Organizing the OCA: Learning Faceted Subjects from a Library of Digital Books

David Mimno, Andrew McCallum
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA
{mimno,mccallum}@cs.umass.edu

ABSTRACT

Large scale library digitization projects such as the Open Content Alliance are producing vast quantities of text, but little has been done to organize this data. Subject headings inherited from card catalogs are useful but limited, while full-text indexing is most appropriate for readers who already know exactly what they want. Statistical topic models provide a complementary function. These models can identify semantically coherent "topics" that are easily recognizable and meaningful to humans, but they have been too computationally intensive to run on library-scale corpora. This paper presents DCM-LDA, a topic model based on Dirichlet Compound Multinomial distributions. This model is simultaneously better able to represent observed properties of text and more scalable to extremely large text collections. We train individual topic models for each book based on the cooccurrence of words within pages. We then cluster topics across books. The resulting topical clusters can be interpreted as subject facets, allowing readers to browse the topics of a collection quickly, find relevant books using topically expanded keyword searches, and explore topical relationships between books. We demonstrate this method finding topics on a corpus of 1.49 billion words from 42,000 books in less than 20 hours, and it easily could scale well beyond this.

Categories and Subject Desriptors: H.3.7 Information Systems : Digital Libraries **General Terms:** Algorithms.

Keywords: Topic models, classification.

1. INTRODUCTION

The past two years has seen the creation of large-scale library digitization projects such as Google Books [9] and the Open Content Alliance (OCA) [17]. At the same time, methods for dividing large document collections into semantically coherent "topics," usually based on Latent Dirichlet Allocation (LDA) [1] or more generally Discrete PCA [2], have gathered substantial attention in the Machine Learning community. These developments are complementary. The digitization projects provide vast amounts of text, but little analysis, while topic modeling provides a method for approaching large, unstructured text corpora. In this paper, we apply a statistical topic model to a portion of the OCA corpus.

A topic model takes as input a collection of short text documents, such as book pages. It outputs a preset number of "topics", which are probability distributions over the words in the collection. Topics are essentially determined by which words occur together on the same page. The most likely words for each topic can then be used to provide human-interpretable keywords for the topic. Examples of the most likely words for topics learned from three books are shown in Figure 1.

The American revolution, 1763-1783; being the chapters relating to America from the author's History of England in the eighteenth century, by William Edward Hartpole Lecky

- act, stamp, colonies, America, parliament, colonial, repeal
- power, sea, war, England, land, France, naval, fleets, navy
- letters, Franklin, person, agent, written, papers
- army, men, war, officers, troops, Washington, states

Mrs. Allen's cook book, by Ida C. Bailey Allen

- pie, pastry, crust, plate, meringue, filling, bake, line, pies
- add, flour, milk, sugar, bake, beat, baking, cupful, salt
- dressing, lettuce, salad, french, mayonnaise, celery, oil
- eggs, egg, omelet, scrambled, hard, boiled, omelets, slip

The assassination of Abraham Lincoln : flight, pursuit, capture, and punishment of the conspirators / by Osborn H. Oldroyd

- Booth, theater, president, box, stage, Ford, door, play
- Baltimore, police, Washington, southern, Virginia
- Lincoln, president, war, people, died, nation, moment
- prisoner, jury, evidence, guilty, examined, throw, sisters

Figure 1: Selected topics from OCA books learned by the DCM-LDA model

The standard LDA topic model has several limitations, which are particularly apparent when applying it to the OCA corpus. First, it scales poorly to large numbers of documents and large numbers of topics. Second, it does not

JCDL'07, June 18–23, 2007, Vancouver, British Columbia, Canada.

take into account the structure of a corpus: intuitively, a page from a book is much more likely to be similar to another page from the same book than a page from a different book. LDA has no ability to represent these connections. Third, it limits the amount of variability in the words that are used by a particular topic. Two books may be about essentially the same topic, but they may discuss that topic in slightly different ways. We would like to both model those differences and examine the resulting topics.

In this paper we present an extension to the LDA model, DCM-LDA. Each book has its own topics, which are not themselves shared between books. Each book-level topic, however, is derived from a more general corpus-level topic that is shared throughout the collection. Each book uses some selection of these corpus-wide topics, but no two books write about a topic in exactly the same way. In this way, we are able to find similarities between books, while still allowing each to have its own distinct perspective. There are several ways in which such a topic-based model can benefit readers:

- **Browsing the collection.** One of the first tasks readers face in approaching a collection is getting an understanding of which subjects are covered. A DCM-LDA model can provide readers with a high-level overview of the subject facets that make up a collection.
- **Searching by keywords.** Although undirected browsing can be useful, readers are often looking for specific topics, and expect the keyword search interface familiar from standard online library catalogs and internet search engines. Topic models have recently been shown to improve information retrieval performance by functioning as a sort of "smoothing" [22]: a relevant document may not contain any words in the query, but may contain many words that are topically similar to words in the query. An interesting aspect of the DCM-LDA model proposed in this paper is that it models the shared topics as probability distributions over sequences of words. As a direct result, for a given query it is simple to rank the topics by the likelihood that they would have generated that query and to integrate topics into existing probabilistic IR systems [18]. Developing new information retrieval benchmarks for the OCA corpus is beyond the scope of this paper, but in future work we plan to evaluate the performance of DCM-LDA based information retrieval on standard IR corpora.
- **Finding related books.** Web sites frequently include a "Find Related" function. This feature can be useful, but often appears to users as a black box. In what way are the results related? A topic-based catalog can provide a substantially more useful listing of related works. Rather than simply returning a single list of related books, the catalog can list the topics that the current book is assigned to. Under each topic, the catalog can then list other books that also share that topic. For example, a book on the American Revolution might touch on many areas that might by themselves be the topic of another book, such as naval battles, Benjamin Franklin, or the British Parliament. A function that simply finds documents similar to the current document as a whole might only return other books specifically about the American Revolution.

