Human-Computer Information Retrieval

Gary Marchionini
University of North Carolina at Chapel Hill
march@ils.unc.edu

CSAIL
MIT
November 12, 2004
Message

• IR and HCI are related fields that have strong (staid?) traditions that have been energized (jolted?) by WWW.
• The intersection of these fields offers interesting new opportunities for high-impact IR R&D.
• Integrating the human and system interaction is the main design challenge: syminfororosis—people continuously engaged with meaningful information

Gary Marchionini, UNC-Chapel Hill
Content-Centered Retrieval as Matching Document Representations to Query Representations

A powerful paradigm that has driven IR R&D for half a century. Evaluation metric is effectiveness of the match. (e.g., recall and precision).

Gary Marchionini, UNC-Chapel Hill
Content Trend

• Content Features (queries too)
  – Not only text
    • Statistics, images, music, code, streams, biochemical
  – Multimedia, multilingual
  – Dynamic
    • Temporal (e.g., blogs, wikis, sensor streams)
    • Conditional (e.g., computed links, recommendations)

• Content Relationships
  – Hyperlinks, new metadata, aggregations
  – Digital Libraries/sharia, personal collections

• Content acquires history=>context retrieval
Responses to Content Trend

• Link analysis
• Multiple sources of evidence (fusion)
  – Authors’ words (e.g., full text IR)
  – Indexer/abstractor words (e.g., OPACs)
  – Authors’ citations/links (e.g., ISI, Google)
  – Readers’ search paths (e.g., recommenders, opinion miners)
  – Machine generated features and relationships
• Two key challenges:
  – What new relationships can we leverage (human and machine)?
  – How can we integrate multiple sources of evidence?
Installed User Base Trend

• Technical advances and technical literacy allows us to leverage information seeker intelligence
  – Rather than sole dependence on matching algorithms, focus on flow of representations and actions in situ as people think with these new tools and information resources

• Web and TV remotes have legitimized browsing as human-controlled information seeking

• To leverage human intelligence and effort, people must assume responsibilities: beyond the two-word, single query

• Aim at understanding rather than retrieval

Gary Marchionini, UNC-Chapel Hill
Responses to People Trend

• Adapt techniques to WWW
  – Relevance feedback
  – Query expansion
  – User modeling/profiles, SDI services
• Recommender systems
  – Explicit and implicit models
• Capture everything (e.g., Lifebits)
• Human tuning of IR systems
• User Interfaces
  – Dynamic queries
  – Agile views
An Expanded Model:

Think of IR from the perspective of an active human with information needs, information skills, powerful IR resources (that include other humans), and situated in global and local connected communities, all of which evolve over time.
HCIR

• Get people closer to the information they need
  – Closer to the backend
  – Closer to the meaning
• Involve information professionals as integral to the IR system
• Increase responsibility as well as control
• Leverage more demanding and knowledgeable installed base
• Consider ubiquity, digital libraries, e-commerce as extended memories and tools (personal and shared)
HCIR: Bringing User Closer to World

Rules
Structures
Context
Labels
Help
Start/Stop

World

Gary Marchionini, UNC-Chapel Hill
Key Challenges

• Linking conceptual interface to system backend
  – metadata generation
  – alternative representations and control mechanisms
• Raising user literacy and involvement
  – Engaging without insulting or annoying
• Adding human intelligence to the system
• Moving beyond retrieval to understanding
  – Context
Relation Browser Example with all EIA pages

Gary Marchionini, UNC-Chapel Hill
RB Goals

• Facilitate exploration of the relationships between (among) different data facets
• Display alternative partitions of the database with mouse actions
• Support string search within partitions
• Serve as an alternative to existing search and navigation tools
Relation Browser Principles

• Architectural Principle: Juxtapose facets
  – Two or more with 5-15 categories per facet
  – Topic is one important facet for most applications

• Interaction Principle: Dynamic exploration of relationships between facets and categories

• Database driven to promote flexible applications (requires systematic metadata)
Key Challenges

• Technical evolutions (Java, metadata to client side)
• User expectations and preparations
• Getting metadata and mapping to RB scheme
  – Given the cost and difficulty with hundreds of thousands of web pages, can we automate this process?
'Automatic' classification works best when its application is supported by humans with knowledge of the domain and the techniques at hand.
Behind the RB: Human-Machine Cooperation

