Procesamiento del Lenguaje Natural, Revista nº 45, septiembre 2010, pp 153-160 recibido 01-05-10 revisado 19-05-10 aceptado 20-05-10

Análisis Comparativo de Investigación Científica con un Modelo Independiente del Lenguaje a través de Múltiples Colecciones

Comparative Scientific Research Analysis with A Language-Independent Cross-Collection Model

Michael Paul and Roxana Girju
University of Illinois, Urbana, U.S.A.
{mjpaul2, girju}@illinois.edu

Resumen: Este artículo trata el problema de análisis de investigación científica a través de múltiples colecciones disciplinarias. Nuestro método se aplica en: (1) descubrimiento sin supervisión de temas científicos a través de múltiples disciplinas; (2) comparación de los temas de misma disciplina; y (3) análisis temporal.
Palabras clave: Evaluación, procesamiento lingüístico, modelos probabilísticos

Abstract: This paper addresses the problem of scientific research analysis across multiple research literature collections. We use topic modeling in three novel comparative tasks: (1) unsupervised discovery and comparison of scientific topics across multiple disciplines; (2) comparison of topics within the same discipline; (3) analysis of topic evolution over time within and across disciplines and trend analysis.
Keywords: evaluation, linguistic processing, probabilistic topic models

1 Introduction

The scale and complexity of today's research problems demand that scientists go beyond the boundaries of their individual disciplines and explore other related areas. Advances in molecular imaging, for example, require knowledge in areas such as radiology, cell biology, physics, and computer vision. Now more than ever, the traditional separation between scientific disciplines needs to be bridged to foster interdisciplinary research.

Although more and more researchers with different backgrounds collaborate on large projects, integrating different disciplines is rather a complex process still largely unexplored. Such integration can open up novel scientific avenues of inquiry and, thus, may give birth to novel insights and correlations which can help answer complex questions. Thus, new research methodologies are required to foster interdisciplinary research.

This paper addresses the problem of scientific research analysis across multiple research literature collections. We employ here *cross-collection Latent Dirichlet Allocation* (ccLDA), a recently-introduced model (Paul and Girju, 2009a), and apply it to three novel comparative tasks in the domain of scientific literature: (1) unsupervised discovery and comparison of scientific topics across multiple disciplines; (2) comparison of topics as they appear in different publications and venues within the same discipline; and (3) analysis of topic evolution over time through modeling topics over different time intervals within and across disciplines. We also experiment with trend analysis and propose a novel measurement of topic influence which measures the temporal correlation of related topics over time. Finally, we provide a quantitative evaluation of ccLDA to supplement previous evaluations of this model. We show that ccLDA can achieve competitive performance when used as a generative classifier for a small number of collections.

2 Previous Work

Most of the work on the analysis of scientific research covers what is known as citation analysis (Rubin, 2004) – the use of citations in scholarly works to build a graph with links between works and researchers. This approach, however is limited in that the citation graphs created are sparse and do not span related fields.

Another possibility is to use topic models which uncover structures used to explore text collections, which has been shown to be useful for analyzing scientific research trends (Griffiths and Steyvers, 2004). For example, the area of computational linguistics has been modeled by Hall et al. (2008) who study the history of ideas using LDA and topic entropy. Topic modeling across multiple collections,

ISSN 1135-5948

however, is a little-studied problem. In previous research, we also used LDA (Paul and Girju, 2009b) to study three research fields using a cosine similarity measure to find related topics. However, one important limitation of LDA alone is that it can not explicitly model interdisciplinary topics.

One possibility is Markov topic models (MTM) (Wang et al., 2009), a family of models which can simultaneously learn the topic structure of a single collection while discovering correlated topics in other collections. These models, however, do not explicitly model the similarities and differences between collections as we propose in this research.

We feel that cross-collection LDA (ccLDA) is a good model for this problem because not only does it find topics that are shared among collections, but it models the per-topic differences between the collections.

3 The Model

In this section we first review the basic probabilistic latent semantic indexing (pLSI) and Latent Dirichlet Allocation (LDA) models. We then review *cross-collection LDA* (ccLDA), an extension of LDA, which we will employ, further analyze, and extend in this research.

