Evaluating Topic Models for Digital Libraries

David Newman[*]
Dept. Computer Science
University of California Irvine,
and NICTA Australia
newman@uci.edu

Youn Noh
Office of Digital Assets
and Infrastructure
Yale University
youn.noh@yale.edu

Edmund Talley
National Institute of Neuro.
Disorders and Stroke
National Institutes of Health
talleye@ninds.nih.gov

Sarvnaz Karimi
NICTA Australia
Melbourne, Australia
sarvnaz.karimi@nicta.com.au

Timothy Baldwin
Dept. Computer Science and
Software Engineering
University of Melbourne
tb@ldwin.net

ABSTRACT

Topic models could have a huge impact on improving the ways users find and discover content in digital libraries and search interfaces, through their ability to automatically learn and apply subject tags to each and every item in a collection, and their ability to dynamically create virtual collections on the fly. However, much remains to be done to tap this potential, and empirically evaluate the true value of a given topic model to humans. In this work, we sketch out some sub-tasks that we suggest pave the way towards this goal, and present methods for assessing the coherence and interpretability of topics learned by topic models. Our large-scale user study includes over 70 human subjects evaluating and scoring almost 500 topics learned from collections from a wide range of genres and domains. We show how a scoring model – based on pointwise mutual information of word-pairs using Wikipedia, Google and MEDLINE as external data sources – performs well at predicting human scores. This automated scoring of topics is an important first step to integrating topic modeling into digital libraries.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering

General Terms

Experimentation

Keywords

topic models, user evaluation

1. INTRODUCTION

The volume of online content accessible through digital libraries and search interfaces is continually growing for all types of collections spanning a wide range of genres and domains. One example is the hundreds of books being scanned every day in projects such as the Google Books Library Project or the Internet Archive. In MEDLINE, thousands of new (bio)medical articles are added each day. To broaden access to their collections, museums and art galleries are producing images of their artifacts and making them available online. For Web 2.0 content, growth can be on a much larger scale – in 2008, over one million images were added to Flickr every day.

In many cases, this growth in collections has outpaced our ability to connect users with specific content. Users searching for resources experience a growing problem with finding interesting, useful and relevant content. Standard information retrieval techniques have limitations: while full-text search allows for efficient Boolean or ranked retrieval, [3] have shown that over the past two decades, little progress has been made in the field, based on standardized evaluation. This issue is exacerbated by digitization. Optical character recognition errors in books, or limited and noisy descriptive data (e.g. for images) tend to worsen overall performance. With the continual growth in collection size, ad hoc retrieval becomes less useful to end users because of the growing size of search results sets, further increasing the need for subject tags.

Labels from ontologies and/or domain specific subject headings can be an aid to document collection navigation. But they are expensive to develop and maintain, both at the level of the class set, and in terms of labeling individual documents. While some digital library content comes with metadata (e.g. Library of Congress Subject Headings for books), collections are increasingly made up of highly heterogeneous content, that is sometimes with little or no metadata, or content from different sources that use different subject metadata standards, topic categorizations or controlled vocabularies.

Topic modeling holds much promise for improving the ways users access text content by automating both the class creation and labeling tasks [11]. The topic model — a recently-developed Bayesian model for text document collections [5, 8, 6] — is considered the state-of-the-art algorithm for extracting semantic structure from text collections. The

[*] Corresponding author.

JCDL'10, June 21–25, 2010, Gold Coast, Queensland, Australia.
Copyright 2010 ACM 978-1-4503-0085-8/10/06 ...$10.00.

topic model automatically learns a set of thematic topics (in the form of lists of words) that describe a collection, and assigns topic labels to each and every document in the collection.

Moreover, topic models provide a way to assemble and describe virtual collections on the fly for unanticipated uses and audiences. This is an improvement upon the largely static digital libraries we have now, with ontologies and/or subject headings that may be difficult to merge.

The potential benefit of applying topic models to improve the ways users find, understand, search and browse content has been demonstrated in various contexts (e.g. [10, 12, 4], `Rexa.info`). For example, the computer science digital library `Rexa.info` extensively uses topic modeling: Rexa learns a consistent set of topics across seven million computer science research articles (that otherwise have no consistent subject metadata); it then allows one to search either by topic or by keyword; and finally it uses topics to analyze areas of research, measure their impact, and show how they relate.

While topics learned by the topic model often look useful — particularly for use in search interfaces and/or digital libraries — they can be less useful for end-use because they have no particular theme. To apply topic modeling to solve real world problems, we need an easy way to identify "junk" topics (e.g. *banana sky canoe furious*), which may be statistically well founded, but of no use to end users. User-based evaluation of automatically-learned topics is the focus of this paper, which is part of an ongoing line of research on comprehensively evaluating topic models for digital libraries. Our aim in this paper is to both propose a sequence of evaluation tasks which we suggest contribute to the ultimate end-goal of enhanced topic modeling, and present concrete results over a subset of those tasks.

2. RELATED WORK

To date, most evaluation of topic models has focused on statistical measures of perplexity or likelihood of test data. But this type of evaluation has limitations. The perplexity measure does not reflect the semantic coherence of individual topics learned by a topic model. Nor does perplexity indicate how well a topic model will perform in some end-user task. In fact, recent research has shown potential issues with perplexity as a measure — [7] suggest that human judgements can sometimes be contrary to perplexity measures.

With this in mind, we pose the following overarching questions relating to evaluating topic models:

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?

While the final question is ultimately the most important, it is appropriate to address these questions in order. It doesn't make sense to talk about the quality of assignments of topics to documents, if one can't agree on what a topic is about. Although topics themselves are not the end goal (the end goal is to use topics to improve some end-user task), the evaluation framework is built on the usability and usefulness of individual topics, and our focus in this paper is thus on the first of the three questions. In the remainder of this section, we review related work on evaluation of topics and topic models.

[1] presented an unsupervised approach to ranking topic significance and identifying what they call "junk" or "insignificant" topics. It is unclear to what extent their unsupervised approach and objective function agrees with human judgements, however, as they present no user evaluations.

