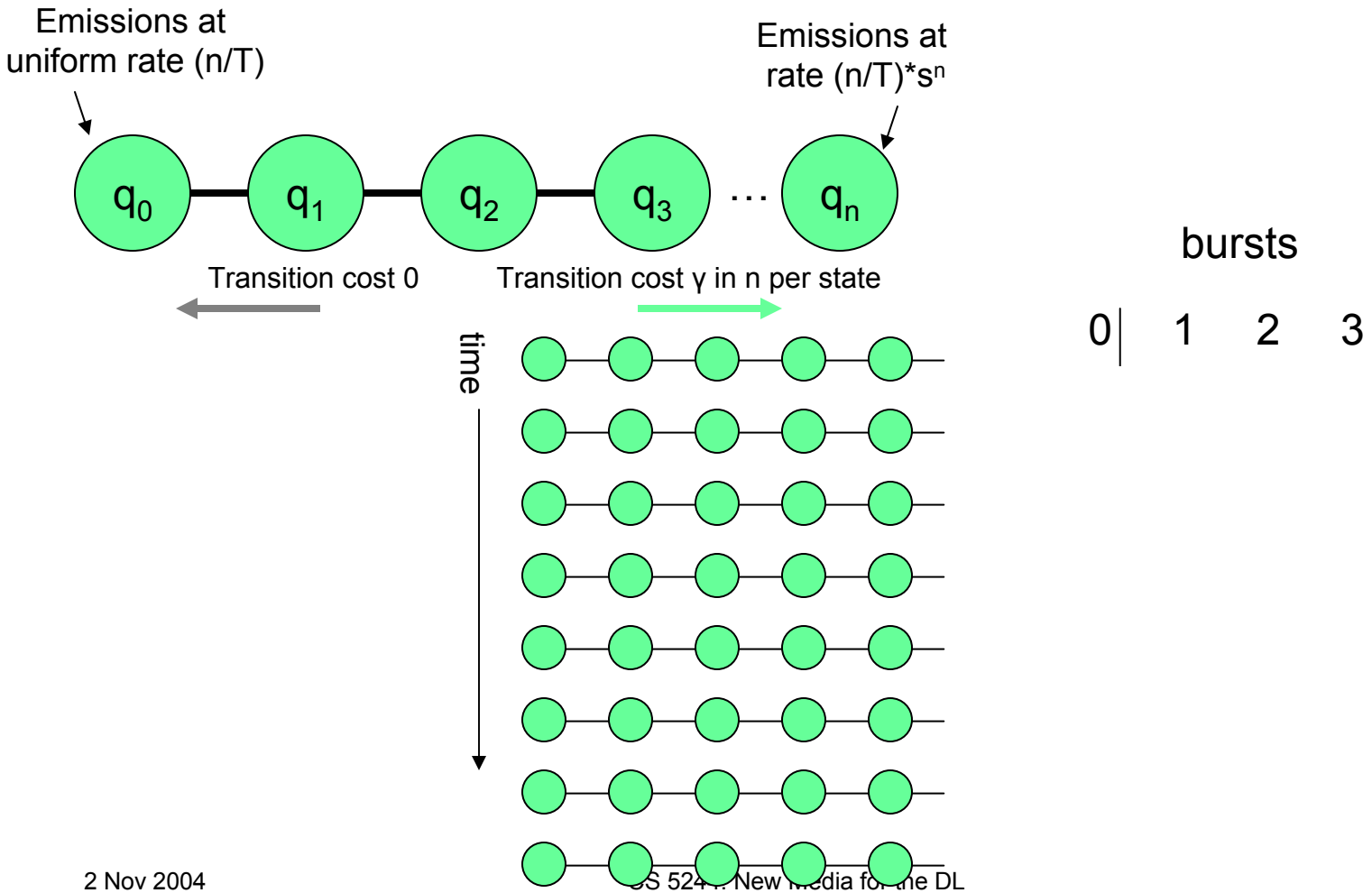
Week 13 Min-Yen Kan


a)
Message #
Minutes since 1/1/97


c)
10/28/99-
2/21/00
10/28-
2/14
10/28-
11/16
11/2-
11/16
11/9-
11/15
1/2-
2/4
1/2-
1/5
7/10/00-
10/31/00
7/10-
7/14


b)
Intensities
0 1 2 3 4 5
10/28/99
10/28
10/28
11/2
11/9
11/15: letter of intent deadline
(large proposals)
11/16
11/16
11/15
1/2/00
1/2
1/5: pre-proposal deadline
(large proposals)
1/5
2/14: full proposal deadline
(small proposals)
2/14
2/4
2/21
4/17: full proposal deadline
(large proposals)
7/10
7/10
7/11: unofficial notification
(small proposal)
7/14
9/13: official announcement
of awards
10/31

Burstiness in Streams



Tracking ideas through blogs

- Strong capabilities of tracking / awareness in blogs
- Gruhl et al. envision a similar model for blog idea tracking: *infection*
  - Threshold model:
    - node adopts idea with probability threshold *t*
    - Iterate at time t
  - Cascade model:
    - If neighbor adopts idea, node adopts with probability *p*

