How Topic Modeling is Useful in Digital Libraries

October 7, 2010

Kat Hagedorn (U Michigan)
David Newman (UC Irvine)
Youn Noh (Yale)

with considerable assistance from Michael Kargela (U Michigan)

Challenges in Making Digital Libraries Useful (1)

- Minimal metadata

Image ID Number	lwlpr15068
Call Number	49 3678
Creator	[Sandby?]
Title	[Strawberry Hill east front]
Physical Description	watercolor : 28 x 43.8
Cite as:	The Lewis Walpole Library, Yale University

Challenges in Making Digital Libraries Useful (2)

- Too much data

"In all the political caricatures by Gillray subsequent to a certain period of the French Revolution, the satirist is merciless in his attacks upon the Man of the People. Even admitting what has been asserted, that the satirist was influenced by the party opposed to this illustrious orator, yet the frantic zeal with which Fox stood forward as the advocate of the direful anarchy of France, not only deprived him of many respectable friends who had hitherto been enrolled with him in the cause of patriotism, but laid him fairly open to much of the graphic libelling which he experienced. The Whig chieftain's pertinacious adherence to his opinions in favour of the inveterate foe, to say the least of it, argued bad taste; for not only the English people, but the whole community of the civilized world, felt alike in their genuine abhorrence of the atrocities perpetrating in France.

"It was this extraordinary conduct on the part of Mr. Fox which seemed to sanction the bitter invectives of the author of the 'Whig Club,' who says, 'That copious stream of words which he pours forth at pleasure, is indeed justly the theme of admiration, but as the viper bears in herself the antidote of her poison, so does his character prevent his abilities from doing all the mischief he otherwise might, by pulling off the mask and showing his plans too soon for their accomplishment.'

"His labours to effect the overthrow of the Ministry and to obtain power have been compared to

Challenges in Making Digital Libraries Useful (3)

- Different users
- Different information needs
- Different search vocabularies

Background

- IMLS R & D Project
  - Improving Search and Discovery Using Topic Modeling
  - Yale (lead), UMich, UC Irvine

- Apply topic modeling to three classes of digital library resources: full-text books, images, and tagged objects
- Build prototypes of user interfaces that make use of topics
- Test the prototypes to assess the value of topic modeling for users

Overview

- At end of 2nd year of 3 year project
- Research subprojects
  - Automated evaluation of topic models
  - Visualization of segmented topic models (ongoing)
  - Regularized topic models (ongoing)
  - Topic modeling image content (ongoing)
  - Usability and user testing studies (ongoing)
- Demonstration subprojects and deliverables
  - Browse application (images)
  - Faceting application (texts)
  - Topic modeling toolkit
  - Usability and user testing reports

Evaluating Topic Models for Digital Libraries

Collections and challenges

- Digitized books
- Images
- Scientific literature
- Web 2.0 content
- … and more

Collections and challenges

- Digitized books
- Images
- Scientific literature
- Web 2.0 content
- … and more

Currently Digitized

- 6,182,629 total volumes
  3,621,100 book titles
  146,505 serial titles
  2,163,920,150 pages
  230 terabytes
  73 miles
  5,023 tons

Collections and challenges

- Digitized books
- Images
- Scientific literature
- Web 2.0 content
- … and more

- Catalog Search
  - Subj: “American Colonial History” 20,000 results
- Full-Text Search
  - “American Colonial History” 1,000,000 results
- Limitation
  - Users don’t have mental model
  - Users don’t trust metadata

Collections and challenges

- Digitized books
- Images
- Scientific literature
- Web 2.0 content
- … and more

YALE UNIVERSITY ART GALLERY

q = “madonna and child”

Madonna and Child, based on Barocci's etching Madonna and Child in the Clouds

Artist/Maker:	Federico Barocci	*Culture:*	Italian
Date:		*Period:*	16th century
Accession #:	1978.105	*Department:*	Prints, Drawings, and Photographs
Location:	Viewable by appointment	*Classification:*	Works on Paper - Drawings and watercolors

Madonna and child

Artist/Maker:	Unknown	*Culture:*	
Date:	20th century	*Period:*	Modern
Accession #:	2003.56.13	*Department:*	Ancient Art
Location:	Not on view	*Classification:*	Sculpture

Holy Family (Madonna and Child with Joseph, John the Baptist, Elizabeth, and Zacharias)

Artist/Maker:	John Trumbull	*Culture:*	American
Date:	1802-1806	*Period:*	19th century
Accession #:	1832.83	*Department:*	American Paintings and Sculpture
Location:	Not on view	*Classification:*	Paintings

Madonna and Child with St. John the Baptist

Artist/Maker:	John Trumbull	*Culture:*	American
Date:	1801	*Period:*	19th century
Accession #:	1832.98	*Department:*	American Paintings and Sculpture
Location:	Not on view	*Classification:*	Paintings

Holy Family (Madonna and Child with Joseph, Elizabeth, and John the Baptist)

Artist/Maker:	John Trumbull	*Culture:*	American
Date:	1839-1840	*Period:*	19th century
Accession #:	1840.4	*Department:*	American Paintings and Sculpture
Location:	Not on view	*Classification:*	Paintings

