Human-Computer Information Retrieval

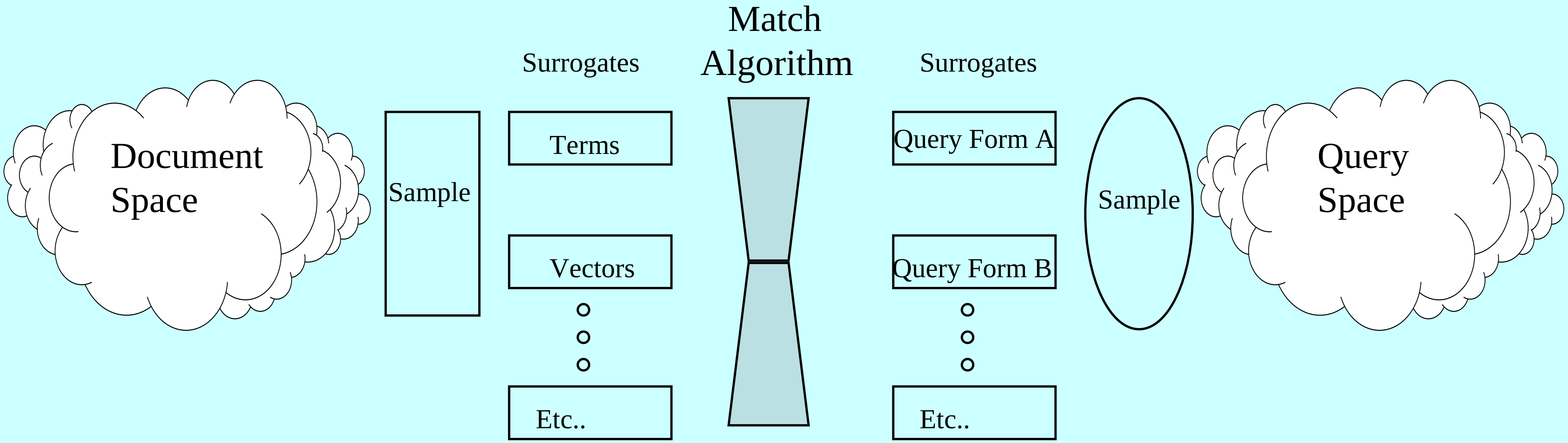
Gary Marchionini
University of North Carolina at Chapel Hill
march@ils.unc.edu

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: syminforosis—people continuously engaged with meaningful information

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).

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

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

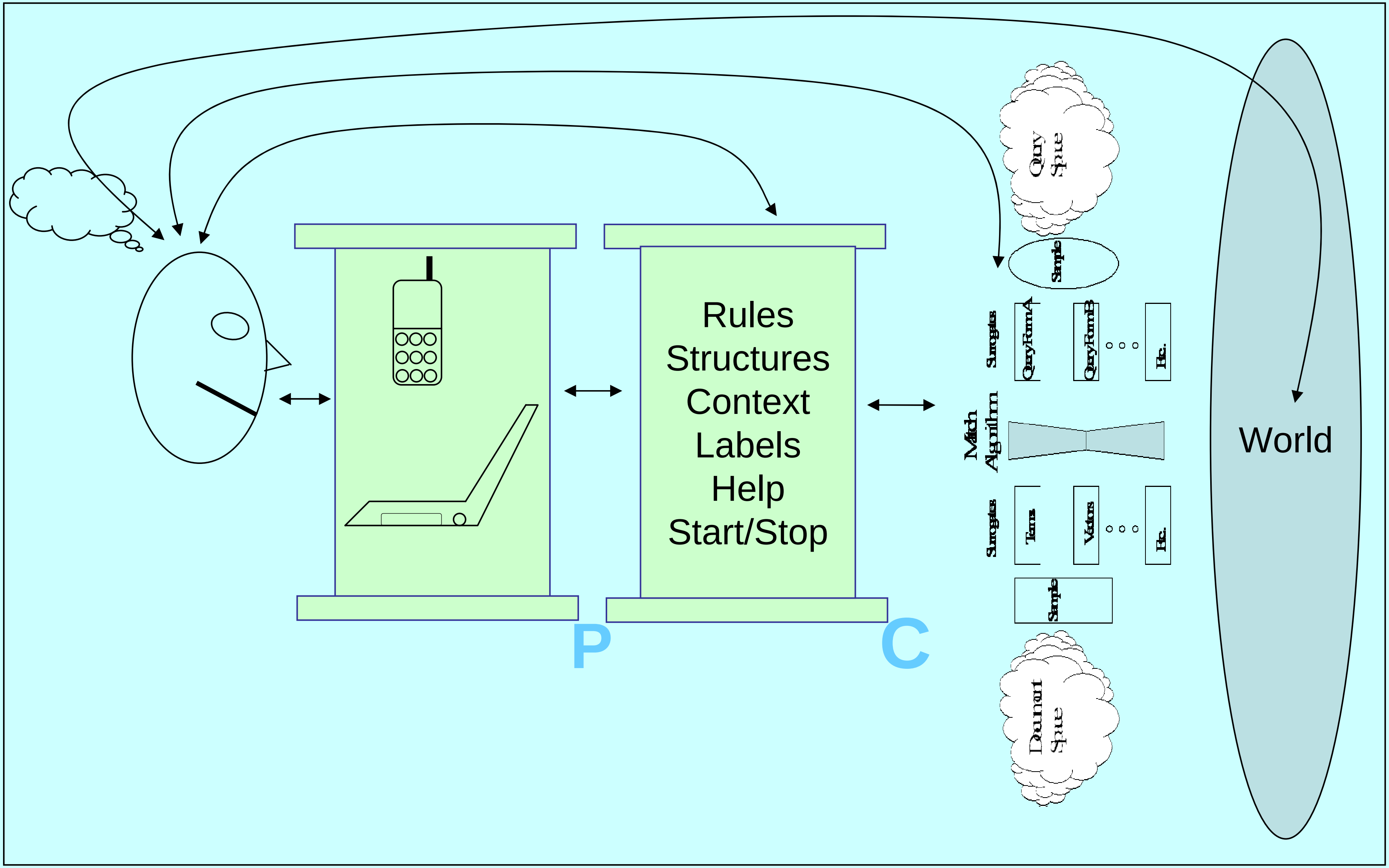
An Expanded Model:

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

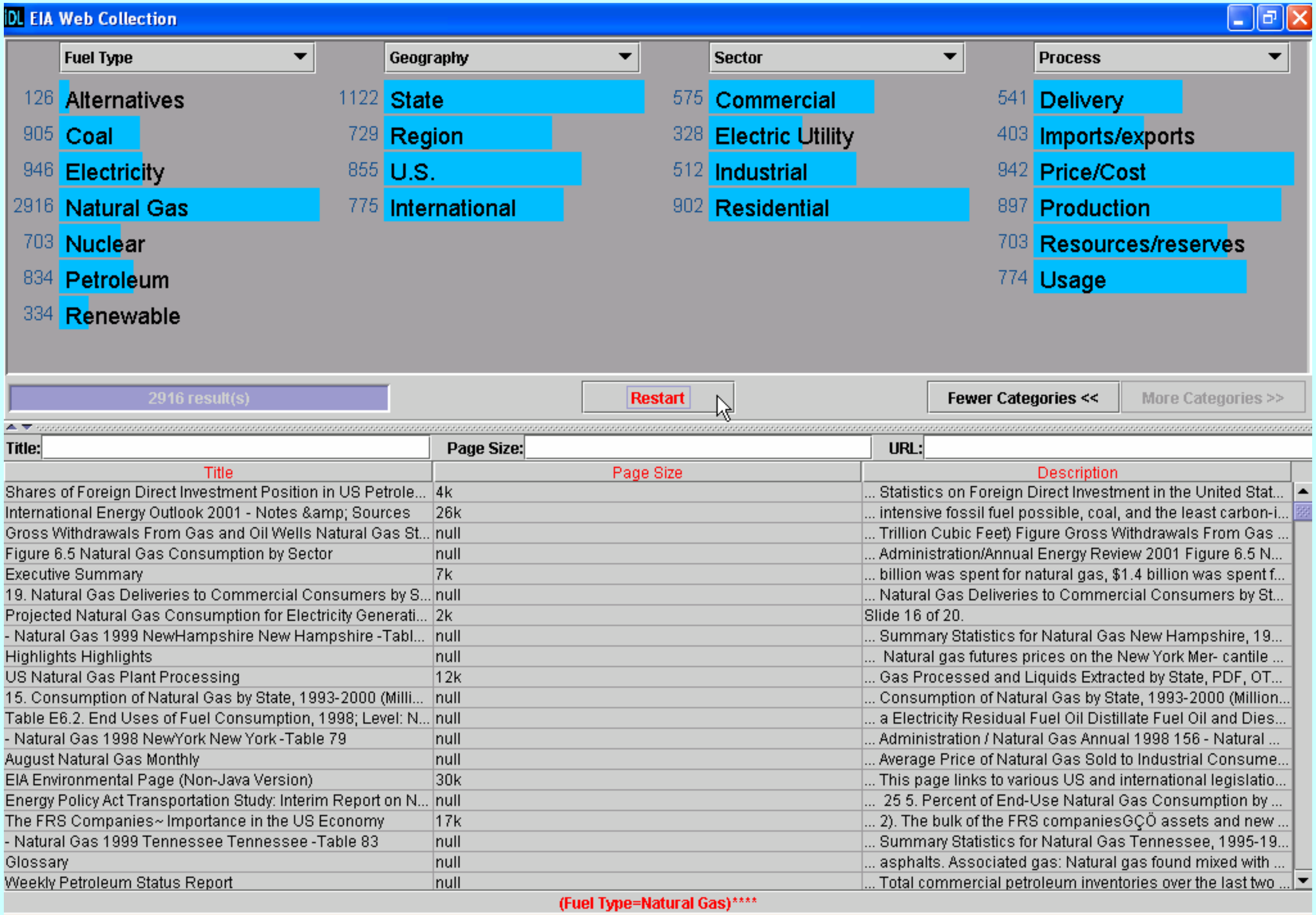


Rules
Structures
Context
Labels
Help
Start/Stop

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



Fuel Type	
126	Alternatives
905	Coal
946	Electricity
2916	Natural Gas
703	Nuclear
834	Petroleum
334	Renewable

Geography	
1122	State
729	Region
855	U.S.
775	International

Sector	
575	Commercial
328	Electric Utility
512	Industrial
902	Residential

Process	
541	Delivery
403	Imports/exports
942	Price/Cost
897	Production
703	Resources/reserves
774	Usage

Title:

Page Size:

URL:

Title	Page Size	Description
Shares of Foreign Direct Investment Position in US Petrole...	4k	... Statistics on Foreign Direct Investment in the United Stat...
International Energy Outlook 2001 - Notes &amp; Sources	26k	... intensive fossil fuel possible, coal, and the least carbon-i...
Gross Withdrawals From Gas and Oil Wells Natural Gas St...	null	... Trillion Cubic Feet) Figure Gross Withdrawals From Gas ...
Figure 6.5 Natural Gas Consumption by Sector	null	... Administration/Annual Energy Review 2001 Figure 6.5 N...
Executive Summary	7k	... billion was spent for natural gas, $1.4 billion was spent f...
19. Natural Gas Deliveries to Commercial Consumers by S...	null	... Natural Gas Deliveries to Commercial Consumers by St...
Projected Natural Gas Consumption for Electricity Generati...	2k	Slide 16 of 20.
