Humanities Research Recommendations via Collaborative Topic Modeling

Nitya Mani and Andy Chen

***Abstract*— We present two novel applications of collaborative topic modeling to the broad datasets of humanities research article recommendations. In the first, we present an adaptation of the semi-supervised collaborative topic regression model to a situation in which no user feedback by simulating users to develop a much better content-based recommendation model (over 95% precision and relevant recall) than several implemented in the status quo, including the recommendations platform implemented by eScholarship, host to the International Journal of Comparative Psychology. In the second, we demonstrate how differential weightings on algorithm parameters can be used to provide relevant recommendations for humanities researchers based on sparse, noisy, varied information and a small dataset.**

I. INTRODUCTION

Currently, academically motivated parties (whether for research, industry purposes, or casual interest) have access to a wealth of scholarly information to meet their information needs. In fact, in the current age with thousands of articles and papers constantly being published in hundreds of journals, most conducting research have the opposite problem of having to sift through too much, often not incredibly relevant information to find what they are looking for. Thus, automated recommendation platforms are becoming increasingly more relevant to uncovering helpful resources and potentially interesting articles that cannot simply be found by following a trail of citations or an unfocused keyword search.

Current article recommendation platforms generally fall under one of two umbrella categories. Content-based models [4] seek to understand the content of an article and compare that to the content of articles a user is interested when making recommendations. Filtering-based models [8] seek to identify users similar to the current user and thus make article predictions without ever modeling the content of the article.

Collaborative topic modeling is a class of recommendation algorithms that combine topic modeling of articles with implicit feedback from users (i.e. information about user's article preferences that is not based on an explicit ranking system) or explicit evaluations of articles. However, the majority of such algorithms currently in place tend to be heavily skewed in favor of one model over another. Collaborative topic modeling has been most prominently used in news article recommendations that focus on filtering and user similarity matrices, only modeling content for broad keywords and for articles released within the hour. On the other hand, scientific article recommendations tend to heavily rely on topic modeling based on the content of the title and abstract, given that these snippets tend to be filled with keywords that give fairly accurate insight into the content of the article [10], [1], [2].

In this paper we seek to apply an adaptation of the collaborative topic regression model to make recommendations for humanities research and nonfiction writing based on a combination of implicit network and user feedback and topic modeling. Non-scientific academic publications and nonfiction writing have topics that can be modeled somewhat effectively, but abstracts and titles tend to be metaphorical in nature and less clustered around a small number of technical terms and clear cut ideas. Further, articles tend to be concentrated around specific niches with respect to readership and reader interest. On the other hand, such publications also tend to come with lower readership and citation count, and thus any recommendation platform needs to be robust with respect to relatively small amounts of implicit feedback. Thus, the different environments of humanities research are not necessarily optimally modeled by either a purely content or filtering-based model or the same parameters and combinations effective for high-volume article sites or technical STEM papers. Finally, we will examine if a similar model can be used to effectively simulate users to iteratively update purely content-based recommendation platforms.

II. THEORETICAL BACKGROUND

A. Topic Modeling via Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [4] is a generative probabilistic model for text corpora and other collections of discrete data. Unlike other information retrieval schemes such as tf-idf, the LDA model reveals aspects of interdocument statistical structure in the corpus. Each document is modeled as a mixture of topics, where each word is assigned to one of those topics. Here, each document represents a "bag of words" i.e. sentence structure does not play a role in the model. More precisely, given M documents, suppose we have K topics $\beta_1, \ldots, \beta_K$, each of which represents a distribution over a fixed vocabulary V. This vocabulary should be free of non topic-specific stop words, such as pronouns or common verbs. Given fixed hyperparameters α, β, ξ, the generative model is then as follows: for each document W_i in the corpus:

1) Choose the word length $N_i \sim \text{Poisson}(\xi)$
2) Assign a topic distribution $\theta_i \sim \text{Dirichlet}(\alpha)$ over the K topics.
3) For each word $w_{ij} \in W_i$:
   a) Choose a topic $z_{ij} \sim \text{Multinomial}(\theta_i)$.
   b) Choose the word $w_{ij} \sim p(w_{ij}|\beta_{z_{ij}})$, which is the multinomial probability conditioned on the topic $\beta_{z_{ij}}$.