It is important to note that the topics generated by the present model are not hierarchical. Although they could be organized into one or more hierarchies as a post-processing step, they are fundamentally faceted. This is an advantage in that it does not force the model to choose the most likely way in which a reader will approach a given book. For example, in traditional classification, it is necessary to choose whether a book on Egyptian art should sit on a shelf next to a book on Aztec art or a book on Egyptian history. In a digital library, we are under no such constraints: we can create multiple "virtual shelves" for a single book, placing it alongside multiple sets of books, each of which are related to the current book in some specific way. An example of such virtual shelves is shown in Figure 4.

DCM-LDA, which is faceted and cluster-based, provides one approach to problems identified with previous methods. Recent work by Hearst [10] compares document clustering with faceted categories in an information retrieval setting. Clustering can be automated and can separate distinct groups of search results, but is hampered by several disadvantages, including "their conflation of many dimensions simultaneously." Faceted subjects, particularly in the form of multiple small hierarchies, are generally found to be useful by users, but such hierarchies need to be known in advance and are most often built by hand. The present approach addresses some of these issues, clustering books without conflating topical dimensions, while also producing meaningful faceted subjects without any human interaction or prior information beyond the texts themselves. The DCM-LDA model does not organize topics into multiple hierarchies, but this could be done separately.

The collection we analyze consists of approximately 42,000 books, totaling 12 million pages and 1.49 billion words. We are able to achieve this scale by taking advantage of the structure of the corpus: each page of a given book is much more likely to be topically similar to another page from the same book than a page from a different, randomly selected book. The learning process, which is an example of Stochastic EM, is as follows. First, we independently train individual topic models for each book. This step divides all the words in a given book between some number of topics based on the cooccurrence patterns of words within pages. Next, we cluster the resulting topics based on their similarity. For each of the 12,000–15,000 clusters, we estimate a distribution. Finally, we retrain the individual topic models for each book using parameters from some selection of these clusters. The process, illustrated in Figure 2 can then be repeated.

The texts in the collection were digitized and made available by the Open Content Alliance (OCA) [17]. The OCA is a major project that seeks to scan and distribute large library collections. As of the time of writing, more than 100,000 books have been scanned, processed by OCR software, and published online at the Internet Archive [11]. The subset used in this study is largely from the Americana subcollection.

We evaluate the model by comparing the automatically extracted topics with manually applied Library of Congress subject headings. We find that the topics that are statistically most related to a given subject heading are highly relevant to that subject heading.

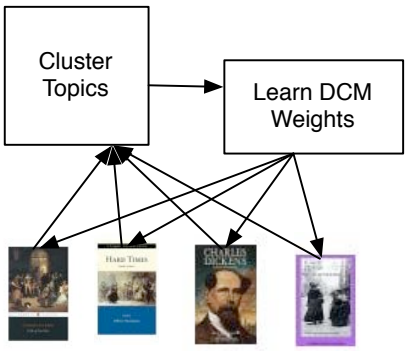

Figure 2: The Stochastic EM training process. Topics from individual book topic models are gathered and clustered. DCM parameters are learned from topic clusters. Finally, new topic models are relearned for each book using DCM parameters from selected clusters as priors.

2. RELATED WORK

Another project making use of Open Content Alliance texts is the Melvyl Recommender project at the California Digital Library [3]. The CDL reports several problems with the texts that we also encounter, such as silently dropped hyphenation. Newman [16] reports on a study on enhancing metadata records from the CDL American West collection using a topic model. The documents consist of the title, description and subject fields from 360,000 metadata records. The model consists of 300 topics. According to a manual evaluation, 78% of the resulting topics are usable and 80% of the metadata records are enhanced by the addition of topic information. Unlike this study, which uses only metadata, our work has access to the full text of each book. The additional data allows us not only to get a more detailed representation of what each book is about, but also to specifically model the particular way in which each book discusses each topic.

Much of the existing work on automatic classification and subject analysis in library collections focuses on applying existing subjects and classifications, often those of the Library of Congress. Recent examples include the OCLC Scorpion project [7] and INFOMINE [6]. Krowne [12] evaluates the application of several automatic clustering methods for organizing digital library collections.

The SOMLib Digital Library [19] organizes text collections into topical hierarchies using Self-Organizing Maps. This model is less flexible than a Dirichlet-based topic model: it assigns whole documents to clusters rather than the individual words within documents. The corpus used, 11,000 newswire articles, is also of much smaller scale than the OCA corpus.

Zhai et al. present a cross-collection mixture model [23], which uses an EM approach to find themes shared between collections. The themes are represented as multinomials over words. The model is tested on a corpus consisting of three collections, each with 30–40 documents. In contrast, the OCA corpus tested here contains more than 8000 collections (ie books), each with an average of several hundred pages.

Another approach to sharing information between document clusters in topic models is in Teh et al. [20], which uses a hierarchical Dirichlet process (HDP) to model the generation of topic mixture distributions. Unlike the present approach, this paper links similar documents through the distributions over topics, so that documents in the same subcorpus draw topic multinomials (the α parameter) from the same Dirichlet distribution. All documents share the same topics, which are represented as multinomials. In contrast, our approach shares information between books through the topic-specific Dirichlet distributions (the β parameters).

A similar Stochastic EM-based approach appears in Goldberger et al. [8]. This paper builds clusters of Gaussian distributions that are derived from Gaussian mixture models.

Veeramachaneni et al. [21] presents a hierarchical Dirichlet model for document classification where each node in the hierarchy is a multinomial over words drawn from a Dirichlet corresponding to the multinomial of its parent multiplied by a constant factor σ. This model differs from the model presented here in that it does not represent documents as mixtures of multinomials (each document has one class) and in that it constrains topics to fall within a single hierarchy.

An example of large-scale topic modeling is presented by Buntine [2], using Discrete Principle Component Analysis (PCA), a more general formulation of LDA. Using a variety of optimizations, the authors are able to train a model with 1000 components (ie topics) on a corpus of 180 million words. In contrast with such simpler topic models, the DCM-LDA approach scales more easily. The individual topic models can be trained independently, so parallelizing the training process by distributing books to a cluster of servers is simple. Furthermore, each model requires a small amount of memory, so no special considerations need to be taken to ensure that they fit within available resources. Finally, the DCM-LDA topics are able to explicitly model the intuition that different authors talk about the same things in different ways, by allowing different levels of variance.

