Latent Topic Feedback for Information Retrieval

David Andrzejewski
Lawrence Livermore National Laboratory
Livermore, CA 94550
andrzejewski1@llnl.gov

David Buttler
Lawrence Livermore National Laboratory
Livermore, CA 94550
buttler1@llnl.gov

ABSTRACT

We consider the problem of a user navigating an unfamiliar corpus of text documents where document metadata is limited or unavailable, the domain is specialized, and the user base is small. These challenging conditions may hold, for example, within an organization such as a business or government agency. We propose to augment standard keyword search with user feedback on latent topics. These topics are automatically learned from the corpus in an unsupervised manner and presented alongside search results. User feedback is then used to reformulate the original query, resulting in improved information retrieval performance in our experiments.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*query formulation, relevance feedback*

General Terms

Algorithms, Experimentation

Keywords

latent topic models, user feedback

1. INTRODUCTION

This work addresses the problem of ad-hoc information retrieval and text corpus navigation under the following conditions. First, document metadata may be limited, unreliable, or nonexistent. Second, the domain of the text documents is specialized, using vocabulary ill-covered by general web documents or lexical resources such as WordNet [10]. Finally, while the text corpus itself may be large, the set of users accessing the corpus is not. This is an important problem because these conditions can preclude the use of effective information retrieval techniques such as faceted search or query log mining. These conditions are different from those encountered in general web or e-commerce search, but are realistic *within* organizations which are trying make sense of large quantities of text, such as private enterprises or government agencies.

A central problem in ad-hoc information retrieval is that users may not be able to formulate the "right" keyword combination in order to retrieve the most relevant documents. Techniques such as real-time query expansion [34] have been developed to directly attack this problem, but often rely upon a dataset of previously submitted queries, which may be sparse without a large user base. Another approach is to solicit alternative types of user input. Faceted document navigation [29] allows users to select documents based on different attributes (e.g., publication venues or hierarchical subject categories) and has emerged as a powerful complement to traditional keyword search. However, the standard assumption is that the facets are manually defined, and that facet values for each document are known.

Because of the challenging scenario we have defined, it is important to exploit all available data. Latent topic models such as Latent Dirichlet Allocation (LDA) [4] provide a means to take advantage of the statistical structure of the corpus itself. LDA assumes that observed documents have been generated by weighted mixtures of unobserved (latent) topics. These topics are learned from the documents and often correspond to meaningful semantic themes present in the corpus. LDA and its extensions have found interesting applications in fields such as natural language processing, computer vision, and social networks analysis [2].

The contribution of this work is a new method for obtaining and exploiting user feedback at the *latent topic* level. Our approach is to learn latent topics from the corpus and construct meaningful representations of these topics. At query time, we then decide *which* latent topics are potentially relevant and present the appropriate topic representations alongside keyword search results. When a user selects a latent topic, the vocabulary terms most strongly associated with that topic are then used to augment the original query. Our experiments with simulated user feedback show improved information retrieval performance. The presentation of relevant topics alongside search results also has the additional benefit of helping the user to understand corpus themes related to the original keyword query.

KDD'11, August 21–24, 2011, San Diego, California, USA.

Table 1: Example learned topic-word multinomials ϕ from three different datasets (see Table 5). For each topic ϕ_t, the five highest-probability words w are shown.

FT - Topic 1		WSJ - Topic 8		LA - Topic 94	
Word w	$P(w\|z)$	Word w	$P(w\|z)$	Word w	$P(w\|z)$
court	0.080	technology	0.094	gun	0.058
case	0.025	research	0.054	weapons	0.052
legal	0.024	high	0.025	assault	0.034
ruling	0.018	development	0.023	guns	0.029
appeal	0.018	cray	0.020	rifles	0.018

2. RELATED WORK

Our approach is partially motivated by the successes of faceted search [35]. Castanet [29] and related systems [9] aim to *automatically* construct facet hierarchies, but these techniques depend crucially on the existence of a rich lexical resource such as WordNet [10]. While specialized ontologies or controlled vocabularies have been constructed for some domains such as Gene Ontology (GO) [31] and Medical Subject Headings (MeSH) [5], the constraints of our setting prohibit us from assuming the existence of such a resource.

In light of this issue, topic models such as LDA have the advantage of relying upon corpus statistics alone. Indeed, previous analysis [23] of the digital library of the Open Content Alliance (OCA) directly posited the analogy between latent topics and faceted subjects, although specific mechanisms for exploiting this insight were not explored. The Rexa academic search engine[1] also displays relevant latent topics as tags for a given research article, allowing further investigation of the topics themselves. Another interesting topic modeling approach uses seed words to learn facet-oriented topics which can then be used to construct informative summaries [18].

LDA has previously been used in information retrieval for both document language model smoothing [33, 20] and query expansion [27]. These techniques both exploit the dimensionality reduction provided by LDA "behind the scenes" in order to improve performance, but do not leverage explicit user feedback in the way that our approach does. The approach we propose in this work can be viewed as complementary to these existing enhancements.