The goal of this LDA algorithm, given the value of the hyperparameter α, is to maximize the likelihood of the corpus data with respect to β and the θ_i. To do so, we use the EM algorithm to learn the K topics $\beta_1, \ldots, \beta_K$ and topic distributions $\theta_1, \ldots, \theta_M$. Note that our unsupervised technique is not a clustering technique on topics, as each document can contain words in different topics.

Note also that the probability of generating each document is $P(N_i)P(W_i|N_i)$, so $\text{argmax}_{\beta,\theta} \log P(W_i) = \text{argmax}_{\beta,\theta_i}(\log P(N_i) + \log P(W_i|N_i)) = \text{argmax}_{\beta,\theta_i} \log P(W_i|N_i)$, as N_i is drawn from a Poisson distribution independent of β and θ_i.

B. Collaborative Filtering via Matrix Factorization

Collaborative filtering [6] to find recommendations for a certain user involves looking at the preferences of other similar users. Suppose there are I users and J articles. If users i_1 and i_2 have similar interests, then for each article j, collaborative filtering can look at whether user i_2 recommends article j and use that information to determine whether user i_1 should read article j.

Matrix factorization for recommendations is a latent factor model, in which manifest variables relate to latent, or nonobservable, variables. In this scenario, we only observe the set of all r_{ij}, which represents the rating that user i gives article j. Notice that while a high rating is unambiguous, a low value can symbolize one of two situations:

- User i has read the article and does not recommend the article to others.
- User i has never seen the article and thus cannot recommend it.

Therefore, the goal of the CTR method is to change zero entries of the second type into predictions about whether user i *would* recommend article j. This distinction is especially critical in light of the fact that the majority of user-document pairs would fall under the latter, rather than former category.

We represent both users and items as latent K-dimensional vectors u_i and v_j, where K is significantly smaller than the number of users or articles. We predict new ratings by computing:

$$\hat{r}_{ij} = u_i^T v_j \tag{1}$$

The goal of the algorithm is to minimize the least squared error over all user-article pairs. If $U = \{u_i\}_{i=1}^{I}$ is the set of all user vectors and $V = \{v_j\}_{j=1}^{J}$ is the set of all article vectors, then the algorithm finds:

$$\underset{U,V}{\operatorname{argmin}} \sum_{i,j} \left((r_{ij} - u_i^T v_j)^2 + \lambda_u ||u_i||^2 + \lambda_v ||v_j||^2 \right) \tag{2}$$

where λ_u and λ_v are regularization parameters.

We use probabilistic matrix factorization (PMF) to model the generation of user and article vectors, as it scales linearly with the number of observations and performs well with large, sparse data. [3] The generative process for producing the user and article data is as follows:

1) For each user $i = 1, \ldots, I$, choose a latent user vector $u_i \sim \mathcal{N}(0, \lambda_u^{-1} I_K)$.
2) For each article $j = 1, \ldots, J$, choose a latent article vector $v_j \sim \mathcal{N}(0, \lambda_v^{-1} I_K)$.
3) For each user-pair (i, j), assign a rating $r_{ij} \sim \mathcal{N}(u_i^T v_j, c_{ij}^{-1})$, where c_{ij} is the precision parameter.

The precision parameters c_{ij} measure the confidence of the rating r_{ij}. As mentioned previously, a high rating r_{ij} unambiguously represents a positive rating that user i gives to article j; therefore, c_{ij} should be high in magnitude. However, a lower rating r_{ij} can symbolize multiple scenarios, so c_{ij} should be lower in magnitude. Specifically, in our algorithm, the input data r_{ij} consists of only binary values, so we assign, for hyperparameters $0 \leq a < b$:

$$c_{ij} = \begin{cases} a & \text{if } r_{ij} = 1 \\ b & \text{if } r_{ij} = 0 \end{cases}$$

(Note, however, that each $u_i^T v_j$ can be a decimal value.) To find the optimal values of U and V given a set of binary r_{ij}, we use gradient coordinate ascent on each u_i and v_j to find U and V in Equation 2. We may then generate a set of $\hat{r}_{ij} = u_i^T v_j$ to use as the prediction ratings.

