Design and Deployment of a Better University Course Search: Inferring Latent Keywords from Enrollments

Matthew Dong
UC Berkeley
Berkeley, CA, USA
mdong@berkeley.edu

Run Yu
Wuhan University
Wuhan, China
run.yu@whu.edu.cn

Zachary A. Pardos
UC Berkeley
Berkeley, CA, USA
pardos@berkeley.edu

ABSTRACT

Liberal arts universities possess a vast catalog of courses from which students can choose. The common approach to surfacing these courses has been through traditional key word matching information retrieval. The course catalog description used to match on may, however, be overly brief and omit important topics covered in the course. Furthermore, even if the description is verbose, novice students may use search terms that do not match relevant courses, due to their catalog descriptions being written in the specialized language of a discipline outside of their own. In this work, we design and user test an approach intended to help mitigate these issues by augmenting course catalog descriptions with topic keywords inferred to be relevant to the course by analyzing the information conveyed by student co-enrollment networks. We tune a neural course embedding model based on enrollment sequences, then regress the embedding to a bag-of-words representation of course descriptions. Using this technique, we are able to predict potentially relevant words that are not in a course's description and surface these words through a real-world recommendation platform.

Keywords

Course search, Inferred keywords, Latent topics, Course2vec, Skip-gram, Higher education, Recommender systems

1 Introduction

The course catalog is often the first resource consulted by current and prospective students when wanting to familiarize themselves with and explore the topical offerings of a university. With many universities offerings thousands of distinct courses over the span of several years, browsing through the description of each is untenable. Instead, classical information retrieval (i.e., search) using keyword matching is now offered at many, but not all, institutions. A keyword matching approach; however, is only as good as the words the description contains and the users' ability to use query terms that will match. Many course descriptions can be overly brief, omitting topical terms from the description that are nevertheless contained in the course. Furthermore, for novice students, it can be difficult to gauge the similarity of courses in different departments because of the superficial differences in how different disciplines describe the same material.

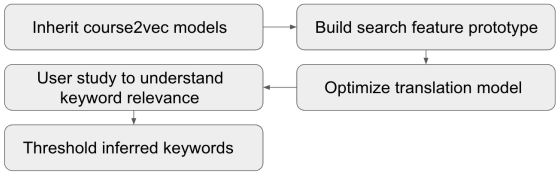

Figure 1: Design process for the enhanced search

In this paper, we seek to mitigate the shortcomings of topic omission and non-standardized keywords across disciplines in catalog descriptions by leveraging the regularizing power of machine learned embeddings. We apply neural embedding models to historic sequences of student course enrollments in order embed courses into a space regularized by abstract features, or concepts, associated with courses. We then regress from this space to the space of course descriptions in order to add semantics to the course vector space. These semantics become the keywords which can be added to an enhanced university course search. Our approach is closer to the user experience of an information system but using machine learning techniques more commonly seen in collaborative-based models. This adding of keywords to an object could be framed as a form of topic modeling. Motz et al. [3] provide an approach in this vein most relevant to ours, in which they use students' course enrollments as a signature with which to learn themes of studying using Latent Dirichlet Allocation (LDA). We substitute LDA with the more contemporary machine regularization of skip-gram models [2] and take the work further by conducting a user study ($N = 75$) in which students at a university were asked to rate relevancy of keywords to courses they had taken, generated from the embedding model and other baselines (e.g., random within and outside of description words). Measuring the degree to which our model's inferred keywords correlate with student perceptions of relevance, the results suggest a probability threshold above which predicted out of course description keywords can be chosen and be expected to be more relevant to students than the random within description baseline. Furthermore, we close the research loop by integrating this modeling process into a larger design scheme (Figure 1) leading to the deployment of this enhanced course

Matthew Dong, Run Yu and Zachary Pardos "Design and Deployment of a Better University Course Search: Inferring Latent Keywords from Enrollments" In: *The 12th International Conference on Educational Data Mining*, Michel Desmarais, Collin F. Lynch, Agathe Merceron, & Roger Nkambou (eds.) 2019, pp. 540 - 543

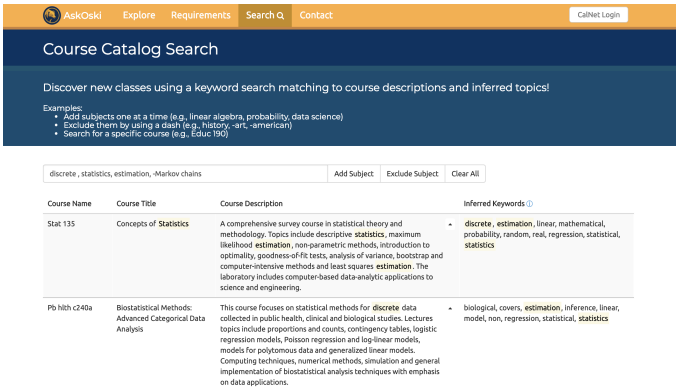

Figure 2: A prototype of the course search feature before model tuning and user testing

search feature in a live course recommendation system linked to by the campus' Office of the Registrar website.