3.1 Basic Topic Modeling

The most basic generative model that assumes document topicality is the standard Naïve Bayes model, where each document is assumed to belong to exactly one topic, and each topic is associated with a probability distribution over words (Mitchell, 1997).

While this single-topic approach can be sufficient for classification tasks – that is, by modeling each document as a single topic or class, we can use the model to predict the class label of new documents – it is often too limiting for unsupervised grouping of semantically related words into topics. A better assumption is that each document is a mixture of topics. For example, a news article about a natural disaster may include topics about the causes of such disasters, the damage/death toll, and relief aid/efforts. Probabilistic latent semantic indexing (pLSI) (Hofmann, 1999) is one such model. In this model, the probability of seeing the ith word in a document d is:

$$P(w_i|d) = \sum_z P(w_i|topic = z)P(topic = z|d)$$

One of the main criticisms of pLSI is that each document is represented as a variable d and it is not clear how to label previously unseen documents. This issue is addressed by Blei et al. (2003) who introduced the Latent Dirichlet Allocation model. Furthermore, the probabilities under this model have Dirichlet priors, which results in more reasonable mixtures and less overfitting. In LDA, a document is generated as follows:

1) Draw $\phi_z \sim \text{Dirichlet}(\beta)$ for each topic z
2) For each document d, draw a topic mixture distribution $\theta^{(d)}$ from Dirichlet(α). Then for each word w_i in d:
   – Sample a topic z_i from $\theta^{(d)}$
   – Sample a word w_i from ϕ_z

The Dirichlet parameters α and β are vectors which represent the average of the respective distributions. In many applications, it is sufficient to assume that these vectors are uniform and to fix them at a value predefined by the user. In this case, the Dirichlet priors simply function as smoothing factors.

3.2 Cross-Collection LDA

Cross-collection LDA (ccLDA) (Paul and Girju, 2009a) is an extension of LDA for comparing multiple text collections. Each topic is associated with two classes of word distributions: one that is shared among all collections, and one that is unique to the collection from which the document comes. For example, when modeling reviews of different laptops, the topic describing the preloaded software contains the words "software", "application", "programs", etc. in its shared distribution with high probability, and the Apple-specific word distribution contains the words "itunes", "appleworks", and "iphoto".

When generating a document under this model, one first samples a collection c (which is observable in the data), then chooses a topic z according to the document's multinomial topic mixture. One then chooses x (either 1 or 0) to determine whether to draw from the shared topic-word distribution or the topic's collection-specific distribution. The probability of x is a binomial that is dependent on the collection and topic of the current token.

The generative process is thus:

1) For each topic z:
   – Draw $\phi_z \sim \text{Dirichlet}(\beta)$
2) For each topic z and each collection c:
   – Draw $\sigma_{z,c} \sim \text{Dirichlet}(\delta)$
   – Draw $\psi_{z,c} \sim \text{Beta}(\gamma_0, \gamma_1)$
3) For each document d, choose a collection c

and draw a topic mixture $\theta^{(d)}$ from Dirichlet(α_c). Then for each word w_i in d:

– Sample $z_i \sim$ Multinomial($\theta^{(d)}$)
– Sample $x_i \sim$ Binomial($\psi_{z,c}$)
– If $x_i = 0$, sample a word w_i from ϕ_z;
else if $x_i = 1$, sample w_i from $\sigma_{z,c}$

Inference of the model can be done with Gibbs sampling, a basic Markov chain Monte Carlo method (Paul and Girju, 2009a).

4 Experimental Results

In this section we describe the corpus employed in this research and the experimental setup. Then we demonstrate the performance of ccLDA on three novel applications in the domain of scientific literature. We experiment with the effects of different parameter settings, the model's performance in the task of document classification, and automatic methods for cleaning the results. We also compare against the relevant literature.

4.1 Experimental Setup

Paul and Girju (2009) experimented with computational linguistics, linguistics, and education papers. Since we would like to try a new but related field, we consider here psychology instead of education. Our corpus consists of approximately 11,100 abstracts from the ACL Anthology (Bird, 2008), 6,000 abstracts from Linguistics journals, and 6,700 abstracts from Psychology journals. The exact distribution is shown in Table 1. We chose to include journals based on the following criteria for each journal: is a top journal, covers topics in areas that are pertinent to this project, and covers a timespan of at least a decade.

