Representation Learning in Course Discovery

Matthew Dong[1,4]; Zachary Pardos[1,2,3] (PI)
[1]University of California, Berkeley, [2]Graduate School of Education,
[3]School of Information, [4]Department of Statistics

CAHL
Computational Approaches to Human Learning

ABSTRACT

Web infrastructure in most schools has remained static in an increasingly technological society; in particular, variety in course enrollment platforms remains lacking in higher education. To provide greater personalized course guidance to individual students, *AskOski*[1] is a state-of-the-art recommender system that lifts machine learning methodologies from other domains to the education field to help improve learning outcomes.

This research project aims to incorporate an intelligent search function in addition to the main features of the site, allowing students to perform a topical search across courses to discover new classes.

What makes this search "intelligent" is that a query will not only be matched against course titles and descriptions, but also to an additional search metric known as "inferred keywords." Inferred keywords are a generalized description produced through machine learning that captures a semantic portrayal of courses beyond university catalog descriptions.

ACKNOWLEDGMENTS

This endeavor would not be possible without the generous support of the Rose Hills Foundation. Thank you to lab members Jenny Jiang and Tiffany Jann for their work integral to completing this project.
(github.com/CAHLR/semantic_model)

CONTACT

Matthew Dong
Computational Approaches to Human Learning
Email: mdong@berkeley.edu
Website: https://github.com/CAHLR

DATASET

Course Information

- Class titles and descriptions sourced from the university course catalog.

Course Vectors

- Abstract representation of courses produced using historic enrollment data.
- Captures semantic relationships between courses through behavioral data, similar to Figure 1.

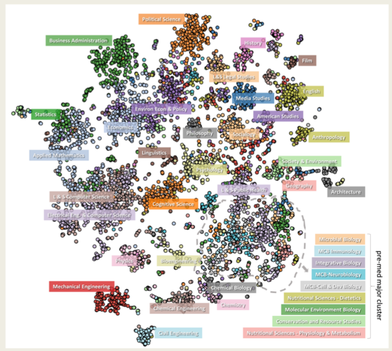

Figure 1. Visualization of student vectors color-coded by department.

METHODOLOGY

1. Convert each course description to its bag-of-words (BOW) representation.
2. Fit a model (Figure 2) mapping BOW descriptions to their corresponding course vector.
3. For each course, let the model predict the top k words that describe it.

There were two possible avenues by which inferred keyword quality could be improved: the input data and the model itself. Our final step was to utilize different combinations of course representations while varying hyperparameters of the model to improve the keywords. The most critical hyperparameter was tf-bias[2], which controls specificity levels of the outputted vocabulary.

$$tf - bias = \left(\frac{number\ of\ occurences\ of\ word}{total\ word\ count}\right)^{-bias}$$

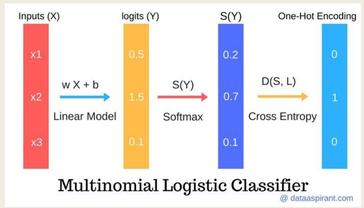

Figure 2. Multinomial regression model used to predict keywords for a course.

RESULTS

Validation of the keyword quality was initially completed by simple semantic evaluations of how related the words were to the actual course content, as well as calculating the metric *average number of unique keywords per course* (Figure 3). This is defined as the size of the set of keywords not already present in the course title or description, and is intended to quantify one dimension of how well the model performed.

We discovered the analogy-optimized course vector set with tf-bias range 0-2 provided the most robust keywords (Table 1).

Table 1. Examples of inferred keywords for selected courses, with corresponding titles and descriptions.

Course Title	Description	Inferred Keywords
Descriptive Cosmology	Non-mathematical description of research and results in modern extragalactic astronomy and cosmology.	astronomy, astrophysics, atmospheres, cosmic, cosmology, electromagnetic, galaxies, kinetic, planets, relativity, rings, stars
Introduction to Archeology	Prehistory and cultural growth. Introduction to the methods, goals, and theoretical concepts of archaeology...	archaeologists, artifacts, enlightenment, ethnographic, ethnography, historic, prehistoric, prehistory
Monetary Theory and the Banking System	Survey of monetary, interest and income theories. Despository institutions, other financial institutions, the Federal Reserve System...	balance, costs, demography, econometric, economists, expectations, fiscal, goods, incentives, macroeconomic, microeconomic, monetary, money, rate, size

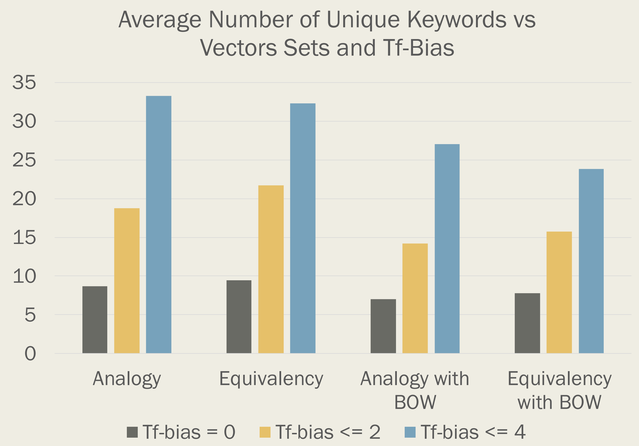

Figure 3. Average number of unique keywords per course with respect to the vector set used and the range of tf-bias values.

DEPLOYMENT

Once quality inferred keywords were obtained, the project then pivoted into utilizing them in production by designing the "intelligent search" feature for *AskOski*.

This feature enables topical and subject based search, with the keywords providing an additional vetting to improve search relevancy, distinguishing it from other course information services. Student users can filter courses by including or excluding certain topics they would prefer or prefer not to see in their classes.

Live Demo: askoski.berkeley.edu:1380/search

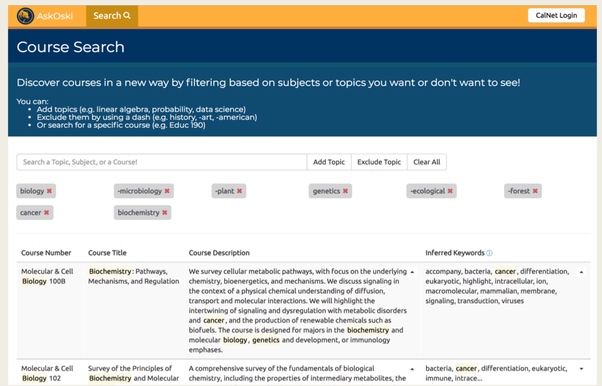

Figure 4. Screenshot demonstrating search feature functionality with inferred keywords.

CONCLUSIONS

The semantic translations of learned course embeddings were able to capture relevant concepts exceeding their catalog descriptions. This immediately translates to enhanced search relevancy but has more powerful implications in terms of leveraging machine learning to gain insight into other educational contexts with potentially broad applications.

FUTURE DIRECTIONS

1. Refine keyword quality by optimizing model.
2. More rigorously validate keyword usefulness to user by obtaining ratings via the search interface.
3. Improve search relevancy algorithm.

REFERENCES

1. Pardos, Z.A., Fan, Z., Jiang, W. (2018) Connectionist Recommendation in the Wild. CoRR preprint, abs/1803.0953.
2. Pardos, Z.A., Nam, A.J.H.: A map of knowledge (in-preparation)