3. METHODS

3.1 Preprocessing

Texts digitized by the OCA are available at the Internet Archive [11]. Each book can be downloaded in several formats. For this work we use the text representation of OCR output. In addition, we download catalog metadata in the form of MARC XML files.

The OCA texts need a small amount of additional preprocessing. As is noted by the Melvyl Recommender project, OCA texts often silently drop hyphens. We do not make a significant effort to rejoin split words, except for common and relatively unambiguous suffixes such as "-ing", "-ment", and "-tion". Applying more sophisticated machine learning methods to this problem would be a useful contribution. Another problem is that the OCR output includes all the words printed on a page, which includes headers. Such repeated segments display unusual statistical properties that can interfere with topic model training, so we avoid them by dropping single-line paragraphs. As is common in text analysis, we remove the most frequently occurring words based on a predefined list. Finally, the text was converted into sequences of numeric features using the Mallet toolkit [14].

3.2 Stochastic EM

The procedure we use for learning DCM topics is an example of an algorithm known as Stochastic EM [4]. Stochastic EM is a method for estimating parameters of a model for which some of the data is missing. In our case, the assignments of words in books to topics is not observed. The algorithm involves alternating between (a) sampling topic assignments using a stochastic method such as Gibbs sampling and (b) estimating the DCM topic parameters given the sampled topic assignments.

3.3 Modeling the content of a library

Latent Dirichlet Allocation (LDA) is a statistical model of the generation of text collections. A topic is modeled as a probability distribution over the words in the vocabulary. Each document, such as a page in a book, is assumed to be a mixture of some number of topics. In order to "generate" a word, a page selects a topic according to its distribution over topics, and then chooses a specific word from that topic's distribution over the vocabulary. Topic models can be learned from unlabeled text by looking at word cooccurrence patterns within pages — no further training data is required. We train topic models using Gibbs sampling, a method in which we first randomly assign every word to a topic, and then repeatedly iterate over every word, choosing a new topic for that word based on the current assignments of every other word on the page and the words assigned to each topic. The probability of assigning a word w on page p to topic t is given by

$$p(t|p,w) \quad \propto \quad \frac{\alpha_t + N_p^t}{\sum_t'(\alpha_t' + N_p^{t'})} \frac{\beta_w + N_t^w}{\sum_w'(\beta_w' + N_t^{w'})}. \qquad (1)$$

where N_p^t is the number of times topic t appears on page p and N_t^w is the number of times words of type w have been assigned to topic t. In a simple topic model like LDA, the hyperparameters α_t and β_w are generally constants, reflecting symmetric, uninformative priors. Intuitively, the model prefers to assign a word to a topic that occurs frequently on the same page and that already has many words of the same type. Therefore within a single book, the topic model uses the cooccurrence patterns of words that occur on the same page to learn topics. Topic models of this nature have a number of advantages, including being able to handle multiple word senses by looking at the topical context of an ambiguous word.

Unfortunately, topic modeling also presents significant computational challenges. In Gibbs sampling, at every iteration we must calculate the weight of every topic for every word in the corpus. The performance of the sampler is therefore proportional to the number of iterations times the number of topics times the number of words in the corpus. Since we generally want the number of topics to be proportional to the number of documents, the computation required for each iteration grows roughly quadratically with the size of the corpus. Other examples of large-scale topic models have limited the number of training iterations to low numbers like 30 [22], as opposed to the 1000 used in this study. Memory is also a problem, as sampling requires access to the number of times each word type has been assigned to each topic as well as the type and current topic assignment of each word token in each document. The current model addresses these problems by learning a topic model for each book independently. Rather than iterating over a large number of documents with a large number of topics, we iterate over small subsets of the corpus with subsets of the topics. As a result, training the independent topic models is trivially parallelizable.

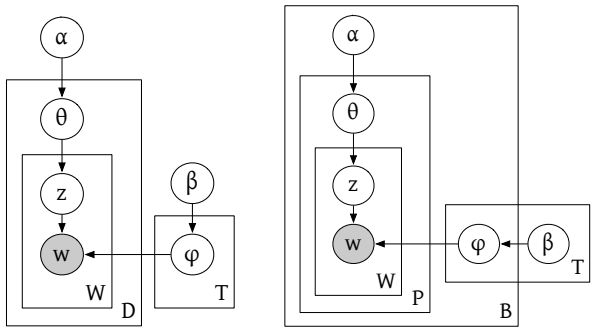

Figure 3: The standard LDA model includes one multinomial (ϕ) for each topic, which are shared by all Documents. The DCM-LDA model groups documents into Pages within Books and uses a distinct Dirichlet distribution (β) for each topic. Each book draws its own topic multinomials from these Dirichlets. DCM-LDA can be seen as a combination of many LDA models, which are linked through the β topic parameters.

The choice of the number of topics is an important factor in topic modeling. For this initial work, we use a number of topics equal to the number of pages in the book divided by 10, plus 10. Improving this aspect of the model is an area for future work. One option is non-parametric priors such as the HDP described in Teh et al [20].

3.4 The Dirichlet Compound Multinomial distribution

The key difference between the general LDA model and the current model is in the representation of topics using the Dirichlet Compound Multinomial (DCM) distribution rather than the multinomial distribution. In LDA, a topic is a single multinomial for the entire corpus. All topic multinomials share a single symmetric (ie uninformative) Dirichlet prior. In the present model, all topics are represented by Dirichlet distributions. Each book draws a multinomial for each topic from that topic's Dirichlet distribution. The DCM distribution results from integrating out the specific multinomial parameters drawn from the Dirichlet.

The DCM distribution has been shown to be a better fit for the statistical properties of text than a simple multinomial [13]. This is largely due to the phenomenon of "burstiness": if a rare word has occurred once in a document, it is much more likely to appear again in that same document than another word with similar overall frequency in the corpus. A multinomial, which models the cooccurrence of two words by simply multiplying the probabilities of the two words, would consider two occurrences of a rare word like "camelid" in the same document to be just as likely as seeing "camelid" and a similarly rare word like "apotropaic" together. In contrast, each time it generates a particular word in a given document, a DCM distribution puts increasing weight on that word.