- Natural Gas 1999 NewHampshire New Hampshire -Tabl...	null	... Summary Statistics for Natural Gas New Hampshire, 19...
Highlights Highlights	null	... Natural gas futures prices on the New York Mer- cantile ...
US Natural Gas Plant Processing	12k	... Gas Processed and Liquids Extracted by State, PDF, OT...
15. Consumption of Natural Gas by State, 1993-2000 (Milli...	null	... Consumption of Natural Gas by State, 1993-2000 (Million...
Table E6.2. End Uses of Fuel Consumption, 1998; Level: N...	null	... a Electricity Residual Fuel Oil Distillate Fuel Oil and Dies...
- Natural Gas 1998 NewYork New York -Table 79	null	... Administration / Natural Gas Annual 1998 156 - Natural ...
August Natural Gas Monthly	null	... Average Price of Natural Gas Sold to Industrial Consume...
EIA Environmental Page (Non-Java Version)	30k	... This page links to various US and international legislatio...
Energy Policy Act Transportation Study: Interim Report on N...	null	... 25 5. Percent of End-Use Natural Gas Consumption by ...
The FRS Companies~ Importance in the US Economy	17k	... 2). The bulk of the FRS companiesGÇÖ assets and new ...
- Natural Gas 1999 Tennessee Tennessee -Table 83	null	... Summary Statistics for Natural Gas Tennessee, 1995-19...
Glossary	null	... asphalts. Associated gas: Natural gas found mixed with ...
Weekly Petroleum Status Report	null	... Total commercial petroleum inventories over the last two ...