Madonna and Child with Saint John

Artist/Maker:	Unknown	*Culture:*	French

Collections and challenges

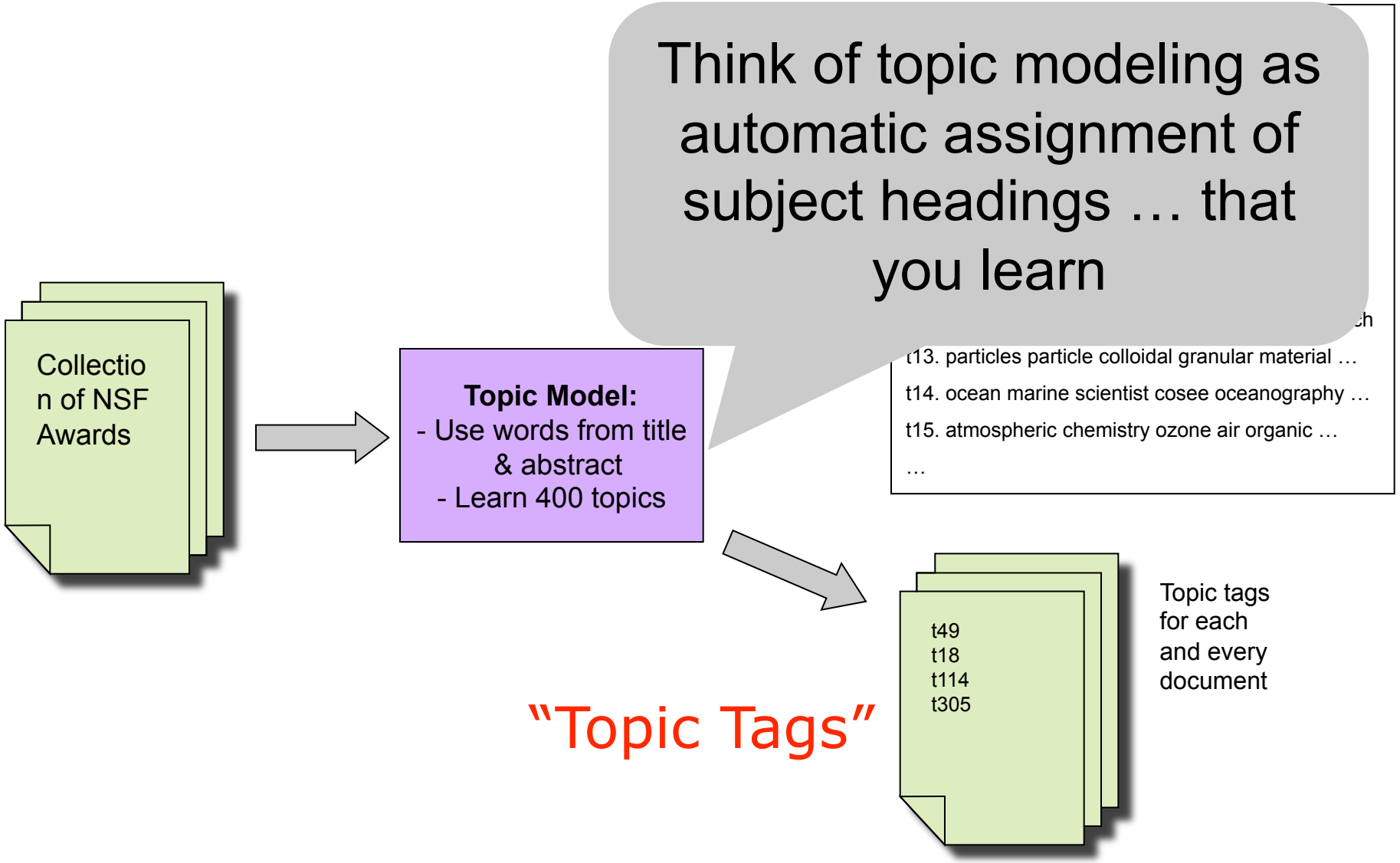
- Digitized books
- Images
- Scientific literature
- Web 2.0 content
- … and more

- 1000 new articles daily
- Indexed using MeSH

What is Topic Modeling?

- Topic Modeling (like text clustering, but better)
- Updated version of Latent Semantic Analysis
- State-of-the-art model for collections of text documents
- Works great on large collections of well written content

Topic model learns topics from co-occurring words



A closer look at one automatically learned topic

topic-6: conflict violence war international military domestic political government terrorism national security civil …

- What is this topic about? Is it a meaningful topic?
- [How] Do we present this to users? … What is a good label for this topic?

Overarching Questions

Q1: Are individual topics meaningful and usable?

Q2: Are assignments of topics to documents meaningful and usable?

Q3: Do topics facilitate better or more efficient document search, navigation, browsing?

Experimental Setup

Collection	Sources	Volume
Books	Internet Archive Hathi Trust	12,000 books 28,000 books
News Articles	LDC Gigaword (NY Times articles)	55,000 articles
Grant Abstracts	National Institutes of Health	60,000 grants
Image Metadata	Yale Library UMich Library	100,000s 100,000s

Experimental Setup

- Topic modeled each document collection (using different topic resolutions). Selected a total of 600 topics for manual coherence scoring

- Have N = 9-15 annotators score each of the 600 topics on a 3-point scale where 3="useful" (coherent) and 1="useless" (less coherent), based on the top-10 topic words
  - also asked annotators to identify "best" topic word … and
  - suggest a short label

Example Coherent and Incoherent Topics

Coherent
(unanimous score=3)

Books

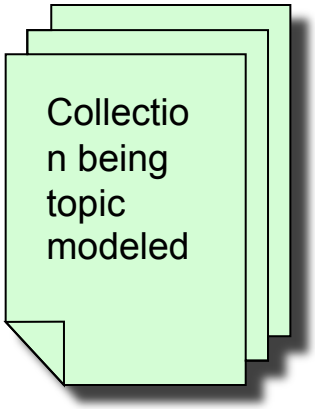
silk lace embroidery tapestry gold embroide
trout fish fly fishing water angler strea

Less coherent
(unanimous score=1)

Incoherent topics are not errors … they are also statistical patterns of word usage seen in the data

Automatic Scoring of Topics?

- Coherence of topic depends on relatedness of all 10-choose-2 pairs of top-10 topic words
- Idea: Use external data to evaluate word pair relatedness (e.g. Wikipedia)



Relatedness of word pairs

Topic: music dance band rock opera …

Pointwise Mutual Information (measure of dependence)

Count co-occurrence in a sliding window

Dance music works often bear the name of the corresponding dance, e.g. waltzes, the tango, the bolero, the can-can, minuets, salsa, various kinds of jigs and the breakdown. Other dance forms include contradance, the merengue (Dominican Republic), and the cha-cha-cha. Often it is difficult to know whether the name of the music came first or the name of the dance.

10-word sliding window

#(dance,music) = 1

Relatedness of word pairs

Topic: music dance band rock opera …

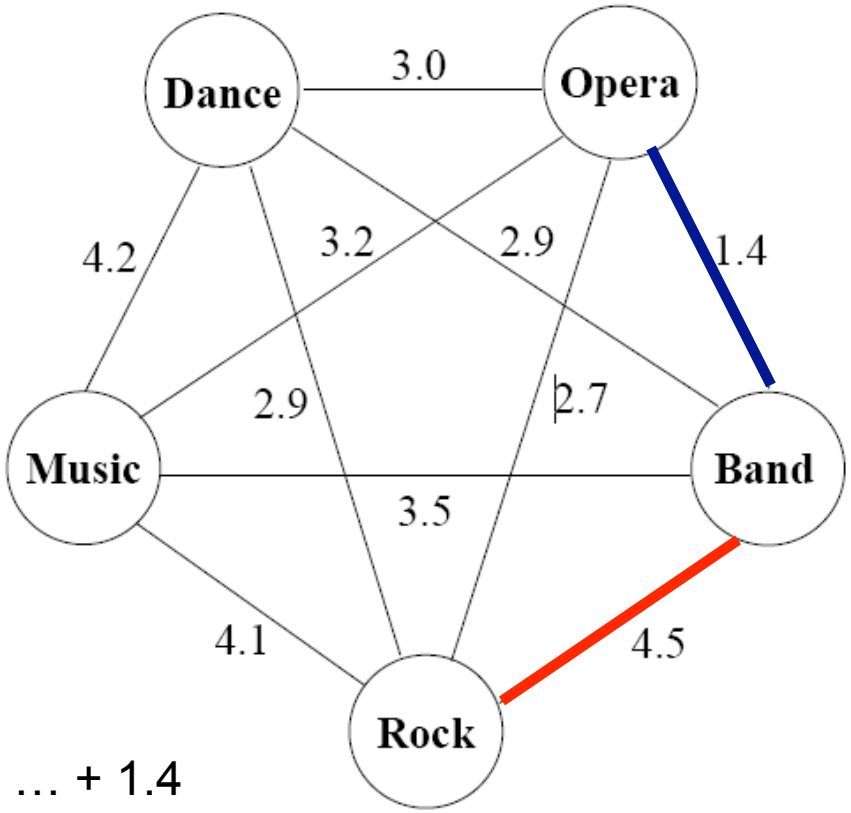
Pointwise Mutual Information (measure of dependence)

