Predicting and Understanding Instructors' Content Preferences with Latent Factor Analysis

Extended Abstract

Jack Z. Wang
Rice University
6100 Main St.
Houston, TX 43005
jzwang@rice.edu

Andrew S. Lan
Princeton University
41 Olden St.
Princeton, NJ 08544
andrew.lan@princeton.edu

Phillip J. Grimaldi
Rice University
6100 Main St.
Houston, TX 43005
phillip.grimaldi@rice.edu

Richard G. Baraniuk
Rice University
6100 Main St.
Houston, TX 43005
richb@rice.edu

ABSTRACT

We propose a latent factor model that analyzes instructors' preferences in explicitly *excluding* particular questions from learners' assignments in a particular subject domain. We formulate the problem of predicting instructors' question exclusion preferences as a matrix factorization problem, and incorporate expert-labeled Bloom's Taxonomy tags on each question as a factor in our statistical model to improve model interpretability. Experimental results on a real-world educational dataset demonstrate that the proposed model achieves superior prediction performance compared to several other baseline methods commonly used in recommender systems. Additionally, by explicitly incorporating Bloom's Taxonomy, the model provides meaningful interpretations that help understand why instructors exclude certain questions.

KEYWORDS

personalized learning, educational data mining, latent factor model, Bloom's Taxonomy

ACM Reference format:
Jack Z. Wang, Andrew S. Lan, Phillip J. Grimaldi, and Richard G. Baraniuk. 2017. Predicting and Understanding Instructors' Content Preferences with Latent Factor Analysis. In *Proceedings of KDD Workshop on Advancing Education with Data, Halifax, Nova Scotia - Canada, August 2017 (KDD2017),* 3 pages.
https://doi.org/10.475/123_4

1 INTRODUCTION

We study the problem of modeling instructors' preferences on educational contents, as a way to 1) understand each instructor's interaction with learning resources, and 2) augment personalized learning action selection systems for learners with instructors' insights. In particular, we focus on a specific instance of instructors' content preferences[1]. We collect instructors' preferences on *excluding* questions from being given to learners in their class via OpenStax Tutor[7], a personalized learning and teaching platform. OpenStax Tutor has a functionality to automatically *select* homework assignment questions for learners from a question corpus. At the same time, it allows instructors to *exclude* questions they do not want OpenStax Tutor to assign to learners in their classes from the corpus. While this exclusion option allows more flexibility for instructors to control homework assignment questions that learners receive, manually selecting questions to exclude from a (potentially huge) corpus is a labor-intensive process. As a result, analyzing instructors' question exclusion behavior has immediate utility in automating the question exclusion process.

2 METHOD

To model instructors' preference on excluding questions, we develop a novel latent factor model that predicts instructors' question preferences in a particular subject domain given previous records of whether instructors choose to *exclude* certain questions from homework assignments. Primarily inspired by SPARFA [6], this approach allows flexible incorporation of prior knowledge in the form of meta-data into the model. Consequently, the model that we develop in this work can be easily extended to include additional information in the form of latent factors to explain instructors' question exclusion preferences, as well as be used in other educational data mining tasks where auxiliary information is available. Additionally, our proposed model incorporates expert-labeled Bloom's Taxonomy tags for each question to explain instructors' question exclusion preferences, based on the conjecture that instructors have varying inclinations towards different Bloom's Taxonomy tags[2].

KDD2017, August 2017, Halifax, Nova Scotia - Canada

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

[1] We will use the phrase "learning resource", "educational content", and the word "content" interchangeably.

[2] Bloom's Taxonomy hierarchically describes questions in terms of one of the six cognitive processes, including remembering, understanding, applying, analyzing, evaluating, and creating, in increasing cognitive complexity [5]. It describes the cognitive processes by which learners encounter and work with knowledge [1].

	Models/Methods			
Metric	**Full Model**	UBCF	IBCF	FSVD
ACC	**0.9033±0.0045**	0.8961±0.0048	0.8895±0.0048	0.8896±0.0045
F-1	**0.6483±0.0128**	0.6007±0.0158	0.5696±0.0137	0.6185±0.0158
Precision	**0.7163±0.0222**	0.7070±0.0214	0.6928±0.0254	0.6964±0.0236
Recall	**0.6153±0.0227**	0.5226±0.0190	0.4954±0.0159	0.5661±0.0248

Table 1: Performance comparison between the proposed model and existing collaborative filtering methods in terms of the four metrics. The proposed model shows superior prediction performance compared to the other methods on all metrics.

