New Tools for Knowledge Management

Hal R. Varian
http://www.sims.berkeley.edu/~hal
School of Information Management and Systems, UC Berkeley

Information retrieval

- H. Simon: “wealth of information…”
- Information retrieval and relevance
- Basic idea
  - characteristics of query and documents
  - probability of relevance
  - logit, neural nets, nonparametric
- New ideas from SIMS (and partners)

Outline

- Controlled vocabulary
- Splitting documents
- Clustering
- Multivalent documents
- Collaborative filtering
- Incentives
- SIMS

Controlled vocabulary

- Sun’s tech support site
  - 60,000 documents
  - 680 Mbytes of info
  - 90% of queries were 1 word!
- Answer
  - classify documents
  - do a 2-level search

Cheshire II

- R. Larson: Cheshire II
- Classify documents
- Save “lead in” terms
- Query user
- Search on controlled vocabulary
- Then search within documents

Problems

- Classification is costly
  - automatic classification works sometimes
- User interface is difficult
  - Marti Hearst’s (March 1997 SciAm)
- Want 2 interfaces
  - naïve
  - experienced
Subtopic splitting

• Nice to have documents similar size
• Can split larger documents along lexical lines into subtopics
• Very helpful in finding relevant documents
• Nice user interfaces possible
• M. Hearst: TextTiles and TileBars

Simple Example Interface

"medical & diagnosis"

Elaborate Interface

After-search clustering

• Disambiguation
• AI approach: hard!
• Pedersen and Hearst: clustering
  – retrieve documents
  – then cluster in order to...
    • disambiguate
    • find themes

NSF Digital Library

• NSF/NASA/ARPA, 4-years
• $4 million
• 6 schools
  – Berkeley, Stanford, Illinois
  – Santa Barbara, Michigan, Carnegie
• http://elib.cs.berkeley.edu

Berkeley

• Partners
  – CS, SIMS, CED
  – Xerox PARC, IBM Almaden, et. al.
• Content
  – California Resources Agency
  – Maps, photos, images, documents
Legacy documents

- How to deal with print?
  - Bit images?
    - not searchable
    - not flexible (can’t copy and paste, manipulate)
  - OCR
    - not reliable
    - destroys structure

Multivalent documents

- Phelps and Wilensky: use both!
  - Bit image linked to OCR ASCII
  - Connected by data structure
  - Users interacts with bit image
  - Computer interacts with ASCII

What can you do?

- Search and highlight
- Copy and paste
- Copy and manipulate
  - references
  - links
  - equations
- Can even manipulate bits!

Other possibilities

- Multiple languages
- SGML and PDF
- Group annotation
- Enliven components
- Geographic layers
- Document recognizers

See it now!

- http://elib.cs.berkeley.edu/ib/about.html
- Need (recent) Java-capable browser
- Capabilities
  - search
  - copy and paste
  - equations
  - references

Collaborative filtering

- Scarcity of attention
- What should I look at?
  - Recommenders
    - friends
    - Consumer Reports
    - Zagats
Automate recommending

- Firefly
- GroupLens
- Workshop
  http://www.sims.berkeley.edu/resources/collab

PICS

- Infrastructure for recommenders
- Inspired by CDA
- But can encode arbitrary metadata
  - Internet
  - Intranet
- Extensions for privacy preferences
- See www.w3.org

Incentives

- How to get people to add to knowledge base?
- Example
- Lesson: worry about incentives
- Possible answer: look at academia

Academic culture

- Publish or perish
- Peer review
- Literature reviews
- Citation analysis
- Reward system for sharing knowledge

SIMS

- Problem: people to make it all work
- Answer: SIMS
  - newest school at Berkeley
  - 12 FTEs, 40 students per year
  - joint programs

Curriculum

- Information transaction
  - Information
  - Communication
  - Users and society
  - Management
**Information**
- Organizing
- Indexing
- Formatting
- DBMS
- Quality assurance

**Communication**
- Complexity management
- Object-oriented design
- Security
- Commerce
- Networking

**Users and Society**
- Needs assessment
- Usability
- Evaluation
- User interfaces
- Groupware
- Law and policy (e.g., IP, privacy)

**Management**
- Planning
- Project management
- Cost accounting
- Organizations

**And more...**
- Continuing education
  - our students
  - others
- Research agenda
  - digital libraries
  - intellectual property
  - economics of information

**SIMS Web Site**
- [http://www.sims.berkeley.edu](http://www.sims.berkeley.edu)
- [http://www.sims.berkeley.edu/~hal](http://www.sims.berkeley.edu/~hal)