2 Models

Our approach to generating inferred course keywords comprises of three fundamental modeling elements: (1) a vector representation of courses learned from enrollment histories (2) a bag-of-words representation of course catalog descriptions (3) a model that translates from the enrollment-based representation to the catalog-based representation. This is essentially a machine translation, not between languages [1], but between a course representation space formed from student enrollment patterns and a semantic space constructed from instructors' descriptions of the knowledge imparted in each course.

2.1 Course2Vec

The course2vec [4] model involves learning distributed representations of courses from students' enrollment records throughout semesters by using a notion of an enrollment sequence as a "sentence" and courses within the sequence as "words", borrowing terminology from the linguistic domain. For each student s, a chronological course enrollment sequence is produced by first sorting by semester then randomly serializing within-semester course order. Then, each course enrollment sequence is trained on like a sentence in a skip-gram model. In language models, two word vectors will be cosine similar if they share similar sentence contexts. Likewise, in the university domain, courses that share similar co-enrollments, and similar previous and next semester enrollments, will likely be close to one another in the vector space. Course2vec learns course representations using a skip-gram model by maximizing the objective function of context prediction over all the students' course enrollment sequences.

It is important to stress that our method of producing a course vector from enrollments (i.e., course2vec) does **not** use any course description information. It is based only on sequences of course IDs, with no natural language used. The generalizing principal is that patterns of student collective course taking can produce representations of courses containing abstract concepts of relevance to student course search. The trick to exploiting this is to associate these abstract concepts with concrete keywords, accomplished by the translation model, explained in the section after the next.

Table 1: Course Keyword Groups Example

Course: STAT 135 - Concepts of Statistics
Course Description: A comprehensive survey course in statistical theory and methodology. Topics include descriptive statistics, maximum likelihood estimation, non-parametric methods, introduction to optimality, goodness-of-fit tests, analysis of variance, bootstrap and computer-intensive methods and least squares estimation. The laboratory includes computer-based data-analytic applications to science and engineering.
Model Sorted (All): regression, statistics, random, statistical, estimation
Model Sorted (Description): statistics, statistical, estimation, variance, tests
Model Sorted (Non-Description): regression, random, real, linear, discrete
Random (Description): course, engineering, includes, methods, computer-based
Random (All): diverse collection, topics problems, year credit, planning research, user interfaces

2.2 Bag-of-Words Representation

We represent course catalog descriptions using the simple but indelible approach of bag-of-words and its variants. To create a course description vector, the length of the number of unique words across all items serves as the dimension of the vector, with a non-zero value if the word in that vocabulary appears in the description. We experiment with the description vector as binary, tf-idf, as well as a custom weighting scheme such as tf-bias that controls the granularity of keywords represented.

2.3 Translation Model

Our premise is that there are useful concepts learned in the embedding of course2vec, but these concepts left in abstract vector form contain no explicit semantics. To associate the patterns learned in course2vec with semantics, we apply a translation from the course2vec vector to its respective natural language course description vector.

We use a multinomial logistic regression to conduct this mapping, where the skip-gram based course vectors are used as input and the corresponding descriptions of every course as bag-of-word encodings are the multi-hot labels being predicted. After this model is trained, the probabilities of each word in the vocabulary belonging to a skip-gram course vector can be computed by consulting the softmax probability distribution over the entire vocabulary. Using this probability distribution, it is now possible to find the high probability words predicted based on course2vec which are NOT in the course description. These words can subsequently serve as inferred keywords in our enhanced course search.

Logistic regression is used to represent translation between languages because the spaces being translated to and from are linear vector spaces (skip-grams have no non-linear activations). However, in case the translation between spaces in the course domain is a non-linear one, we also evaluate a single hidden layer neural network with non-linear activation as a candidate translation model in our optimization experiments.