CTR on its own, however, cannot make accurate recommendations for articles that few or no users have seen. Therefore, we must complement the CTR method with LDA topic modeling in our model.

III. The Regression Model

Collaborative topic regression [10] models users as having interests based on implicit article "recommendations" and models documents with topic proportions θ_j naively learned from Latent Dirichlet Allocation. Thus CTR is as a regression model is able to differentiate document topics that characterize content from those that might characterize readership interest of a body of academics. CTR uses an algorithm in the vein of expectation maximization to indirectly learn the MAP estimates of the log likelihood function of all the parameters being estimated U, V, θ given the initial conditions on the LDA and matrix factorization models.

Matrix factorizations learns short feature vectors to represent each user and document, and predicts the recommendation status of the pair user i and document j as $\hat{r}_{ij} = u_i^T v_j$. When incorporating content modeling of the documents, we continue to predict $u_i^T v_j$, but here $v_j = \theta_j + \epsilon_j$, where θ_j is the topic proportion learned via LDA and ϵ_j is a latent error variable to offset θ_j, the purely content based proportion that enables the document's latent vector to diverge from θ_j. Thus as more users rate an article, the prediction becomes increasingly dependent on the recommendations of users and less so on the LDA model of the document proportions. The CTR model has very strong similarities to collaborative filtering as can be seen in the generative process that characterizes the model and its assumptions about how these articles and feedback are generated. Assume that we begin with K topics derived from an LDA analysis of the documents $\beta_1 .... \beta_k$:

1) For each user $i = 1, \ldots, I$, choose a latent user vector $u_i \sim \mathcal{N}(0, \lambda_u^{-1} I_K)$.
2) For each article $j = 1, \ldots, J$ choose a latent article vector $v_j \sim \mathcal{N}(\theta, \lambda_v^{-1} I_K)$.
3) For each document j, select word $w_n^{(j)} \sim \text{Mult}(\beta_{\alpha_n^{(j)}})$ where $\alpha_n^{(j)} \sim \text{Mult}(\theta_j)$
4) For the user-document pair (i, j), assume the rating can be modeled as $r_{ij} \sim \mathcal{N}(u_i^T v_j, c_{ij}^{-1})$

Given this generative model, the expectation of r_{ij} is (similar to collaborative filtering) $\mathbb{E}(r_{ij}) = u_i^T v_j$, with the primary difference in how we model the latent document vector, incorporating the content-based proportion: $v_j = \theta_j + \epsilon_j$ where $\epsilon_j \sim \mathcal{N}(0, \lambda_v^{-1} I_K)$. Thus, learning each of these parameters can be done by finding a maximum a posteriori estimate of u_i, v_j, θ_j, and r_{ij}, where he can find the MAP estimates of U, V, R using coordinate ascent. We compute the MAP estimates by maximizing the log likelihood of the data (in particular the overall log likelihood of each of $U, V, R, \theta_1 ... \theta_j$), minimizing the least squared error of our eventual prediction with regularization:

$$\begin{aligned} & LL(U, V, \theta_{1:J}, R; \beta, \lambda_u, \lambda_v) \\ & = -\frac{\lambda_u}{2} \left(\sum_{i=1}^{I} ||u_i||_2^2 + \sum_{j=1}^{J} ||v_j - \theta_j||_2^2 \right) \\ & \quad + \sum_{j=1}^{J} \sum_{n} \log \left(\sum_{k=1}^{K} \theta_k^{(j)} \beta_{k, w_n^{(j)}} \right) \\ & \quad - \frac{1}{2} \sum_{i=1}^{J} \sum_{j=1}^{J} c_{ij} (r_{ij} - u_i^T v_j)^2 \end{aligned}$$

Then, we can iteratively maximize this function using coordinate ascent by setting the gradient of the log likelihood to 0 and determining the the new optimal values for each user and document latent vector u_i and v_j as:

$$\begin{aligned} u_i &:= (V \text{diag}(c_{ij}|_{1:J}) V^T + \lambda_u I_K)^{-1} V \text{diag}(c_{ij}|_{1:J}) R_i \\ v_j &:= (U \text{diag}(c_{ij}|_{1:I}) U^T + \lambda_v I_k)^{-1} (U \text{diag}(c_{ij}|_{1:I}) R_j + \lambda_v \theta_j) \end{aligned}$$

Here $R_i = (r_{ij})|_{j=1}^{J}$ and $R_j = (r_{ij})|_{i=1}^{I}$. For the moment, we will fix θ_j as the original LDA proportions and treat the proportion vectors as constants.

