

Using Topic Modeling in Digital Libraries: University of Michigan, Oct 7, 2010

David Newman (UC Irvine)

Kat Hagedorn (UMich)

Youn Noh (Yale)

William Dueber (Umich)

Roger Espinosa (Umich)

Background

- IMLS Research 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

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”

	<p>Madonna and Child, based on Barocci's etching Madonna and Child in the Clouds</p> <p><i>Artist/Maker:</i> Federico Barocci <i>Culture:</i> Italian <i>Date:</i> <i>Period:</i> 16th century <i>Accession #:</i> 1978.105 <i>Department:</i> Prints, Drawings, and Photographs <i>Location:</i> Viewable by appointment <i>Classification:</i> Works on Paper - Drawings and watercolors</p>
	<p>Madonna and child</p> <p><i>Artist/Maker:</i> Unknown <i>Culture:</i> Modern <i>Date:</i> 20th century <i>Period:</i> Modern <i>Accession #:</i> 2003.56.13 <i>Department:</i> Ancient Art <i>Location:</i> Not on view <i>Classification:</i> Sculpture</p>
	<p>Holy Family (Madonna and Child with Joseph, John the Baptist, Elizabeth, and Zacharias)</p> <p><i>Artist/Maker:</i> John Trumbull <i>Culture:</i> American <i>Date:</i> 1802-1806 <i>Period:</i> 19th century <i>Accession #:</i> 1832.83 <i>Department:</i> American Paintings and Sculpture <i>Location:</i> Not on view <i>Classification:</i> Paintings</p>
	<p>Madonna and Child with St. John the Baptist</p> <p><i>Artist/Maker:</i> John Trumbull <i>Culture:</i> American <i>Date:</i> 1801 <i>Period:</i> 19th century <i>Accession #:</i> 1832.98 <i>Department:</i> American Paintings and Sculpture <i>Location:</i> Not on view <i>Classification:</i> Paintings</p>
	<p>Holy Family (Madonna and Child with Joseph, Elizabeth, and John the Baptist)</p> <p><i>Artist/Maker:</i> John Trumbull <i>Culture:</i> American <i>Date:</i> 1839-1840 <i>Period:</i> 19th century <i>Accession #:</i> 1840.4 <i>Department:</i> American Paintings and Sculpture <i>Location:</i> Not on view <i>Classification:</i> Paintings</p>
	<p>Madonna and Child with Saint John</p> <p><i>Artist/Maker:</i> Unknown <i>Culture:</i> French <i>Date:</i> n.d. <i>Period:</i> 19th Century</p>





Collections and challenges

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

UM, History of Art, VRC

images with captions | images with record | captions only

check all unche

<p>view record</p> <p>UM HistArt VRC <input type="checkbox"/></p> <p>Dura-Europos, Late Mithraeum</p> <p>Int., south wall, portrait of Zoroaster</p> <p>100</p> <p>Creation location: Salahiyeh (Dayr az-Zawr, Syria); Repository: Yale University Art Gallery (New Haven, Connecticut, USA); Repository: 1c Iranian Decorative Arts Museum (Tehran, Iran)</p> <p>: no sort value</p>	<p>view record</p> <p>UM HistArt VRC <input type="checkbox"/></p> <p>Palazzo del Te</p> <p>Interior, Sala di Psiche, east wall, Jupiter and Olympia fresco, portrait of Giulio Romano</p> <p>architect: Giulio Romano (Italian, 1499-1546)</p> <p>designer: Giulio Romano (Italian, 1499-1546)</p> <p>1520</p> <p>Site: Mantua (Italy)</p> <p>: no sort value</p>	<p>view record</p> <p>UM HistArt VRC <input type="checkbox"/></p> <p>Vatican Palace, Sistine Chapel</p> <p>Portrait of Sixtus IV from medal by Antonio Guazzalotti da Prato</p> <p>Roman</p> <p>1400</p> <p>: no sort value</p>	 <p>UM HistArt VRC <input type="checkbox"/></p> <p>Issac Jefferson</p> <p>photographer: John Plumbe Jr. (American, 1809-1857)</p> <p>1840</p> <p>: no sort value</p>
 <p>UM HistArt VRC <input type="checkbox"/></p> <p>Portrait identified as Attalos I</p> <p>Two views</p> <p>Greek</p> <p>-210</p> <p>Discovery location: Pergamon (Denizli Ili, Aegean Region, Turkey); Repository: Staatliche Museen zu Berlin--Preussischer Kulturbesitz (Berlin, Germany)</p> <p>: no sort value</p>	 <p>UM HistArt VRC <input type="checkbox"/></p> <p>Portrait identified as Attalos I</p> <p>Front view</p> <p>Greek</p> <p>-210</p> <p>Discovery location: Pergamon (Denizli Ili, Aegean Region, Turkey); Repository: Staatliche Museen zu Berlin--Preussischer Kulturbesitz (Berlin, Germany)</p> <p>: no sort value</p>	<p>view record</p> <p>UM HistArt VRC <input type="checkbox"/></p> <p>Portrait identified as Attalos I</p> <p>Front view</p> <p>Greek</p> <p>-210</p> <p>Discovery location: Pergamon (Denizli Ili, Aegean Region, Turkey); Repository: Staatliche Museen zu Berlin--Preussischer Kulturbesitz (Berlin, Germany)</p> <p>: no sort value</p>	 <p>UM HistArt VRC <input type="checkbox"/></p> <p>Portrait Identified as Attalos I</p> <p>Two views</p> <p>Greek</p> <p>-0210</p> <p>Discovery location: Pergamon (Denizli Ili, Aegean Region, Turkey); Repository: Staatliche Museen zu Berlin--Preussischer Kulturbesitz (Berlin, Germany)</p> <p>: no sort value</p>