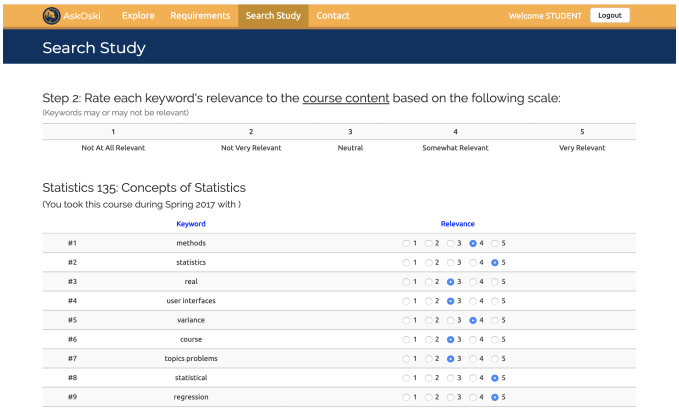

Figure 3: Keyword rating form for given course

3 USER STUDY

Through offline model selection based on 144 experiments attempting to optimize heuristics expected to correlate with relevancy ratings, we settled on a regression model and corresponding hyperparameters that maximized our custom model evaluation metric. Following the experiment driven model selection, we follow up with a human judgment evaluation to better gauge how the model results are aligned with students' perception of relevance. A user study was conducted during which students were asked to rate keywords belonging to five different groups:

1. *Model Sorted (All)*: Top five overall keywords as predicted by the model.
2. *Model Sorted (Description)*: Top five words in the description in order of likelihood as predicted by the model.
3. *Model Sorted (Non-Description)*: Top five words not in the description in order of likelihood as predicted by the model.
4. *Random (Description)*: Five random words from within the description.
5. *Random (All)*: Five random words across all collective descriptions.

An example of these keyword groups for a particular course are shown in Table 1. The Random (All) words represent a baseline relevancy score. We expect the description groups to perform much better than this baseline and desire that the model predicted non-description words are also better than randomly selected words. The random (description) group provides the second benchmark to compare our model sorted non-description group to, quantifying how much value our enhanced search proposes to add on top of the catalog description.

3.1 Study Design

Undergraduates were recruited from popular university associated Facebook groups to participate remotely in exchange for a $10 Amazon gift certificate. Study participants logged into the main *AskOski* recommender site using their university credentials in order to access the survey. Figure 3 shows the rating form for one course with its corresponding unique keywords from each of the five groups randomly shuffled. We intentionally did not show the description and

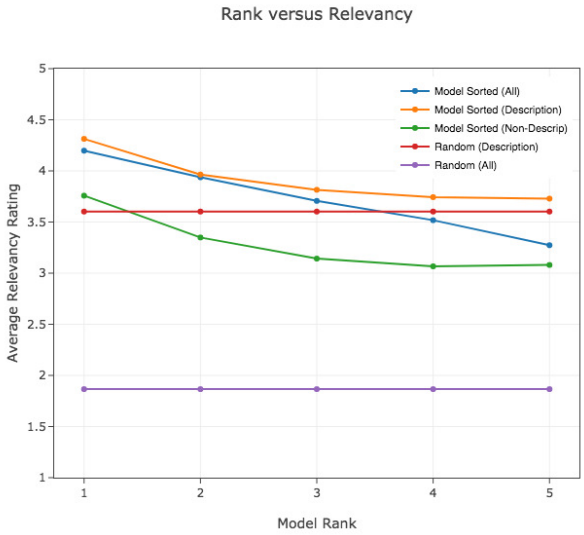

Figure 4: Keyword group rank vs relevancy

requested students to not look them up and rate solely on their experience with the class to prevent bias in keyword ratings whereby a student may be tempted to simply rate a word as relevant only if it appeared in the description.

For every keyword, students were asked for their five point Likert scale agreement with the following statement: *This keyword is relevant to the course*, where a score of 1 corresponded with *Not Relevant At All* and a score of 5 corresponded with *Very Relevant*. A total of 75 students participated in our study, rating a total of 8,355 keywords.

3.2 Results

The average student relevancy ratings of keywords from each of the five groups is shown in Figure 5. All three Model Sorted groups, and the Random (Description) group, scored between a 3 (neutral) and 4 (relevant) in keyword relevance. Selecting keywords at random from the entire vocabulary, Random (All), scored a 1.836 (below "Not Very Relevant"), representing students' lower bound for perception of relevance. All pairwise differences between keyword groups were statistically significantly reliable at $p < 0.005$, after applying a Bonferroni correction for multiple (10) Wilcoxon rank sum tests, except between Model Sorted (All) and Random (Description) groups which was not statistically separable ($p = 0.019$).

The benefit of the model-based approach in terms of improving relevance of chosen keywords can be quantified by the difference in ratings between the random within-description selection group, Random (Description) - 3.612, and the model-based within-description selection group, Model Sorted (Description) - 3.916. A breakdown of the proportion of each rating level by group can be seen in Figure 5. The majority (51%) of Model Sorted (Description) keywords received a 5 rating (Very Relevant), compared to Random (Description), for which 42.1% were Very Relevant. Model Sorted (Non-Description) has a much lower proportion of Very Relevant ratings (31.5%), but still considerably higher than the Random (All) baseline, with 7.3%, and with 62.3% of keywords in its group receiving the lowest relevancy rating as compared with Model Sorted (Description), that received 20.6% Not Relevant ratings.