IV. Making Recommendations

For the moment, we predict the expected rating $\hat{r_{ij}} = u_i^T v_j$ where $v_j = \theta_j$ if there is no user information about document j. We will eventually employ an algorithm in the style of [7] to rank our recommendations.

V. Empirical Study: Simulating Recommendations with eScholarship

Collaborative topic regression and similar models have been studied in a wide variety of largely scientific contexts ([10], [5], [9]) as a mechanism by which to incorporate implicit and explicit user feedback about documents into a recommendation system. However, even in situations in which little or no user data has been collected, collaborative topic regression and similar models can be applied to iteratively update existing recommendations. Here, we reimagine the model in a novel context, one in which no real user-feedback is present, but where we can simulate users to improve traditional content-based recommendations. We consider by way of example the International Journal of Comparative Psychology, hosted on the site eScholarship with the journal issues. For each journal article, the website provides a listing of approximately 20 ”similar articles” generated by a content modeling process that selects other journal articles most similar in content/keywords to the one being viewed.

However, many of these recommendations are largely if not completely irrelevant to the content of the article. Thus, we sought to improve these recommendations, by using CTR on a set of simulated recommendations to iteratively update these ”similar article” recommendations. We simulated users by considering each list of 20 similar articles corresponding to a particular article to be recommendations of a user and then applying CTR on this set of articles and ”user” (note that we discard the article from which we generated the simulated user) to arrive at a new list of recommendations for this user, which hopefully present a better set/ranking of similar articles to the original article.

In order to test this, we gathered data from 580 ”users” and 4827 articles from the International Journal of Comparative Psychology. Each ”user” had a total of 20 similar articles, from which we isolated the odd numbered ones as user recommendations. This gave a total of 5800 user-item observed pairs. Experimentally, we considered a variety of values for the model hyperparameters to optimize precision and recall using a restricted grid search, settling on $K = 100$, $\lambda_u = 0.01$, $\lambda_v = 0.1$, $a = 1, b = 0.01$. Some example topics yielded have top words ’species patterns california populations habitat’, ’public policy states issues economic’, and ’expression gene genetic function levels’. Some of the hyperparameter search statistics for K are pictured below:

Some sample data for hyperparameter configurations illustrates more completely the relative performance and confidence of our model in predicting training, witheld testing, and new articles in its top recommendations.

As desired, the learning model predicted all of the user-recommended articles with high ratings, and 100% of the time recommended the original article from which the ”user” was created with a rating at least 0.75 or in the top 20 recommendations (note that we withheld these articles for testing purposes from the training data set). Additionally, 100% of the time, the top 3 withheld recommendations were recommended by the CTR algorithm. Overall, the recall on the relevant withheld and provided recommendations was 95%. As the data suggest, not all of withheld articles were predicted, in part because the content-based model used by the International Journal of Comparative Psychology often made predictions irrelevant to the bulk of the other recommended articles that were thus discarded. In this sense, our model performed far better in recommending relevant articles.

To compute the precision quantitatively, we took a random sample of 20 of the 580 users and classified each of the original recommendations from the International Journal of Comparative Psychology as + (relevant) or − (irrelevant) by hand, where only articles clearly about a different subject entirely were marked as −. Then, we saw that in aggregate, our model correctly recommended (with a rating at least 0.75 or in the top 20 recommendations) over 90% of the + articles, and less than 5% of the − articles. As a