Collections and challenges

- Digitized books
- Images
- **Scientific literature**
- Web 2.0 content
- ... and more



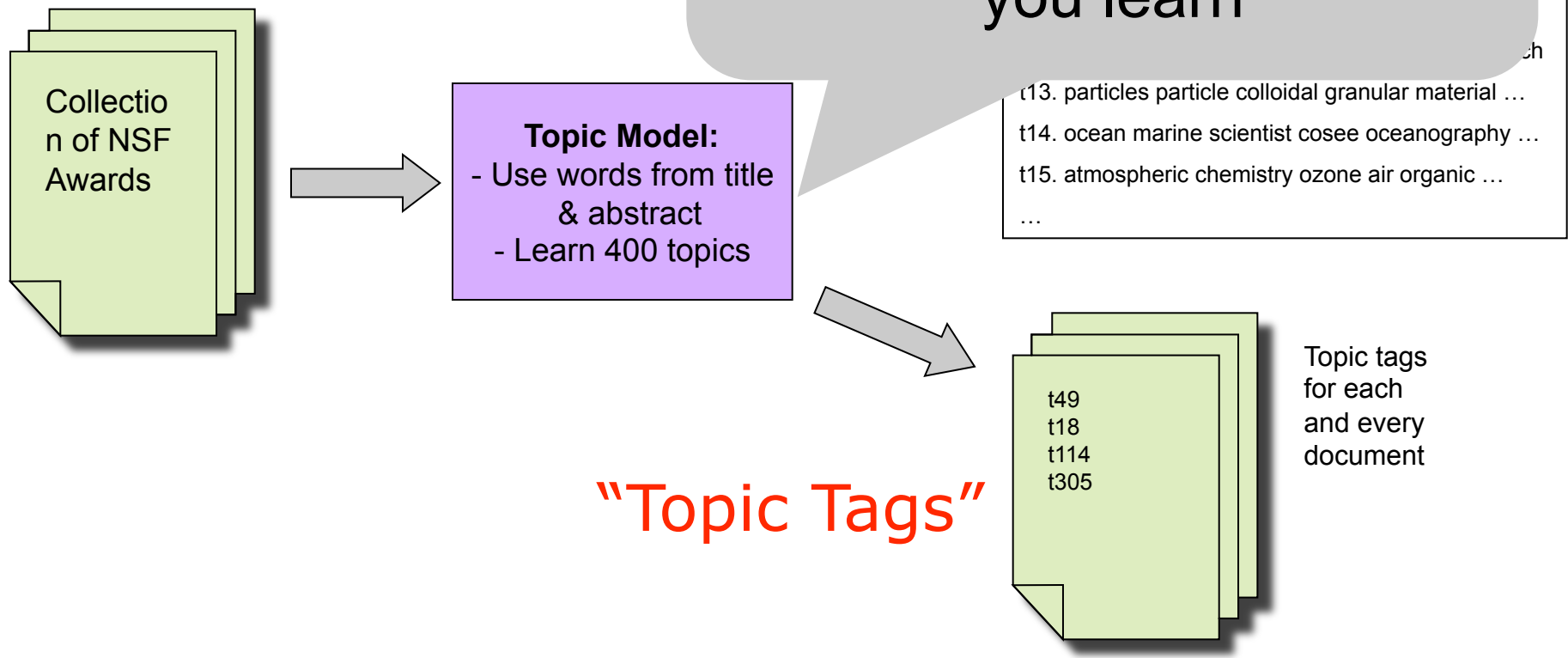
- 1000 new articles daily
- Indexed using MeSH

What is Topic Modeling?

- Topic Modeling (aka Latent Dirichlet Allocation)
- 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

Think of topic modeling as automatic assignment of subject headings ... that you learn