The BYU Topic Browser [11] provides an environment for rich explorations of learned LDA topics and how they relate to words and documents within a corpus. However, the tasks supported are more appropriate for advanced analysis by a relatively sophisticated user, as opposed to a general search setting.

3. OUR APPROACH

We propose to present automatically learned topics alongside keyword search results, allowing the user to provide feedback at the latent topic level. While it is well-known that we can learn latent topics with LDA, incorporating these topics into an information retrieval system requires us to address several questions. First, how should these topics be presented? Previous user studies [29] have found that users can become frustrated by raw LDA output. Second, which topics should be presented for a given query? To avoid overwhelming the user, we clearly cannot present all latent topics (potentially hundreds or greater) for every query. Furthermore, not all learned topics truly correspond to meaningful semantic concepts, and the presence of these incoherent topics will not be appreciated by users either. Third, how can we incorporate user latent topic feedback into search results? Ideally, the mechanism used should be simple and easy to integrate with existing search technologies. Finally, can this type of feedback improve information retrieval performance, as measured by standard metrics?

We now describe our approach, beginning with a brief review of latent topic modeling concepts and moving on to address the above questions. All examples shown are actual learned topics from the experimental datasets described in Table 5. In Section 4, experimental results demonstrate that our approach can indeed achieve performance gains.

3.1 Latent Dirichlet Allocation (LDA)

In LDA [4], it is assumed that observed words in each document are generated by a document-specific mixture of corpus-wide latent topics. We define our corpus of length N with the flat word vector $\mathbf{w} = w_1 \dots w_N$. At corpus position i, the element d_i in $\mathbf{d} = d_1 \dots d_N$ designates the document containing observed word w_i. Similarly, the vector $\mathbf{z} = z_1 \dots z_N$ defines the hidden topic assignments of each observed word. The number of latent topics is fixed to some T, and each topic $t = 1 \dots T$ is associated with a topic-word multinomial ϕ_t over the W-word vocabulary. Each ϕ multinomial is generated by a conjugate Dirichlet prior with parameter β. Each document $j = 1 \dots D$ is associated with a multinomial θ_j over T topics, which is also generated by a conjugate Dirichlet prior with parameter α. The full generative model is then given by

$$P(\mathbf{w}, \mathbf{z}, \phi, \theta \mid \alpha, \beta, \mathbf{d}) \propto \left(\prod_t^T p(\phi_t|\beta)\right)\left(\prod_j^D p(\theta_j|\alpha)\right)\left(\prod_i^N \phi_{z_i}(w_i)\theta_{d_i}(z_i)\right),$$

where $\phi_{z_i}(w_i)$ is the w_i-th element in vector ϕ_{z_i}, and $\theta_{d_i}(z_i)$ is the z_i-th element in vector θ_{d_i}. Given an observed corpus $(\mathbf{w}, \mathbf{d})$ and model hyperparameters (α, β), the typical modeling goal is to infer the latent variables $(\mathbf{z}, \phi, \theta)$.

While exact LDA inference is intractable, a variety of approximate schemes have been developed [24, 4, 30]. In this work, we use Markov Chain Monte Carlo (MCMC) inference, specifically collapsed Gibbs sampling [12]. This approach iteratively re-samples a new value for each latent topic assignment z_i, conditioned on the current values of all

[1] http://rexa.info/

Table 2: Features used to determine "best topic word" labels for each topic. The topic-word posterior $P(z = t|w)$ is computed using Bayes Rule and a uniform prior over topics.

Description	Score
Word probability	$f_1(w) = P(w|z = t)$
Topic posterior	$f_2(w) = P(z = t|w)$
PMI	$f_3(w) = \sum_{w' \in W_t \setminus w} PMI(w, w')$
Conditional 1	$f_4(w) = \sum_{w' \in W_t \setminus w} P(w|w')$
Conditional 2	$f_5(w) = \sum_{w' \in W_t \setminus w} P(w'|w)$

other z values. After running this chain for a fixed number of iterations, we estimate the topic-word multinomials ϕ and the document-topic mixture weights θ from the final $\mathbf{z}$ sample, using the means of their posteriors given by

$$\phi_t(w) \propto n_{tw} + \beta$$
$$\theta_j(t) \propto n_{jt} + \alpha$$

where n_{tw} is the number of times word w is assigned to topic t, and n_{jt} is the number of times topic t is used in document j, with both counts being taken with respect to the final sample $\mathbf{z}$. The topic-word multinomials ϕ_t for each topic t are our learned topics; each document-topic multinomial θ_d represents the prevalence of topics within document d.

3.2 Topic representation

Typically, each learned topic-word multinomial ϕ_t is presented as a "Top N" list of the most probable words for that topic, as shown for three example learned topics in Table 1. We define the k-argmax operator to yield the k arguments which result in the k largest values for the given function. We use this operator to define the ten most probable words for topic t as W_t, given by the following expression with $k = 10$

$$W_t = \text{k-argmax}_w \, \phi_t(w)$$

We apply techniques from recent topic modeling research to improve on this basic representation. Our post-processing of the learned topics has three components: label generation, n-gram identification, and capitalization recovery.