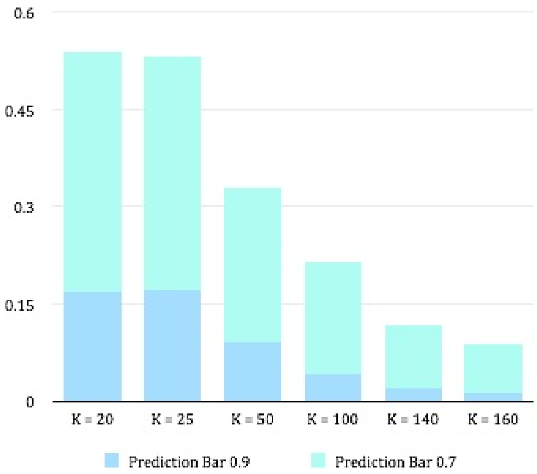


Fig. 1. This graph shows the accuracy of model on the withheld testing data for a variety of choices for the number of topics/size of latent feature vectors K. Given the irrelevance of many of the articles recommmended by the journal, selecting $K = 100$ optimized the recall of relevant articles with respect to the precision of the overall model, and was high enough to ensure user-provided recommendations were maintained with high confidence.

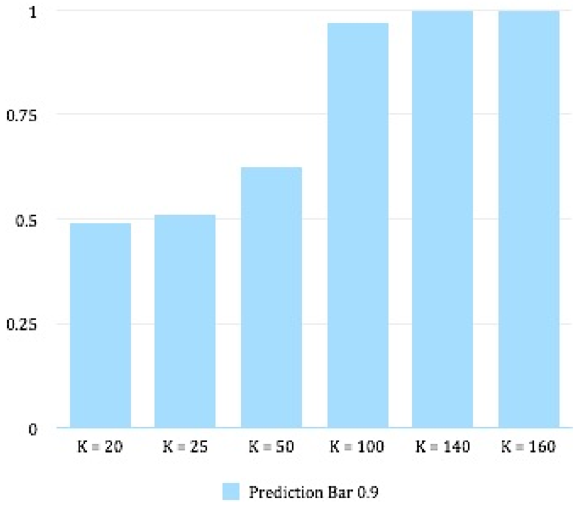


Fig. 2. This graph shows the model accuracy on the training data with respect to the number of topics/size of feature vectors K when conducting the hyperparameter search. Note that for the purposes of low computation costs, parameters were treated as independent when conducting this search and were sequentially optimized although a grid search would have likely been more optimal.

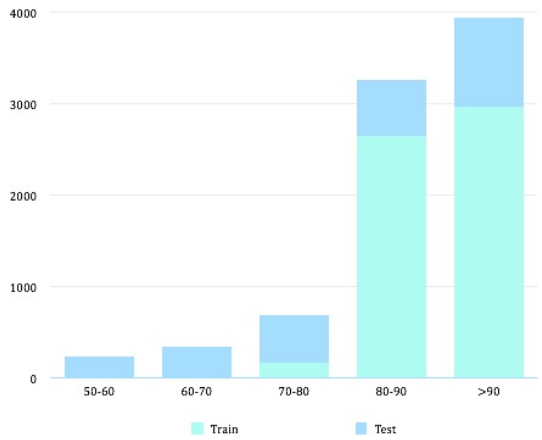


Fig. 3. This graph shows the proportion of documents that received predicted rating r_{ij} in the intervals depicted on the x-axis for $K = 25$, where a higher rating corresponds to a prediction that user i is more likely to like document j. The blue bars represent training data and the green bars represent withheld "user" information. As the data suggest, the provided recommendations received high scores (as they should given that a user has expressed interest already) and many of the withheld documents (around 50%) occurred as a top 20 prediction for the article.