(Fuel Type=Natural Gas)****

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

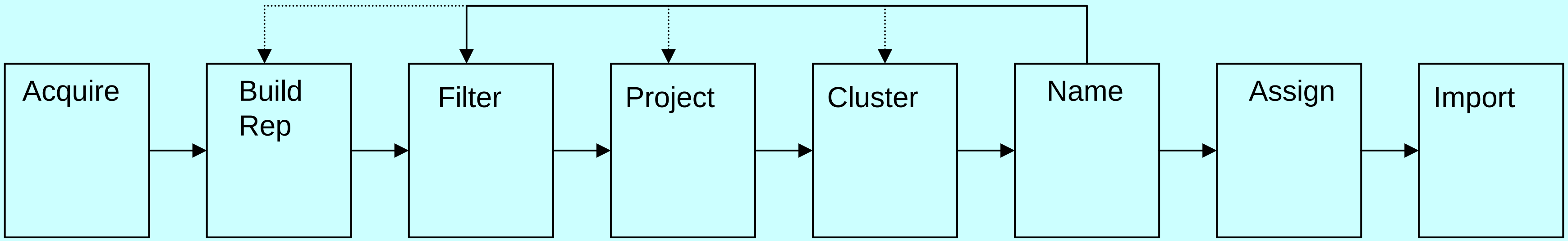
- Architectural Principle: Juxtapose facets
  - Two or more with 5-15 categories per facet
  - Topic is one important facet for most applications
- Interaction Principle: Dynamic exploration of relationships between facets and categories
- Database driven to promote flexible applications (requires systematic 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



Acquire → Build Rep → Filter → Project → Cluster → Name → Assign → Import

Crawl

mirror

[HTML]

Term/Doc matrix

Titles, anchor text, metadata tags

Stop words

infrequents

Reduce dimensionality to 50-100 dim

PCA

LSA

ICA

K-means

EM

Yields prob model

Human effort

Frequencies

Log-odds

Cataloging (binning) based on model

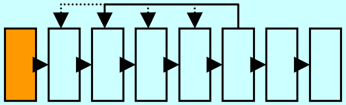
Pipe to RB

Add other facets

A Metadata Mining Toolkit is Available

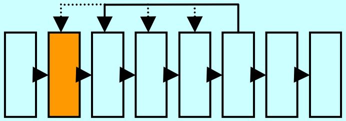
www.ils.unc.edu/govstat/demos.html

Acquire Data



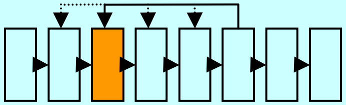
- Crawl Site (sites)
- Currently HTML only
- Mirror locally
  - – E.g., BLS yields 23,530 pages
- Clean data
  - – Remove non-alphas
  - – Lower case all
  - – WordNet validate words
  - – Stem or not stem

Build Representation



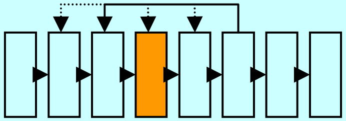
- Select data to include
  - Pages to include/exclude (e.g. BLS ED, 1279 pages)
  - ASCII text from
    - Titles
    - Link anchors
    - Metadata tags
- Build raw term-document matrix
  - Pages as rows (observations)
  - Terms as columns (variables) (e.g., BLS 26,772 terms)
  - Frequencies or TF-IDF weights in cells

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)

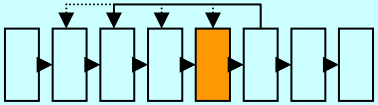
Project data onto Lower Dimensional Space(s)



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

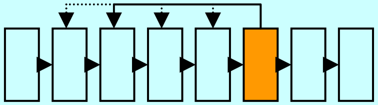
- Reduces to ‘narrower’ term-doc matrix
  - Note: we are experimenting with this at this time

Cluster Documents (pages)



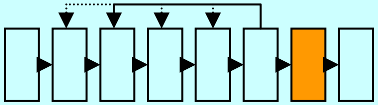
- K-means, e.g., with k<<100
- EM yields a probability distribution for each document over the clusters (so a document has some probability of belonging to each cluster)

Evaluate Clusters and Name Topics



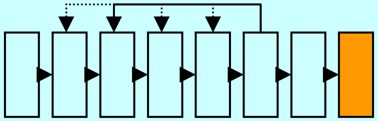
- Create usable output
  - A web page with the clusters and number of documents in each
  - For each cluster, a list of the top 10 most frequently occurring terms; a list of the top 10 log-odds ratio terms; and links to all the pages in that cluster
  - Eyeball the terms, pick a cluster (topic) name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