	Bloom's Taxonomy tag					
Instructor	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
$i = 3$	0.9%	1.6%	0.5%	1.8%	0.0%	0.0%
	0.058	0.083	0.038	0.216	0.075	0.084
$i = 5$	16.9%	16.3%	19.0%	5.5%	21.1%	33.3%
	0.441	0.448	0.501	0.360	1.000	0.858
$i = 9$	63.1%	67.8%	72.4%	67.3%	42.1%	33.3%
	0.826	1.000	0.985	0.924	0.583	0.215

Table 2: Comparison between $\mathbf{p}_{ik}$ (second row for each instructor) and the percentage of questions they actually excluded under each Bloom's taxonomy tag k (first row for each instructor), for selected instructors. The values of $\mathbf{p}_i$ estimated by the proposed model closely resemble the actual number of questions each instructor excluded.

We emphasize that our proposed model is not limited to analyzing instructors' question exclusion preferences; it can be easily modified to analyze instructors' preferences on a broader range of learning resources. Therefore, our work serves as an initial investigation into extending the capability of existing PLSs with the analysis of instructor learning resource interaction data.

3 LATENT FACTOR MODEL

Let N, Q, K denote the total number of instructors, the total number of questions, and the total number of distinct Bloom's Taxonomy tags, respectively. Let $\mathbf{Y}$ be the binary-valued matrix of dimension N by Q that represents instructors' preference for a particular course, where $Y_{ij} = 1$ indicates instructor i explicitly denotes preference to exclude question j, and $Y_{ij} = 0$ indicates no preference. Also let $\mathbf{a}_j$ be a vector of dimension K that represents the question–Bloom's Taxonomy tag association for question j, where a_{jk} denotes the kth component of $\mathbf{a}_j$. $a_{jk} = 1$ indicates an association of question j with Bloom's Taxonomy tag k, and $a_{jk} = 0$ indicates no association.

With the above setup, we model $\mathbf{Y}$ as Bernoulli random variables:

$$Y_{ij} \sim \text{Ber}\Big(\phi(\mathbf{p}_i^T \mathbf{a}_j + \mathbf{g}_i^T \mathbf{h}_j)\Big), \tag{1}$$

Where the function $\phi(\cdot)$ is the sigmoid function:

$$\phi(x) = \frac{1}{1 + e^{-x}}.$$

In the model, $\mathbf{p}_i \in \mathbb{R}^K$, $\mathbf{g}_i \in \mathbb{R}^M$, $\mathbf{h}_j \in \mathbb{R}^M$ are model parameters to be estimated, where M is the dimension of $\mathbf{g}_i$ and $\mathbf{h}_j$ (we select the value of M via cross validation). Intuitively, the latent factor $\mathbf{p}_i$ represents the instructor Bloom's Taxonomy tag preference vector that reveals instructors' different preferences on each Bloom's Taxonomy tag. The latent factors $\mathbf{g}_i$ and $\mathbf{h}_j$ model additional factors that also contribute to explaining the observed data matrix $\mathbf{Y}$.

We use block coordinate descent to efficiently find the locally optimum set of parameters by iteratively updating each parameter in turn.

4 EXPERIMENTS AND RESULTS

We collect from OpenStax Tutor [7] 20 instructors' preferences on all 896 questions of the textbook "Concepts of Biology" that these instructors use in their classes, resulting in a fully observed data matrix $\mathbf{Y}$ of dimension 20 by 896. We also collect the Bloom's Taxonomy tag for each question, labeled by domain experts, as meta-data on the questions. Since there are 6 distinct Bloom's Taxonomy tags in total, the dimension of the question–Bloom's Taxonomy tag association vector $\mathbf{a}_j$ is $K = 6$. The entries of $\mathbf{a}_j$ correspond to Bloom's Taxonomy tags in increasing levels of cognitive complexity, i.e., $k = 1$ represents "remembering", $k = 2$ represents "understanding", etc. Additionally, each question is only associated with one Bloom's Taxonomy in our dataset. Therefore, the values of $\mathbf{a}_j$ satisfy $a_{jk} \in \{0, 1\}$ and $\sum_k a_{jk} = 1$ for all j.

We first compare our model and its variants against three methods frequently used in recommender systems: user-based collaborative filtering (UBCF), item-based collaborative filtering (IBCF),

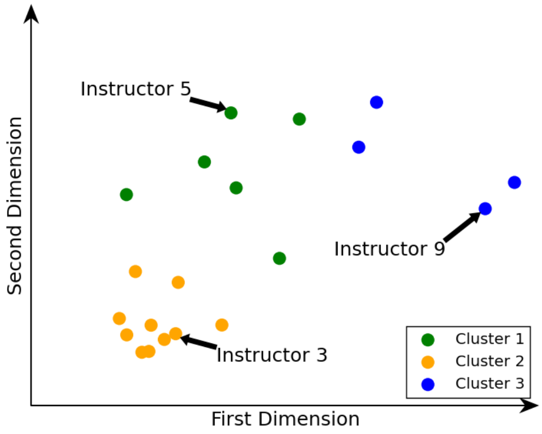

Figure 1: 2D projection of instructor Bloom's Taxonomy tag preference vectors using multidimensional scaling and clustering using k-means that shows instructors' diverse question exclusion preferences. Notice that instructors 3, 5, and 9 that we show to have very different question exclusion preferences also appear far apart in the plot.