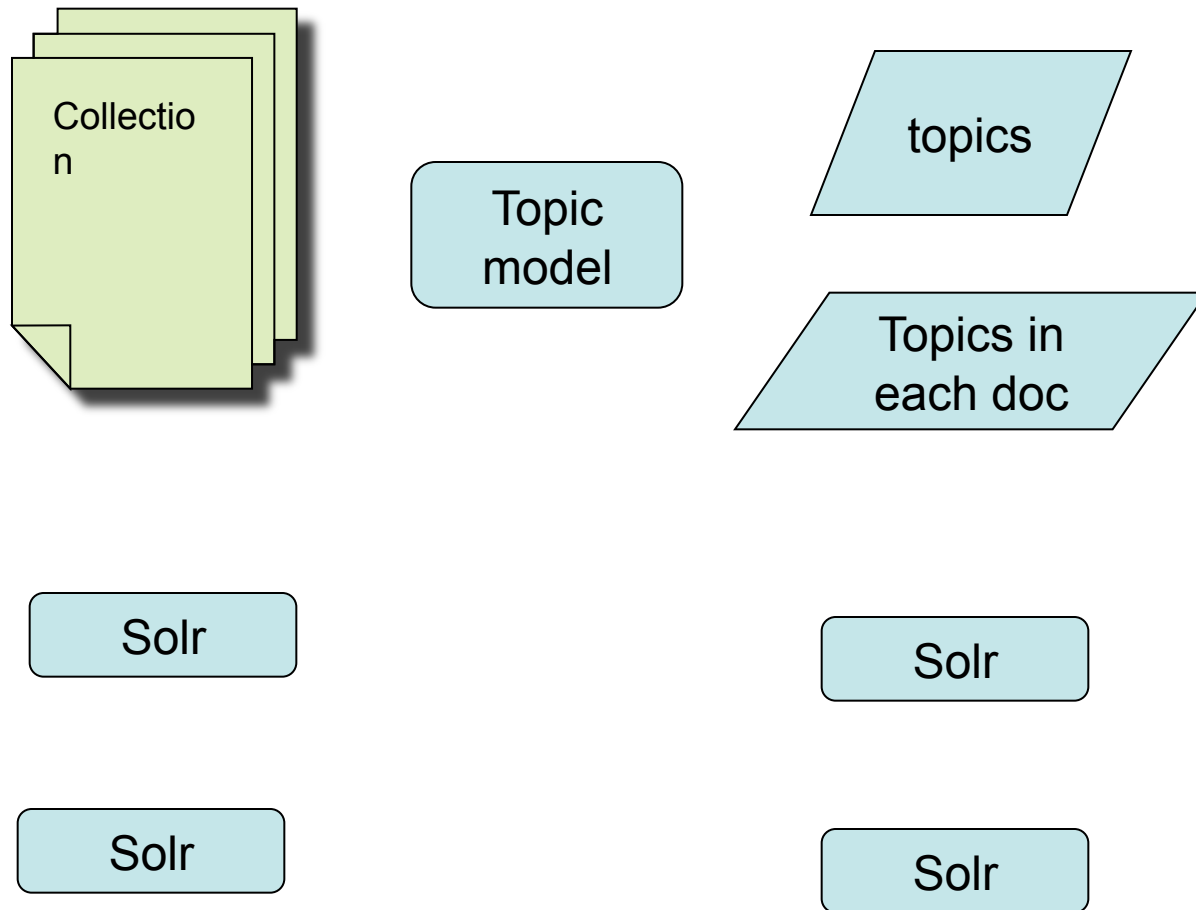
Tutorial

- Gain an understanding of topic modeling
- Understand where topic modeling might add value
- Learn how to integrate topics into the search back-end
- See how to use topic modeling user interfaces

Collections

- UMich History of ART
 - example
 - 100,000s images
- Books from Internet Archive
 - link
 - 100,000s books

Workflow



Topic Modeling

- Download MALLET
 - <http://mallet.cs.umass.edu/>
- Follow instructions to build
- Topic model collection
 1. import-dir
 2. train topics

Topic Modeling

1. `mallet import-dir`

```
--input /data/mycoll  
--output mycoll.mallet.in  
--keep-sequence  
--remove-stopwords
```

2. `mallet train-topics`

```
--input mycoll.mallet.in  
--num-topics 20  
--output-topic-keys topics.txt  
--output-doc-topics topicsindocs.txt
```


Topic Modeling

1. `mallet import-dir`

```
--input /data/mycoll  
--output mycoll.mallet.in  
--keep-sequence  
--remove-stopwords  
--extra-stopwords mystopwords.txt
```

2. `mallet train-topics`

```
--input mycoll.mallet.in  
--num-topics 20  
--output-topic-keys topics.txt  
--output-doc-topics topicsindocs.txt
```


- 
- Review topics.txt
 - Review topicsindocs.txt

Solr

- Download Solr
 - <http://lucene.apache.org/solr/>
- Follow instructions to build
- Ingest collection into solr
 1. Convert to “xml”
 2. Use post.sh

Solr

- Follow example in `exampledocs` directory
- Review `solr.xml`
- `Post.sh solr.xml`
- Check in solr admin panel
- Run searches in solr admin panel

Solr

- Solr is the popular, blazing fast open source enterprise search platform from the Apache Lucene project. Its major features include powerful full-text search, hit highlighting, faceted search, dynamic clustering, database integration, and rich document (e.g., Word, PDF) handling. Solr is highly scalable, providing distributed search and index replication, and it powers the search and navigation features of many of the world's largest internet sites.