In terms of parameters, the difference between the LDA topic model and the DCM-DLA topic model is in the setting of the β topic priors. Rather than using a single uninformative constant, DCM-LDA learns specific values for each word type for each topic. In LDA, the model has no prior expectation about what words will be assigned to topic 17. In DCM-LDA, we learn specific prior distributions over words for each topic. As a result, if a book topic model is retrained using a Dirichlet prior distribution for topic 17, we know a great deal about what words will be assigned to that topic. In this way, books have access to information such as cooccurrence patterns learned from the entire corpus in a summarized form. This summarization enables the substantial performance gains for DCM-LDA relative to LDA. Figure 3 shows the distinction between the standard LDA topic model and the DCM-LDA model in terms of their graphical models. In LDA, topics are represented by a single multinomial ϕ_t, shared by all documents in the corpus. In DCM-LDA, each book draws a separate multinomial for each topic from that topic's Dirichlet prior β_t.

3.5 Clustering topics

Once we have computed topic models for each book, the next step is to cluster topics from the independent book models into "corpus-wide" topics and then learn DCM parameters from those clusters. This process completes the Stochastic EM procedure by maximizing the topic parameters given a pseudo-complete sample consisting of the observed words and the sampled topic assignments.

The choice of clustering method is not a fundamental part of the process. As a first stage, we have used a greedy agglomerative method with hard cluster assignments (that is, each topic is assigned to one and only one cluster). An EM-based clustering method with weighted soft cluster assignments such as that described by Elkan [5] is another alternative.

Greedy agglomerative clustering repeatedly uses a distance metric to find and merge the closest pair of instances. Given a topic t represented as a bag of words, we can define a probability distribution p such that $p(w)$, the probability of word w, is N_t^w/N_t. The metric we use for clustering topics is Jensen-Shannon (JS) divergence. For two distributions p and q, we can define a mean distribution $(p+q)/2$. Jensen-Shannon divergence is the average of the Kullback-Leibler (KL) divergence between p and the mean distribution and the KL divergence between q and the mean distribution. There are two reasons for using JS divergence rather than simple KL divergence. First, unlike KL divergence, JS divergence is symmetric, so the order in which topics are compared is not significant. Second, KL divergence becomes undefined if there is a word w for which p has mass but q does not, because the term $p(w)\log q(w)$ contains the log of zero. This is an important consideration, since we expect any two arbitrary topics to share very few words.

In fact, the small expected word overlap between any two given topics can be used to calculate the full distance matrix efficiently. A naive approach would be to calculate each of the $T(T-1)/2$ possible topic pairs, comparing the proportion of every word that occurs in the first topic ($p(w)$) to its proportion in the other topic ($q(w)$) and vice versa. However, it is possible to rewrite the expression for Jensen-Shannon divergence considering only the probabilities of words that occur in both topics:

$$1 + \frac{1}{2}\sum_{w\in\{p\cap q\}} p(w)\log p(w) + q(w)\log q(w) - (p(w)+q(w))\log\frac{p(w)+q(w)}{2} - p(w) - q(w) \quad (2)$$

It can be shown that if two topics share no common words, the JS divergence between p and q is equivalent to summing over two probability distributions and dividing by two. All distributions sum to one, so this value is one. In Equation 2, we start with one. Then, for every word that occurs in both distributions, we subtract the probabilities of the word in both distributions and add the terms from JS divergence that include those probabilities.

Given that we only need to consider shared words when calculating topic similarities, an efficient method for calculating the full distance matrix involves essentially constructing an inverted index mapping word types to sets of topics. As each topic is indexed, we examine every distinct word type for which the topic has a non-zero count. For each of the previously indexed topics that also contains this word, we update the distance between that topic and the new topic using Equation 2. Finally, we add the new topic to the index of topics that contain that word.

Once we have identified clusters of topics, the next step is to use the topics to learn DCM parameters for each cluster. This process is fairly straightforward. A thorough treatment of methods for estimating the parameters of a DCM distribution, also described as a Dirichlet-Multinomial or Polya distribution, is provided by Minka [15]. For this work we use Minka's fixed point iteration method as stated in that paper.

There is one final step. In order to train the independent book topic models efficiently, we select a subset of the learned cluster DCMs to propagate to each book. We do this by performing one round of Gibbs sampling. Once that round has completed, we select the n most frequently used topics, where n is currently 10 plus the number of pages divided by 10. Improving this selection process is an area for future refinement.

4. RESULTS

We were able to process the 1.49 billion word corpus efficiently, using between 12,000 and 15,000 topics. For the 42,000 books, training topic models and clustering topics each take under 10 hours, using 30 CPUs. As an LDA model with similar specifications is currently beyond our technical limitations, it is difficult to compare these results with a standard model. For comparison, the largest book in our current collection contains 730,000 tokens. A standard LDA model with 1000 topics takes slightly more than 7.5 seconds per iteration. If we linearly scale the number of tokens by a factor of 2000 to one and a half billion and the number of topics by 12, we can estimate that a single Gibbs sampling iteration would take 50 hours. For the same number of iterations (1000), such a model would take almost six years to train.

Another method of training topic models, variational EM [1], is more amenable to parallelization than Gibbs sampling but still much less scalable than DCM-LDA. In parallelized variational EM, individual nodes must communicate with a

head node between every iteration. DCM-LDA trains topic models for each book completely independently given only the β_{tw} parameters. No communication is required between the individual book topic models until the entire corpus has been processed. DCM-LDA does not require any complicated synchronization and is thus much easier to program.

4.1 From topics to virtual shelves

Tables 1, 2 and 3 list topics selected from models generated for three books. For each topic, the table lists the most probable words for the topic under its DCM parameters along with the words in the book most frequently assigned to the topic. For the topics derived from *Chancellorsville and Gettysburg*, we also include relevant (manually selected) passages from the Wikipedia articles "Battle of Chancellorsville" and "Battle of Gettysburg"[1] for historical context.

