Collaboration: Towards Automated Study Guides for MOOCs

Jose Hernandez
Stanford University
josehdz@stanford.edu

Andrew Lamb
Stanford University
lamb@cs.stanford.edu

Andreas Paepcke
Stanford University
paepcke@cs.stanford.edu

Jeffrey Ullman
Stanford University
ullman@stanford.edu

1 Introduction

One of the biggest challenges of delivering education with massively open online courses (MOOCs) is the high student to instructor ratio, which is further compounded in courses where students don't arrive at traditional term boundaries. In the extreme, *untended* courses may not have any active teaching staff. Without support, students fend for themselves when they don't understand a concept or need to review for a test. Our team is building a system to autonomously help students enrolled in unattended courses; this will enhance, not replace, human instruction and improve the support for all learners.

For example, when a forum post indicates that the posting learner is confused, we wish to offer automatically identified related learning resources. We define learning resources broadly — they might be pointers into relevant video snippets, Wikipedia articles, past homework assignments, or forum answers from learners who took the course in the past. As online learning is enriched beyond the current video lecture model, we expect more types of resources to emerge. Finding such related resources requires that (a) the many topics that comprise the course are automatically identified and (b) resources are associated with each topic. Finally (c), the forum post in question must be associated with one or more of those topics, so that appropriate resources are identified.

We break (a) - the identification of topics - into two steps. First we have worked to identify topics constituent to an online course on compilers from the closed caption files of the associated instructional videos. We began by automatically identifying words and phrases that primarily describe concepts important to the course. The result is much like the index at the end of a book, though referencing video snippets rather than pages.

In the second step (b), we have been attempting to group the index words into clusters large enough to each describe a topic with which we can associate learning resources in the future. While this work is not mature, and we do not yet have results, we briefly describe our approaches to both these steps.

2 Keyword Extraction

Most previous work on keyword extraction has used datasets of journal abstracts or newspaper articles [3], [5], [8]. However, course material has very different characteristics from those collections. For example, a course's lectures are more semantically coherent than a collection of newspaper articles. Similarly, journal abstracts are written by different authors, while a course is usually delivered by a single instructor.

Creating an index of keywords has a number of difficult cases. For example, in a compilers course, "first set" should be attached to a topic, but "first" is too general to be included in the index. An algorithm must have a good strategy for cleanly separating out overly broad phrases. Many methods for keyword extraction are based on the distribution of phrases over the corpus, which may not work as well with an online course. In a collection of abstracts, key technical terms likely only appear in relevant documents, while in a series of lectures an instructor might mention a concept in passing long before they cover it in depth.

Thus far, we have mostly experimented with techniques from classical Information Retrieval. For instance, we separated the lecture videos into 10-minute segments and then ranked phrases with *term frequency - inverse document frequency* (*tf-idf*), which assigns a score to words proportional to the number of occurrences of the word in the corpus and inversely proportional to the logarithm of the number of documents in which the word appears. We have tried a number of variants of *tf-idf*, such as filtering keywords based on part of speech tags.

Our system currently generates a reasonable set of keywords using a tf-idf based algorithm, but we are hoping to apply other algorithms from the keyword extraction literature to the online education domain. A few of the algorithms we have considered are Rapid Automatic Keyword Extraction (RAKE), which focuses on smart delineation of phrases by stopwords [8]; Matsuo and Ishizuka's algorithm that uses distributions of word co-occurrences, and is able to operate on a single document [4]; and Hulth's strategy of using parts of speech [3].

3 Topic Clustering

There has not been much previous research on keyword clustering into topics. Many clustering algorithms have been researched in the context of journals [2], and social networks [1].

Since our clustering context differs from those contexts, we need a new strategy. Agglomerative, or "bottom-up", algorithms create a hierarchy of topics by initializing each topic as a singleton cluster and repeatedly merging topics together. Similarly, divisive, or "top-down" algorithms split topics on each iteration. While this family of hierarchical clustering algorithms encode useful information, they create disjoint topics. Thus, no word can be included in more than one topic, which is clearly overly restrictive.

We encountered this problem during preliminary experiments with hierarchical clustering. We created vector representations of our keyword set using the full text of a course-relevant textbook, generated with Google's *word2vec* [6]. We then created clusters from these numeric representations of keywords, using different parameterizations of the hierarchical algorithms. For example, the distance between topics can be measured using Euclidean distance, cosine distance, or another metric. The range of stopping criterion is also quite varied; we experimented with stopping once a topic's number of constituent words reached a minimum, or when the distance between merged clusters reached a certain threshold.

One way to solve the disjoint topics problem is a graph-based approach, in which we define vertices as keywords, and identify topic clusters, represented by dense subgraphs, such as k-cliques. There are a variety of ways to define edges, such as the distance between vector representations of words or how often they co-occur in a sentence. Below are an example of words that, respectively, should and should not be part of a "soup" topic cluster.

Initial steps in clustering have created mixed quality topic clusters, but have not reached pedagogically useful quality. We plan to experiment with integrating rule-based criteria in our hierarchical clustering strategy. This would require a data-centered feature set. Lectures tend to follow a pedagogical structure, with definitions, motivations, and examples, so we can create stronger features using the structure of the lecture data. Leveraging Wikipedia's knowledge-graph may provide rich contexts [2]. With Wikipedia, we can identify connections between keywords using untapped, external resources, by analyzing the articles with keyword mentions.

4 Collaboration Opportunities

4.1 Experiments with Other Algorithms

Creating topic models and other useful structured representations of online courses is an area that has the potential to improve the experience of millions of learners, but has received relatively little attention thus far. We welcome discussion of ideas and sharing of results among interested researchers.

4.2 Creating Datasets

We have been conducting experiments using the closed caption files from a freely available undergraduate course on compilers. Because we don't have a clear gold set of keywords, we have been evaluating our algorithms based on overlap with the (manually created) index of the course's accompanying textbook. As mentioned above, most publicly available datasets for keyword extraction are based on abstracts or keywords, for example the NUS Keyphrase Corpus [7]. In addition to providing a better method for evaluation, a gold standard set would make supervised approaches to keyword extraction possible. We would like to create an analogous dataset for MOOCs, ideally with professors and teaching assistants for the course annotating lectures with keywords. Sharing of digital textbooks and indexes would also be quite useful.

5 Available Resources

Our team is at Stanford University, which has delivered hundreds of courses online. Data from those courses, and the associated several million learners are available in easy-to-analyze form.

A senior staff member of our team is Director of Data Analytics, and is in charge of the accumulated data. Our group owns several large-memory machines, two with a terabyte of RAM. The equipment is easily adequate for the most demanding computations over our data.

References

See

References

[1] Vincenzo Fioriti and Marta Chinnici. Identifying sparse and dense sub-graphs in large graphs with a fast algorithm. *Europhysics Letters*, 108(5), 2014.

[2] Thomas Hu, Xiaodan Zhang, Caimei Lu, E. K. Park, and Xiaohua Zhou. Exploiting wikipedia as external knowledge for document clustering. In *Proceeding KDD '09 Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 389–396. Knowledge Discovery and Data Mining, 2009.

[3] Anette Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 216–223. Association for Computational Linguistics, 2003.

[4] Yutaka Matsuo and Mitsuru Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01):157–169, 2004.

[5] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. Association for Computational Linguistics, 2004.

[6] Thomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*. Neural Information Processing Systems, 2013.

[7] Thuy Dung Nguyen and Min-Yen Kan. Keyphrase extraction in scientific publications. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, pages 317–326. Springer, 2007.

[8] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic keyword extraction from individual documents. *Text Mining*, pages 1–20, 2010.