$$PMI(w_1, w_2) = \log \frac{\Pr(w_1, w_2)}{\Pr(w_1)\Pr(w_2)}$$

$$PMI - Score(w) = \sum_{ij} PMI(w_i, w_j), ij \in 1 \ldots 10, i < j$$

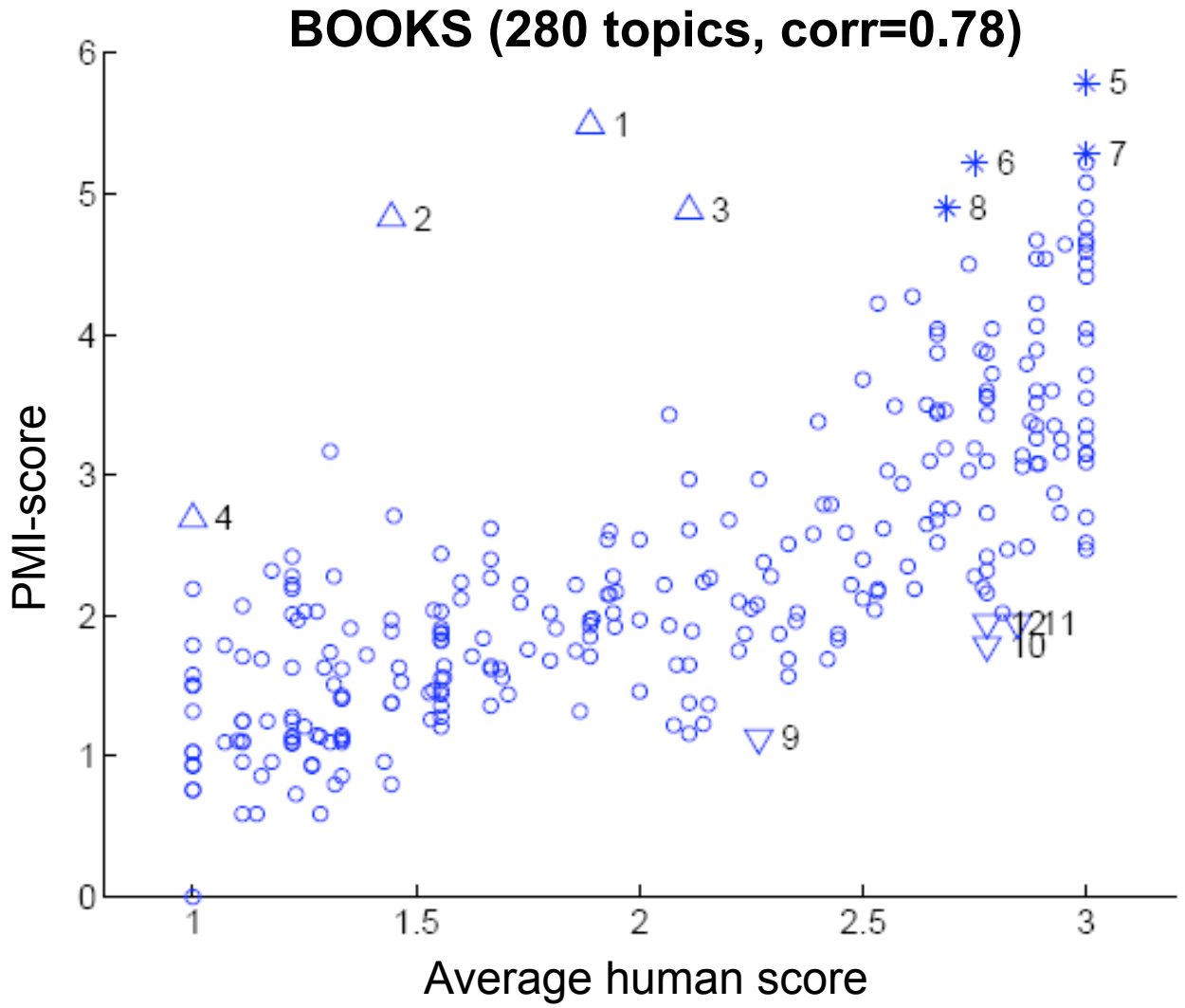
Relatedness of word pairs

Topic: music dance band rock opera …

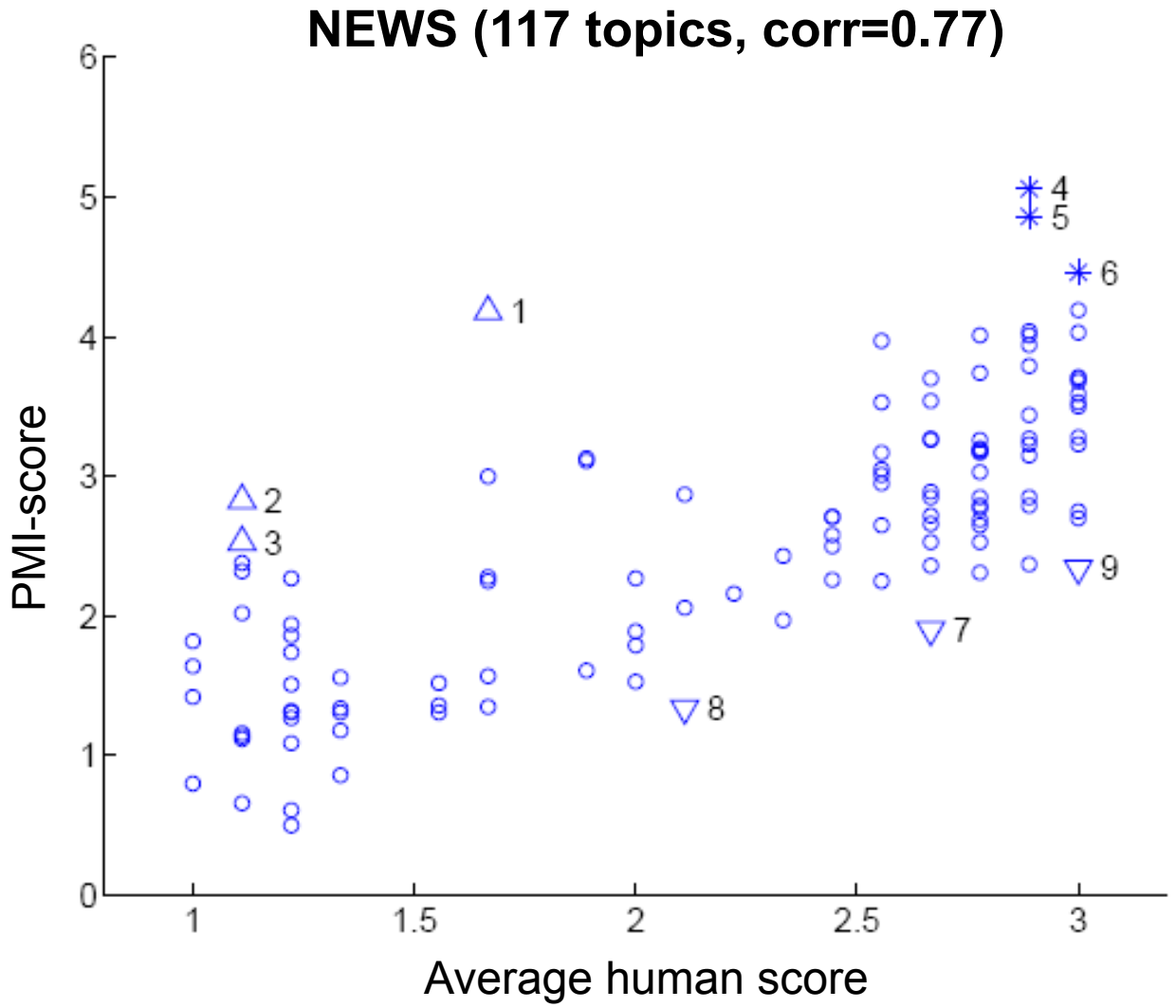


PMI-Score = 4.5 + 4.2 + … + 1.4

PMI-score achieves human-level performance



PMI-score achieves human-level performance



Best topic word and suggested Label

Topic	Suggested Label
trout fish fly *fishing* water angler stream rod flies ...	fly fishing
space earth moon science scientist light nasa ...	space exploration
race car *nascar* driver racing …	nascar racing

Suggested Label … from Wikipedia (work in progress)

Topic	Wiki Article Titles
trout fish fly fishing water angler stream rod flies ...	fly fishing fishing angling trout …
space earth moon science scientist light nasa ...	space exploration space space science space colonization nasa missions …

Large-Scale User Studies