Solr

Try it yourself! Design Techniques in Architecture

- Researchers in the C2 (Creative Consilience of Computing and the Arts) Program at Yale are investigating *Sketching and Alternative Design Techniques* in architecture:
- *Computer graphics plays a major role in the architecture profession. For example, modeling and rendering systems have proven to be invaluable aids in the visualization process, allowing designers to walk through their designs with photorealistic imagery. However, computer graphics techniques are typically employed at the conclusion of the design process. In fact, most of the artistic and intellectual challenges of an architectural design have already been resolved by the designer sits down in front of a computer.*
- <http://topics.catalog.hathitrust.org>
- <http://quod.lib.umich.edu/cgi/i/image/image-idx?c=hart4topics>



SPARE SLIDES

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	12,000 books
	Hathi Trust	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	100,000s
	UMich Library	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)

Less coherent
(unanimous score=1)

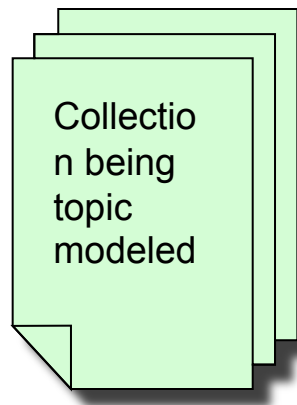
Books

silk lace embroidery tapestry gold embroidery
trout fish fly fishing water angler stream ..

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

$\#(\text{dance}, \text{music}) = 1$

Relatedness of word pairs

Topic: music dance band rock opera ...

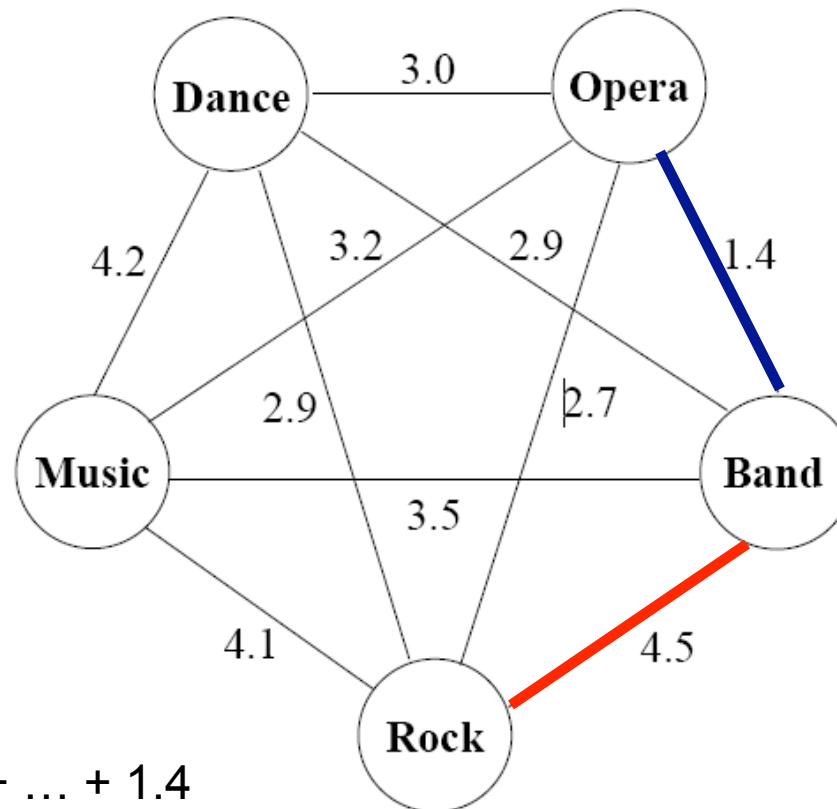
Pointwise Mutual Information (measure of dependence)