The second topic in Table 1 contains an example of burstiness. The DCM topic is about the battle of Gettysburg, which occurred shortly after the battle of Chancellorsville. The book-specific topic prominently includes Gen. Hooker, but the DCM topic does not. Hooker commanded the Union army until shortly before the battle, when he was replaced by Gen. Meade. Therefore, in most descriptions of the battle of Gettysburg Hooker plays almost no part, while in this book, which follows events from the earlier battle of Chancellorsville through Gettysburg, he features more prominently.

Table 4 demonstrates the variability that can exist for each book's version of a given DCM topic. The most probable words given the β_t prior parameters are listed at the top of the table, while the most frequent words that each book assigns to its own version of the topic are listed along with the title. In most cases the book-specific words are quite similar to the topic parameters, although there are a few book-specific topics that seem to be connected to the DCM topic only by the presence of Generals Hooker and Lee.

Table 4 and Figure 4 illustrate the use of the topic model in recommending books that are related to a given book. In the table, we show books that share the same topic, organized in descending order by the proportion of the book assigned to the topic. For the topic about the battle of Chancellorsville, the book *Chancellorsville and Gettysburg* is most prominent. Figure 4 shows several of these "virtual shelves" for a single book, *A short history of the United States* by John Spencer Basset (1913). This book contains many topical facets, including information about naval history, early American statesmen, the New England region, and the construction of the Panama canal. In each case, the topic model has discovered these topical facets, and for each facet, we can provide readers with links to other books in the collection that share that specific subject area.

4.2 Topics and manual subject headings

In order to evaluate the performance of the topic model, we compare the resulting DCM topics to the Library of Congress Subject Headings contained in the MARC metadata distributed by OCA with the texts. We measure the affinity between a topic and a subject heading using mutual information, an information theoretic tool that measures the predictability of two random variables, one given the other.

[1] http://en.wikipedia.org

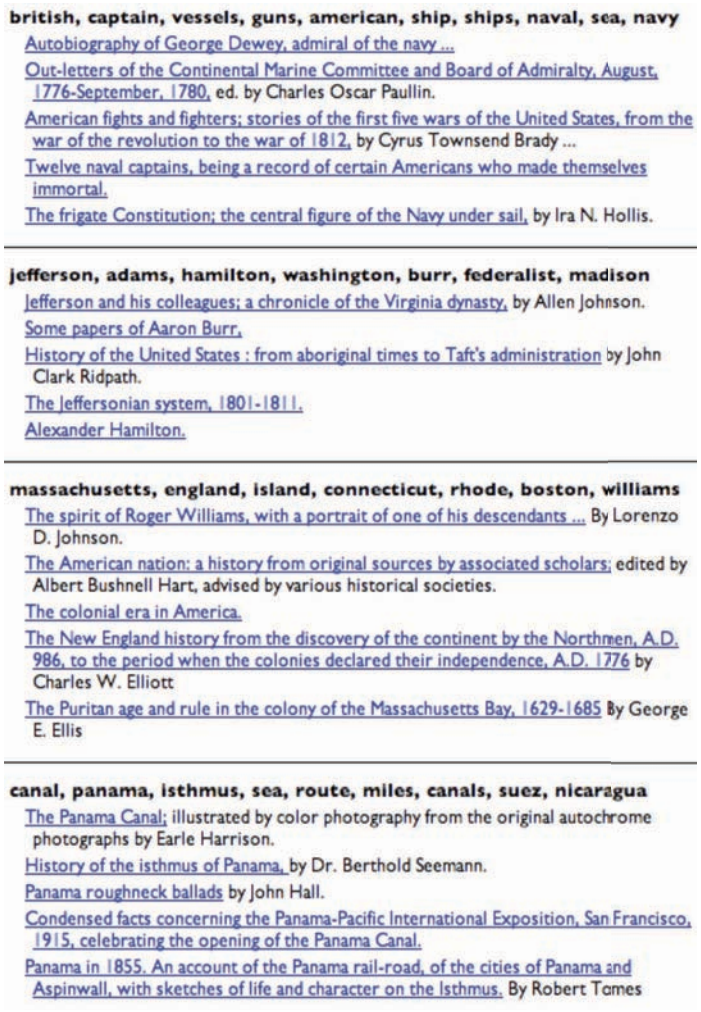

Figure 4: Four "virtual shelves" for a single book, *A short history of the United States* by John Spencer Basset (1913). The model has discovered multiple subject facets within the book, and has suggested books specifically related to those facets.

Table 5 contains the topics with the highest mutual information for several common subject headings. The contrast between learned topic clusters and the manually applied subject headings is indicative of the difference between these two approaches. In the case of a fairly concrete subject such as "Evolution", the topic model finds a single topic with relatively high mutual information with the subject heading. In the case of "Lincoln, Abraham", the topic model finds several topics that have roughly the same mutual information and all place high probability on the name Lincoln. In addition, the topics have divided Lincoln into several aspects: Lincoln debating Douglas, Lincoln enacting the emancipation proclamation, Lincoln being chief executive in Washington. Finally, for the subject heading "Authors, American", only two of the topics with highest mutual information to this subject heading contain the words "Author" or "American" in any position of prominence. Although the topic model successfully finds words about authorship

Table 1: Topics extracted from *Chancellorsville and Gettysburg*, by Abner Doubleday, with comparisons to their DCM topic parameters. The topics place substantial weight on proper nouns. This is typical of historical documents, which emphasize people and places.