- Developed prototype user interfaces for Image Collections and Book Collections that use topics
- Large scale user studies at Yale and UMich underway
- Qualitative and quantitative assessment of the value of topics

What else we can do with topics?

http://datalab-3.ics.uci.edu/elsevier/climate/

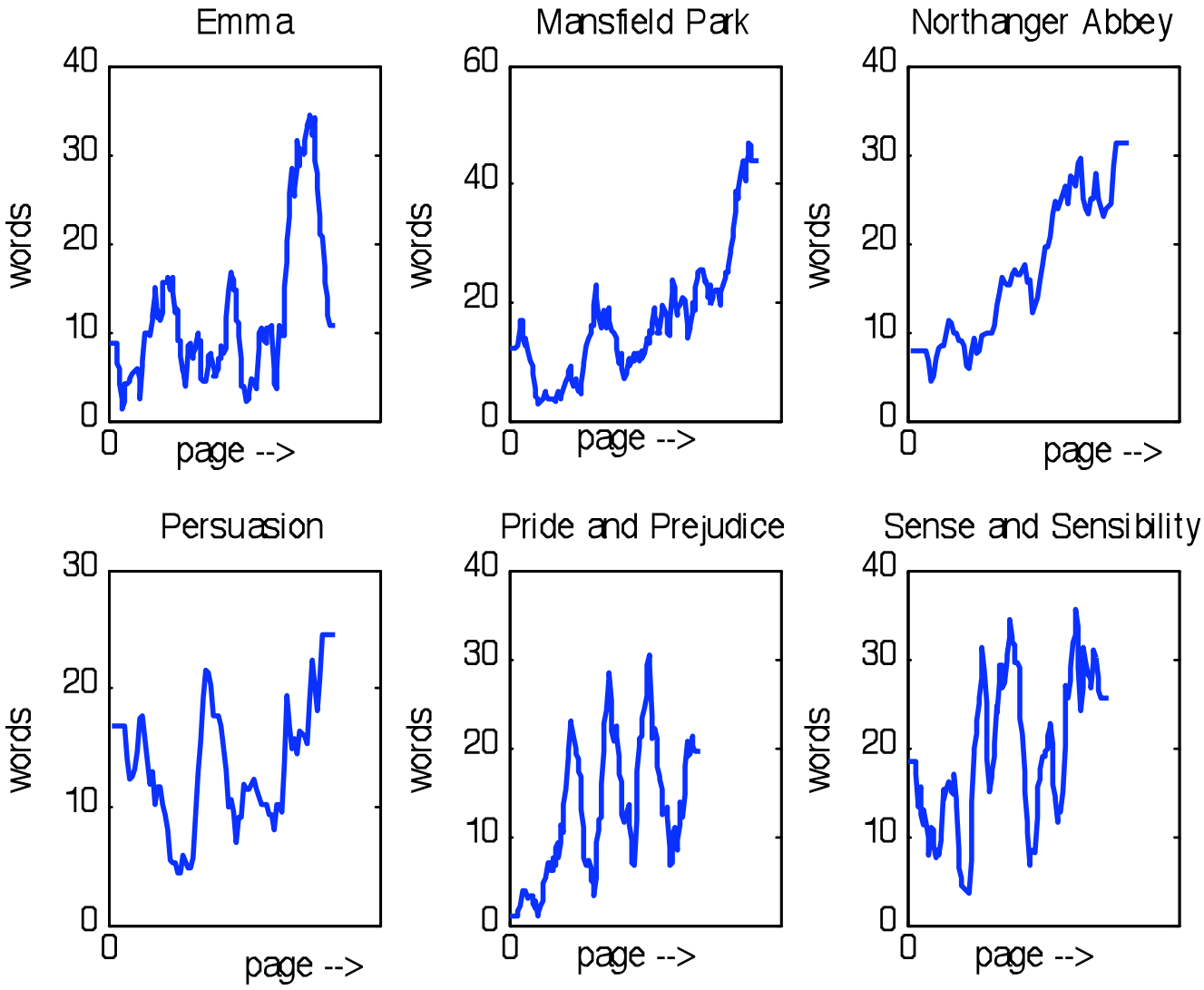
climate (5572 articles)

- 20% [t3] energy cost emission electricity fuel power market production ...
- 15% [t1] sediment sea ice level water record basin event ...
- 12% [t6] model water soil temperature data surface parameter climate ...
- 12% [t5] species model population data distribution effect mean forest ...
- 11% [t7] group animal treatment infection patient health human control ...
- 11% [t8] data system space information research change process set ...
- 11% [t2] plant effect cell treatment activity concentration experiment level ...
- 8% [t4] samples water content concentration organic material sample formation ...