We removed a standard set of stop words as well as words with a corpus frequency less than 10. All punctuation was treated as a word separator.

The hyperparameters of the Dirichlet/Beta priors must be either learned or specified by the user. Our implementation of ccLDA uses a non-uniform α_c for each collection which is estimated automatically using the approach in (Li and McCallum, 2006). We leave the other parameters as predefined constants. We follow the heuristic that $\beta = \delta = 0.01$ is a good value for smoothing topic-word distributions (Griffiths and Steyvers, 2004) and we use standard Laplace smoothing factors such that $\gamma_0 = \gamma_1 = 1.0$.

Unless otherwise specified, in each experiment we ran the Gibbs sampler for a burn-in period of 2000 iterations, then we collected and averaged 15 samples, each separated by a 100-iteration lag.

Field	Venue	# Docs	Years
CL	ACL Journal	943	80-06
CL	ACL Workshops	4,122	80-07
CL	ACL	1,826	79-08
CL	EACL	517	83-06
CL	NAACL	543	01-07
CL	Applied NLP	262	83-00
CL	COLING	1,549	65-08
CL	HLT	872	86-05
CL	IJCNLP	471	05-08
CL	**Total**	**11,105**	**65-08**
LING	Language	379	93-08
LING	Linguistics	152	97-08
LING	Linguistic Inquiry	448	99-08
LING	Journal of American Ling.	449	93-08
LING	Journal of Sociology of Lang.	1,778	76-08
LING	Language & Speech	1,385	58-08
LING	Natural Lang. & Ling. Theory	558	83-08
LING	Ling. & Philosophy	847	77-08
LING	**Total**	**5,996**	**58-08**
PSYC	Applied Cognitive Psychology	1,470	87-09
PSYC	Basic & Applied Social Psych.	849	80-08
PSYC	Cognitive Psychology	475	90-00
PSYC	Eur. Journal of Social Psych.	1,804	71-08
PSYC	Psychological Inquiry	892	90-08
PSYC	Journal of Social Issues	896	90-08
PSYC	Social Cognition	321	96-08
PSYC	**Total**	**6,707**	**71-09**
Total	**Total**	**23,808**	**65-09**

Table 1: Dataset - number of tokens and documents per field and publication venue. CL - Computational Linguistics, LING - Linguistics, PSYC - Psychology.

4.2 Interdisciplinary Research Topic Discovery

Automatic discovery of scientific topics is an important part of modern literature analysis, and topic models like LDA can be used to aid trend analysis and browsing of related literature (Griffiths and Steyvers, 2004). Our contribution is to discover scientific topics that cross disciplines and to see how they compare and differ across fields.

We modeled our corpus of 3 scientific fields – computational linguistics (CL), linguistics (LING), and psychology (PSYC) – with 15 topics[1]. This approach discovers some topics that are at least somewhat related across all three fields. For example, within the topic of words/lexicology, the CL word distribution is largely about word sense disambiguation, the LING distribution is about phonology, and the PSYC distribution is about language acquisition and reading comprehension.

[1]We tried various numbers of topics, but 15 seemed to work the best, as there are not many topics that belong to all three collections. For this task of inducing clusters representing scientific topics, there is no metric with which to optimize this parameter, so tuning the number of topics is largely a "trial and error" approach using our qualitative judgments.

Since the three collections are largely different, however, the model struggled to find topics that fit nicely across all of the collections, and the clusters are fairly noisy. We also tried modeling only two collections at a time (i.e., CL – LING, CL – PSYC, LING – PSYC), which gave much better results. We used 20 topics, determined after some empirical experimentation. If the number of topics is too large, then most of the discovered topics are not shared across both collections.

Tables 2 and 3 show an example of a topic related to communication. We see that in CL, this is strongly relevant to dialogue systems; in linguistics and psychology, this topic is more focused on human behavior and social interaction.

speech information task recognition interaction understanding human users using speaker communication language	
CL	**PSYC**
dialogue	face
spoken	communication
user	cues
systems	facial
utterances	verbal
utterance	presence
dialogues	expression
input	stimuli
domain	spatial
natural	expressions

Table 2: The topic of communication.