Book-specific topic words	DCM topic words	Manually selected passages from Wikipedia, "Battle of Chancellorsville" and "Battle of Gettysburg"
Hooker, corps, Jackson, Chancellorsville, Sickles, Howard, eleventh, Lee, army, plank	Hooker, corps, Lee, Chancellorsville, Jackson, Sedgwick, ford, Fredericksburg, Stuart, Sickles	"... the battle pitted Union Army Maj. Gen. Joseph Hooker's Army of the Potomac against an army half its size, Lee's Confederate Army of Northern Virginia. ... The Confederate victory was tempered by the mortal wounding of Lt. Gen. Stonewall Jackson..." (Chanc.)
Lee, Hooker, Meade, Longstreet, Potomac, army, hill, Gettysburg, Ewell, union	Lee, Meade, Gettysburg, corps, cavalry, hill, july, Ewell, Potomac, Buford	"Maj. Gen. Joseph Hooker moved his army in pursuit, but was relieved almost on the eve of battle and replaced by Meade." (Gett.)
eleventh, Howard, Buford, hill, Gettysburg, Reynolds, seminary, Wadsworth, ridge, Cutler	hill, Gettysburg, ridge, seminary, Ewell, Buford, Reynolds, cemetery, Doubleday, eleventh	"... the incompetent commander of the Union XI Corps, Maj. Gen. Oliver O. Howard." (Chanc.) "General Buford realized the importance of the high ground directly to the south of Gettysburg" (Gett.)
top, Sickles, orchard, peach, round, Crawford, de, ridge, Birney, Vincent	round, top, orchard, peach, Sickles, den, devil, Hood, de, Birney	"Sickles was quite bitter about giving up this high ground; his insubordinate actions at the Peach Orchard in the Battle of Gettysburg two months later were probably influenced strongly by this incident." (Chanc.) "Lee launched a heavy assault on the Union left flank and fierce fighting raged at Little Round Top, the Wheatfield, Devil's Den, and the Peach Orchard." (Gett.)
cavalry, Stuart, Pleasonton, infantry, Middleburg, Aldie, Gregg, station, Kilpatrick, Brandy	cavalry, Stuart, Gregg, infantry, Kilpatrick, Hampton, Buford, horse, confederate, station	"Confederate cavalry under Maj. Gen. J.E.B. Stuart..." (Chanc.)

Table 2: Topics extracted from *The horse, in the stable and the field : his management in health and disease* / by J. H. Walsh, with comparisons to their DCM topic parameters. The topics found divide the text between words about horsemanship and words about medicine. A recommender system that does not recognize topical distinctions might only be able to suggest other books about equestrian medicine, but a topic-based recommender system will be able to find books about medicine but not horses and vice versa.

Book-specific topic words	DCM topic words
saddle, left, hand, reins, rider, mouth, rein, riding, stirrup, seat	left, foot, saddle, leg, seat, position, body, ground, stirrup, knee
shoulder, muscles, blade, arm, bones, action, upright, joint, oblique, hip	muscles, muscle, bone, shoulder, arm, action, power, bones, thigh, weight
good, merit, select, possessed, desirable, animals, essential	boar, breeding, bred, pure, good, animals, herd, animal, boars, sire
treatment, inflammation, skin, cases, generally, applied, part, case, pain, swelling	part, inflammation, skin, treatment, applied, cases, case, generally, water, pain

Table 3: Topics extracted from *A treatise on the differential calculus with numerous examples* / by I. Todhunter, with comparisons to their DCM topic parameters. Common topics include technical language such as the mathematical topics found in this book. This language is very uniform across books, so there is little difference between book-specific and DCM topics.

Book-specific topic words	DCM topic words
dx, dy, dz, du, fdu, equations, variables, dv, independent, df	dx, dy, dz, du, dt, dv, da, dp, df, dr
differential, coefficient, respect, function, coefficients, art, suppose, result, functions, denote	differential, coefficient, function, coefficients, exist, suppose, equal, continuous, generally, exists
limit, indefinitely, unity, infinite, increases, small, greater, finite, made, increase	limit, infinite, approaches, finite, increases, indefinitely, ratio, number, unity, sum
sin, cos, tan, log, shew, sm, result, ic, tt, cot	cos, sin, tan, sm, ft, angle, tt, cot, angles
theorem, term, series, powers, expression, taylor, true, expansion, result, remainder	theorem, series, powers, taylor, functions, expansion, coefficients, ascending, terms, maclaurin

Table 4: The books with the highest proportion of the topic "Hooker, corps, Lee, Chancellorsville, Jackson, Sedgwick, ford, Fredericksburg, Stuart, Sickles." In DCM-LDA, each book has its own book-specific topics, each drawn from a corpus-wide topic. These book-specific topics show the different perspectives each book brings to the general topic.

Book-specific topic words	Title and author
Hooker, corps, Jackson, Chancellorsville, Sickles, Howard, eleventh, Lee, army, plank	*Chancellorsville and Gettysburg*, by Abner Doubleday ...
army, Lee, general, Potomac, command, river, part, Hooker, crossed, crossing	*The One Hundred and Twentieth Regiment New York State Volunteers. A narrative of its services in the war for the Union.* By C. Van Santvoord... Pub. by the One Hundred and Twentieth N. Y. Regimental Union.
Jackson, commission, Stuart, battle, soldiers, field, Stonewall, Christian, dead, nation	*The Americans at home* : pen-and-ink sketches of American men, manners and institutions.
Hooker, Lee, Chancellorsville, Jackson, Fredericksburg, Sedgwick, roads, Anderson, corps, ford	*The strategy of Robert E. Lee*,
Hooker, Jackson, corps, Lee, Sedgwick, Chancellorsville, army, sixth, eleventh, Fredericksburg	*The photographic history of the Civil war ...* / Francis Trevelyan Miller, editor-in-chief; Robert S. Lanier, managing editor. Thousands of scenes photographed 1861-65, with text by many special authorities.
Hooker, Lee, Chancellorsville, Jackson, Howard, received, corps, Sedgwick, couch, pillow	*Union portraits,* by Gamaliel Bradford.
states, union, united, Hooker, south, southern, state, government, northern, party	*John Ashton: a story of the war between the states.* By Capers Dickson...
Hooker, Lee, Jackson, Chancellorsville, corps, Sedgwick, Anderson, ford, army, common	*The life of General Robert E. Lee*, by G. Mercer Adam; the life-career and military achievements of the great Southern General, with a record of the campaigns of the Army of northern Virginia.

("poems, poet, literary, poem, poetry, volume") and examples of American authors (Longfellow, Lowell, Whittier, Hawthorne, Thoreau and Poe along with Howells, who was editor of the *Atlantic Monthly*), it cannot, in this case, generalize to the more abstract class.