The ability to come up with a short topic label for a learned topic is at the heart of the question of topic coherence and interpretability. [9] proposed various approaches for automatically suggesting labels for topics that had some success for this broadly unsolved problem. However, open questions remain. For example, their framework was unable to produce labels that are hypernyms of topic words, which is often the most appropriate type of label.

[2] addressed the fact that the topic model can learn topics that contain one or more words that are clearly not associated with the theme of the topic. By incorporating domain knowledge into the topic model (in terms of correlated and uncorrelated pairs of words, encoded as "must-links" and "cannot-links"), they were able to penalize pairs of words from occurring in topics, and emphasize pairs of words to occur in topics. However, it appears that the incorporation of the domain knowledge is a manual task, and it is unclear how one would anticipate which unrelated pairs of words might turn up in learned topics.

Choosing the number of topics — a long-discussed subject — has a huge effect on the learned topics. While the arrival of the non-parametric topic model [15] provided a neat mathematical solution to this problem, it was less widely adopted in practice. [16] proposed that the optimization of a probability prior (one of the model parameters) is critical, and that the correct optimization of this prior may alleviate the problem of exactly picking the "right" number of topics to be learned. While this work did not address evaluating individual topics, it did provide one approach to simplifying the configuration of learning a topic model.

The first works to report on human scoring of topics were [7], [13] and [14]. The first study used a novel but synthetic intruder detection task where humans evaluate both topics (that had an intruder word), and assignment of topics to documents (that had an intruder topic). The second and third studies had humans directly score topics learned by a topic model. This latter work introduced the pointwise mutual information score to model human scoring that we further develop and much more extensively report on here.

These works on evaluating topic models highlight an important need in the topic model community, and broaden the definition of evaluation. The first overarching question of quality and coherence of individual topics is still largely unexplored, and is the focus of this paper.

3. EXPERIMENTS

3.1 Collections

To evaluate the effectiveness of our proposed methods under a range of settings which is truly representative of the actuality of digital libraries, we experiment with a set of document collections which varies widely over document type, document length, genre and domain. Documents in the collections vary from 1000-page books down to 10-word metadata records associated with an image. Domains range from medieval art to the latest biomedical research. The col-

Collection	Source(s)	Avg. Doc. Len. (word)	No. Topics	Total No. Evals.
Books	Hathi Trust (28,000 books) Internet Archive (12,000 books) Yale Center for British Art (500 books)	∼10,000	280	3,680 (N=11-15)
News	NY Times from LDC Gigaword (55,000 news articles)	∼1,000	117	1,053 (N=9)
Abstracts	NIH/NINDS (60,000 grant abstracts)	∼100	60	600 (N=10)
Metadata	Yale/VRC (200,000 images) Yale/YUAG (90,000 images) UMich/HART (290,000 images)	∼10	40	440 (N=11)

Table 1: Collections used for topic modeling and human evaluation. N is the number of users evaluating each topic.

lections were organized into four categories: Books, News, Abstracts and Metadata, as shown in Table 1, where each category is made up of one or more sub-collections obtained from different sources.

We intentionally selected collections where learned topics could potentially improve the way users find and browse content, or where one may want to create a thematic virtual sub-collection on the fly. One common feature of all the collections is that there is little or no pre-existing subject metadata to help users find items. For example, a 1911 book on Naval Architecture by Cecil Peabody (in Hathi Trust, digitized by Google), only included the subject heading *Naval Architecture*, which obviously does not provide any additional information to the user about relevance, nor does it help in faceted browsing.

Our collection of over 40,000 books came from multiple sources including Hathi Trust, Internet Archive and the Yale Center for British Art. While full-text search is currently (or soon to be) available across all these collections, the huge and growing collection sizes make it difficult for users to obtain the best search results. For example, a search for *naval architecture* returns 154 books in the Internet Archive search interface, and 350 books in the Hathi Trust search interface. Beyond the time consuming task of browsing through each book, the user has no guidance in quickly seeing what each book is about, and how books differ.

The News and Abstracts collections are arguably easier candidates for topic modeling. In both cases — news articles written by journalists, and NIH grant abstracts written by researchers and scientists, respectively — documents usually attempt to clearly convey information to the reader. Nevertheless, these documents are excellent candidates for topic modeling because manual subject tagging or classification/categorization by editors is expensive, unscalable and often inaccurate.

The Metadata collection consisted of a huge volume of relatively short metadata records that accompany individual images from three different image collections at Yale University and University of Michigan. Typically, these text records have very limited descriptive metadata (on occasion listing just the artist and title), and keyword search will often return too many or too few results. For image collections like these, creating useful and meaningful topic labels could potentially have the biggest benefit to users, since effective image search is a particularly difficult problem.

3.2 Topic Modeling

The input to the topic model is a bag-of-words representation of the collection of text documents, where word counts are preserved, but word order is lost. While this sounds straightforward, the choice of tokenization strategy can have a large effect on the topics learned. One can make any combination of decisions around stemming, filtering out terms (e.g. only keeping nouns), removing stopwords, and identifying and replacing collocations or multiword expressions. We opted for the fairly standard approach of not stemming, removing stopwords, and not tokenizing collocations. Stemming is not usually used in topic modeling when there is ample data, e.g. more than 10^6 total number of words in the corpus, because the unstemmed terms generally result in topics that are more easily interpreted.

Similar procedures were used to create the bag-of-words representation for each collection. The biggest difference was whether or not to filter out proper nouns. We can argue that proper nouns are easily found by keyword search, and are thereby not needed in learned topics. Furthermore, without proper nouns, one may expect to learn topics that are more general and broad in nature. While we may want to exclude proper nouns to learn more general topics, we'll see that keeping proper nouns sometimes makes learned topics more coherent and interpretable.

For Books, we made the decision to remove proper nouns. This is in anticipation of using topics for faceted search, in conjunction with keyword search. For example, a user may start by issuing a keyword search for *Picasso*, but then proceed by filtering the search results using the topic facets. For all other collections (News, Abstracts and Metadata), proper nouns were kept. For News and Abstracts, one could argue either way on whether or not to include proper nouns. However, for Metadata, the paucity of data meant there was no choice but to include proper nouns, to give the topic model the best chance of learning meaningful topics.