Topic diffusion in blogs

- Topic = keyword
- Need to track relevant words w.r.t. time
  - tf $\times$ cidf (cumulative idf); corpus is a moving window
- Find three distributions of topics
  - Chatter: topics continuously discussed (e.g., alzheimers)
  - Spike: topic exhibiting a usage spike, then inactivity (e.g., chibi)
  - Spiky Chatter: Topics (e.g., microsoft)
    - Overlay of above two types (multiple spikes possible)
    - Spike removal possible with spike model

Conclusions

- New media allow us to rethink and repackage knowledge and its transmission
- Themes of collaboration, informality, recency and ubitiquity throughout along with uncertainty

To think about:

- The Virtual Reference Desk is organized as an email triage center. Do you think new media can improve this initiative?
- How do the new media types handle the different *patterns of use* exhibited by scholars? Which tasks are well-supported? Which are not?

References

- Bellotti et al. (2003) Integrating tools and tasks: Taking email to task: the design and evaluation of a task management centered email tool Proc. CHI 2003
- Kleinberg (2003) Bursty and Hierarchical Structure in Streams *Data Mining and Knowledge Discovery*, 7(4)
- Gruhl et al. (2004) Information diffusion through blogspace Proc. WWW 2004.
- Jackson et al. (2003) Understanding email interaction increases organizational productivity CACM
- Christopher Campbell et al. (2003) Expertise identification using email communications Proc. CIKM 2003.