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

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

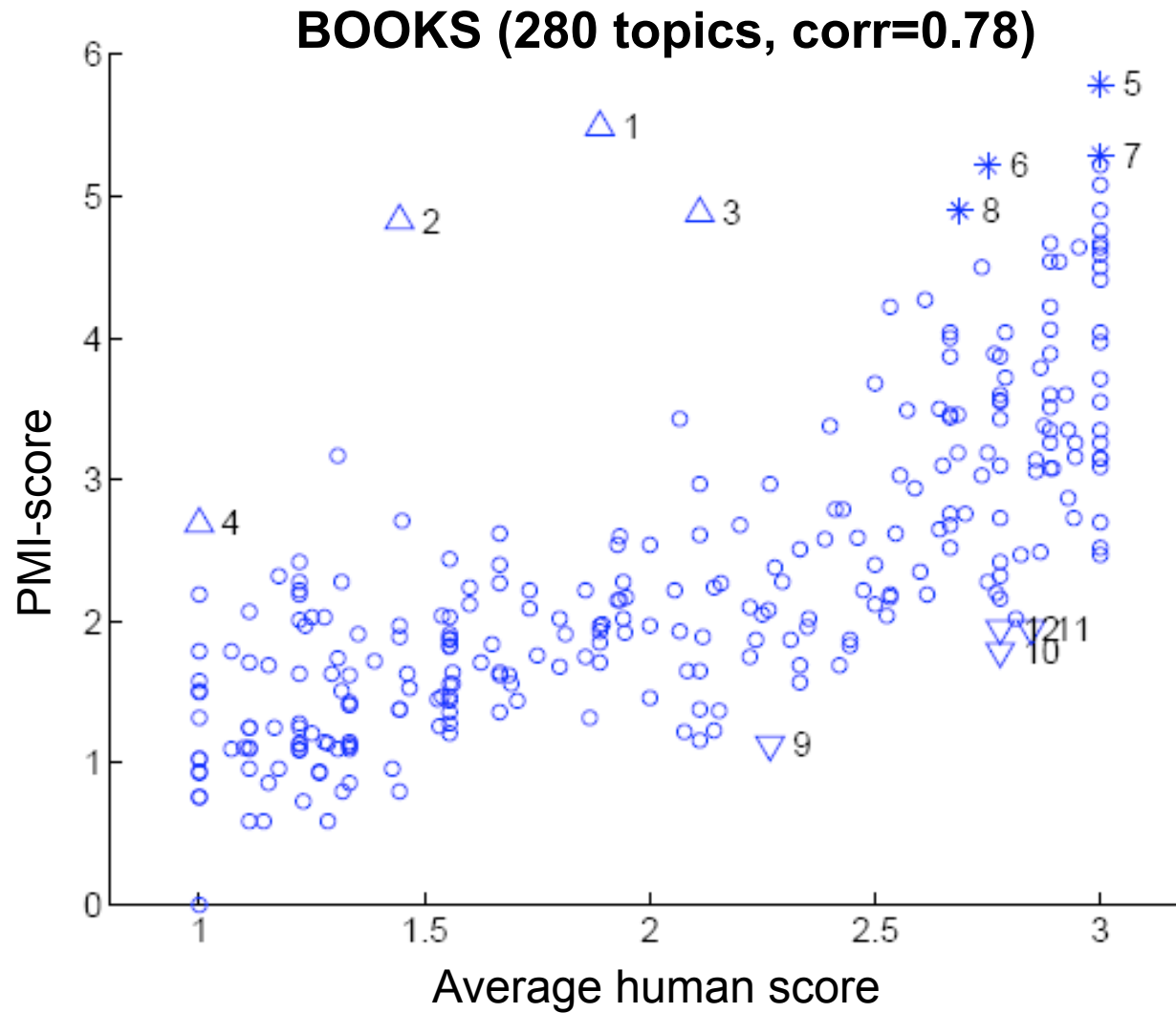
Relatedness of word pairs

Topic: music dance band rock opera ...