Visualization of search results (each dot is a search result)

Topic trends throughout books

[SENTIMENT] felt comfort feeling feel spirit mind heart point moment ill letter beyond mother state never event evil fear impossible hope time idea left situation poor distress possible hour end loss relief dearest suffering concern dreadful misery unhappy emotion …



Conclusion

- Topic models seem to be useful in digital libraries for creating additional metadata …
- … but learned topics can vary in usefulness and coherence
- We developed model to automatically evaluate topic coherence that matches human judgments
- This is a step in integrating topic modeling into digital libraries

Ideas for testing our models

How were we going to get a significant sample of users at each library? We knew testing 7-8 users wasn't going to be sufficient for this project.

Yale chose TechSmith's Morae software, which was untested at either library.

- ...unmoderated, meaning there isn't anyone giving the test
- ...in part automated, meaning the screen is captured and stored on a web server for later access
- ...in part not automated, meaning incentives and analysis are still manually processed

Interfaces tested: HathiTrust & UM History of Art

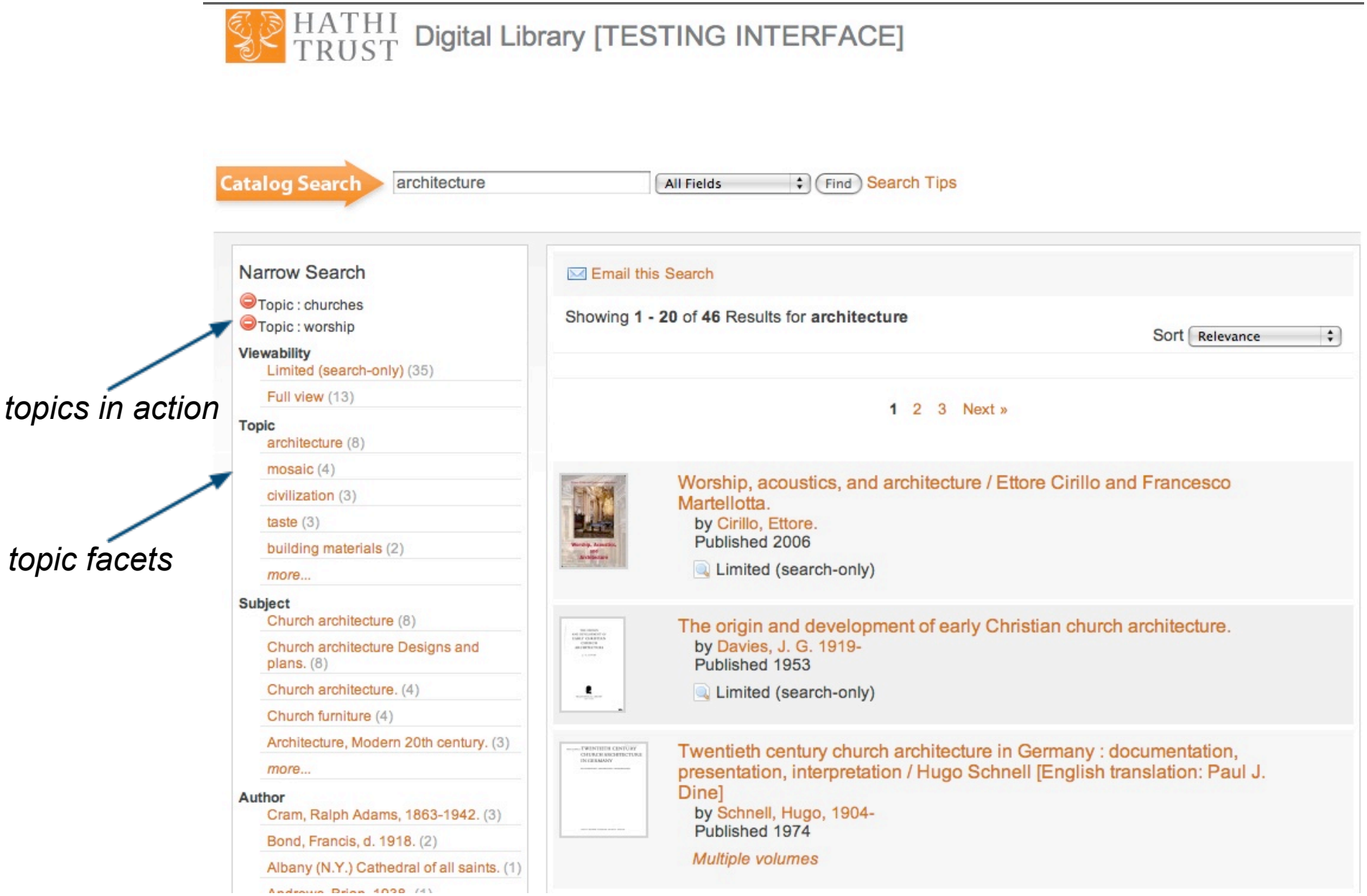
HathiTrust:

- Subset of HathiTrust taken in Fall 2009: only art, architecture and art history texts, totaling 27,688 volumes.
- Built a specific instance of the HathiTrust interface that we could add the topic clusters to. Included both search and browse.
- http://topics.catalog.hathitrust.org/

UM History of Art (HART):

- Snapshot of the HART collection (same time, same subjects), totaling 66,924 images.
- Search was disabled so we could test only browse.
- http://quod.lib.umich.edu/cgi/i/image/image-idx?c=hart4topics