For topic labeling, we assume the availability of a *reference corpus* containing themes similar to the target retrieval corpus. Since only raw text is required, this should be considerably easier to obtain than a full ontology, even for specialized domains. For example, a user exploring a corpus related infectious disease outbreaks could obtain a suitable reference corpus by crawling web resources from the United States Centers for Disease Control and Prevention. Since our experiments use general newswire corpora for evaluation, we use Wikipedia[2] as our reference corpus.

We label each topic using a simplified variant of the "Best Topic Word" [15] method. For a given topic t, this method selects a single word label from the top ten most probable words W_t, using features designed to test how representative each word is of the topic as a whole. We deviate slightly from Lau et al. to avoid relying upon WordNet, selecting the label word by majority vote among five features shown in Table 2 where each feature f_i casts its vote for the highest scoring word and ties are broken arbitrarily. Several of these features are computed from co-occurrence frequencies among words in W_t, counted within ten-word sliding windows taken over the reference corpus. Specifically, we compute the pointwise mutual information (PMI) and conditional occurrence probabilities between each pair of words (w, w') as

$$PMI(w, w') = \log \frac{P(w, w')}{P(w)P(w')}$$
$$P(w|w') = \frac{P(w, w')}{P(w')}$$

where $P(w, w')$ is the probability of jointly observing w and w' within a given sliding window, and $P(w)$ is the probability of observing w within a sliding window. Several example labels can be seen in the "label" column of Table 3.

We then identify statistically significant bigrams and trigrams (e.g., "White House", "President Barack Obama") for each topic using an approach based on the Turbo Topics [3] algorithm. This approach considers adjacent word pairs (w_i, w_{i+1}) occurring in the same document and assigned to the same topic (i.e., $d_i = d_{i+1}$ and $z_i = z_{i+1}$) and identifies pairs which occur much more often than we would expect by chance alone, proceeding similarly for trigrams. For each topic, we show the topic label along with the most significant trigram, the two most significant bigrams and the four most probable unigrams. Example latent topic representations are shown in Table 3.

Finally, we restore capitalization to the topic n-grams before presenting them to the user. As a pre-processing step, all text is converted to lower-case before doing LDA inference. However, the information conveyed by capitalization can ease user interpretation of topics (e.g., by making proper names obvious). For each n-gram, we simply count all occurrences of each possible capitalization occurring in the original documents, and present the most frequent version to the user.

Table 3: Topic representations for example high-PMI (coherent) and low-PMI (incoherent) topics.

PMI	Label	n-grams
3.09	jurors	Deputy Dist Atty cross examination, closing arguments trial, jury, case, testified
1.68	Petroleum	state oil company North Sea, natural gas production, exploration, field, energy
-0.09	things	(no trigrams found) pretty good, years ago ve, ll, time, don
-0.03	sales	(no trigrams found) year earlier, Feb Feb December, March, month, rose

[2]http://www.wikipedia.org

3.3 Topic selection

It will typically be necessary to learn at least hundreds of latent topics in order to get suitably fine-grained topics for user feedback. This makes it impractical to present all topics to the user after every query; we therefore must decide which topics to present.

We use the idea of pseudo-relevance feedback [6] by assuming that the top two documents returned by the original query q, which we call D_q, are relevant. For each of these documents, we consider the top $k = 2$ topics as determined by the topic weights θ to be *enriched* topics for the user query. This constitutes a natural set of candidates for latent topic feedback, and can be defined as

$$E = \bigcup_{d \in D_q} \text{k-argmax}_t \, \theta_d(t).$$

However, we also show the user topics that are *related* to the enriched topic set E, but which may themselves not be present in the highly ranked documents. We identify related topics by looking for topics highly likely to co-occur with the enriched topics E, using the $T \times T$ topic covariance matrix Σ of the estimated $D \times T$ document-topic θ matrix. Letting $\Sigma(t_1, t_2)$ be the covariance between $P(z = t_1|d)$ and $P(z = t_2|d)$ computed over all documents $d = 1, \ldots, D$, we take the $k = 2$ topics with the highest covariance with each of our enriched topics in E. We define this *related* topic set as

$$R = \bigcup_{t \in E} \text{k-argmax}_{t' \notin E} \, \Sigma(t, t').$$

The candidate topics for feedback are the union of the enriched and related topics $E \cup R$, but we perform a final filter before presenting these topics to the user.

One hazard of presenting automatically discovered latent topics to the user is the threat of incoherent "junk" topics which do not seem to have a single clear theme. We filter out these topics using a recently developed topic evaluation method [26, 25] which has been shown to predict human topic quality judgments at nearly the inter-annotator agreement rate. Similar to the topic labeling technique, this method uses PMI values computed over a reference corpus (again, we use Wikipedia), except that we now apply these scores to the topics themselves. We compute the PMI score of a topic t as the average PMI between all pairs of words within the top $k = 10$ most probable words W_t

$$PMI(t) = \frac{1}{k(k-1)} \sum_{(w,w') \in W_t} PMI(w, w').$$

Table 3 shows example high-PMI (coherent) and low-PMI (incoherent) latent topics.