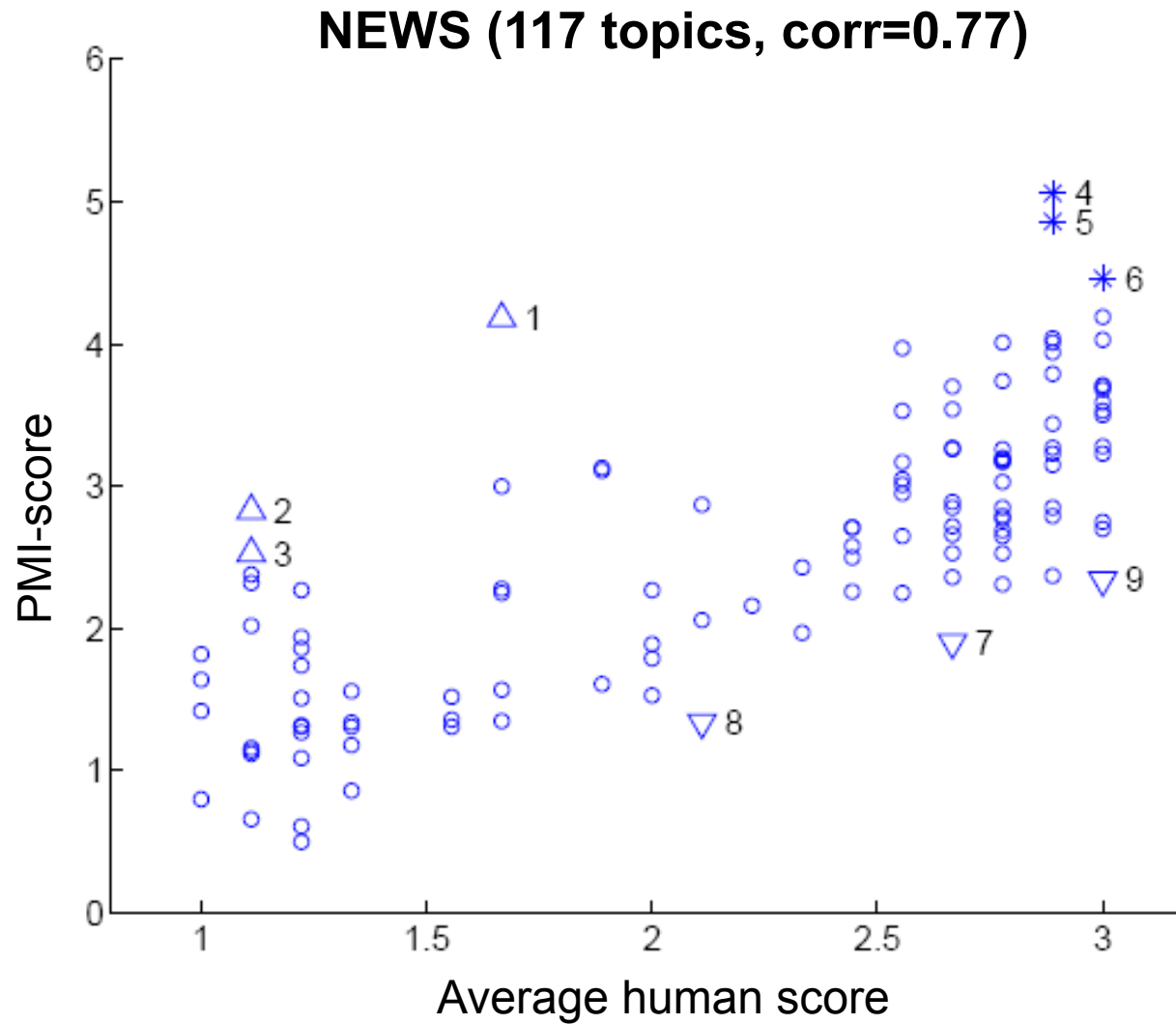


$$\text{PMI-Score} = 4.5 + 4.2 + \dots + 1.4$$

PMI-score achieves human-level performance



PMI-score achieves human-level performance



Correlation Results

Collection	Human-human correlation	PMI-Human correlation
Books	0.76	0.78
News Articles	0.73	0.77
Grant Abstracts	0.48	0.63
Image Metadata	0.51	0.53

Outliers

- PMI-score over-predicts coherence
 - thou thy thee hast hath thine mine heart god heaven
 - viii vii xii xiii xiv xvi xviii xix xvii main
 - century fifteenth thirteenth fourteenth twelfth sixteenth middle .
 - want look going deal try bad tell sure feel remember
- PMI-score under-predicts coherence
 - public government america policy nation political issues ...
 - british britain england country united national foreign nation ...
 - account cost item profit balance statement sale credit loss ...

Best topic word and suggested Label

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

Best topic word task

Topic

trout fish fly fishing water angler stream rod flies ...

space earth moon science scientist light nasa ...

race car nascar driver racing ...

Features

- PMI(word1, word2)
- Prob(word | *) ... word is evoked by other words ... e.g. space
- Prob(* | word) ... evokes other words ... e.g. nascar

SVM-rank using above features beats baseline of first topic word
(Lau, Newman, Karimi, Baldwin COLING 2010)

Suggested Label ... from Wikipedia (work in progress)

Topic

trout fish fly fishing water angler stream rod flies ...

Wiki Article Titles

fly fishing
fishing
angling
trout
...

space earth moon science scientist light nasa ...

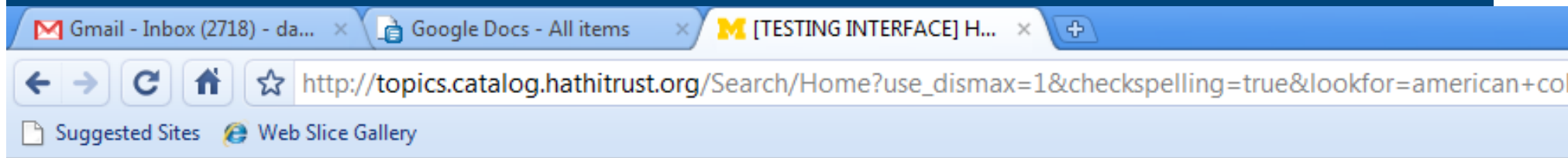
space exploration
space
space science
space colonization
nasa missions
...

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
- Assessing qualitative and quantitative value of topics



Prototype with topic facets



Catalog Search

All Fields

[Search Tips](#)

Narrow Search

Viewability

Limited (search-only) (523)

Full view (31)

Topic

architecture (84)

cultural identity (66)

slavery (51)

furniture (44)

politics (30)

[more...](#)

Subject

Architecture United States History 20th century. (29)

United States (26)

Art American 20th century Exhibitions

[Email this Search](#)

Showing 1 - 20 of 547 Results for **american colonial history**

Sort

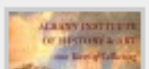
1 2 3 4 5 6 7 8 9 10 11 Next » [28]



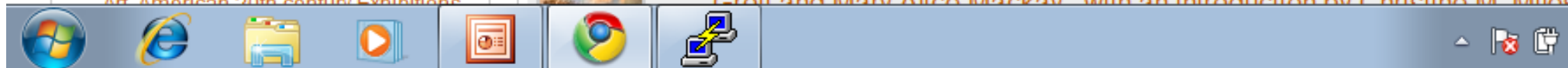
Treasures of the Old West : paintings and sculpture from the Thomas G
Institute of American History and Art / Peter H. Hassrick.

by Thomas Gilcrease Institute of American History and Art.
Published 1984

[Limited \(search-only\)](#)



Albany Institute of History & Art : 200 years of collecting / edited by Tam
Croft and Mary Alice Mackay ; with an introduction by Christine M. Milo



What else we can do with topics?