These tables also highlight the differences in modeling the collections pairwise compared to modeling all three collections at once. The communication topics formed modeling CL–LING and CL–PSYC are actually quite similar, but the topic is a bit different when formed with all three collections. The topics we get with only two collections tend to be more semantically coherent.

information recognition using interaction task time based communication context current interactive real processing		
CL	**LING**	**PSYC**
user	speech	performance
speech	focuses	memory
systems	spontaneous	task
spoken	utterances	cognitive
dialogue	speaker	effects
input	function	tasks
users	spoken	recall
utterances	discourse	learning
human	utterance	better
utterance	relationship	verbal

Table 3: The topic of communication.

4.2.1 Trend Analysis and Topic Influence

Much work has been done with trend analysis of LDA-induced research topics (Griffiths and Steyvers, 2004) (Hall, Jurafsky, and Manning, 2008). Plots and regression can be used to detect "hot" and "cold" topics by measuring the frequency of documents containing various topics over time.

We are particularly interested in the temporal trends of multiple, related topics across disciplines. For example, if the topic of *semantics* rises or falls in linguistics, does the topic follow suit in computational linguistics? Does interest/disinterest in a topic carry over across research fields? To investigate this problem, we propose a novel measurement of *topic influence* based on temporal correlation.

We say a topic t *influences* a topic T if there is a temporal correlation between them such that the frequency of T rises or falls within 0-2 years of a similar change in t. We define this below as an accumulation of the product of the change in each topic's frequencies over time intervals in which the frequencies are changing either in the same or opposite direction. This measure changes with the size and strength of these similar/dissimilar intervals.

Of course, these statistics alone can not tell us that research in one topic is really influencing the other – that would require an in-depth citation analysis which we leave for future research. However, they are very useful since the result of this measure can direct one's attention to the topics that show a strong statistical influence, whereupon background knowledge and further research can be used to determine if the correlation is causative.

Work has been done to compare the similarity of histograms, but these approaches mainly compare the distance between segments or the similarity of peaks (Strelkov, 2008). They are thus sensitive to the entire shape of the data, which is not appropriate for this research since we do not expect topics from different fields to be highly similar. We also do not use standard correlation coefficients because these compare changes relative to the distribution of the data; instead we want to compare how the frequencies change relative to the previous year.

Thus, we formulate the problem such that

we look for time intervals where the the topic frequencies are changing in the same or opposite directions. Specifically, we compare the derivative of a topic t's frequencies to the second-order derivative of topic T at a given year[2]. We would also like the measure to be influenced more by larger changes, and we would like to assign exponential weight to the contiguous length of these time intervals, since we believe the correlation is much more likely to be causal if it spans many consecutive years. For a contiguous time interval $[i, j]$ in which topics change in either the same or opposite direction over the entire interval, we assign an influence score:

$$C(i, j, d) = \alpha^{j-i} \times \int_i^j \left(\frac{d}{dy} t(y - d) \frac{d^2 T}{dy^2} \right) dy \qquad (1)$$

where $t(y)$ is the frequency $P(t|y)$ and d is an offset by which T lags t. We estimate $\frac{dt}{dy}(y_i)$ as $t(y_i) - t(y_{i-1})$. α is a user-defined parameter that determines how much weight is given to the length of the interval. We compare (the discrete equivalent of) the first derivative of t to the second derivative of T to capture patterns such as when topic t increases and topic T is still decreasing but the rate at which it decreases slows down. (Certainly, if T instead increases, then the correlation will be even stronger.)

If we sum the C values of all such intervals across the timespan we want to compare, we get an influence measure that satisfies the preferences stated above. However, the time-series should first be smoothed, otherwise the year-to-year fluctuations that are natural in this kind of data might prevent similarities from being discovered. We did this by taking a weighted average of the yearly frequencies in overlapping intervals spanning 3 years.