We learned topic models for each of the eight source sub-collections presented in the second column of Table 1. In all cases, topic models were learned for different topic number settings. After a quick review of the learned topics, one particular setting of number of topics was selected for each sub-collection, and used in all experiments presented in this paper. For each selected topic model, we used the conventional approach of presenting the topic via its top-10 words (which on average contain about 30% of the topic mass). From over 1000 learned topics from the different topic models, we selected 497 topics for human evaluation. Our selection of topics was either random, comprehensive, or stratified to over-represent good and bad topics based on an initial assessment.

4. HUMAN SCORING OF TOPICS

The 497 topics selected for human scoring were each evaluated by between 9 and 15 users (from a combined pool of 70 users), for a total of 5,773 topic evaluations. In a typical survey, a user was asked to evaluate anywhere from 60 to

120 topics. The instructions varied slightly for surveys from different collections, but in general, we asked, for each topic in the survey:

- Score topic for "usefulness" on a scale of 1 to 3
- Select the single best word that exemplifies the topic (for topics with score=3)
- Select one or more worst words (this task was only specified for Metadata topics)
- Suggest a topic label (for topics with score=3)

We provided guidelines on how to judge whether a topic was useful or less useful/useless/junk. In addition to showing several examples of useful and useless topics, we gave the following instructions to people performing the evaluation:

> *The topics learned by a topic model are usually sensible, meaningful, interpretable and coherent. But some topics learned (while statistically reasonable) are not particularly useful for human use. To evaluate our methods, we would like your judgment on how "useful" some learned topics are. Here, we are purposefully vague about what is "useful" ... it is some combination of coherent, meaningful, interpretable, words-are-related, subject-heading-like, something-you-could-easily-label, etc.*

Our human evaluators came from a range of populations. Since Books and News topics were familiar to a fairly general audience, we recruited from a general pool of students, faculty and staff on three different university campuses. For the NIH grant abstracts (Abstracts), our users included 10 experts from the National Institutes of Health. For Metadata, we recruited students more familiar with the various image collections to take our survey. The average number of users evaluating each topic is listed in Table 1 (as N in the final column).

Figure 1 shows a selection of topics with high and low human scores, from topics for the Books, News, Abstracts and Metadata collections. High-scoring topics are exemplified by a set of words that has a clear semantic thread, which could easily be labeled, and which could be used to find and access content (i.e. is useful as a subject heading). For example, the News topic *space earth moon science scientist light nasa mission planet mars* is obviously about space exploration, and documents tagged with this topic would likely be relevant to space exploration.

The low scoring topics exhibit a variety of problems, from inclusion of typographical errors (*httle*) to being a seemingly useless list of first names (*john william james henry robert*). Other low scoring topics include a list of fairly general words that may be found in prose (*soon gave returned replied told appeared arrived*). One common aspect of the low scoring topics is that we would not expect them to be particularly useful information when searching or browsing. Note that these low scoring topics are not artifacts produced by the topic model, but are in fact stable and robust statistical features in the text data. It is also useful to keep in mind that topic models have no internal way to learn semantically sensible topics: they are just learning patterns of co-occurring terms.

To make sense of our human-scoring of topics, we need a reasonable level of inter-rater agreement, as measured by the Spearman rank correlation of individual users with the mean of the remaining users, computed on a leave-one-out basis. Before we proceed to presenting models to predict the human scoring of topics, and comparing these to the human judgements, we removed some outlier human evaluators and topics. Out of all 497 topics, we removed 11 topics with the highest variance in human scores (4 out of 60 of the Abstracts topics, and 7 out of 40 Metadata topics), and just two evaluators (one from Abstracts, and one from Metadata). This removal of outliers improved the inter-rater correlation for Abstracts and Metadata up to around $\rho = 0.5$. No removal of topics or users was done for Books or Abstracts, since the inter-rater correlation was already high at $\rho = 0.76$ and $\rho = 0.73$, respectively.

Selected high-scoring topics:	
Books	silk lace embroidery tapestry gold embroidered ...
	ware porcelain pottery potter ceramic glaze ...
	trout fish fly fishing water angler stream rod flies ...
News	space earth moon science scientist light nasa ...
	health drug patient medical doctor hospital care ...
	car ford vehicle model auto truck engine sport ...
Abstracts	epilepsy seizures seizure epileptic epileptogenesis ...
	pain chronic neuropathic migraine sensitization ...
	spinal cord injury sci recovery regeneration injured ...
Metadata	japan scroll kamakura ink hanging oyobe hogge silk ...
	persian iran manuscript folio firdawsi century ...
	drawing plan architecture 20th elevation building ...
Selected low-scoring topics:	
Books	able bring eye hfe hght hke httle lost power turn
	head friend fellow captain open sure turn human ...
	soon gave returned replied told appeared arrived ...
News	oct sept nov aug dec july sun lite adv globe
	dog moment hand face love self eye turn young ...
	art budget bos code exp attn review add client sent
Abstracts	research clinical training career candidate neurology ...
	cns mechanisms repair nervous determine ...
	conditions specific effects systems proposal ...
Metadata	abstraction sculpture united female english tabbaa ...
	print 19th etching 18th french engraving manuscript ...
	oil john canvas william james england henry robert ...

Figure 1: Selected high-scoring and low-scoring topics from each of the four collections (each line is one topic, with fewer than ten topic words displayed because of limited space). High-scoring topics are exemplified by a set of words that have a clear semantic meaning, and which could potentially be used to find and access content. Low-scoring topics exhibit a variety of problems making them less useful.

5. PMI MODEL TO SCORE TOPICS

The intuition behind the scoring model comes from the idea that the coherence of a set of ten words implies relatedness of most or all pairs of words taken from the set. This leads to the idea of a scoring model based on word association between pairs of words, for all word pairs in a topic. But instead of using the collection itself to measure word association — which could reinforce noise or unusual word statistics — we use a large external text data source.

Specifically, we measured co-occurrence of word pairs from large external text datasets such as all articles from English Wikipedia, the Google n-gram data set, or all of MEDLINE, as

External		
Collection	Source	Description
Wikipedia	`en.wikipedia.org` (articles)	2M articles, 1G words
Google	LDC Web 1T (2-grams)	Statistics from 1T words
MEDLINE	`pubmed.gov` (abstracts)	19M abstracts, 1G words

Table 2: External text data collections used for PMI scoring of topics.