We can use these PMI values to avoid confusing users with incoherent topics. Letting PMI_{25} be the 25^{th} percentile PMI score among all learned topics, we define our set of "dropped" topics D as

$$D = \{t | t \in E \cup R \text{ and } PMI(t) < PMI_{25}\}.$$

We present the topics in $\{E \cup R\} \setminus D$ to the user alongside the returned documents for the original keywords query q. Note that the union operations and final filtering mean that the number of topics actually presented to the user may vary from query to query. Since we consider the top two topics within the top two documents, along with each of their top two related topics, we will present a maximum of $(2 \times 2) + (2 \times 2 \times 2) = 12$ topics, minus set overlaps and PMI-filtered topics.

3.4 Query expansion

If the user selects a topic as relevant, we reformulate the query by combining the top ten most probable words W_t for that topic with the original query q. To preserve the intent of the original query, we use the Indri [22] `#weight()` operator to form a weighted combination of the original query keywords and the highly probable latent topic words. The weight parameter $\gamma \in [0, 1]$ controls the trade-off between the original query keywords and the latent topic words. A larger γ value places more weight on the new latent topic words, while setting $\gamma = 0$ is equivalent to the original keyword query.

Each of the N_q words in the original query is given weight $(1 - \gamma)/N_q$ and each new topic t word w is given weight $\gamma * \tilde{\phi}_t(w)$, where $\tilde{\phi}$ is the re-normalized topic-word probability

$$\tilde{\phi}_t(w) = \frac{\phi_t(w)}{\sum_{w' \in W_t} \phi_t(w')}.$$

While our implementation uses the Indri query language, it would be straightforward to achieve similar results in other information retrieval systems and frameworks (e.g., by using term boosting in Apache Lucene[3]).

3.5 Example

We now walk through an example query for a corpus of news articles from the Financial Times (FT). The query is "euro opposition", and it targets documents discussing opposition to the introduction of the single European currency. The corpus, query, and relevance judgments used here are drawn from our experimental dataset which will be used in Section 4. The number of topics used is $T = 500$.

The *enriched* topics E shown in Table 5a consist of three distinct topics: two topics related to the euro debate within the United Kingdom and Denmark, and a confusing topic vaguely centered around "business" which is *dropped* by our PMI filtering. Within this topic, the interesting trigram "PERSONAL FILE Born" arises from brief biographies sometimes found at the bottom of the articles.

High θ covariance with topics in E is then used to identify the five *related* topics R shown in Table 5b, which deal with various aspects of business and politics. However the appearance of "economic monetary union" and "Europe" in the topic 79 representation appear highly related to the euro currency union, and indeed selecting this topic as feedback improves retrieval results. Selecting topic 79 as user feedback and setting the feedback weight $\gamma = 0.25$, our approach produces an expanded query containing the most probable words from topic 79

```
#weight(0.375 euro, 0.375 opposition,
        0.031 European, ..., 0.015 Emu).
```

[3]http://lucene.apache.org/

Table 4: A detailed example of our approach for the query "euro opposition" on the Financial Times (FT) corpus. The strikethrough topic 466 is not presented to the user due to low PMI coherence score. The bolded topic 79 results in improved information retrieval performance versus the baseline query: NDCG15 increases 0.22, NDCG increases 0.07, and MAP increases 0.02. The prominent term "Emu" appears to be an alternate form of the acronym "EMU" commonly used in Financial Times articles.

Enriched topic	Terms
196 (debate)	Tory Euro sceptics social chapter, Liberal Democrat mps, Labour, bill, Commons
404 (ratification)	ratification Maastricht treaty Poul Schluter, Poul Rasmussen Danish, vote, Denmark, ec
~~466 (business)~~	PERSONAL FILE Born years ago, past years man, time, job, career

(a) Enriched topics E.

Related topic	Terms
79 (Emu)	economic monetary union Maastricht treaty, member states European, Europe, Community, Emu
377 (George)	President George Bush, White House Mr Clinton, administration Democratic, Republican, Washington
115 (powers)	de regulation bill Sunday trading, Queen Speech law, legislation, government, act
446 (years)	chairman chief executive managing director, finance director Sir, board, group, company
431 (cabinet)	Mr John Major prime minister, Mr Major party, tory, government, Conservative

(b) Related topics R.

Using ground truth document relevance judgments, we can see that documents returned by this expanded query have superior performance on standard information retrieval measures as described in the caption of Table 4. Figure 1 shows the receiver operating characteristic (ROC) curves for the baseline query (dotted) and the expanded topic 79 query (solid). Points on the ROC curve correspond to the true positive rates (TPR) and false positive rates (FPR) for sets of documents returned at different ranking thresholds. Here we consider the true positive (TP) set to be the union of relevant documents found within the top 500 documents returned by both queries. This plot visually depicts a clear improvement in the ranking of relevant documents. An additional benefit is that users are given the opportunity to see and explore different aspects of "euro opposition" such as the political dimension with respect to the United Kingdom.