This formula allows the topics to change after a lag d, but what if we do not expect this offset to be constant across the entire timespan? This is natural taken into consideration the time necessary for the information to spread after publication. Thus, we define a lag range $L = [0, 2]$ and look to maximize the possible influence of disjoint intervals. The optimal solution up to point j can be solved with the recurrence:

[2]The intuition is that if the frequency of t is increasing, we want to know if the rate of the frequency of T also increases (that is, even if T is still decreasing, if the rate at which it declines slows down, then the increase from t could be said to be an influence.

speech word phonological words phonetic english prosodic acoustic	
CL	**LING**
recognition	perception
vocabulary	production
recognizer	vowel
continuous	examines
spoken	listeners
news	identification
synthesis	consonants
automatic	frequency
transcription	vowels

Table 4: The topics of speech recognition in CL and speech in LING, which are related.

$$Opt(j) = \max_{1 \le i \le j; d \in L} \Big(\sum_{[a,b] \in [i,j]} C(a, b, d) + Opt(i - 1 - d) \Big) \qquad (2)$$

$[a, b] \in [i, j]$ refers to each subinterval $[a, b]$ in the time range $[i, j]$ such that $\frac{d}{dy} t(y-d) \frac{d^2 T}{dy^2}$ has the same sign for every $y \in [a, b]$.

We say that $Inf(T, t) = Opt(n)$ is the *influence* of t on T, where n is the last year in the timespan considered.

This can be efficiently solved with dynamic programming. The complexity is dominated by precomputing the C values for different time ranges $[i, j]$ and different d values. Instead of explicitly finding each interval $[a, b]$ in the time range, we simply iterate from $y = i$ to j, summing $\frac{d}{dy} t(y - d) \frac{d^2 T}{dy^2}$ along the way. When the sign of $\frac{d}{dy} t(y-d) \frac{d^2 T}{dy^2}$ changes, we multiply this sum by α^l where l is the length of this segment, then we add this to the total and reset the sum and l to 0. We can treat d as a small constant: there are $O(n^2)$ intervals $[i, j]$ for which we must compute C in this manner (i.e., $O(n)$ time). Thus we can naïvely calculate $Inf(n)$ in $O(n^3)$. There is a less naive way to compute this in $O(n^2)$, but we save the details of this for future research.

Figure 1 shows a pair of topics (as in Table 4) with one of the top influence scores.

To evaluate the influence measure, we presented unlabeled plots over time of the 10 pairs with the highest influence scores and asked two judges to determine ("yes" or "no") if they show a clear correlation over some time interval(s). The judges agreed "yes" on 70% of the examples.

To summarize, in this subsection we introduced the measure of influence which finds disjoint intervals such that the change in fre-

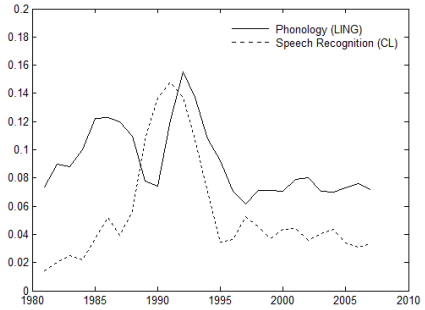

Figure 1: Speech recognition has a strong influence on phonetics which tends to rise and fall about 2 years after a similar trend in speech recognition (early 90s). The small rises in speech recognition in the 2000s are also mimicked by phonetics after a short delay. This corresponds to the tremendous market opportunities which emerged for speech recognition in the 90s.

quency of topic T has either the same or opposite sign as that of t 0-2 years earlier, where the year offset is variable. The measure sums the product of the changes and multiplies the summation over each interval by a factor that is exponential in the length of the interval.

4.3 Comparing Publication Venues

Partitioning documents by venue, we may be able to detect editorial differences between different journals and conferences. Since workshops often focus on very specific topics, we thought it would be interesting to compare documents from the ACL workshops with the ACL main conference.

We modeled these two collections with 50 topics, and while none of the topics showed stark differences, we did see some specialized topics that were more prevalent in workshops. For example, bioinformatics is likely to appear more in the workshops than the main conference, as indicated by the information extraction topic, whose workshop-specific distribution contains words like *protein*, *genes*, and *biomedical*.