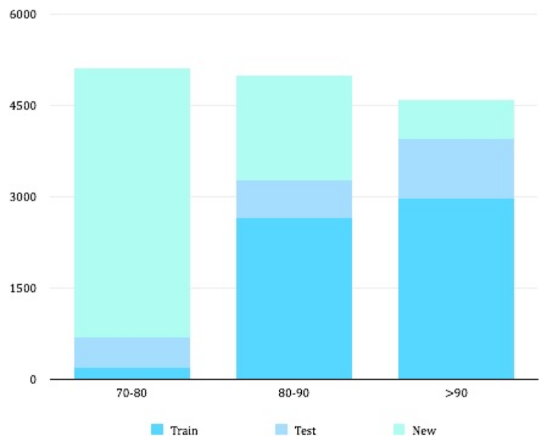
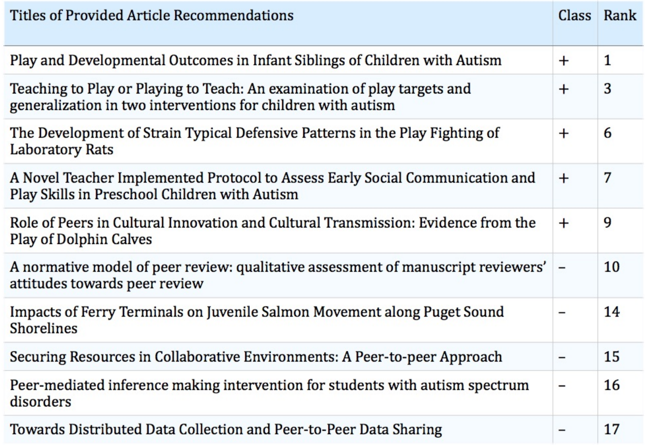
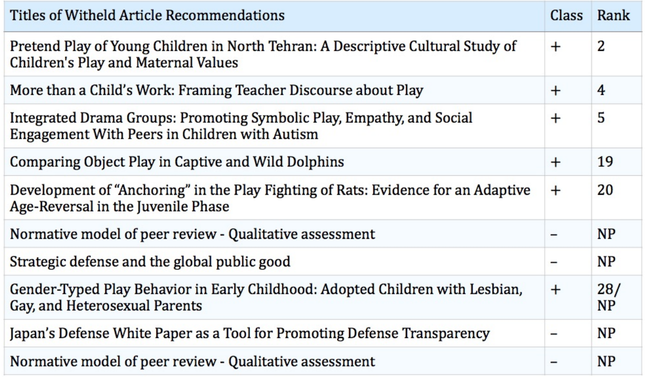


Fig. 4. This graph shows the proportion of documents that received predicted rating r_{ij} in the intervals depicted on the x-axis for $K = 25$, including new predictions of articles not originally predicted by the Comparative Psychology journal model with the yellow bars. Given the sparseness of the training matrix and also the high confidence in predicting original "user" recommendations, there are a large number of new articles predicted, especially as the confidence decreases.



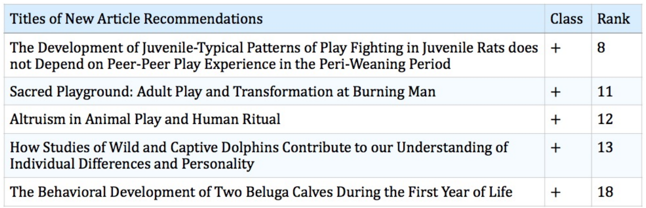
Titles of Provided Article Recommendations	Class	Rank
Play and Developmental Outcomes in Infant Siblings of Children with Autism	+	1
Teaching to Play or Playing to Teach: An examination of play targets and generalization in two interventions for children with autism	+	3
The Development of Strain Typical Defensive Patterns in the Play Fighting of Laboratory Rats	+	6
A Novel Teacher Implemented Protocol to Assess Early Social Communication and Play Skills in Preschool Children with Autism	+	7
Role of Peers in Cultural Innovation and Cultural Transmission: Evidence from the Play of Dolphin Calves	+	9
A normative model of peer review: qualitative assessment of manuscript reviewers' attitudes towards peer review	–	10
Impacts of Ferry Terminals on Juvenile Salmon Movement along Puget Sound Shorelines	–	14
Securing Resources in Collaborative Environments: A Peer-to-peer Approach	–	15
Peer-mediated inference making intervention for students with autism spectrum disorders	–	16
Towards Distributed Data Collection and Peer-to-Peer Data Sharing	–	17

Fig. 5. This table shows the rankings of the provided article recommendations, where as illustrated irrelevant articles received lower rankings than more relevant articles



Titles of Witheld Article Recommendations	Class	Rank
Pretend Play of Young Children in North Tehran: A Descriptive Cultural Study of Children's Play and Maternal Values	+	2
More than a Child's Work: Framing Teacher Discourse about Play	+	4
Integrated Drama Groups: Promoting Symbolic Play, Empathy, and Social Engagement With Peers in Children with Autism	+	5
Comparing Object Play in Captive and Wild Dolphins	+	19
Development of "Anchoring" in the Play Fighting of Rats: Evidence for an Adaptive Age-Reversal in the Juvenile Phase	+	20
Normative model of peer review - Qualitative assessment	–	NP
Strategic defense and the global public good	–	NP
Gender-Typed Play Behavior in Early Childhood: Adopted Children with Lesbian, Gay, and Heterosexual Parents	+	28/ NP
Japan's Defense White Paper as a Tool for Promoting Defense Transparency	–	NP
Normative model of peer review - Qualitative assessment	–	NP