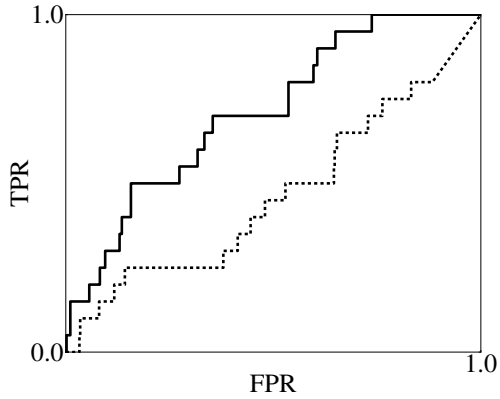


Figure 1: ROC curve of baseline (dashed) versus topic 79 feedback (solid) for the example query "euro opposition".

4. EXPERIMENTS

To our knowledge there has been no attempt to use latent topics as a user feedback mechanism in the way we have described. To determine whether our approach could be genuinely useful in practice, we must answer several questions. First, can query expansion with latent topic feedback improve the results of actual queries? While previous work has found that latent topics align well with existing document subject categories [23], it may be that these categories are more "topically coherent" than the relevant result sets for ad-hoc queries, and therefore more suitable for topic modeling. Second, assuming that for a given query there exists some latent topic which would improve retrieval results, will the topic selection approach described in Section 3.3 present it to the user? Finally, there is a third question which we do not address in this work: if presented with a helpful topic, will a user actually select it? For the following experiments we make the simplifying assumption that the user will always select the most helpful topic (with respect to the information retrieval measure of interest) among those presented. If no topic feedback will improve the set of returned documents, we assume the user will not provide topic feedback.

4.1 Experiment setup

While the ultimate goal of this work is to improve search and navigation under the specialized conditions described in Section 1, we evaluate our approach by conducting information retrieval experiments on several benchmark datasets from the Text REtrieval Conference (TREC) [32], using Wikipedia as a reference corpus. Each datasets consists of a corpus of text documents, a set of queries, and relevance judgments for each query. For each query, the individual words in the the title field are used as the baseline keyword query (e.g., "Industrial Espionage" is broken up into "Industrial", "Espionage"). Table 5 shows dataset details.

For each corpus, we first apply the LDA model to learn a set of latent topics, using the MALLET topic modeling toolkit [21]. We pre-process documents by downcasing, removing numbers and punctuation, applying a standard stopword list, and finally filtering out rarely occurring terms to yield vocabulary sizes of between 10,000 and 20,000 terms. We run parallelized collapsed Gibbs inference for 1,000 samples, re-estimating the document-topic hyperparameter α

Table 5: TREC datasets and queries (known within TREC as "topics") used in experimental evaluations. The number of documents D is given in thousands and Q denotes the number of queries.

Corpus	Abbrev	D	Q	TREC topics
Associated Press	AP	243	100	51-150
Financial Times	FT	210	200	251-450
Los Angeles Times	LA	128	150	301-450
Wall Street Journal	WSJ	173	100	51-100
				151-200
Federal Register	FR	37	150	301-450
Foreign Broadcast Information Service	FBIS	127	150	301-450

every 25 samples. We learn $T = 500$ topics for each corpus in our experimental dataset, except $T = 250$ for the significantly smaller Federal Register (FR) corpus.

For all queries, we use the Galago [8] information retrieval system with default settings to retrieve 500 documents. Galago uses a query language and retrieval model based on Indri [22]. For the topic-expanded queries we set $\gamma = 0.25$, based on trial-and-error experimentation on held-aside preliminary development datasets.

4.2 Results

We calculate improvement over the baseline query with respect to three information retrieval measures [8]: mean average precision (MAP), normalized discounted cumulative gain (NDCG), and NDCG calculated with the first 15 results only (NDCG15). These quantitative results are shown in Table 6, along with the average number of feedback candidate topics shown to the user by our topic selection technique (fewer than eight topics per query).

We now return to the experimental questions we had set out to answer. These results demonstrate that latent topic feedback can indeed improve information retrieval results. Across evaluation measures, the results of approximately 40% of queries can be improved by latent topic feedback. However, these gains are irrelevant if we cannot identify potentially helpful topics and present them to the user. Again across measures, we see that our topic selection approach is able present a helpful topic for more than 40% of the queries for which there exists at least one helpful topic. Doing the rough arithmetic, this means that for about 16% of the queries in our experiment the user would be presented with at least one latent topic which would improve the relevance of the returned documents. Furthermore, we stress that even for the "missed" queries where presented topics do not provide quantitative relevance improvement, the corpus theme information conveyed may still be beneficial.

To give a better feel for the nature of these results, Figure 2 shows six queries along with helpful topics which were selected for presentation by our approach. In all cases, the connection between the topic and the query is fairly clear, resulting in gains across retrieval performance measures and visible improvement on ROC curves.