Table 5 shows the topic of word sense disambiguation (WSD). There are not many words assigned to the collection-specific distribution for the main conference, which covers the topic broadly. However, from the second collection-specific distribution, we see that WSD-related shared tasks and competitions are much more likely to take place at workshops than the main ACL conference.

We also compared two different conferences, and thus modeled ACL and COLING, again with 50 topics. We again did not find prominent differences in most of the topics, but there were some. For example, in the *user interfaces* topic, we find that few tokens are assigned to the collection-specific distribution for COLING and thus the distribution seems somewhat arbitrary. However, the ACL distribution contains top words such as *audio*, *video*, *captions*, and *restaurants*. From this, one might infer that ACL publishes more technical details and business applications in this topic than COLING.

disambiguation sense word words sense lexical context	
Conference	**Workshops**
thought	semeval
avoided	participated
diab	subtask
suggest	competition
counting	tries
finer	tagging
counting	team
heuristics	participation

Table 5: Comparing the word-sense-disambiguation topic between the ACL main conference and the ACL workshops.

4.4 Topic Evolution Over Time

Analyses such as the temporal correlation above measure the *frequency* of a topic over time, and not the *character* of a topic over time, which is also important.

Mei and Zhai (2005) model how topics *change* over time, by partitioning the data into time periods and modeling topics in each time period. They discover related topics across time periods using KL-divergence, a measure of the similarity of two probability distributions. Dynamic topic models (Blei and Lafferty, 2006) also model topics as they appear in different time partitions, but instead of capturing topic relations using a post-hoc similarity measure, the topic similarity is modeled directly using a logistic normal prior over the word distributions, resulting in a smooth topic evolution over time.

However, both of these models will only give us a snapshot of how the topics appear in a given interval, and they do not explicitly model the main differences between the intervals, which would be useful for qualitatively evaluating the key changes over time. ccLDA can be applied to this task in a similar manner – by partitioning the data into time intervals – but since ccLDA explicitly models what is unique to each interval, we think it is

parsing grammar tree parser grammars free context syntactic parse structure	
Old	**New**
number	dependency
result	probabilistic
corresponding	stochastic
networks	treebank
known	pcfg
binding	constraint
lr	lexicalized
introduce	ccg
consider	projective
transformational	robustness

Table 6: The grammar topic compared across two time intervals.

better suited for this problem.

Here we partitioned the computational linguistics documents by their decade of publication. We discarded papers published before 1980, and thus we could compare three decades: the 80s, the 90s, and the 2000s. We also experimented with partitioning the data in two: "new" publications (year ≥ 2000) and "old" publications (everything else).

The collection-specific clusters of ccLDA show terms that are more likely to appear in that collection than the others. In this experiment, that means topical words that are unique to that time period – in the most recent time period, this will give us which terms are novel or newly hot within a topic; in the other time periods, this will give us once-popular terms which are no longer largely researched.

For example, we see *webster* in the topic of lexical resources in the 1980s, but by the 2000s, the top words in this topic are *web*, *ontology*, *wordnet*, etc. We find once again, however, that the clusters are somewhat noisy with three collections, and we see better results when we use only two collections, "new" and "old". A sample topic from this set is shown in Table 6. Other examples are the *learning* topic, which has *svm* as the top word in the new distribution and *neural networks* at the top of the old distribution. In *machine translation*, we see that comparable corpora and automatic alignments are prominent research areas in the more recent time interval.

4.5 Evaluation

In our previous work, we evaluated ccLDA with human judgments of cluster coherence and by measuring the model log-likelihood of held-out data compared against other models. We would like to take this opportunity to perform another evaluation of ccLDA, this time by applying it to a prediction task.

The main thing we would like to glean from each experiment is the set of terms within each topic that are good descriptors of what is unique to each collection (e.g., a research field, a publication venue, or a time period). We can quantitatively evaluate the model's ability to do this by applying it to the task of collection prediction, which will give us a measure of how discriminative the collection-dependent word distributions are. In this subsection, we will use ccLDA to classify a document according to its time period.