HathiTrust test interface

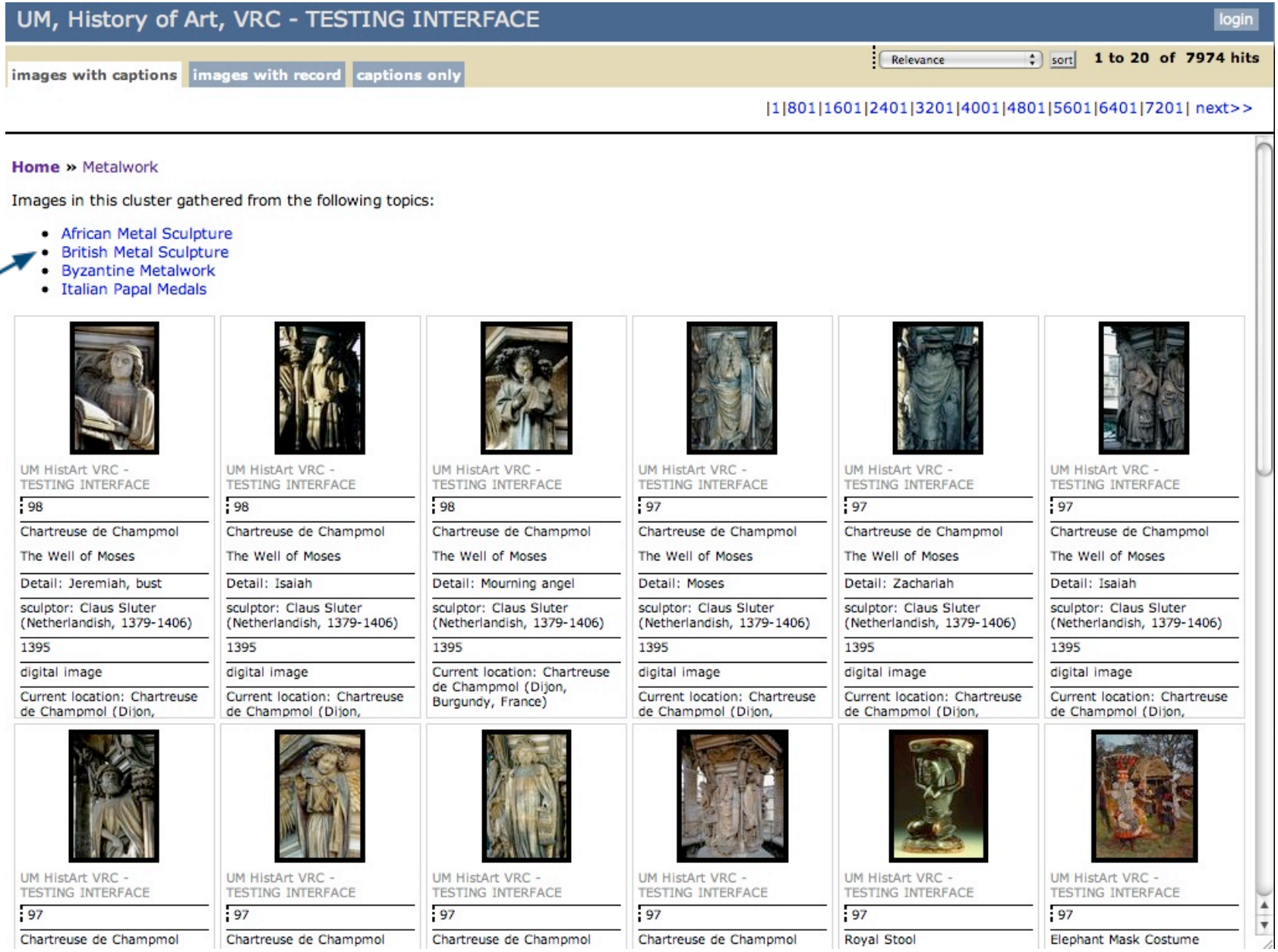


topics in action

topic facets

HART test interface

two levels of topics



Morae testing

For unmoderated testing, needed...

- Considerable assistance from Library IT to install Morae, build web server, and troubleshoot problems
- Daily assistance during the test from Library Finance to purchase gift certificates
- Special assistance from Library Media Relations to create alluring posters to draw testers to the machines
- Plus a lot of help from everyone in Library Information Technology

Special note: $15 gift certificates really, really work.

Thanks Panagis, DJ, Gary, Liene, Hongyun, Mary & Karen!

Morae testing

text located next to the testing station to provide additional information

PLEASE LOGON TO START THE TEST.

In the following evaluation, you will be asked to perform 3 tasks using a web site. We will log and record the screen during the session. The total evaluation should take no longer than 15 minutes. **You will receive your $15 Amazon.com gift certificate via email to your UM (uniqname) account within 10 days**. If you do not receive it in that time, please email topicmodel@umich.edu.

Participation is voluntary. You may stop the test at any time by clicking "Exit Session."

THANK YOU for your participation!

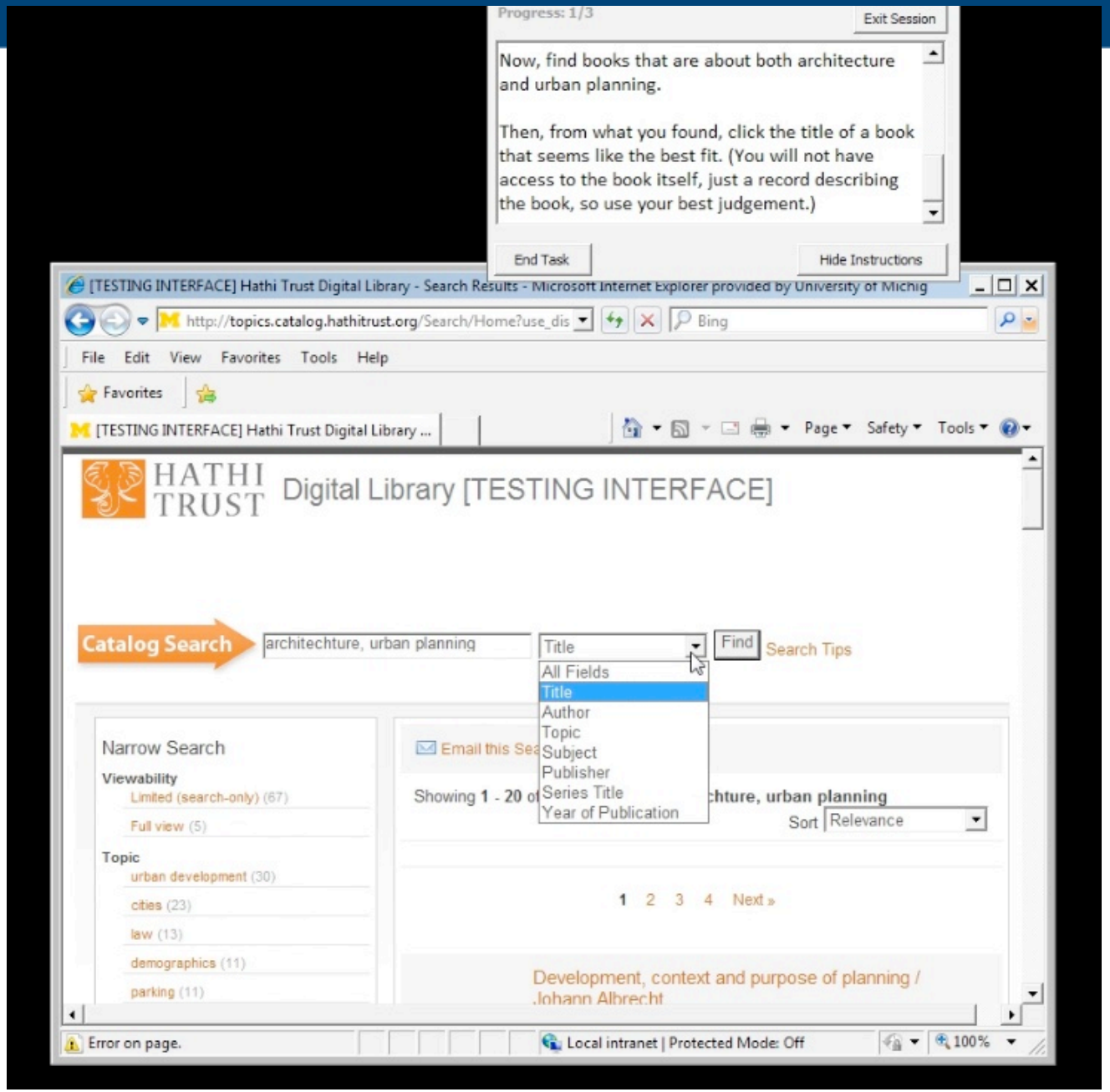
To start the test, **please click "Begin Study" below, and then the large RED button on the following screen**. (This may take a few minutes to load.) You may see a "Tips" pop-up, which you will need to close to access the RED button.