← → ↻ ☆ <http://datalab-3.ics.uci.edu/elsevier/climate/>

climate (5572 articles)

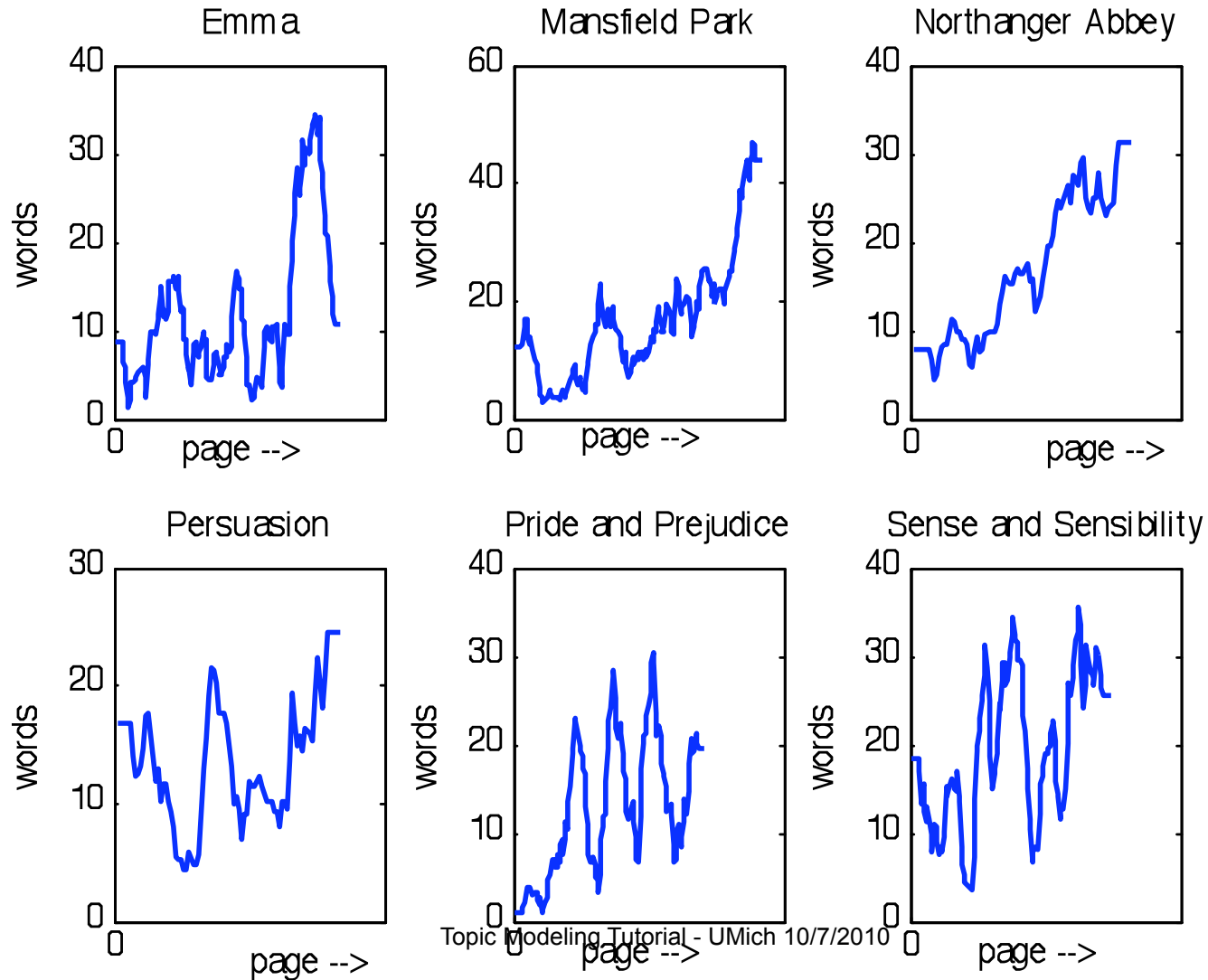
- 20% [t3] energy cost emission electricity fuel power market production ...
- 15% [t1] sediment sea ice level water record basin event ...
- 12% [t6] model water soil temperature data surface parameter climate ...
- 12% [t5] species model population data distribution effect mean forest ...
- 11% [t7] 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



Thank You

Goal

- Improve Search and Discovery
 - Improve user experience

- How
 - Improve search results
 - Improve ranking of search results
 - Improve display of search results

Collections and problems

- The internet (web pages)
- Scientific literature
- News articles
- Digitized books
- Images



- Uses PageRank to rank search results
- We've learned to use it
- We're used to it
- Limitation: One size fits all

Collections and challenges

- The internet (web pages)
- **Scientific literature**
- News articles
- Digitized books
- Images

CiteSeer^x_{beta}

Collections and problems: FIXME: DELETE

- The internet (web pages)
- Scientific literature
- News articles
- Digitized books
- Images
- How to get cross-cutting topics?
- Limitation: cross-cutting topics?