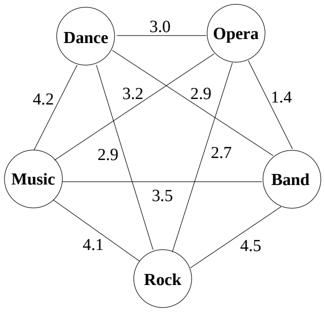

Figure 2: Illustration of pointwise mutual information between word pairs.

	Inter-rater	PMI-Human Correlation		
Collection	Correlation	Wikipedia	Google	MEDLINE
Books	0.76	0.78	0.79	—
News	0.73	0.77	0.72	—
Abstracts	0.48	0.55	0.55	0.63
Metadata	0.51	0.53	0.41	—

Table 3: Spearman rank correlation between PMI score and mean human score. In most cases, the correlation between the PMI and the human score is better than the inter-rater correlation.

shown in Table 2. For Wikipedia and MEDLINE we counted a co-occurrence as words w_i and w_j co-occurring in a 10-word window in any article/abstract, and for Google n-grams, we counted co-occurrence as w_i and w_j co-occurring in the list of 2-grams.

Following [13] and [14], we use pointwise mutual information (PMI) as the measure of word association, and define the following scoring formula for a topic $\mathbf{w}$:

$$\text{PMI-Score}(\mathbf{w}) = \text{mean}\{\text{PMI}(w_i, w_j), ij \in 1 \ldots 10, i \neq j\},$$

$$\text{PMI}(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)},$$

where the top-10 list of words in a topic is denoted by $\mathbf{w} = (w_1, \ldots, w_{10})$, and we exclude the self-PMI $i = j$. The PMI score for each topic is the mean PMI for all pairs of words in a topic (so for a topic defined by the top-10 words, the PMI score is the mean of 45 PMIs). PMI is a simple method which can tap into large data sets such as Wikipedia. Additionally, in related work comparing a wide array of scoring methods [14], it was found to consistently be the best performer out of all methods trialled.

The PMI scoring model is illustrated in Figure 2 for a topic of five words: *music band rock dance opera* (here we demonstrate using 5 words instead of 10 for simplicity). Using co-occurrence frequencies from Wikipedia, we see unsurprisingly high-scoring word pairs, such as PMI(*rock*,*band*)=4.5 and PMI(*dance*,*music*)=4.2. Some pairs of words exhibit more independence, such as PMI(*opera*,*band*)=1.4. The PMI score for this topic is the simple arithmetic mean of all the PMIs, or PMI-Score=3.2.

Table 3 lists the correlation between the PMI scores and the mean human scores for topics in the four collections. The first column — which can can be considered as an upper bound for the task — is the inter-rater correlation, i.e. a measure of how much the human evaluators agree with the average of the scores from the remaining human evaluators for that topic. As the gold standard score used in evaluating the PMI scores is similarly calculated by averaging the human scores, any score at or above the inter-rater correlation can be interpreted as that method having achieved human performance levels.

Starting with Wikipedia as our external data for PMI scoring, we see excellent agreement between the PMI score and the mean human score, with the correlation exceeding the inter-rater correlation in all cases. Particularly impressive are Books and News, where the correlations are close to 0.8 for a total of 397 topics spanning a huge range of subjects. While not shown in the table, the correlation for the subset of topics from the Internet Archive is a very high $\rho = 0.85$.

The correlations between PMI and the average human score for Abstracts and Metadata are lower, possibly due to domain specific content and noisy topics. As Abstracts is sourced from NIH grant data, Wikipedia is not the best choice for external data. For example, the term *apoptosis* occurs 8712 times in Wikipedia, but 630,093 times in MEDLINE, so we expect better PMI statistics from MEDLINE for these biomedical topics.

Using Google's 2-grams to compute PMI scores produces somewhat similar PMI-human correlations as those from Wikipedia-based PMI scores. This result is a good indication that the PMI scoring model is robust. Again, we see the most consistency in Books and News, which include topics that are well reflected by these external data sources. Metadata had more of a drop in correlation when PMIs were computed using Google's 2-grams, in large part due to the lower counts and the thresholding used in the creation of the Google n-gram data sets. Our implementation of PMI scoring using the Google's 2-grams is possibly overly simplistic, and better results have been obtained by instead using Google's 5-grams [13].

The final column shows the correlation result for Abstracts with PMIs computed using MEDLINE, which is arguably the most appropriate external data for the Abstracts collection of NIH grants. Here, we see a marked improvement to $\rho = 0.63$, again which is well above the inter-rater correlation for those 56 topics of $\rho = 0.48$.

Figure 3 shows scatterplots of PMI score vs. mean human score for topics from the four collections, where the PMI score was computed using Wikipedia for Books, News and Metadata, and using MEDLINE for Abstracts. While we see broad agreement between PMI scores and human scores,