Only a sample of the LC subject headings are shown here, but the ones shown are illustrative of the patterns found in the rest. Of the more than 100 subject headings that appear more than ten times in the corpus, for almost all of them the topics with high mutual information were found to be relevant. Exceptions included topically vague headings such as "Online resources. — local" and subject headings where topics were dominated by very frequent words that are not in our Academic English stoplist. These include "American wit and humor," which makes heavy use of "dialect" forms, and "Mathematics," which in this collection is largely in French.

Table 6 contains a selection of the topics that occur most often in the corpus. Specifically, the table is sorted by the number of documents for which a given topic is assigned, regardless of the proportion of that document that is actually in the topic. Therefore, this table indicates the breadth of use of a topic, and not necessarily its proportion of the overall corpus.

The topics in this table demonstrate the use of the DCM-LDA topic model in providing a summary of the contents of the collection. The books were downloaded from the OCA "Americana" collection, which includes a substantial number of oral history transcriptions collected at the University of California as well as US history, California history, Mathematics, Military history, and books on cookery. Several geographic regions are frequently the subject of books in the collection: Russia, Athens and Sparta, the Pacific Rim, and Italy. The topic model also identifies portions of the collection that include text in French, Latin, and German.

Again, the topics in this table highlight the role of stoplists—sets of words that are deemed too frequent to have any distinct meaning. Because of the large number of these words, they are typically removed so that they do not overwhelm less frequent but more meaningful words. Not shown, but very frequently used, are several topics similar to the topic indicated by "thy, thou, thee, hast, art". In this case, our stoplist, which was developed for modern English, does not include the archaic pronoun "thou" and its most common verb forms. As a result, topic models for texts that use this pronoun must account for many instances of these words. The resulting topics are not always similar enough that our clustering algorithm merges them. Note, however, that the topic model recognizes that "thou" and "thy" are frequently used in religious language, as in the topic "god, love, thou, thy, father, evil". Another example is the most frequently assigned topic, "time, made, man, day". These words are not exactly true stopwords, which serve primarily syntactic functions, but their meaning and usage is very general.

The model is robust in the face of substantial variation in the data. For example, related words like "poet, poetry, poems, poem, poets" appear prominently in the same topic. Even though the topic model is unaware of any similarity between these words (it represents them as arbitrary integers), the statistical patterns in the text are sufficient to group them together. No additional stemming or other preprocessing is necessary. Another source of variation is im-

Table 5: Do DCM topics match up with Library of Congress Subject Headings? Here we list the topics that are most associated with selected LC subject headings. The column on the left lists the mutual information between the event that a book is assigned to the topic and the event that it is assigned to the subject heading. Lower mutual information implies more statistical independence.

Lincoln, Abraham, — 1809-1865	
0.01419	lincoln, douglas, speech, illinois, debate
0.01310	lincoln, lin, coln, john, abraham, henry
0.01278	slavery, lincoln, emancipation, war, proclamation
0.01266	slavery, lincoln, mr, free, douglas, state
0.01211	washington, mansion, executive, lincoln
0.01210	lincoln, mrs, edward, president, madam
Authors, American	
0.00336	poems, poet, literary, poem, poetry, volume
0.00319	longfellow, whittier, lowell, england, poet
0.00301	hawthorne, author, book, books, work, read
0.00295	life, poems, published, literature, american
0.00270	thoreau, concord, henry, channing, house
0.00266	editor, howells, wrote, atlantic, mr, york
0.00249	poe, hawthorne, tales, mr, pp, vol, told
Canada — Description and travel	
0.00241	french, canada, english, canadians, canadian
0.00225	canada, anglais, quebec, montreal, france
0.00202	general, army, st, george, command, niagara
0.00182	timber, canada, trade, canadas, country
0.00182	upper, canada, province, assembly
0.00166	st, quebec, canada, lawrence, laurence
0.00165	canada, canadian, quebec, montreal, toronto
Calculus	
0.00599	dx, bx, ax, cx, exponent, du, au, adx, fx
0.00558	dx, dy, dz, du, dt, dv, da, dp, df, dr
0.00387	integral, dx, integration, integrals, function
0.00380	derivative, function, derivatives, differentiation
0.00373	function, derivative, ordinate, interval
0.00351	differential, variable, constant, differentials
0.00334	cos, sin, tan, sm, ft, angle, tt, cot, angles
United States – History – Revolution, 1775-1783	
0.00960	burgoyne, arnold, british, americans, army
0.00755	america, king, house, lord, parliament
0.00699	british, washington, york, americans, island
0.00601	greene, island, rhode, british, providence
0.00598	lord, cornwallis, rawdon, camden, general
0.00597	colonies, british, lord, america, parliament
0.00594	act, colonies, parliament, america, house
Evolution	
0.01162	theory, evolution, facts, doctrine, origin
0.00910	relation, embryology, compared, eye
0.00882	theory, evolution, fact, organic, facts, nature
0.00851	germ, cells, characters, plasm, weismann
0.00846	mr, darwin, species, selection, views
0.00791	selection, natural, theory, elimination
0.00741	species, plants, animals, distinct, sterility

Table 6: A selection of frequently occurring topics in the Americana collection. This view provides a high level overview of the content of the collection.