Assign Pages to Topics



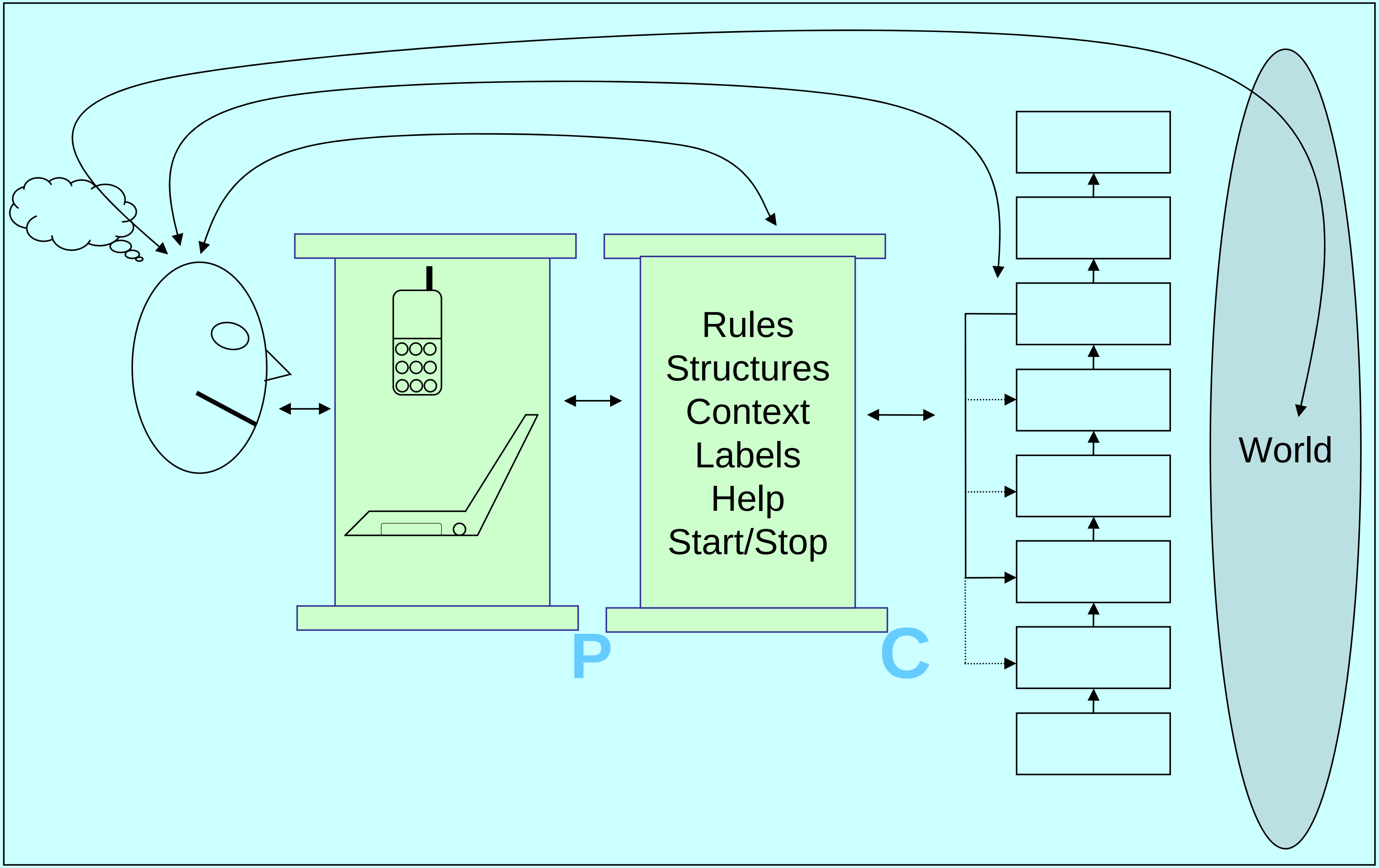
- For every page, compute the probability distribution (using EM model) over each cluster/topic
- Select a threshold for placing pages into topics (most easily go into only one topic)