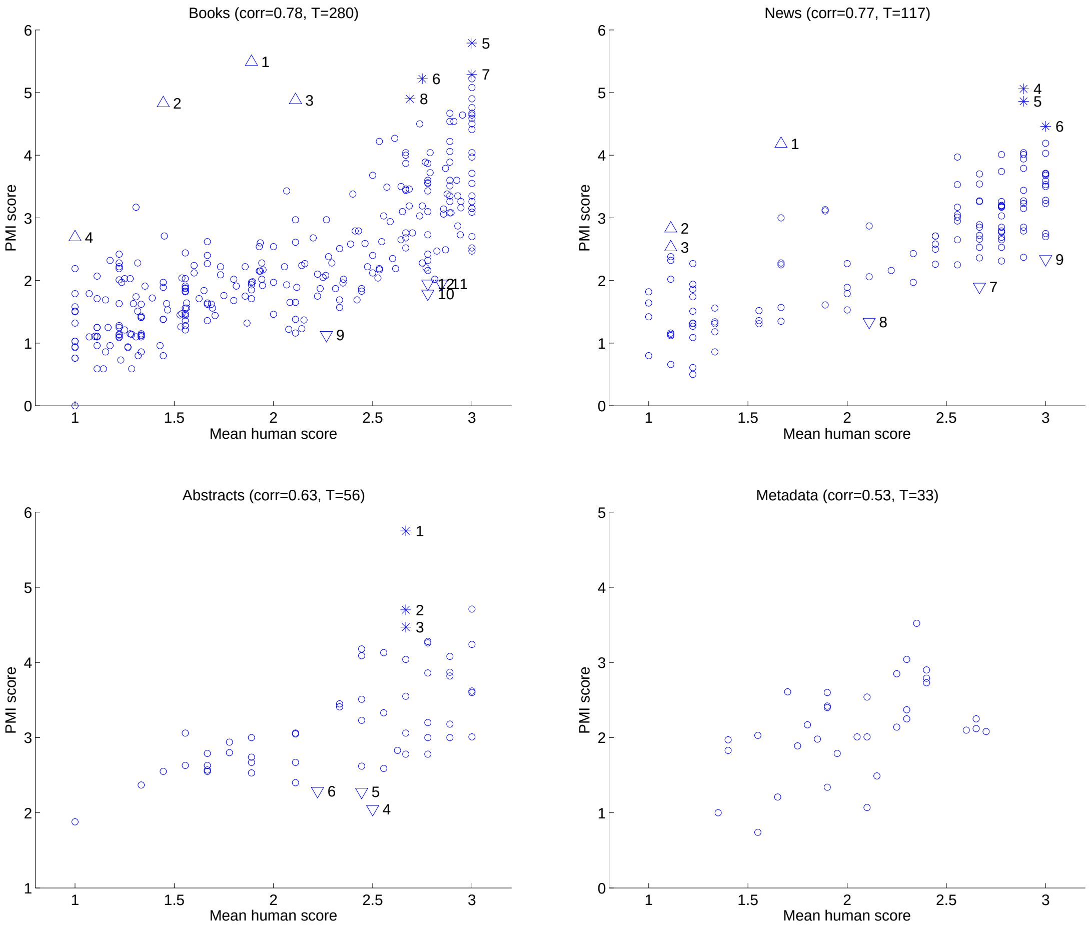

Figure 3: Scatter plots showing PMI score vs. mean human score for topics in the four collections. Annotated topics show outliers that either overpredict PMI at low scores (△), overpredict PMI at high scores ($*$), or underpredict PMI at high scores (▽). We see relatively good agreement between the PMI score and the human score (as indicated by the correlation values) across a wide variety of text collections.

we do see patterns of outlier topics. Figure 4 contain topics marked in Figure 3 as △, $*$ or ▽. We see that over-prediction at low scores (△) happens for words that go together, but are uninteresting from a topical or subject heading perspective. Over-prediction at high scores ($*$) happens for topics that have low-frequency words, where PMI is notorious for over-estimation of the score. Finally, under-prediction at high scores (▽) happens for topics that contain fairly general words. In this case the topic is useful as a broad subject heading, but the PMI score is low because the words are frequent and typically have many senses.

In terms of using the PMI score to rank topics, over-prediction at high scores is not a real concern (we are more interested in over-prediction at low scores and under-prediction at high scores, and most interested in accurate identification of junk topics). This somewhat inconsequential over-prediction at high scores could be reduced by smoothing counts, since PMI tends to overstate PMI of pairs of rare words. Specifically, the maximum PMI possible in a corpus of N words is when word w_1 and w_2 occur just once, and occur together. In that case $\mathrm{PMI}(w_1, w_2) = \log \frac{1/N}{1/N1/N} = \log N$. Smoothing by adding pseudo counts will reduce this upper bound on PMI, and tend to correct the over-prediction at high scores.

We checked the sensitivity of the PMI-human correlation to the window size used to compute individual PMI scores and found that PMI-human correlation was relatively insensitive to window size. Generally, we observed that a 10-word window worked well – unless noted, all correlations reported in this paper are computed with a window of 10 words.

6. HUMAN-SELECTED BEST TOPIC WORD

As part of our proposed roadmap for evaluating topic models for use in digital libraries, we devised the tasks of selecting the best topic word and suggesting a short label. In our user-studies, human subjects were asked to select the single best word contained in a given topic that exemplified the topic. For some surveys, we instructed the evaluators to only complete this task for topics that they scored highly (score=3). Here we compare that human-selected best topic word to the PMI score for that word, as defined by:

$$\text{PMI-Score}(w_i) = \mathrm{mean}\{\mathrm{PMI}(w_i, w_j), j \in 1 \ldots 10, j \neq i\}.$$

NEWS	**space** $\boxed{\text{earth}}$ moon science scientist light $\boxed{\text{nasa}}$ mission planet mars
	health drug $\boxed{\text{patient}}$ **medical** doctor hospital care **cancer** $\boxed{\text{treatment}}$ disease
	cell $\boxed{\text{human}}$ animal scientist **research** gene $\boxed{\text{researcher}}$ brain university science
	$\boxed{\textbf{car}}$ ford **vehicle** model **auto** truck engine sport wheel $\boxed{\textbf{motor}}$
	$\boxed{\textbf{health}}$ care **insurance** patient hospital medical cost $\boxed{\textbf{medicare}}$ coverage doctor
ABSTRACTS	$\boxed{\text{children}}$ **pediatric** exercise physical $\boxed{\text{childhood}}$ **intervention** syndrome developmental adults problems
	enzyme acid synthesis enzymes cholesterol $\boxed{\textbf{lipid}}$ products $\boxed{\text{fatty}}$ specific storage
	treatment clinical patients therapy $\boxed{\text{efficacy}}$ $\boxed{\textbf{trial}}$ phase safety dose drug
	$\boxed{\text{hiv}}$ infection virus viral $\boxed{\text{infected}}$ cns brain **aids** replication macrophages
	estrogen **hormone** sex effects androgen $\boxed{\textbf{hormones}}$ male female $\boxed{\textbf{steroid}}$ estradiol

Figure 5: Illustration of the best word task. The bolded words are the words selected by humans as the best words, and the words in boxes are the top-2 words ranked by PMI.