Docs	DCM topic words
1423	time, made, man, day, great, long, good, found, make, back
650	thy, thou, thee, hast, art, heart, thine, eyes, wilt, love
373	de, la, les, des, le, en, du, qui, par, dans
304	history, oral, california, university, library, office, berkeley
229	est, ad, ut, quod, cum, qui, sed, de, quam, si
222	oct, june, jan, sept, july, dec, aug, nov, feb, april
213	states, united, state, government, american, union, citizens
188	eyes, face, girl, love, moment, turned, hand, man, suddenly
187	san, francisco, california, bay, city, pacific, coast, santa
179	russia, russian, soviet, revolution, government, moscow, world, communist, party, revolutionary
172	god, love, thou, thy, father, evil, child, hath, thee, lord
172	cos, sin, tan, sm, ft, angle, tt, cot, angles, equation
167	greece, athens, greek, greeks, athenian, athenians, ancient, city, sparta, war
165	major, general, captain, colonel, lieutenant, brigadier
164	university, california, berkeley, faculty, campus, students
163	troops, regiment, infantry, artillery, cavalry, st, confederate, battery, union, regiments
146	die, der, und, von, den, zu, auf, mit, das, des
143	japan, japanese, china, east, chinese, asia, russia, pacific
139	italy, italian, florence, medici, tuscany, pisa, venice, milan
138	god, lord, thy, love, thou, thee, christ, life, holy, jesus
133	persian, egypt, king, empire, babylon, asia, persia
127	church, bishop, bishops, st, rome, ecclesiastical, pope
125	warren, earl, california, introduction, attorney, john, james
121	poet, poetry, poems, literature, style, poem, poets, author
117	add, butter, cook, stir, milk, hot, flour, salt, serve, minutes

proper hyphenation. In many cases, pseudo-words like "lin" and "coln" appear in topics along with the original words, as in Table 5.

One unexpected benefit of the DCM-LDA topic model is its ability to identify multiple copies of the same book. This occurs in two ways. When the same edition of a book has been scanned multiple times, the topics discovered are very similar. As one might expect, identical pages result in highly similar topics. The topic model is also capable of finding different editions of the same text. For example, the top books for one topic that places high probability on the words "Grenada" and "Alhambra" are either editions of Washington Irving's *The Alhambra* or editions of his collected works. It is not unlikely that algorithms specifically designed for identifying copies of the same work might not perform better at this task, but it is nevertheless an interesting side effect.

5. DISCUSSION

We have presented a topic model appropriate for large-scale libraries of digital books. Our goals in designing this model were that it be able to scale to large document corpora and large numbers of topics, that it take advantage of the observed structure of the corpus, and that it identify common topics while allowing books some flexibility in their language.

DCM-LDA is capable of discovering the rich topical structure of the OCA collection while also handling its vast scale. The current 1.49 billion word corpus does not significantly stress any component of the system—there is no reason at

this point to expect that it could not scale up to the billions of words and hundreds of thousands of books that will soon be available through the OCA. Furthermore, the approach is generalizable to any collection that can be broken into distinct, semantically coherent sections. For example, a collection of scientific literature such as the Medline corpus of biomedical abstracts could be divided into journals and news corpora could be divided by time periods.

Large-scale topic models can benefit readers in many ways. We have outlined three primary applications. First, the model generates browsable summaries that accurately reflect the content of the collections. Second, topic models can support improved information retrieval by leveraging the contextual patterns of billions of words to expand user queries [22]. Finally, the model provides recommendations for similar items based on specific topical facets that we have shown match closely with manual subject headings.

Beyond the specific model in question, this work highlights an important issue. Collections of digital books have their own challenges and their own opportunities, distinct from existing digital libraries of shorter documents. The digitization of whole libraries is unquestionably a milestone. The materials in collections like the OCA represent both a substantial investment in time and money and a resource of lasting cultural value. Traditional library technology and practices, however, are not structured to take advantage of full-text collections. It is imperative that we focus on developing tools like DCM-LDA, which thrive on vast data sets, in order to bring the full benefits of this investment to readers.

6. ACKNOWLEDGMENTS

The authors would like to thank everyone involved in the creation and development of the OCA.

This work was supported in part by the Center for Intelligent Information Retrieval, in part by The Central Intelligence Agency, the National Security Agency and National Science Foundation under NSF grant #IIS-0326249, in part by NSF grant #CNS-0551597, and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[2] W. Buntine, S. Perttu, and H. Tirri. Building and maintaining web taxonomies. In *XML Finland 2002*, 2002.

[3] California Digital Library. The Melvyl Recommender project full text extension supplementary report. http://www.cdlib.org/inside/projects/melvyl_recommender/report_docs/mellon_extension.pdf.

[4] G. Celeux, D. Chauveau, and J. Diebolt. On stochastic versions of the EM algorithm. Technical Report RR-2514, INRIA.

[5] C. Elkan. Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution. In *ICML 2006*, 2006.

[6] E. Frank and G. W. Paynter. Predicting library of congress classifications from library of congress subject headings. *J. Am. Soc. Inf. Sci. Technol.*, 55(3):214–227, 2004.

[7] C. J. Godby and J. Stuler. The Library of Congress Classification as a knowledge base for automatic classification. In *IFLA Preconference*, 2001.

[8] J. Goldberger and S. Roweis. Hierarchical clustering of a mixture model. In *NIPS 2004*, 2004.

[9] Google Books. http://books.google.com.

[10] M. Hearst. Clustering versus faceted categories for information exploration. *Communications of the ACM*, 49(4):59–61, 2006.

[11] Internet Archive. http://www.archive.org/texts.

[12] A. Krowne and M. Halbert. An initial evaluation of automated organization for digital library browsing. In *JCDL 2005*, 2005.

[13] R. Madsen, D. Kauchak, and C. Elkan. Modeling word burstiness using the Dirichlet distribution. In *ICML 2005*, 2005.

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

[15] T. Minka. Estimating a Dirichlet distribution, 2000.

[16] D. Newman. American west metadata enhancement feasibility study, 2005. http://www.cdlib.org/inside/projects/amwest/cluster.pdf.

[17] Open Content Alliance. http://www.opencontentalliance.org/.

[18] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR*, 1998.

[19] A. Rauber and D. Merkl. Text mining in the SOMLib digital library system: the representation of topics and genres. *Applied Intelligence*, 18:271–293, 2003.

[20] Y. W. Teh, M. Jordan, M. Beal, and D. Blei. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *NIPS 2004*, 2004.

[21] S. Veeramachaneni, D. Sona, and P. Avesani. Hierarchical Dirichlet model for document classification. In *ICML 2005*, 2005.

[22] X. Wei and B. Croft. LDA-based document models for ad-hoc retrieval. In *SIGIR 2006*, 2006.

[23] C. Zhai, A. Velivelli, and B. Yu. A cross-collection mixture model for comparative text mining. In *KDD 2004*, pages 743–748, 2004.