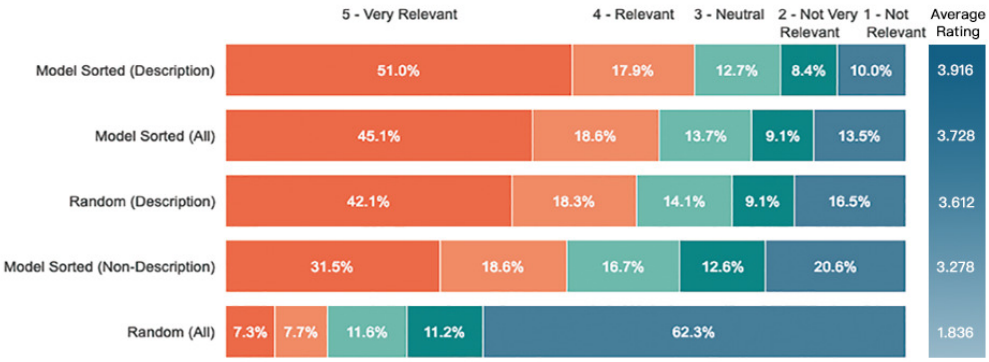

Figure 5: User study relevancy ratings by keyword group

The way in which student relevancy ratings played out with respect to the within-group ranking of the keyword based on model probability is shown in Figure 4. The average relevancy rating (y-axis) by rank (x-axis) is plotted for each of the three model-based approaches. Since the two random models do not involve any model probabilities, they also are not associated with a rank. Therefore, they are represented in the plot as horizontal lines corresponding to their averages. The Model Sorted (All) trend shows the highest average ratings at rank 1, followed by an apparent asymptote down to just above the average random within-description level. Differences in ratings between these two at each rank level are statistically significantly reliable except at ranks 3 and 4. The Model Sorted (Non-Descrip) trend is initially above Random (Description) at rank 1, but then dips down and asymptotes to a Neutral average ratings value of 3.

A premised benefit of the predictive model was to surface relevant keywords that are not in a course's description (Non-Descrip). If we were to highlight inferred keywords, we would like to show only keywords that are "better" than words chosen randomly from the description, or at least not show words statistically significantly worse. The Model Sorted (All) ratings are statistically reliably higher than Random (Description) at ranks 1 and 2. We use this information to tailor our strategy for when and how many inferred keywords to display in the production version of our enhanced course search feature.

3.3 Selecting keywords to display in search

With an improved understanding of the relevancy of the model's predicted keywords, we discuss how to leverage this information towards improving the search feature by updating our inferred keyword selection criteria. In the prototype, the criterion was to always display the top 10 model keywords, which did not exclude words in the description. We continue to not restrict the display of keywords from the description, as showing them could serve the added benefit of a topic category source for reference. Thus, we choose Model Sorted (All) as the focus of this analysis.

We leverage the observation that Model Sorted ratings correlate with rank to investigate how well the underlying model probabilities of those words correlate with student relevancy ratings. If there is a correlation, then the probabilities, along with a threshold, could be used to dynamically determine which words should be included as inferred keywords on a per course basis. To conduct an analysis comparing model probabilities to user ratings, we normalize these two sets of ratings using Z-scores and then average them by Model Sorted rank. We find a substantive correlation between probability and rank and would like to choose a threshold of probability from Model Sorted (All), such that all keywords with that probability or above can generally be expected to produce keywords perceived by students to be more relevant, on average, than a word chosen at random from the description. The analysis in the previous section (Fig 4) found that user relevancy ratings for Model Sorted (All) were significantly higher than Random (Description) at ranks 1 and 2. Therefore, we use the probability at rank 2 as the cut-off. Using this probability cut-off, we find 4.32 total words on average expected to be displayed for each course, with 2.33 within-description words and 2.00 non-description words surfaced on average within these semantics.

4 Conclusion

We explored surfacing novel, searchable semantics of a course using an embedding of courses informed by course selection histories, and supported our methodology through a user study to evaluate the relevancy of these keywords. Our experiment contributes both methodologically to the use of embeddings to surface latent semantics and to the design of data-driven information systems in educational settings. Our process of interface prototyping, followed by offline model optimization, user testing, and incorporation of study findings into the production software system can also serve as a design model and guide for other technologies to tune EDM analyses towards better student experiences.

5 References

[1] T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168, 2013.

[2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[3] B. Motz, T. Busey, M. Rickert, and D. Landy. Finding topics in enrollment data. In *Proceedings of the 11th International Conference on Educational Data Mining*, 2018.

[4] Z. A. Pardos and A. J. H. Nam. A map of knowledge. *CoRR*, abs/1811.07974, 2018.