BOOKS

Over-prediction at low scores ($\triangle$)

1 thou thy thee hast hath thine mine heart god heaven
2 viii vii xii xiii xiv xvi xviii xix xvii main
3 century fifteenth thirteenth fourteenth twelfth sixteenth middle ...
4 want look going deal try bad tell sure feel remember

Over-prediction at high scores ($*$)

5 ware porcelain pottery potter ceramic glaze decoration vases ...
6 cloth cotton loom machine mill thread warp weaving wool yarn
7 gothic vault vaulting nave rib romanesque aisle pier arches church
8 aisle arch arches chapel nave pointed roof tower transept window

Under-predictions at high scores ($\bigtriangledown$)

9 assessment associa association common covenant development ...
10 british britain england country united national foreign nation ...
11 account balance cost credit item loss profit sale shown statement
12 account cost item profit balance statement sale credit shown loss

NEWS

Over-prediction at low scores ($\triangle$)

1 mexico los mexican que del las una para por mas
2 bad doesnt maybe tell let guy mean isnt better ask
3 johnson jones miller scott robinson george lawrence murphy ...

Over-prediction at high scores ($*$)

4 kosovo serb nato war albanian milosevic yugoslavia serbian ...
5 food restaurant chef recipe cooking meat meal kitchen eat chicken
6 health disease aids virus vaccine infection hiv cases infected ...

Under-predictions at high scores ($\bigtriangledown$)

7 public government america policy nation political issues ...
8 www mail internet online page list search user message free
9 space earth moon science scientist light nasa mission planet mars

ABSTRACTS

Over-prediction at high scores ($*$)

1 myelin axonal oligodendrocyte myelination oligodendrocytes ...
2 synaptic synapses plasticity synapses mechanisms transmission ...
3 cortical cortex sensory somatosensory neurons specific input ...

Under-predictions at high scores ($\bigtriangledown$)

4 agents exposure nerve agent chemical toxin htt therapeutic ...
5 biomarkers specific detection samples levels diagnosis mass ...
6 respiratory motor sma smn disease patients ganglioside ...

Figure 4: Outliers to correlation between PMI and human score.

	Match with top-n PMI words					
Collection	$n = 1$	≤ 2	≤ 3	≤ 4	≤ 5	≤ 10
BOOKS	19	36	46	59	63	100
NEWS	23	37	49	62	76	100
ABSTRACTS	32	52	62	76	84	100

Table 4: Ability of PMI to predict the human-selected best word. $n = 1$ column is percent of time that best word is ranked 1st in terms of PMI score (i.e. PMI scoring predicts human best word). $n \leq 2$ column is percent of time that human-selected best word is ranked 1st or 2nd in terms of PMI score, and so on. METADATA **was excluded because of insufficient responses.**

Note that the PMI score for the topic is simply the mean of the PMI scores for each word in the topic.

Figure 5 illustrates the best-word analysis using topics from NEWS and ABSTRACTS. In the five example NEWS topics and five example ABSTRACTS topics, the two words with the highest mean PMI are indicated by the boxes, and the human-selected best words are shown in bold. We see some matches between the high-scoring PMI words, and the human-selected best words. In the first NEWS topic, humans unanimously selected *space* as the word that best exemplified the topic, but highest PMI scores went to *earth* and *nasa*. In the last NEWS topic, the two highest PMI scores matched two out of three of the human-selected best words.

Table 4 shows the ability of the PMI score to predict the human-selected best word. The columns show the percent of time that the best word is ranked in the top-n words ranked by PMI score (so the first column is the percent of time the highest-scoring PMI word matches the human best word). We see that only approximately half the time, the human best word is one of the top three words ranked by PMI. This result is perhaps lower than expected. One possible reason is that words with a high PMI score tend not to correlate well with words chosen by humans as the best word to exemplify the topic. For example, in the topic *china government trade chinese country countries economic foreign political nation*, humans generally agreed that *china* was the best word to exemplify the topic. However, the words in this topic arranged by PMI order are *countries foreign* *economic political china trade nation government country chinese*. More general words such as *countries* have a better PMI with all other words than a more specific word such as *china*. Thus in this case, the highest scoring words by PMI do not match the best word chosen by users.

Also note that the variability in humans choosing the best word reduces the overall ability of PMI to match the best word. For the ABSTRACTS topics, when we restricted to the 16 topics where there was a consensus of at least 4 users all selecting the same best word, and when we took that as the best word, we observed a much higher accuracy of PMI predicting the best word. In this case, the best PMI word matched the human best word 10 times out of 16, or 60%, more than double the figure in Table 4.

6.1 Match with Suggested Label

Interpreting the meaning of individual topics is often challenging. If an appropriate label can be assigned to a topic, it can be easier for humans to interpret and understand the meaning of the topic. Moreover, short labels are essential and necessary in practice for using learned topics in user interfaces. We define a label as a single word or short phrase that concisely and completely conveys the semantic meaning of the topic. Research on automatic labeling of topics is limited, with the exception of [9], who used a probabilistic model to extract meaningful phrases as labels. They define a label as "understandable" if it is relevant, understandable for humans, has good coverage, and is discriminative across topics.

We investigate, for the task of automatic labeling, whether it is sufficient to use topic words as the main source for labeling, or whether additional terms and resources need to be included. For the selected topics from the collections in Table 1, we asked the human subjects to suggest a short topic label (of one to three words). They were instructed that the label does not necessarily have to contain any topic words. Example topics from each collection and their corresponding human labels are shown in Figure 6. General information about the labels assigned by the subjects can be found in Table 5. Topic labels tended to be two words long. On average, 58% of the label words were taken from the topic words, leaving 42% of the label words *not* from the topics. This statistic suggests that the top-10 topic words are reasonably good candidates for constructing part of the labels, but they are not in general sufficient for the entire label. There was a tendency for our human subjects to make use of their best word, with 38% of their suggested label words being their best word. The match with the top PMI words was smaller: 28% of the time the label words included one of the top-2 PMI words. Here, we have just started to scratch the surface of investigating labeling of topics, one of our key tasks for evaluating topic models for digital libraries. In future work we will more extensively explore this problem of automatically suggesting topic labels.