BE SURE TO LOGOFF when you've finished the study.

The University of Michigan - in partnership with Yale University and the University of California, Irvine - is conducting this study funded by the Institute of Museum and Library Services to evaluate improving the ability to find research materials (IMLS award no.: LG-06-08-0057-08).

Morae testing

snapshot of a tester's screen



The test itself

To date, we have tested and started analyzing the HathiTrust interface at UMichigan.
250 recordings at UMichigan, 150 recordings at Yale.
We've processed around 80 recordings so far...

The test script: we asked users to search in the interface three different ways

- *Task 1*: without any specific instruction
- *Task 2*: asked to use the facets on the left
- *Task 3*: asked to use the topic facet, in particular, on the left

The test itself

Task 1: The professor in your introductory architecture class is asking you to do a general overview of urban planning as it relates to architecture... Find books that are about both architecture and urban planning... Click the title of a book that seems like the best fit.

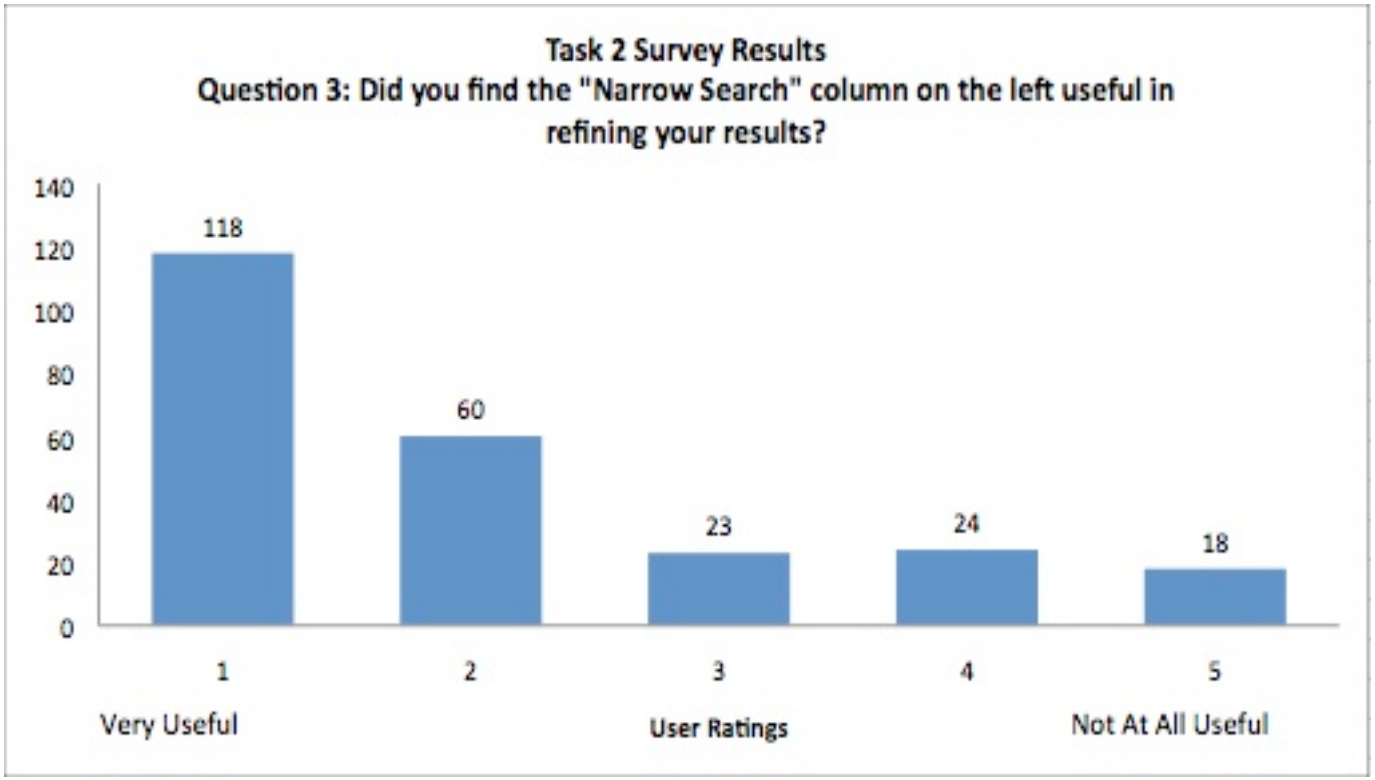
Task 2: A couple weeks later, your professor asks you to get a different overview-- the modernism movement and how it relates to architecture. Do a search for modernism... View a list of books related to architecture. Use the "Narrow Search" column on the left to do this... Click the title for a book that seems like the best fit.

Task 3: At the end of the semester, your professor asks for a final overview-- the architecture of religious buildings. Search for architecture first. You'd like to dive a bit further into architecture to just view a list of items related to religious buildings. Use the "Topic" section of the "Narrow Search" column on the left to do this... Click the title for a book that seems like the best fit. *

* these are abbreviated to fit the slide, and don't include the end-of-task surveys

