CS/INFO 431

Architecture of Web Information Systems

Spring 2006
Beyond Search and Access

• There is more to life than Google
• Content (data) is just one dimension
  – information
  – knowledge
  – wisdom
• Integrate static content with computational services
• Capture, reuse, and benefit from the “wisdom of crowds”
• Personalize information at various levels of community
  – value of the Amazon metaphor
Library Tradition

• Functions
  – Selection
  – Collection
  – Organization
  – Reference
  – Preservation

• Characteristics
  – Standardized
  – Professionalized
  – Service-oriented
  – In it for the long-haul
  – Conservative
Web “Tradition”

- Decentralized/Anarchic/Illegal
- Agreements are technical (at best)
- Roles are undefined and fluid
- You don’t have to be an expert (or “no one knows you are a dog”)
- Immediate
- Ephemeral
Finding the Appropriate Blend

• There are many points on this spectrum

  • *Evolutionary* perspective: preserve traditional information institutions such as libraries but adapt them to digital context

  • *Revolutionary* perspective: ubiquitous computing and networks render many traditional practices irrelevant
Building systems to add value

- At their core libraries add value to content (organize, select, preserve)
- The Web and Internet is the largest collection of data known to humans
  - Traditional “library-like” content
  - Informal content
  - Artifacts of social interaction
  - Integration of content (data) and services (computation)
- How can we build relationships and integrate this content to add value to it
  - create that data->information->knowledge->wisdom continuum
  - information network overlay
Many facets of the problem/solution

technology

law

economy

human/social factors
What we’ll talk about in this course

- Technical Basics
  - XML
  - XSLT
  - RDF
- Knowledge Organization
  - Cataloging
  - Metadata
  - Ontologies
- Documents
  - Identity
  - Types
- Service-oriented architectures
  - integrating content and computation
- Semantic Web
- Information Preservation
  - Traditions
  - New Models
- Intellectual Property
  - Copyright
  - Rights Management
- Scholarly Publishing
Technical Trade-offs

Cost vs. Functionality
Course Web Resources

http://www.cs.cornell.edu/Courses/cs431/2006sp
Code of Academic Integrity

http://cuinfo.cornell.edu/Academic/AIC.html
Some Pet Peeves
Lagoze’s general course philosophy

- A course is a collaborative experience
- Instructor provides the structure and foundation for learning
- Student engages, contributes, challenges
- We learn from each other
And now for some history...
Library of Alexandria

- Established by Ptolemy I in 290 BC
- 532K papyrus rolls
- Acquisition by copying mandate
- Destroyed in 490 AD during burning alive of Hypatia, the last keeper of the library
- New library at Alexandria
Melvil Dewey

- “Father of modern librarianship”
- Frustrated by dedicated shelving method
- Invented method of classifying into 10 categories
- 21st edition of Dewey Classification system now published
- Started ALA
S. R. Ranganathan

- Colon Classification System
- 42 main classes
- Subject classification by appending facets within class: who, what, when, where
Vannevar Bush

- “As We May Think” Atlantic Monthly 1945
- Pivotal landmark in hypertext research
- “This is the essential feature of the memex. The process of tying two items together is the important thing”
Claude Shannon

- “Father of Information Theory”
- Seminal “The Mathematical Theory of Communication”
- Data vs. Information
Henriette Avram

• “Mother of MARC”, “Melvil Dewey of the 20th Century”
• Developed MAchine Readable Cataloging (MARC)
• Allows standardization and sharing of bibliographic records
J.C.R. Licklider

- “Man-Computer Symbiosis”
- Developed the idea of the “universal network” and interactive computing
- Developed and led ARPANET funding initiative
Inventors of Internet

- Cerf, Kahn, Metcalfe, etc.
- Packet rather than circuit switching
- Layered protocols (TCP/IP, telnet, ftp...)
Ted Nelson

- Inventor of the notion of “non-sequential writing” and term “hypertext” and “hypermedia” circa 1960
- Founder of Project Xanadu
Gerard Salton

- Preeminent figure in modern information retrieval
- SMART information retrieval system: basis of many well-known IR concepts
- Among founders of Cornell CS department
Tim Berners-Lee

- First client and server 1990
- Directory of World Wide Web Consortium and faculty at MIT
Sergey Brin and Larry Page

- Two Stanford students who failed to get their Ph.Ds.
CS 431 Student
Who am I?

- Member of Information Science Program
- Research areas: interoperability architecture, metadata, document architecture, Scholarly Publishing
- Publications, Personal, etc.
  - http://www.cs.cornell.edu/lagoze/