Evaluation

- Statistical
- User
 - How to eval
 - TREC-style

Selected high-scoring topics (unanimous score=3):

[News] *space earth moon science scientist light nasa mission planet mars ...*

[News] *health disease aids virus vaccine infection hiv cases infected asthma ...*

[Books] *steam engine valve cylinder pressure piston boiler air pump pipe ...*

[Books] *furniture chair table cabinet wood leg mahogany piece oak louis ...*

Selected low-scoring topics (unanimous score=1):

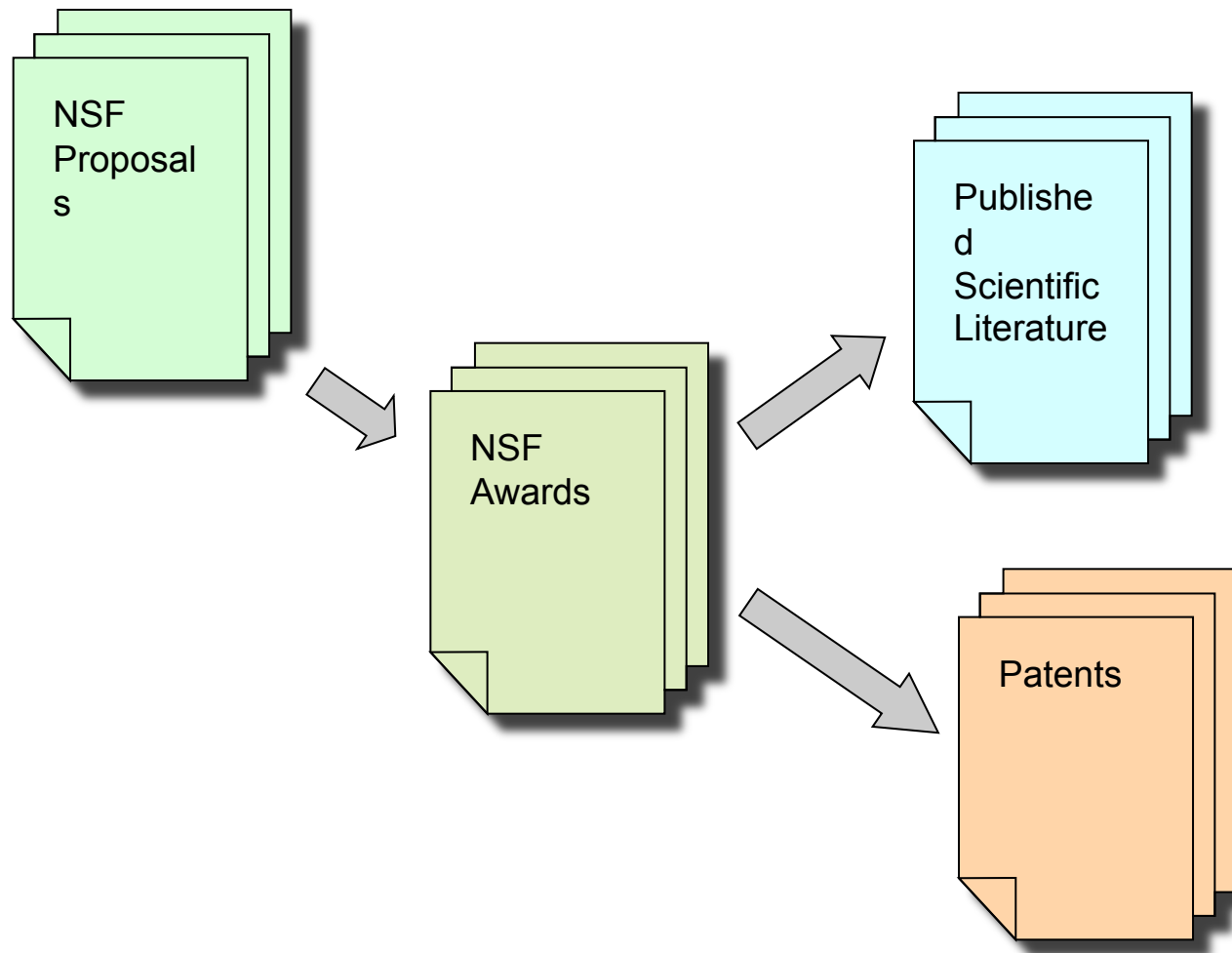
[News] *king bond berry bill ray rate james treas byrd key ...*

[News] *dog moment hand face love self eye turn young character ...*

[Books] *soon short longer carried rest turned raised lled turn allowed ...*

[Books] *act sense adv person ppr plant sax genus applied dis ...*

Topic modeling can relate and unify documents from different collections



Topic Model:

- Automatically learns set of topics for any collection of documents
- Labels each document with a few topics
- Provides a way to relate documents from different collections
- Reveals flow of ideas
- Is mature technology and considered state-of-art
- Is basis for this tool!

- Measure Spearman rank correlation between different methods and the average of the user ratings (reversing the sign for distance-based methods)
- Upper bound = average inter-annotator correlation, as measured by leave-one-out cross validation between the annotators