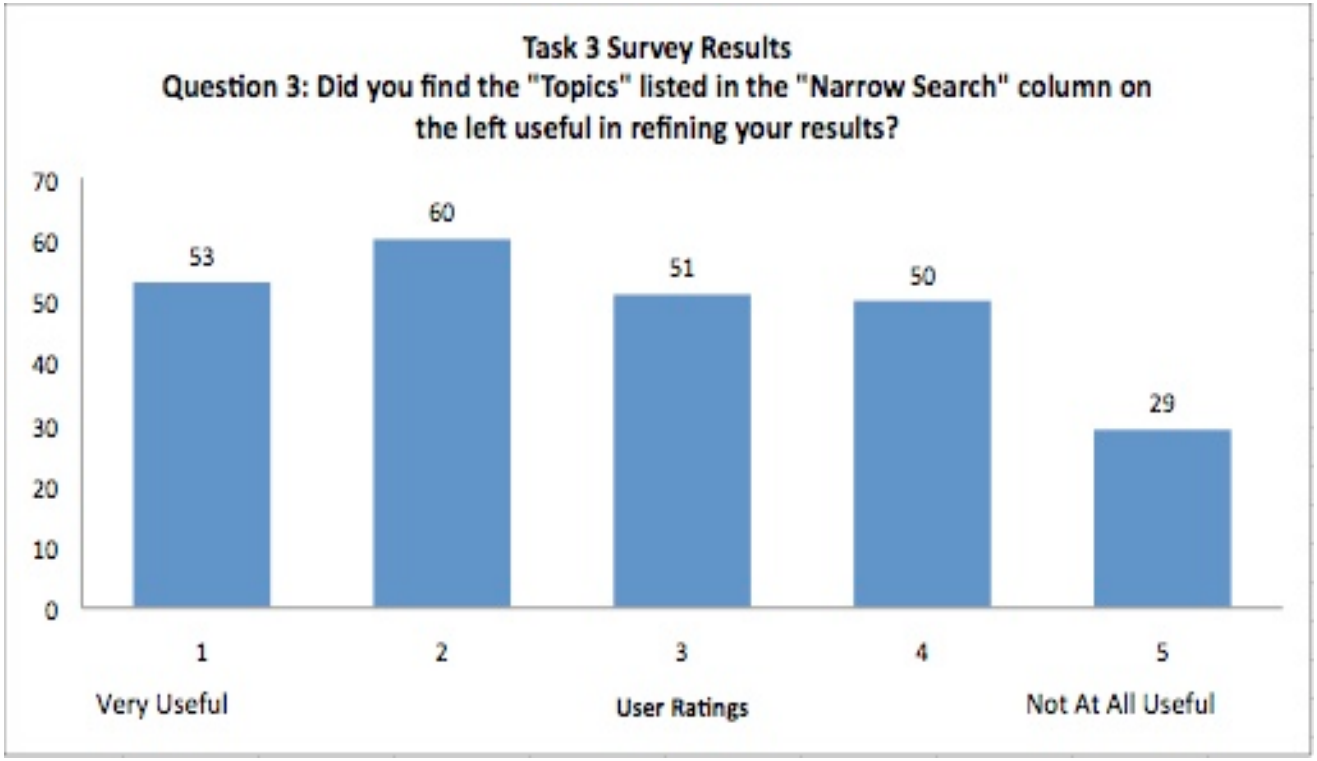
Results: use of facets

Tasks 2 and 3 showed that testers were by and large satisfied with their results, and also found the facet column useful.



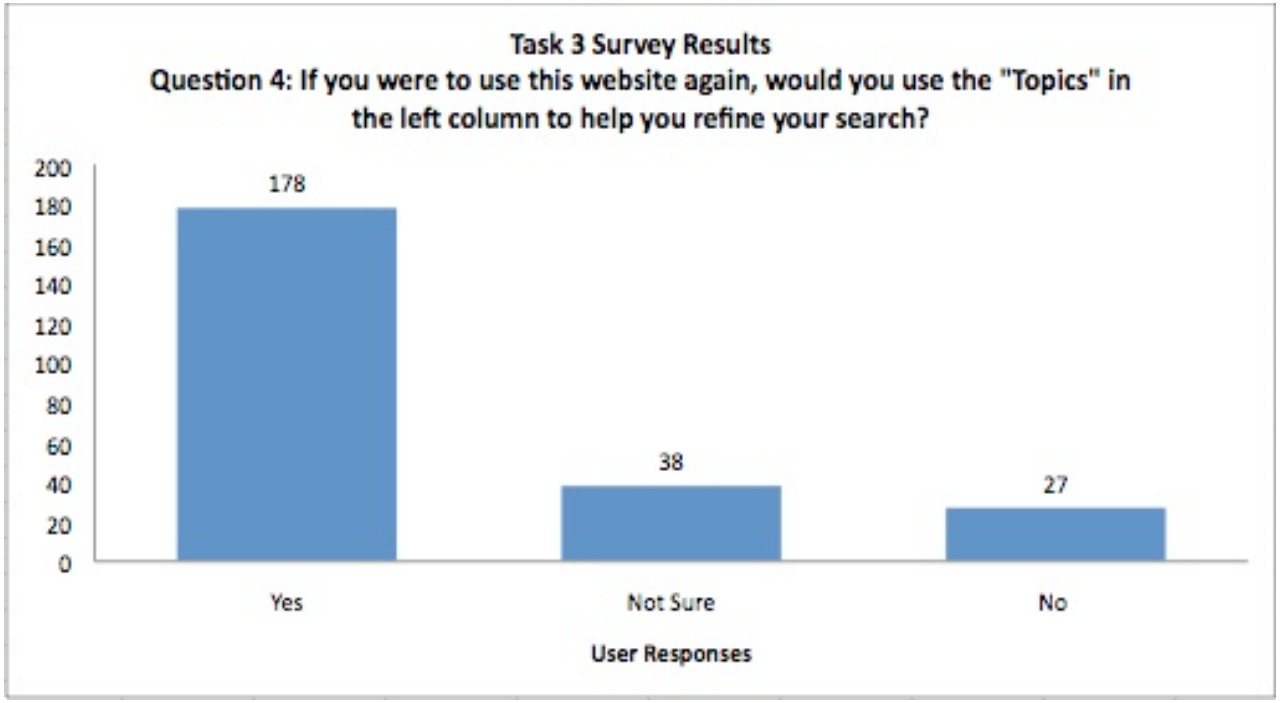
Results: use of facets

But ratings for the topic facet itself showed some ambivalence.



Results: use of facets

Of course, testers often try to please...
Even in an unmoderated test, they were still trying to say the "right" thing! 73% indicated they would use topics to narrow search results.



Results: use of facets

- For both undergraduates and graduate students... the use of facets seems to have hindered them somewhat in choosing the text they wanted from the results list. (Choice was farther down the list.)
- When users were **not** asked to look at topics (i.e., in Task 1), they did so regardless 50% of the time.
- When users were asked to look at the facets, but **not** specifically the topic facets, they chose a topic facet 50% of the time.

Results: searches used

In general, the same search terms were used for each task, consistently across testers. This is not surprising, as our test described each task using specific words.

- Task 1: 64 of 72 (89%) used the term “architecture” in their search
- Task 2: 66 of 71 (92%) used the term “modernism” in their search
- Task 3: 67 of 71 (94%) used the term “architecture” in their search *

* 72 searches were captured, 8 had capture errors

Results: comments from testers

- "The topic categories don't always start broadly enough."
- "It would be helpful if there were a more robust advanced search option up front, rather than forcing the user to click through the 'narrow search' options..."
- "I think that I deliberately chose books with the appropriate key words in the title, rather than those I would actually pursue for research purposes."
- "The reason that I chose to say that the sidebar was not helpful was that despite my attempts, I could not appropriately target results that we[re] both specific enough and yet general overviews."

Future Work

- *Year 3 Activities*
  - Usability and user studies: further face-to-face testing, data analysis, reporting & presentations
  - Topic modeling toolkit

- *Ongoing Research*
  - Topic segmentation and visualization
  - Regularized topic models

Questions and Contact Information

Thanks for coming!

Questions?

Contact Kat Hagedorn (khage@umich.edu),
David Newman (newman@uci.edu), or
Youn Noh (youn.noh@yale.edu).