4.3 Analysis

First, we observe that for most queries (roughly 60%), there did not exist a single latent topic for which feedback would enhance information retrieval results. From manual inspection, this can occur because either no learned topic is well-aligned with the relevant documents, or because the results of the original query are good and difficult to improve upon.

Second, for queries where there exists one or more topics which *would* improve results, roughly 60% of the time our topic selection approach fails to select them. Minor variations on our topic selection method (i.e., showing more topics) did not correct this – many of the "missed" topics are not even close to making the cutoff. Manual investigations reveal that, interestingly, these topics often appear to be helpful *because* they are somewhat "distant" from the original query and the top few baseline documents returned. Attempts to predict topic feedback gain using linear or logistic regression and features such as $P(\text{query}|\phi_t)$ were unsuccessful, although more sophisticated approaches or richer features could possibly be applied.

It is also instructive to further examine the impact of two key aspects of our topic selection procedure: the inclusion of related topics and the exclusion of incoherent topics. For simplicity we will discuss NDCG15 measurements, but similar results hold for MAP and NDCG. Our selection approach recovers helpful topics for 133 out of 850 queries (15.6%) while presenting an average of 7.76 topics to the user for each query.

If we do *not* use PMI to filter out topics suspected of being incoherent, the number of topics shown per query rises to 9.79, but the number of queries for which helpful topics are presented only increases to 143 out of 850 (16.8%). The presence of incoherent topics may also impose cognitive burdens on the user, and it is uncertain whether users would be able to successfully identify incoherent topics for feedback.

If we were to omit the related topics R, it would decrease the average number of topics shown to 2.70, but it would decrease substantially the number of queries for which a helpful topic is presented, down to 93 out of 850 (10.9%). Also, we note that the presentation of related topics is potentially useful for exploratory corpus search, giving the user information about corpus themes "adjacent" to the topics present in returned documents.

Taken together, these findings suggest that our topic selection procedure is reasonable. The inclusion of related topics considerably increases the number of queries for which we present helpful topics while presenting novel and possibly interesting corpus themes. The filtering of suspect low-PMI topics does not discard many helpful topics, and should spare users the ordeal of interpreting ill-defined topics.

5. DISCUSSION

In this work we have developed a novel technique for improving text corpus search and navigation in difficult settings where we do not have access to metadata, rich lexical resources, or large user populations. This is an important problem because these conditions make information retrieval more difficult, and are applicable within organizations that have large quantities of *internal* text documents which they wish to explore, analyze, and exploit.

Table 6: Improvement from simulated latent topic feedback calculated only over queries where feedback improves performance. The "avg shown" column indicates the average number of topics actually shown to the user as a result of the topic selection procedure described in Section 3.3. For each query and evaluation measure, the "imprv" column shows the number of queries for which there *exists* at least one latent topic which improves performance, "found" shows the number of queries for which a helpful topic is *actually presented* to the user by our selection scheme, and "avg gain" shows the mean improvement when a helpful topic is presented to the user.

			NCDG15			NCDG			MAP		
Corpus	Q	avg shown	imprv	found	avg gain	imprv	found	avg gain	imprv	found	avg gain
AP	100	7.79	32	16	0.165	32	21	0.093	31	20	0.037
FT	200	7.47	97	43	0.238	138	80	0.134	137	72	0.041
LA	150	8.65	79	27	0.090	81	27	0.070	82	29	0.027
WSJ	100	7.73	29	16	0.205	30	18	0.050	29	18	0.026
FR	150	7.22	26	10	0.131	39	13	0.034	39	11	0.024
FBIS	150	7.78	62	21	0.163	64	25	0.037	67	29	0.024

To enhance search and exploration capabilities in this scenario, we have developed an approach that gives users the ability to provide feedback at the latent topic level. We leverage recent advances in latent topic modeling in order to construct meaningful representations of latent topics while filtering out incoherent "junk topics". We propose a mechanism for deciding on a manageably small set of topics to present to the user, as well as a method for constructing expanded queries based on user topic feedback. Quantitative results on benchmark TREC datasets show that this technique can result in major improvements for a non-trivial proportion of queries. Furthermore, the presentation of enriched and related topics alongside search results can help to deliver insights about corpus themes, which may be beneficial for knowledge discovery as well.

One potential obstacle to this approach is the scalability bottleneck presented by LDA topic inference. However, two factors act to ameliorate these concerns. First, topics can be inferred "offline" in advance; we do not need to do any expensive inference at query-time. Second, there have been significant recent advances along multiple fronts in scalable LDA inference. A distributed system developed at Yahoo! is reported to process 42,000 documents per hour [28]. Alternatively, an online inference algorithm for LDA [13] promises both improved scalability and a principled means of updating topics to reflect new documents. In practice, a hybrid system could update topics in an online fashion as documents are received, periodically performing distributed batch inference to refresh the learned topics.

6. FUTURE WORK

There are several promising directions in which to extend this approach. Two obvious areas for improvement are increasing the proportion of queries for which a helpful topic exists and improving the selection method for presenting helpful topics to the user.