Because ccLDA gives a document likelihood that depends on the document's collection or class, it is naturally suited for this task. Classification of an unlabeled document d thus becomes the problem of choosing the c that maximizes the formula:

$$P(c) \prod_{w \in d} \sum_{z} P(z|d) \quad [P(x=0|c,z)P(w|z,x=0) + P(x=1|c,z)P(w|z,c,x=1)]$$

These probabilities are obtained when the model is learned on a training set, except for $P(z|d)$, which depends on the new document. We can learn this through another Gibbs sampling procedure, treating the document as if c is known and doing this for all values of c, however, the ability to quickly label a new document is necessary for many classification tasks, so we instead use a simple approximation from the learned Dirichlet prior for each collection, which represents the average topic mixture within that collection. That is, $P(z|d) \approx \frac{\alpha_{cz}}{\sum_z \alpha_{cz}}$.

To see how important $P(z|d)$ is to the performance, we also experimented with approximating this as a uniform constant, $P(z|d) = \frac{1}{Z}$ where Z is the number of topics.

Table 7 shows the 5-fold cross-validation accuracy of the old vs. new set from the previous experiment, compared against that of an optimally tuned SVM. In each cross-validation iteration, the data is partitioned in the same way for each classifier; that is, they are evaluated with the same training/test sets. We used the SVM^{light} and $SVM^{multiclass}$ kits [3] with the regularization factor (that is, the trade-off between margin

[3] http://svmlight.joachims.org

size and training error) set to the default[4] $\frac{1}{x^2}$. ccLDA was run with 50 topics.

	SVM1	SVM2	ccLDA1	ccLDA2
P	0.793	0.754	0.792	0.781

Table 7: The precision obtained by various classifiers during 5-fold cross-validation on the "new" vs. "old" dataset. SVM1 refers to a support vector machine using the regularization parameter $C = \frac{1}{x^2}$; SVM2 uses $C = 1.0$. ccLDA1 uses the method described above with an approximation for $P(z|d)$ based on α_c. ccLDA2 uses a topic-independent, uniform approximation of $P(z|d)$.

Due to its ability to separate out less-discriminative words by way of the collection-independent model, ccLDA achieves comparable performance to the SVM.

5 Conclusions

This paper addresses the problem of scientific research analysis across multiple research literature collections. We employ a recently-introduced model, *cross-collection Latent Dirichlet Allocation* (ccLDA), and apply it to three novel comparative tasks in the domain of scientific literature. We also experiment with trend analysis and propose a novel measurement of topic influence which measures the temporal correlation of related topics over time. We evaluate ccLDA on the task of document classification, which yields performance comparable to an optimally-tuned SVM. Ultimately, we show that ccLDA has potential for a variety of applications in the important domain of scientific research analysis, and could be a valuable component in fostering interdisciplinary research. As future work, we will experiment with other collections and languages.

References

Bird, S. 2008. Association for Computational Linguistics Anthology. In *http://www.aclweb.org/anthology-index/*.

Blei, D. and J. Lafferty. 2006. Dynamic topic models. In *The 23rd International Conference on Machine Learning (ICML)s*, pages 113–120.

Blei, D., A. Ng, and M. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3.

Griffiths, T. and M. Steyvers. 2004. Finding scientific topics. In *The National Academy of Sciences of the United States of America*.

Hall, D., D. Jurafsky, and C. Manning. 2008. Studying the history of ideas using topic models. In *Empirical Natural Language Processing Conference*.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *The 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

Li, W. and A. McCallum. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *International Conference on Machine Learning*.

Mei, Q. and C. Zhai. 2005. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *The 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 198–207.

Mitchell, T. 1997. *Machine Learning*. McGraw-Hill, Boston.

Paul, M. and R. Girju. 2009a. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *The Conference on Empirical Methods in Natural Language Processing*, pages 1408–1417.

Paul, M. and R. Girju. 2009b. Topic modeling of research fields: An interdisciplinary perspective. In *The International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Rubin, R. 2004. *Foundations of Library and Information Science*. 2nd ed. New York: Neal-Schuman.

Strelkov, V.V. 2008. A new similarity measure for histogram comparison and its application in time series analysis. *Pattern Recognition Letters*, 29(13):1768–1774, October.

Wang, C., B. Thiesson, C. Meek, and D. Blei. 2009. Markov topic models. In *The 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 583–590.

[4] We did try many other settings and found this to consistently give the best accuracy.