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

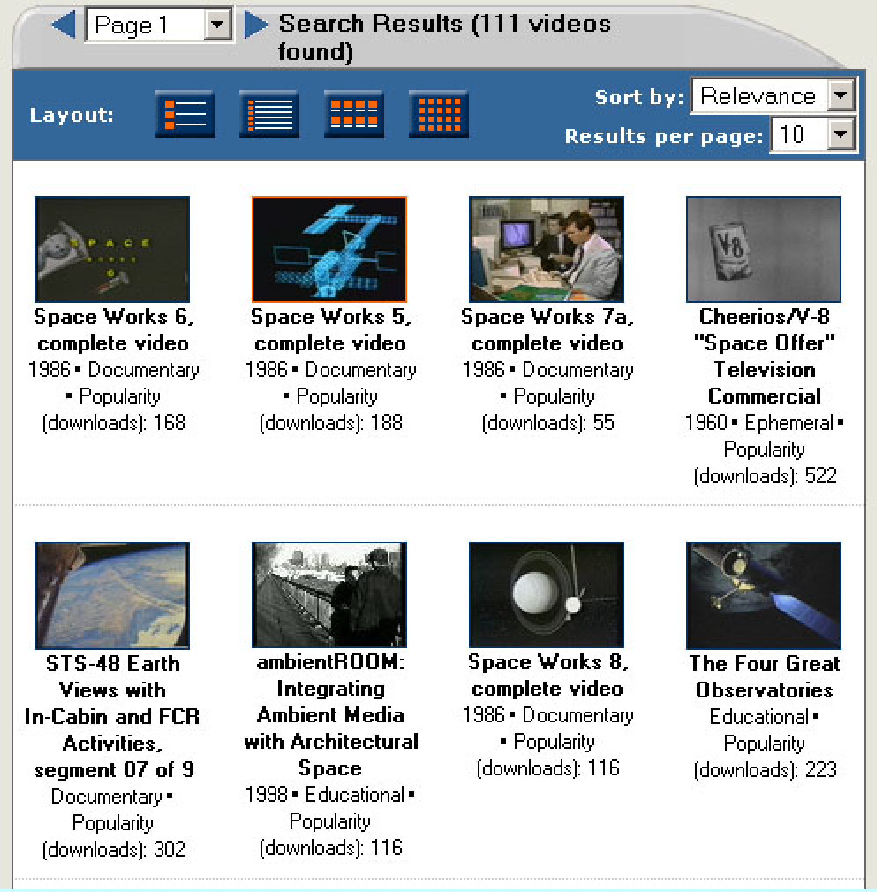


Rules
Structures
Context
Labels
Help
Start/Stop

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

Alternative Overviews of Result Sets



Search Results (111 videos found)

Layout:

Sort by: Relevance

Results per page: 10

Space Works 6, complete video
1986 • Documentary • Popularity (downloads): 168

Space Works 5, complete video
1986 • Documentary • Popularity (downloads): 188

Space Works 7a, complete video
1986 • Documentary • Popularity (downloads): 55

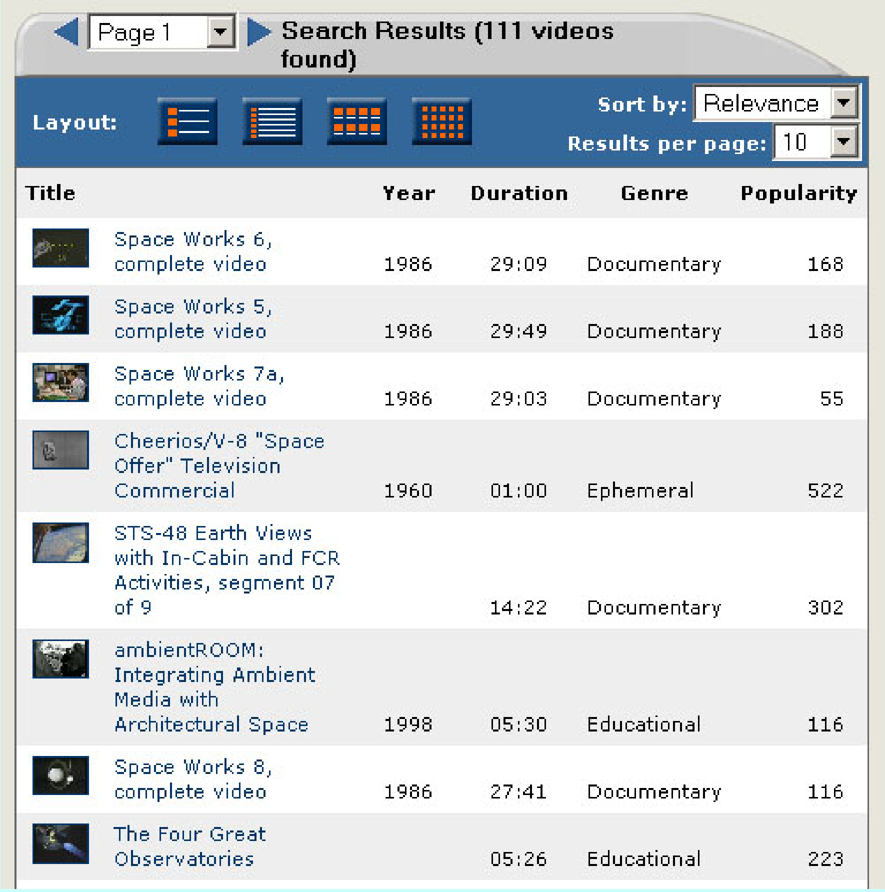
Cheerios/V-8 "Space Offer" Television Commercial
1960 • Ephemeral • Popularity (downloads): 522

STS-48 Earth Views with In-Cabin and FCR Activities, segment 07 of 9
Documentary • Popularity (downloads): 302

ambientROOM: Integrating Ambient Media with Architectural Space
1998 • Educational • Popularity (downloads): 116