7. WORST TOPIC WORD

In our first round of user-evaluations on METADATA topics, in addition to asking users to score each topic and pick the single best word, we also asked users to identify one or more words that did not fit the topic. While this task was designed for later comparisons with PMI scoring, the variable quality of the METADATA topics made this a frustrating and confusing task for users. This task was discontinued in subsequent rounds of user evaluations on topics from the other collections.

Collection	No. Topics	No. Lbls	Avg. Lbl Size
BOOKS	200	1,673	1.4
NEWS	40	433	1.8
ABSTRACTS	56	396	1.9
METADATA	29	318	2.3
Avg.	—	—	1.8

Table 5: General statistics on the human-suggested labels. Label size is measured in words.

	Human		
Collection	Topic Word Usage	Best Word Usage	PMI (Top-2)
BOOKS	40	22	6
NEWS	64	54	64
ABSTRACTS	74	33	28
METADATA	54	42	14
Avg.	58	38	28

Table 6: Percentages of the usage of topic words and best topic words in human suggested labels. The two highest scoring words (Top-2) from PMI were also compared with the human suggested label words.

Nevertheless, we simulated the evaluation of "worst topic word" by using the word intrusion task introduced by [7]. In the word intrusion task, a random word in a given topic is replaced by a random word selected from another topic from the same topic model (where the replacement word is also from the top ten words from some topic). Instead of asking a human to predict the intruder word, we have our PMI model predict the intruder word. Or expectation is that the intruder word will have the lowest average PMI with other words in the topic, and thereby be a proxy for "worst topic word".

Figure 7 illustrates the intrusion task using topics from NEWS and ABSTRACTS. In the five example NEWS topics, the word with the lowest mean PMI (boxed) perfectly predicts the random intruder words (bolded). For ABSTRACTS, the PMI correctly predicts three out of five intruder words. In some cases the intruder word is obvious, for example the term *patient* in the first topic. However, unlike in the original work, we did not require that the intruder word have low probability in the current topic, so in some cases (e.g. *university* in the 2nd topic, or *humans* in the 8th topic), it is not immediately obvious which word is the intruder.

	Match with bottom-n PMI words					
Collection	$n = 1$	≤ 2	≤ 3	≤ 4	≤ 5	≤ 10
BOOKS	44	58	69	79	82	100
NEWS	70	85	87	89	92	100
ABSTRACTS	57	84	89	95	96	100
METADATA	27	55	70	82	88	100

Table 7: $n = 1$ is percent of time that intruder word is ranked last in terms of PMI score (i.e. PMI scoring predicts intruder word). $n \geq 2$ column is percent of time that intruder word is ranked in bottom two in terms of PMI score, and so on.

	Topic	Label
Books	king prince queen **royal** (11) court **crown** (1) palace princess majesty throne	monarchy (2), royalty (4), royal family, royal courts, royal
News	**space** (11) earth moon science scientist light **nasa** (1) mission planet mars	space exploration (4), space universe, space travel, space mission, outer space, automotive, space, space missions, space program
Abstracts	**hiv** (5) infection virus viral infected cns brain **aids** (1) replication macrophages	hiv neuropathology, aids (2), neuroaids (3), hiv (3)
Metadata	**holland** flander van **portrait** paul peter jan drawing ruben oil **amsterdam** (2) der **rembrandt** (4) rijn landscape	Dutch painters (2), portrait, Rembrandt's portrait, Dutch artists, Rembrandt, Dutch painting (2), Dutch art (2), Dutch artists and terms

Figure 6: Examples of topics and their suggested human labels. Human-selected best words are indicated in bold. The number in brackets is the number of people choosing the word or suggesting the label.

News	space earth moon science scientist light nasa $\boxed{\textbf{patient}}$ planet mars
	health drug $\boxed{\textbf{university}}$ medical doctor hospital care cancer treatment disease
	cell human animal scientist research gene researcher brain $\boxed{\textbf{motor}}$ science
	car ford vehicle model auto truck engine sport wheel $\boxed{\textbf{health}}$
	$\boxed{\textbf{market}}$ care insurance patient hospital medical cost medicare coverage doctor
Abstracts	children pediatric exercise physical childhood intervention $\boxed{\textbf{cellular}}$ developmental adults problems
	enzyme acid synthesis enzymes cholesterol lipid $\boxed{\textbf{epileptogenic}}$ fatty specific storage
	hiv **humans** virus viral infected cns $\boxed{\text{brain}}$ aids replication macrophages
	treatment clinical patients therapy efficacy trial phase $\boxed{\textbf{cells}}$ dose drug
	estrogen hormone sex $\boxed{\text{effects}}$ androgen hormones **receptors** female steroid estradiol

Figure 7: Illustration of intruder task. Bold word is intruder word (random word selected from a different topic), and boxed word is 10th (last) word ranked by PMI.

We see that the PMI model performs well at predicting the intruder word, particularly for News and Abstracts, where the intruder word has the lowest or second lowest mean PMI 85% and 84% of the time. The lower performance of Books and Metadata stem from several sources. The Books vocabulary contained no proper nouns, and therefore tended to produce lower PMI scores overall. Metadata topics were both noisier and more domain specific, so detecting intruders was possibly more difficult. There was a small positive correlation between high human topic score and high accuracy of PMI predicting the intruder word. Note that the intruder task is much easier than the best word task in Section 6.

In [7], the authors conclude that it would be useful to develop models that are a computational proxy for human judgements, which is what we have demonstrated with the PMI scoring model.

8. DISCUSSION

The PMI scoring model has high correlation with human scores for a wide variety of topics. Across four completely different genres and domains including books (Books), news articles (News), grant abstracts (Abstracts) and image metadata (Metadata), the PMI-human correlation exceeds the inter-rater correlation, so the PMI score is no worse than the scores from any one human. One surprising result is that the PMI scoring model can predict semantic coherence of topics, despite it being a purely distributional technique, and one that is not semantic- or concept-based such as WordNet.