Fig. 6. This table shows the rankings of the withheld articles, with NP indicating that the withheld article in question was not predicted as a similar article of interest by our CTR algorithm. Note that all of these withheld articles with the possible exception of 1 that were not recommended were classified as — or irrelevant to the user.



Titles of New Article Recommendations	Class	Rank
The Development of Juvenile-Typical Patterns of Play Fighting in Juvenile Rats does not Depend on Peer-Peer Play Experience in the Peri-Weaning Period	+	8
Sacred Playground: Adult Play and Transformation at Burning Man	+	11
Altruism in Animal Play and Human Ritual	+	12
How Studies of Wild and Captive Dolphins Contribute to our Understanding of Individual Differences and Personality	+	13
The Behavioral Development of Two Beluga Calves During the First Year of Life	+	18

Fig. 7. This table shows the rankings of the new recommendations (within the top 20) provided by our recommendation platform that were not present in the list of similar items generated by the IJCP content-based recommendation.

case study, consider user 19, created from the recommendations generated for the article 'The Development of Juvenile-Typical Patterns of Play Fighting in Juvenile Rats does not Depend on Peer-Peer Play Experience in the Peri-Weaning Period'. Below is a table summarizing the results we obtained as far as the journal recommendations. Our model predicted 8 new articles not included in this set, the highest rank of which (rank 8) was the original article from which we drew the simulated user data, and which included all 7 other articles scored as a +, like 'Altruism in Animal Play and Human Ritual' and 'How Studies of Wild and Captive Dolphins Contribute to our Understanding of Individual Differences and Personality'.

Thus our final model was able to achieve around 95% precision and relevant recall on the dataset, making it a far better article recommendation platform than the existing content-based platform employed by the International Journal of Comparative Psychology, that recommended at least 30% irrelevant () articles for each of the randomly sampled papers.

VI. Empirical Study: Humanities Research with CiteULike

After simulating user data, we implemented our algorithm on user-given ratings data. For this scenario, the user-article interactions are much sparser and noisier than in the first scenario; while the simulated users for eScholarship each had at least ten recommendations, the average number of recommendations in our

CiteULike dataset is roughly 6.4, with most users recommending fewer than 5 articles. In addition, users rarely recommended articles that were all in the same topic.

We also expanded the diversity of our article corpus by not limiting our articles to one journal; while the eScholarship articles primarily originated from *International Journal of Comparative Psychology*, our CiteULike articles were found in journals ranging from *Latin American Research Review* to *Asian Theatre Journal*. Some were even written in foreign languages (see more details in the discussion section). These articles, compared to scientific articles, collectively had fewer abstracts. These humanities abstracts also tended to be less summary-focused.

We collected all 2115 of the user profiles whose declared research areas lay in European, Eastern, Asian, African, American, or Australasian language or literature studies. Of these, 223 had at least one article in her personal library; collectively, the set of users had 1269 articles. Like in the previous empirical study, we withheld half of our collected user-article interactions (ratings) to reserve for the test set. Therefore, the training set of user-article interactions consisted of 715 instances of a user recommending an article.

As before, we used grid search to find the hyperparameter values that maximized our precision and recall. For each set of hyperparameters, we ran LDA-CTR on the training data and produced recommendations for our set of users. For this study, the hyperparameter values that optimized our precision and recall were $K = 40, \lambda_u = 0.01, \lambda_v = 100$, and $c_{ij} = 1, 0.01$, where $c_{ij} = 1$ when user u_i recommended v_j and $c_{ij} = 0.01$ otherwise.