Space Works 8, complete video
1986 • Documentary • Popularity (downloads): 116

The Four Great Observatories
Educational • Popularity (downloads): 223



Search Results (111 videos found)

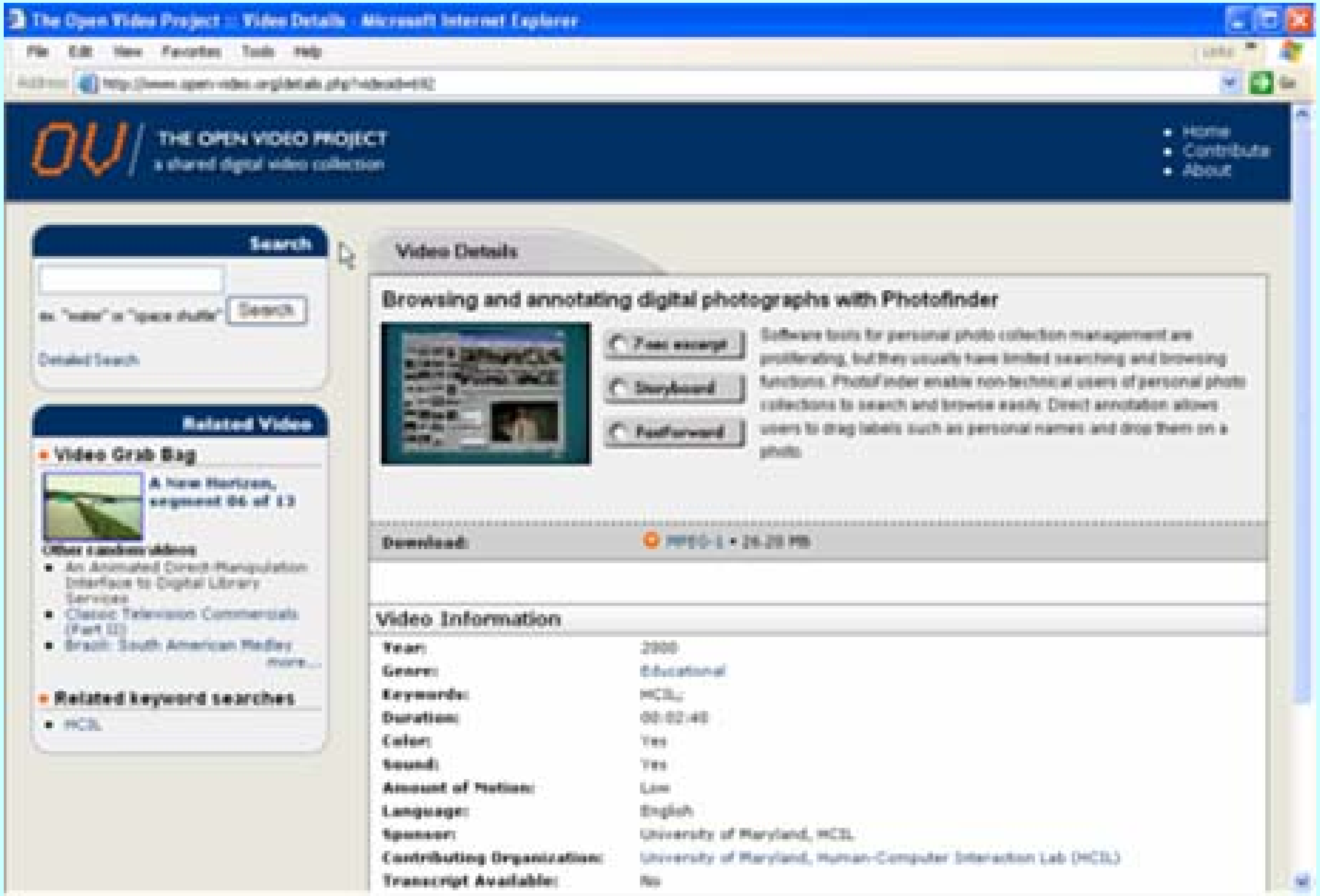
Layout:

Sort by: Relevance

Results per page: 10

Title	Year	Duration	Genre	Popularity
Space Works 6, complete video	1986	29:09	Documentary	168
Space Works 5, complete video	1986	29:49	Documentary	188
Space Works 7a, complete video	1986	29:03	Documentary	55
Cheerios/V-8 "Space Offer" Television Commercial	1960	01:00	Ephemeral	522
STS-48 Earth Views with In-Cabin and FCR Activities, segment 07 of 9		14:22	Documentary	302
ambientROOM: Integrating Ambient Media with Architectural Space	1998	05:30	Educational	116
Space Works 8, complete video	1986	27:41	Documentary	116
The Four Great Observatories		05:26	Educational	223

Alternative Previews for a Specific Video Segment


The Open Video Project :: Video Details - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links
Address http://www.open-video.org/details.php?videoid=692
Go
THE OPEN VIDEO PROJECT
a shared digital video collection
Home
Contribute
About
Search
Search
Detailed Search
Related Video
Video Details
○ 7 sec excerpt
○ Storyboard
○ Fastforward

ex. "water" or "space shuttle"

Video Grab Bag

A New Horizon, segment 06 of 13

Other random videos

- An Animated Direct-Manipulation Interface to Digital Library Services
- Classic Television Commercials (Part II)
- Brazil: South American Medley

more...