It may be possible to improve the alignment between learned topics and user queries by the use of more sophisticated topic models such as the Pachinko Allocation Model (PAM) [16]. While these models were *not* found to be helpful for document smoothing [36], rich hierarchical topics may be beneficial when combined with the *explicit* user feedback present in our approach. Our approach could also exploit prior information such as predefined concepts by using topic model variants which can incorporate domain knowledge [7, 1].

However, learning finer-grained topics can only increase the importance of carefully choosing which topics to show the user. Here it may be instructive to consider the large body of research on "learning to rank" [19], as well as recent work in facet selection [17, 14].

The query expansion mechanism is another potential target for extension. If our underlying information retrieval system supports phrase search terms (e.g., "White House"), it may be helpful to directly use discovered n-grams as well.

Further work could also compare the use of topics for explicit feedback in this work versus the implicit use of topics to improve document language models in prior work [33]. It may be that the two techniques could be combined profitably, with some topics being more suitable for explicit feedback while others are better used for smoothing.

Finally, another important step is to validate our user model assumptions. One approach may be to directly evaluate information retrieval performance using actual user feedback, for example via Amazon Mechanical Turk [37]. It may also be interesting to explore the relationship between topic presentation (e.g., topic labeling strategies, whether to display n-grams) and user behavior.

7. ACKNOWLEDGMENTS

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 (LLNL-CONF-471258).

8. REFERENCES

[1] D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *ICML*, pages 25–32. Omnipress, 2009.

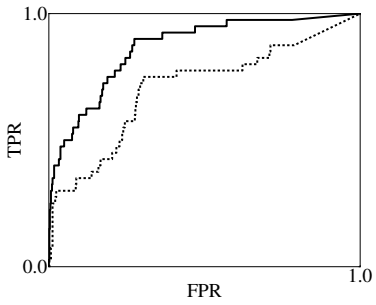
[2] D. Blei, L. Carin, and D. Dunson. Probabilistic topic models. *Signal Processing Magazine, IEEE*, 27(6):55 –65, 2010.

[3] D. Blei and J. Lafferty. Visualizing topics with multi-word expressions. Technical report, 2009. `arXiv:0907.1013v1 [stat.ML]`.

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

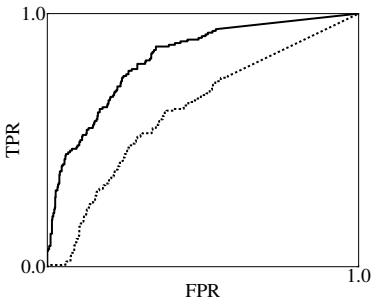
FBIS query 426 "law enforcement dogs"
Topic 321 (heroin) seized kg cocaine drug traffickers, kg heroin police, arrested, drugs, marijuana

NDCG15	NDCG	MAP
+0.299	+0.065	+0.046



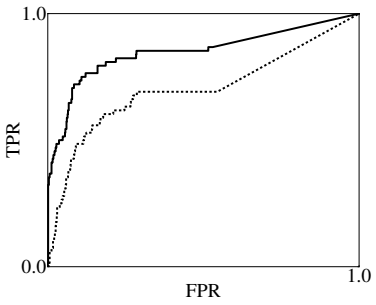
FBIS query 450 "King Hussein, peace"
Topic 293 (Amman) Majesty King Husayn al Aqabah, peace process Jordan, Jordanian, Amman, Arab

NDCG15	NDCG	MAP
+0.708	+0.175	+0.171



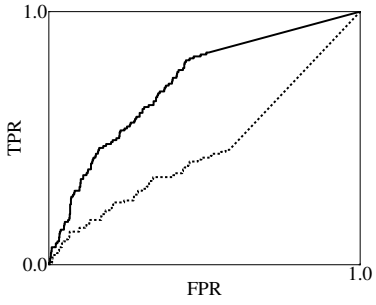
WSJ query 86 "bank failures"
Topic 444 (FDIC) Federal Deposit Insurance William Seidman, Insurance Corp banks, bank, FDIC, banking

NDCG15	NDCG	MAP
+0.602	+0.121	+0.110



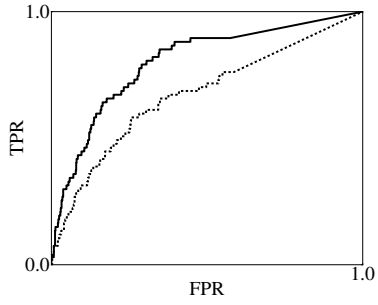
AP query 127 "U.S.-U.S.S.R. Arms Control Agreements"
Topic 232 (missile) Strategic Defense Initiative United States, arms control treaty, nuclear, missiles, range

NDCG15	NDCG	MAP
+0.296	+0.209	+0.105



AP query 135 "Possible Contributions of Gene Mapping to Medicine"
Topic 325 (called) British journal Nature immune system, genetically engineered cells, research, researchers, scientists

NDCG15	NDCG	MAP
+0.147	+0.040	+0.019



AP query 113 "New Space Satellite Applications"
Topic 237 (communications) European Space Agency Air Force, Cape Canaveral satellite, launch, rocket, satellites