As previously mentioned, we hide half of the users' ratings and use them to evaluate our algorithm's recall performance; to calculate recall, for each user u_i, we compute how many articles in the entire dataset that user u_i rated positively, both in the hidden and training halves of the recommendation data. We then take an average of our results. Our algorithm then has a 64% recall rate, which means that our algorithm predicts at least 28% of the hidden articles (note that it is possible for our algorithm to choose to not recommend articles associated with u_i supplied in the training set). Though the nominal value is low, our algorithm performs relatively well compared to other recommendation systems.

To consider our algorithm's performance in the context of CiteULike's current recommendations, we analyze our algorithm's accuracy only for users who have rated at least 20 articles; users with accounts can only receive recommendations after adding at least 20 articles to their libraries. We calculated precision in a similar manner as in the previous empirical study; for each user in a random sample, we classified the recommendations for which the algorithm-provided rating were above 0.75 for that user. We then manually classified the recommendations as relevant (+) or irrelevant (-) in a similar manner as in the previous experiment. With this metric, 89% of the algorithm's recommendations fell into the (+) category, giving us an 89% precision value.

VII. Discussion

Based on our analysis, we conclude that composing LDA topic modeling with collaborative filtering significantly improves the existing recommendations from eScholarship's *International Journal of Comparative Psychology*. For each psychology article, our algorithm not only adds relevant "similar articles", but also removes irrelevant articles from the original set of given recommendations. This means that our LDA-CTR algorithm can augment eScholarship's existing recommendation system. When we apply our LDA-CTR model to the CiteULike humanities articles database, given the sparse and noisy user data, we achieve precision and recall results that compare to those of previous recommendation algorithms.

An interesting observation was that the LDA algorithm on the CiteULike data categorized words from foreign languages (besides English) into their own topic. This phenomenon has a theoretical explanation; articles written in Spanish, French, and Italian comprised a significant portion of the articles retrieved from the CiteULike database, and within these documents, foreign word tokens appear together. Therefore, to effectively assign topics to foreign documents, we must employ a machine translation model in the future.

For our other future progress, we are looking to implement our algorithm in practice; we are in the process of communicating with both eScholarship and CiteULike to inform them of our suggest improvements to their recommendation algorithms. In terms of improving our current model, we plan to employ the following changes to our algorithm:

- Accounting for documents with the same author as previous recommendations: given a user u_i and document v_j, we set a different document v_k with the same author as v_j to have $c_{ik} = 0.1$. Implementing this change would allow the algorithm to have less sparse data concerning users who do not rate many articles.
- Extending the LDA to run on introductions rather than only abstracts: as many humanities articles lack legitimate abstracts, introductions would expand the dataset to include more articles.
- Incorporating citation sources into the learning model: given a user u_i and document v_j, we set a different document v_k that cites or is cited by v_j to have $c_{ik} = 0.1$.

Acknowledgment

Thank you to Professor Duchi for advice on the project and to Chong Wang and David Blei for sample data.

References

[1] Agarwal, D., and Chen, B.-C. flda: matrix factorization through latent dirichlet allocation. In *Proceedings of the third ACM international conference on Web search and data mining* (2010), ACM, pp. 91–100.

[2] Blei, D. M., and Lafferty, J. D. A correlated topic model of science. *The Annals of Applied Statistics* (2007), 17–35.

[3] Blei, D. M., Ng, A., and Jordan, M. I. Latent dirichlet allocation. vol. 3, ACM, pp. 993–1022.

[4] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *the Journal of machine Learning research 3* (2003), 993–1022.

[5] Bogers, T., and Van den Bosch, A. Recommending scientific articles using citeulike. In *Proceedings of the 2008 ACM conference on Recommender systems* (2008), ACM, pp. 287–290.

[6] Hu, Y., Koren, Y., and Volinsky, C. Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on* (2008), Ieee, pp. 263–272.

[7] Keener, J. P. The perron-frobenius theorem and the ranking of football teams. *SIAM review 35*, 1 (1993), 80–93.

[8] Koren, Y., Bell, R., and Volinsky, C. Matrix factorization techniques for recommender systems. *Computer*, 8 (2009), 30–37.

[9] Parra-Santander, D., and Brusilovsky, P. Improving collaborative filtering in social tagging systems for the recommendation of scientific articles. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on* (2010), vol. 1, IEEE, pp. 136–142.

[10] Wang, C., and Blei, D. M. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (2011), ACM, pp. 448–456.