A Metadata Mining Toolkit is Available

www.ils.unc.edu/govstat/demos.html

Gary Marchionini, UNC-Chapel Hill
Acquire Data

• Crawl Site (sites)
• Currently HTML only
• Mirror locally
  – E.g., BLS yields 23,530 pages
• Clean data
  – Remove non-alphas
  – Lower case all
  – WordNet validate words
  – Stem or not stem

Gary Marchionini, UNC-Chapel Hill
Build Representation

• Select data to include
  – Pages to include/exclude (e.g. BLS ED, 1279 pages)
  – ASCII text from
    • Titles
    • Link anchors
    • Metadata tags

• Build raw term-document matrix
  – Pages as rows (observations)
  – Terms as columns (variables) (e.g., BLS 26,772 terms)
  – Frequencies or TF-IDF weights in cells

Gary Marchionini, UNC-Chapel Hill
Filter Data

• Stop word lists
  – General terms
  – Domain specific terms
  – Web and navigation terms
  – Iteratively developed/refined
• Term discrimination filters (various)
  – .01-.1 doc frequency interval
  – Interval augmented by 100 top freq
  – Empirical threshold (e.g., > 5 docs)
Project data onto Lower Dimensional Space(s)

- First N principal components
- 50-100 latent semantic dimensions
- 50-100 independent components

- Reduces to ‘narrower’ term-doc matrix
  - Note: we are experimenting with this at this time

Gary Marchionini, UNC-Chapel Hill
Cluster Documents (pages)

- K-means, e.g., with k<<100
- EM yields a probability distribution for each document over the clusters (so a document has some probability of belonging to each cluster)
Evaluate Clusters and Name Topics

• Create usable output
  – A web page with the clusters and number of documents in each
  – For each cluster, a list of the top 10 most frequently occurring terms; a list of the top 10 log-odds ratio terms; and links to all the pages in that cluster
  – Eyeball the terms, pick a cluster (topic) name (names); else iterate previous steps
Assign Pages to Topics

• For every page, compute the probability distribution (using EM model) over each cluster/topic

• Select a threshold for placing pages into topics (most easily go into only one topic)
Create Other Facets and Pipe to RB

- Use a set of heuristic rules to place pages into geographic categories
- Use a set of heuristic rules to place pages into temporal categories (ad hoc at present)
- Map the files onto the RB relational scheme
RB is Embedded in Larger Process of Information Seeking

Gary Marchionini, UNC-Chapel Hill
Open Video Example

www.open-video.org

- Open access digital library of digital video for education and research
- 2000+ video segments: MPEG1, MPEG-2, MPEG-4, QuickTime
- Multiple visual surrogates
- Agile Views Design Framework
  - Different types of views
    - Overviews, previews, shared views
  - Multiple examples of views
  - Dynamic control mechanisms

Gary Marchionini, UNC-Chapel Hill
Alternative Overviews of Result Sets
Alternative Previews for a Specific Video Segment

Gary Marchionini, UNC-Chapel Hill
Some Interaction Principles and Caveats in These Examples

• Principles
  – Look ahead without penalty
  – Minimize scrolling and clicking
  – Alternative ways to slice and dice
  – Closely couple search, browse, and examine
  – Continuous engagement—useful attractors
  – Treasures to surface

• Caveats
  – Scalability (getting metadata to client side)
  – Metadata crucial
    • We are working on automatically creating partitions
  – Increasing expectations about useful results (answers!)
Long Term Paradigm: Information Interaction as Core Life Process

Examples represent early ways to get the information seeker more involved in the information seeking process—there is plenty more to do. Like eating we have varying expectations, invest different levels of effort, and use diverse and ubiquitous infrastructures. Key challenge is to span boundaries between cyberinfrastructure and the ‘real’ world.
Coda

• Our hopes that we can create systems (solutions) that ‘do’ IR for us are unreasonable.

• Our expectations that people can find and understand information without thinking and investing effort are unreasonable.

• We aim to develop ‘systems’ that involve people and machines continuously learning and changing together. Google would not work as well next month if there were not a large group of employees tuning the system, adding new spam filters, and crawlers checking out pages and links continuously.
Thank You!

Questions and Discussion
march@ils.unc.edu

www.ils.unc.edu/govstat
www.open-video.org
NSF Grants EIA 0131824 and IIS 0099638