NDCG15	NDCG	MAP
+0.237	+0.033	+0.007

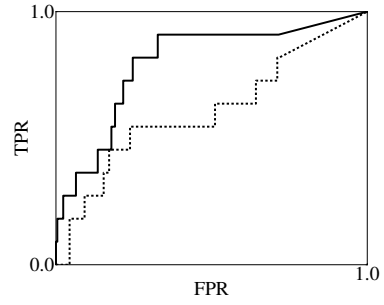


Figure 2: Six example queries with helpful topics and ROC curves. For each ROC curve, the set of true positive (TP) relevant documents is considered to be the union of the relevant documents discovered (i.e., ranked within the top 500) by the baseline query (dashed line) and the expanded query that incorporates latent topic feedback (solid line).

[5] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl 1):D267–D270, 2004.

[6] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using SMART: TREC 3. In *TREC*, pages 69–80. NIST, 1994.

[7] C. Chemudugunta, A. Holloway, P. Smyth, and M. Steyvers. Modeling documents by combining semantic concepts with unsupervised statistical learning. In *ISWC*, pages 229–244. Springer, 2008.

[8] B. Croft, D. Metzler, and T. Strohman. *Search*

Engines: Information Retrieval in Practice. Addison-Wesley, 2010.

[9] W. Dakka, P. G. Ipeirotis, and K. R. Wood. Automatic construction of multifaceted browsing interfaces. In *CIKM*, pages 768–775. ACM, 2005.

[10] C. Fellbaum. *WordNet : an Electronic Lexical Database*. MIT Press, 1998.

[11] M. J. Gardner, J. Lutes, J. Lund, J. Hansen, D. Walker, E. Ringger, and K. Seppi. The topic browser: An interactive tool for browsing topic models. In *NIPS Workshop on Challenges of Data Visualization*. MIT Press, 2010.

[12] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(Suppl 1):5228–5235, 2004.

[13] M. Hoffman, D. Blei, and F. Bach. Online learning for latent Dirichlet allocation. In *NIPS*, pages 856–864. MIT Press, 2010.

[14] J. Koren, Y. Zhang, and X. Liu. Personalized interactive faceted search. In *WWW*, pages 477–486, New York, NY, USA, 2008. ACM.

[15] J. H. Lau, D. Newman, S. Karimi, and T. Baldwin. Best topic word selection for topic labelling. In *Coling 2010: Posters*, pages 605–613. Coling 2010 Organizing Committee, 2010.

[16] W. Li and A. McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In *ICML*, pages 577–584. ACM, 2006.

[17] S. Liberman and R. Lempel. Approximately optimal facet selection. In *(submission)*, 2011.

[18] X. Ling, Q. Mei, C. Zhai, and B. Schatz. Mining multi-faceted overviews of arbitrary topics in a text collection. In *KDD*, pages 497–505. ACM, 2008.

[19] T.-Y. Liu. Learning to rank for information retrieval. In *SIGIR Tutorials*, page 904, 2010.

[20] Y. Lu, Q. Mei, and C. Zhai. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval*, pages 1–26, 2010.

[21] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[22] D. Metzler and W. B. Croft. Combining the language model and inference network approaches to retrieval. *Inf. Process. Manage.*, 40:735–750, 2004.

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

[24] T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *UAI*, pages 352–359. Morgan Kaufmann, 2002.

[25] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *HLT-NAACL*, pages 100–108, Morristown, NJ, USA, 2010. ACL.

[26] D. Newman, Y. Noh, E. Talley, S. Karimi, and T. Baldwin. Evaluating topic models for digital libraries. In *JCDL*, pages 215–224. ACM, 2010.

[27] L. A. Park and K. Ramamohanarao. The sensitivity of latent Dirichlet allocation for information retrieval. In *ECML PKDD*, pages 176–188. Springer-Verlag, 2009.

[28] A. Smola and S. Narayanamurthy. An architecture for parallel topic models. *Proc. VLDB Endow.*, 3:703–710, 2010.

[29] E. Stoica, M. Hearst, and M. Richardson. Automating creation of hierarchical faceted metadata structures. In *HLT-NAACL*, pages 244–251, Rochester, New York, April 2007. ACL.

[30] Y. W. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *In NIPS*. MIT Press, 2007.

[31] The Gene Ontology Consortium. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[32] E. M. Voorhees and D. K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[33] X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *SIGIR*, pages 178–185. ACM, 2006.

[34] R. W. White and G. Marchionini. Examining the effectiveness of real-time query expansion. *Inf. Process. Manage.*, 43:685–704, 2007.

[35] K.-P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. In *CHI*, pages 401–408. ACM, 2003.

[36] X. Yi and J. Allan. A comparative study of utilizing topic models for information retrieval. In *ECIR*, pages 29–41. Springer-Verlag, 2009.

[37] L. Zhang and Y. Zhang. Interactive retrieval based on faceted feedback. In *SIGIR*, pages 363–370. ACM, 2010.