and funk singular value decomposition (FSVD)[3, 4] using five metrics for model evaluation, as shown in Table 1. Comparing across columns, we see that the performance of the full model, regardless of the choice of metric, is significantly better than the rest of the models, showing promise for the proposed latent factor model in predicting instructors' question exclusion preferences.

We then demonstrate that the factor $\mathbf{p}_i$'s, the instructor Bloom's Taxonomy association vectors, fairly accurately characterize instructors' question exclusion preferences. Table **??** presents a comparison between the numerical values of entries in the instructor Bloom's Taxonomy tag preference vector $\mathbf{p}_i$ and the percentage of questions that the corresponding instructor excludes with each Bloom's Taxonomy tag, for a selected subset of instructors $i \in \{3, 5, 9\}$. Comparing the values in the two rows for each instructor i in the table, we observe that higher values of $\mathbf{p}_{ik}$ correspond to a higher percentage of the questions of Bloom's Taxonomy tag k that the instructor excludes. Therefore, $\mathbf{p}_{ik}$ reflects the degree to which instructor i prefers to exclude questions with Bloom's Taxonomy tag k.

Furthermore, the instructor Bloom's Taxonomy tag preference vectors uncover differences and patterns in instructors' Bloom's Taxonomy tag preferences. Figure 1 plots each $\mathbf{p}_i$ as a point in the 2-dimensional space using multidimensional scaling [2], and colors each point using K-means algorithm with $k = 3$. The figure shows obvious clustering patterns, which means that instructors exhibit only a few patterns on their Bloom's Taxonomy tag preferences. Note that instructors 3, 5 and 9 are far apart in the figure and belong to different clusters.

5 CONCLUSIONS AND FUTURE WORK

We have presented a latent factor model that predicts instructors' question preferences, and explicitly incorporates questions' Bloom's Taxonomy tags to improve model interpretability. Evaluated on a real-world educational dataset, our proposed model shows superior prediction performance over popular collaborative filtering methods frequently used in recommender systems. Additionally, we demonstrated model interpretability by showing that the Bloom's Taxonomy captures each instructor's question preferences reasonably well, and also visualized different Bloom's Taxonomy preference patterns across instructors. These encouraging results show the promise of using latent factor approach for instructors' content preferences modeling to 1) potentially automate the question exclusion process in OpenStax Tutor, and 2) more broadly, to improve various aspects of personalized learning systems such as intelligent content recommendation that takes into account of instructors' preferences.

To achieve these goals, the following avenues of future research seem appropriate. First, we used only one source of meta-data, i.e., Bloom's Taxonomy tags, in the proposed model. We have shown that the proposed model is easily extendable to accommodate additional meta-data; moreover, the performance comparison between the P Model and the GH Model shows the need to incorporate additional factors. Therefore, we plan to extend the proposed model to include other sources of meta-data, such as the textbook chapter or section that each question belongs to, to improve both prediction accuracy and model interpretability. Second, we focused on instructors' preferences in a very specific content, i.e., question exclusion. We are interested to see how well the proposed modeling approach can be adapted to analyze instructors' preference for other learning resources. Third, we also plan to expand our experiments from a single textbook to multiple textbooks and domains, in order to validate the proposed approach for analyzing instructor preferences on a wide range of contents and across different subject domains.

REFERENCES

[1] Patricia Armstrong. 2014. *Bloom's Taxonomy.* https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy/.

[2] Ingwer Borg and Patrick JF Groenen. 2005. *Modern Multidimensional Scaling: Theory and Applications.* Springer.

[3] Michael D Ekstrand, John T Riedl, and Joseph A Konstan. 2011. Collaborative filtering recommender systems. *Foundations and Trends in Human–Computer Interaction* 4, 2 (Feb. 2011), 81–173.

[4] Simon Funk. 2006. *Netflix update: Try this at home.* http://sifter.org/~simon/journal/20061211.html.

[5] David R Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice* 41, 4 (Nov. 2002), 212–218.

[6] Andrew S Lan, Andrew E Waters, Christoph Studer, and Richard G Baraniuk. 2014. Sparse factor analysis for learning and content analytics. *Journal of Machine Learning Research* 15, 1 (Jan. 2014), 1959–2008.

[7] OpenStax Tutor. 2016. https://tutor.openstax.org/. (2016). https://tutor.openstax.org/