Related keyword searches

- HCIL

Browsing and annotating digital photographs with Photofinder

Software tools for personal photo collection management are proliferating, but they usually have limited searching and browsing functions. PhotoFinder enable non-technical users of personal photo collections to search and browse easily. Direct annotation allows users to drag labels such as personal names and drop them on a photo.

Download: MPEG-1 • 26.20 MB

Video Information

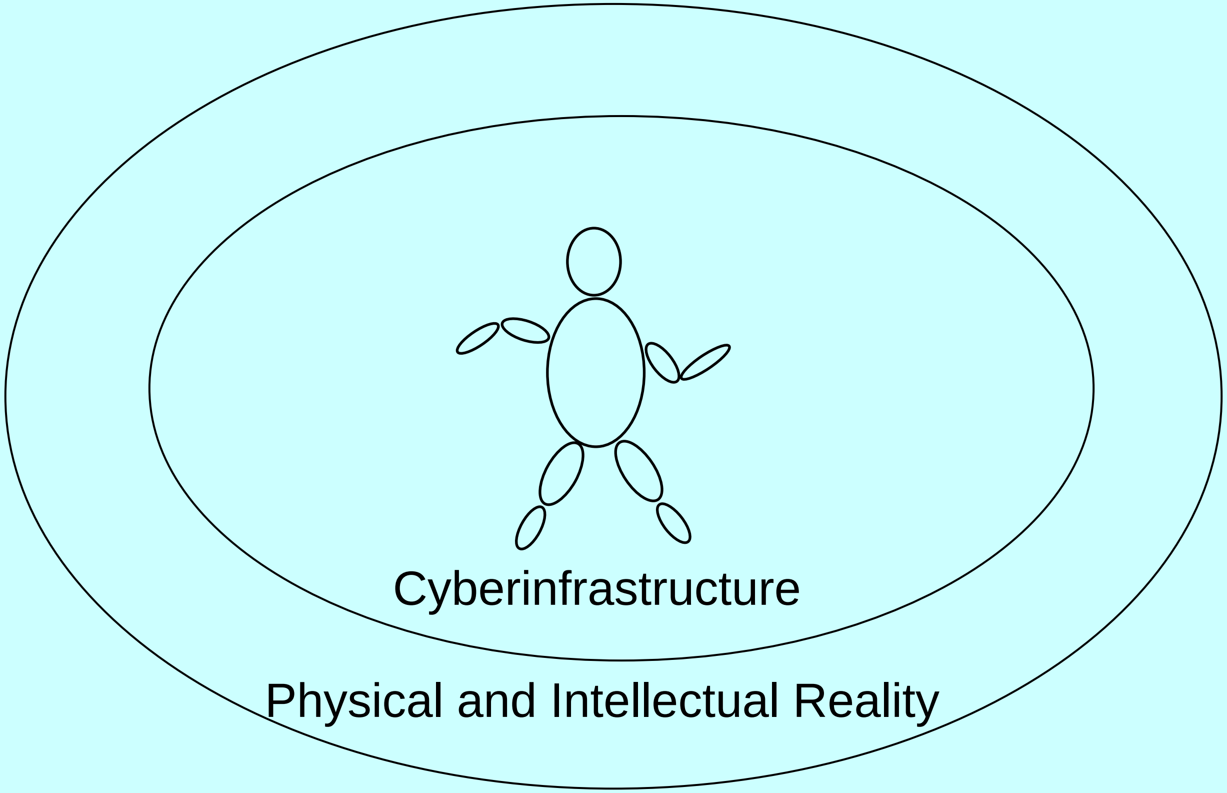
Year:	2000
Genre:	Educational
Keywords:	HCIL;
Duration:	00:02:48
Color:	Yes
Sound:	Yes
Amount of Motion:	Low
Language:	English
Sponsor:	University of Maryland, HCIL
Contributing Organization:	University of Maryland, Human-Computer Interaction Lab (HCIL)
Transcript Available:	No

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.



Gary Marchionini, UNC-Chapel Hill

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