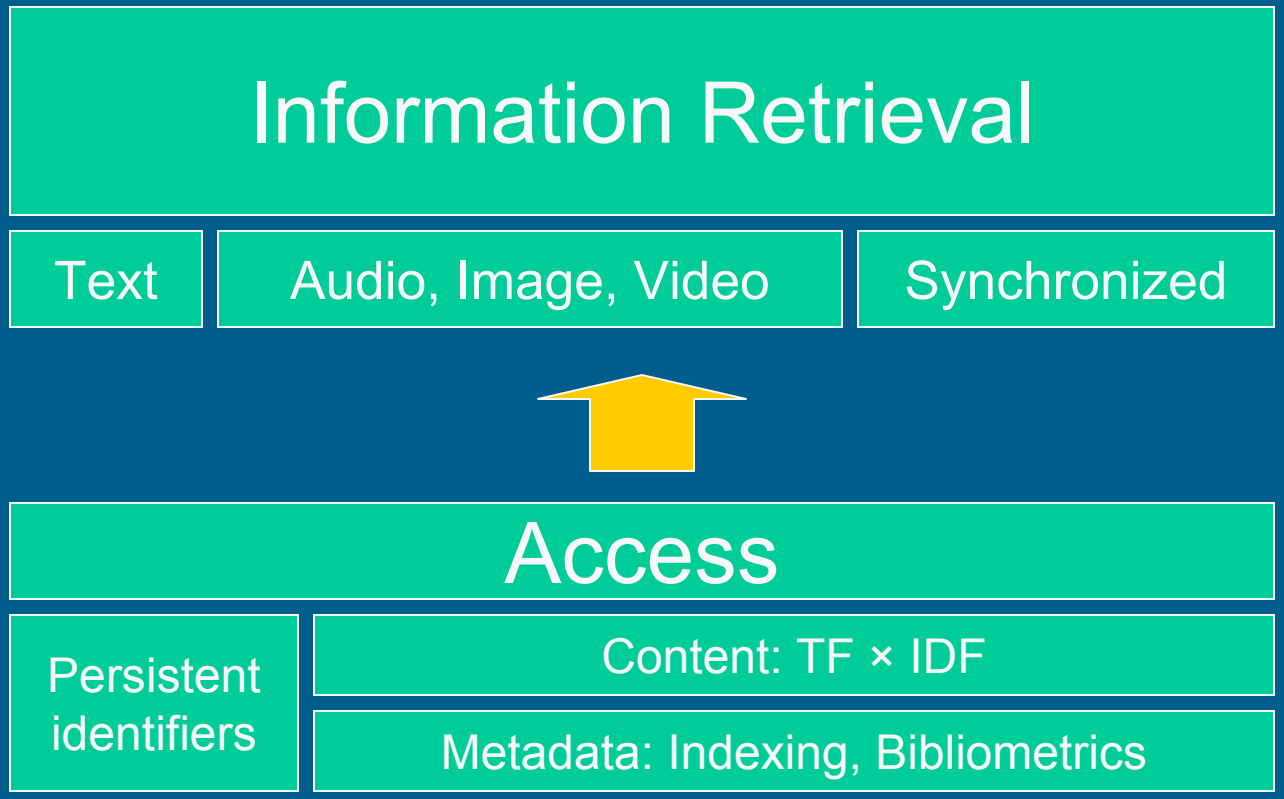
Water break

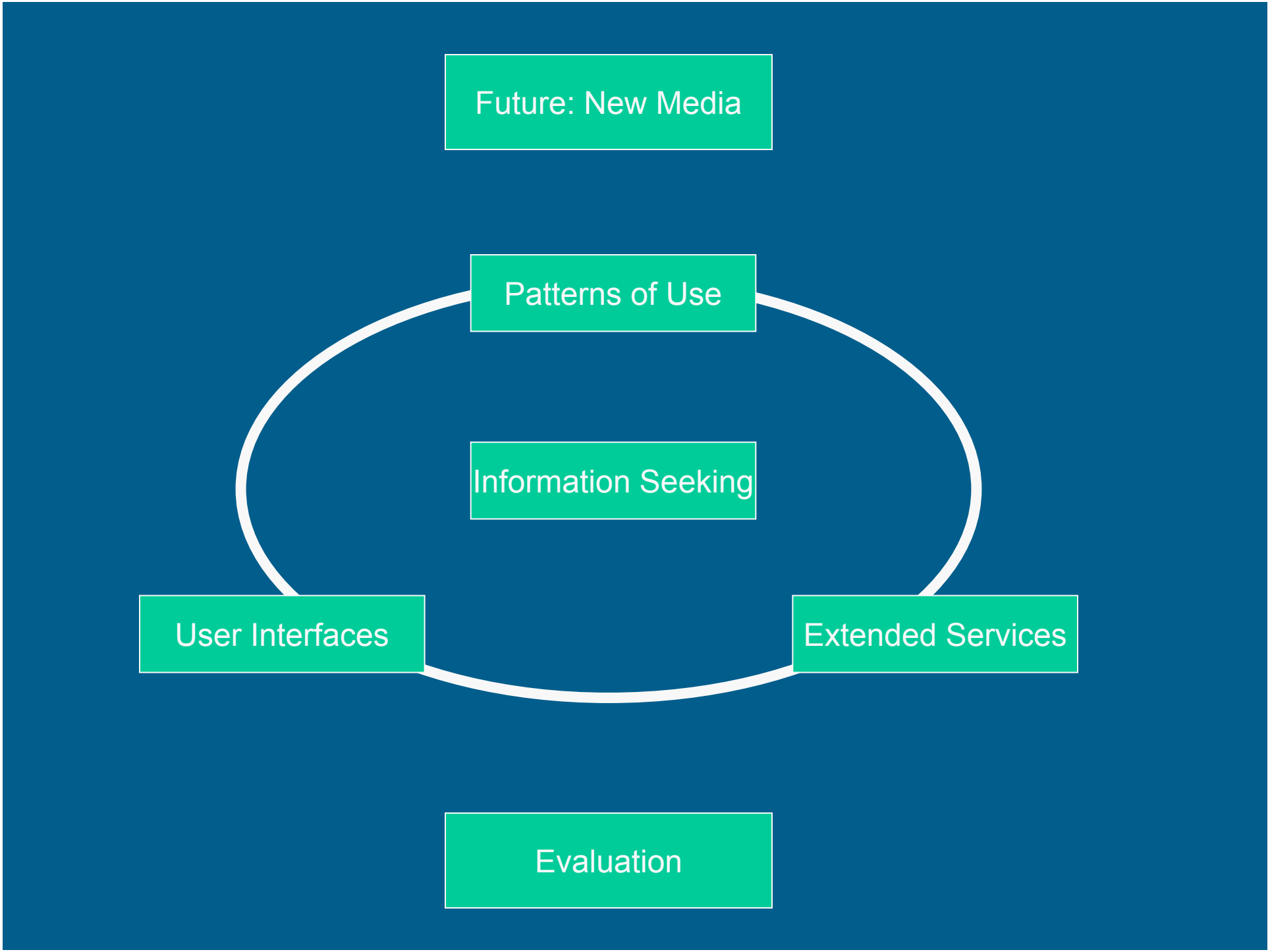
- Last break of the year. See ya!

Digital Libraries

Revision

Week 13 Min-Yen Kan


Information Retrieval
Text
Audio, Image, Video
Synchronized
Access
Persistent identifiers
Content: TF × IDF
Metadata: Indexing, Bibliometrics


Future: New Media
Patterns of Use
Information Seeking
User Interfaces
Extended Services
Evaluation

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

New Media

- IM, Email, Blogs to Wikis: User based
  - Purpose and salient characteristics
  - How do they play a role in the future of the article and the scholar?
- Semantic Web: Agent based
  - Allowing agents autonomy
  - The web as a giant database
  - RDF: representing knowledge as triples
  - OWL: language to map different ontologies

Evaluation

- IR based metrics
  - P / R / Sn / Sp and compound metrics
- Library metrics
  - Use centered vs. materials centered
  - Micro vs. macro evaluation

Final Exam

- 1 ½ hours, 20% of final grade
- Same format as midterm exam
  - Definitions
  - Calculation
  - Critical essay

- Slightly longer (in length) than midterm, questions of higher weight
- Emphasizes second half of course
- First half still fair game
  - some questions may need to refer to first half material

Digital Libraries

Presentation Guidelines*

Week 13 Min-Yen Kan

* These are cribbed from in-class presentation of survey papers, but still apply here.

Also, I will be out next week (7-15 Nov) so I may not be able to proofread much

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

- From Russ Flegal's class notes

- **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

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…