Despite the fact that the PMI model achieved human performance levels, there was a relatively large range of PMI-human correlations across the four categories of Books, News, Abstracts and Metadata. The PMI scores for Metadata did not match human scores as well as the others. For Metadata topics, humans themselves were having difficulty with agreeing on scoring of the topics. For Abstracts, the PMI-human correlation was potentially lower due to possibly confusing instructions to evaluators who were asked to give a low score to topics that were coherent, but not related to a specific biomedical research area.

Observing and analyzing the patterns of outliers, where the PMI score was somewhat different to the mean human score, provided insight into the possible limitations (and potential improvements) to the PMI scoring model. For example, the PMI model gives a high score to the topic of Roman numerals *viii vii xii xiii xiv xvi xviii xix xvii* .... Despite the terms being obviously related, this is clearly a junk topic for human use. One potential improvement to the outliers which over-predict PMI at high human score, is smoothing of counts from the external data source. Smoothing of counts will attenuate over-estimates of PMI for the lower frequency terms that are typically found in topics belonging to this group of outliers. In future work we hope to refine and improve the PMI scoring of topics.

Automatic identification of junk topics is a highly useful capability, and is directly useful to our overarching goal of improving searching and browsing. It would be particularly interesting to use the PMI score, in conjunction with other techniques, to identify specific classes of junk topics, such as: mix of two distinct concepts; prose-type topics; topics with just one or two bad words; topics of boilerplate text; and topics that are a seemingly miscellaneous list of unrelated words.

The closest related work to our work is the work of [7]. Our use of their word intrusion task in Section 7 showed that the PMI score performs very well at that task. The other work that aims to directly rank topics, the topic significance ranking of [1], is still primarily an internal method and therefore may be limited in its wider application. We argue that the use of external data (such as Wikipedia, Google or MEDLINE) is key to the strength of the PMI method, and appears to mimic humans' internal sense of word-associations likely used in their evaluation of topic coherence.

9. CONCLUSIONS

This work addresses evaluating topics learned by topic models. In our large-scale user study, we ask humans to evaluate and score almost 500 individual topics for semantic coherence across a wide variety of genres and domains. We then present results showing that the pointwise mutual information (PMI) scoring model provides relatively good predictions of human scoring. Having an automated scoring model of learned topics is an important step in integrating topic models into digital libraries.

We observed relatively high correlations between the PMI score and the human score that generally exceeded the human inter-rater correlation. For topics learned from more general content such as those in our BOOKS or NEWS collections, we measured PMI-human correlations of $\rho \approx 0.8$. For our collection of biomedical grant abstracts (ABSTRACTS), we measured a correlation of $\rho \approx 0.6$. And for our most challenging collection of very short metadtata records, we we measured a correlation of $\rho \approx 0.5$. Our analysis of outliers explained patterns of disagreement between the PMI scores and the human scores.

Having an automatic scoring model of topics helps automate the process of learning topic models, and obtaining an initial assessment of useful and junk topics. Automatic scoring of topics is critical, because it provides an external evaluation framework — if consistent with human judgements — that could be used to guide the selection of: number of topics; different types of topic models; model hyperparameter selection; and preprocessing treatment to create the bag-of-words. The PMI model is highly flexible in that it can work with any type of topic model that can express a topic as a list of words. The use of external data is a strength in the model, as it further provides an external basis to evaluate topics. Furthermore, our PMI approach is higly scalable, and can work efficiently on any sized collection.

In addition, topic modeling in the context of digital libraries is highly compatible with Web 2.0 technologies. Web 2.0 opens up the possibility for collaborative labeling of topics, and collaborative evaluation of topic tags to items, and even learning topics of human contributed tags, the last of which is part of our ongoing research.

Topic modeling could potentially have a huge impact on improving search and discovery in digital libraries, by automatically learning and applying topic tags to individual resources in a highly scalable, consistent and economical fashion. Being able to automatically evaluate topics, and identify junk topics, is a key to the success of this enterprise. In our work we have both broadened the definition of topic model evaluation, and suggested specific sub-tasks to address evaluation, including an accurate scoring model. Ultimately we see this work as a building block towards the goal of realizing the potential of topic models in digital libraries.

Acknowledgement

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. This work has also been supported by Institute of Museum and Library Services grant LG-06-08-0057-08. We thank Miriam Leenders (NIH/NINDS) and Kat Hagedorn (UMich) for their help with conducting surveys.

References

[1] L. AlSumait, D. Barbará, J. Gentle, and C. Domeniconi. Topic significance ranking of LDA generative models. In *ECML/PKDD (1)*, pages 67–82, 2009.

[2] D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *ICML*, page 4, 2009.

[3] T. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In *CIKM*, pages 601–610, 2009.

[4] D. Blei and J. Lafferty. Dynamic topic models. In *ICML*, pages 113–120, 2006.

[5] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[6] W. Buntine and A. Jakulin. Applying discrete PCA in data analysis. In *UAI*, pages 59–66, Banff, Canada, 2004.

[7] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei. Reading tea leaves: How humans interpret topic models. In *NIPS*, pages 288–296, 2009.

[8] T. Griffiths and M. Steyvers. Finding scientific topics. In *PNAS*, volume 101, pages 5228–5235, 2004.

[9] Q. Mei, X. Shen, and C. Zhai. Automatic labeling of multinomial topic models. In *SIGKDD*, pages 490–499, 2007.

[10] D. Mimno and A. McCallum. Organizing the OCA: learning faceted subjects from a library of digital books. In *JCDL*, pages 376–385, 2007.

[11] D. Newman, T. Baldwin, L. Cavedon, S. Karimi, D. Martinez, and J. Zobel. Visualizing document collections and search results using topic mapping. *Journal of Web Semantics*, to appear.

[12] D. Newman, K. Hagedorn, C. Chemudugunta, and P. Smyth. Subject metadata enrichment using statistical topic models. In *JCDL*, pages 366–375, 2007.

[13] D. Newman, S. Karimi, and L. Cavedon. External evaluation of topic models. In *ADCS*, pages 11–18, 2009.

[14] D. Newman, J. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *NAACL HLT 2010*, Los Angeles, USA, to appear.

[15] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *JASA*, 101(476):1566–1581, 2006.

[16] H. Wallach, D. Mimno, and A. McCallum. Rethinking LDA: Why priors matter. In *NIPS*, pages 1973–1981, 2009.