Academic Curriculum Generation using Wikipedia for External Knowledge

Anurag Reddy Muthyala
IIIT Hyderabad
India
anurag.reddy@research.iiit.ac.in
&
Vikram Pudi
IIIT Hyderabad
India
vikram@iiit.ac.in

November 15, 2022

Abstract

In this paper, we address the problem of automatic academic curriculum generation. A curriculum outlines definitive topics with their sub-topics and enables teachers and students to form an overall idea of the course outcomes and goals, and a plan of what to teach and learn to achieve those goals. Automatic curriculum generation is relevant in modern times with the ever increasing, rapidly changing, digitally-available academic content, that is too large for manual processing by human teams. Using Wikipedia as an external knowledge-base, along with a pipeline of standard components, we show that it is possible to generate human-interpretable 2-level topic hierarchies. We show that our approach works on publicly available textbooks, by first removing their title-structure, and then automatically regenerating a 2-level title structure that is on-par.

1 Introduction

We address the problem of automatic academic curriculum generation. We treat a curriculum as one that outlines definitive topics with their sub-topics, in order to enable teachers and students to form an overall idea of the course outcomes and goals, along with a plan of what to teach and learn to achieve those goals.

Automatic curriculum generation is relevant in modern times with the ever increasing, rapidly changing, digitally-available academic content, that is too large for manual processing by human teams. The need for automation is crucially felt in interdisciplinary fields [Jacobi(2014)], and to personalize content and presentation for individual student needs and flow [Katuk and Ryu(2010)].

We formulate the problem as generating a 2-level human-interpretable topic hierarchy consisting of module titles and the topics within those modules. This caters to the most common requirement of most academic curricula. However, we add that this formulation is not restrictive as it is possible, when needed, to devise methods to generate deeper hierarchies using the base method for 2-level hierarchies through recursive application.

We aim to implement a subject-centered generative model that generates topics based on the domain knowledge instead of the learner's ability. This ensures that we generate a uniform structure for all learners, which is the typical goal of a curriculum.

Our model is an unsupervised approach based on the probability distribution of words for topic generation. We incorporate the salient features necessary for generating curriculum from given set of documents.

The primary objective is to generate a 2-level module-topic hierarchy following a data-driven approach that does not depend on the academic domain and discipline. Our model is a simple pipeline of standard components. In order to create a se-

mantic structure (titles) from the candidates that are generated, we use Wikipedia for external knowledge and links to Wikipedia pages as learning objects to enhance learner's curiosity. Employing the previously mentioned approaches, we generate module-topic hierarchies that are on par with human generated ones, by using ideas (described in Section 3) of semantic structure, maximum coverage, relationship sanity and curriculum ambiguity.

2 Related Work

There is very little work that focuses on academic curriculum generation. In [Jacobi(2014)], the authors propose an approach for interdisciplinary fields that is based on how curricula may be designed manually in the real world. For instance, it contains steps to generate a consensus on the topics chosen. Several steps of this method require manual input by domain experts, who may be hard to find for novel interdisciplinary fields. Inputs include a core skill levels list, application skill levels list, etc. Our system aims to overcome this limitation and extend the capability by being entirely data-driven. The area of topic modelling has been widely studied over the years for its extensive applications in diverse fields [J. Boyd-Graber and Mimno(2017)]. Topic models help the reader to understand the general theme of the given document. This is achieved by associating each topic in the document to generated key phrases which best represent them. Although there are several topic modelling algorithms like LDA [David M. Blei and Jordan(2003)] and its variants, they are designed to derive a fixed set of topics from a corpus. The intuition behind LDA is based on reverse-engineering the process of creating a document using keywords occurring in it. LDA generates a set of keywords which are not structured into hierarchies, and hence cannot be directly used for our task. A variant of LDA was implemented in [P. Liu and Wang(2012)] which generates a *hierarchy* of topics. Unfortunately both LDA and this variant produce topics that are mathematical representations suitable for machine-processing but not for human readability. Aside from LDA and its derived methods, graph-based ranking algorithms similar to PageRank Algorithm [Page et al.(1998)Page, Brin, Motwani, and Winograd] have been implemented for the task of topic modelling. The TextRank Algorithm [Mihalcea and Tarau(2004)] was the first one to generate keyphrases pertaining to the topics by creating a graph using the words, and their edge relations were derived based on the offset in the document. However, this algorithm doesn't consider the hierarchical relationships between topics that is necessary for curriculum generation. Similar drawback can be observed in the SingleRank Algorithm [Wan and Xiao(2008)] which considers different documents to enrich the topics generated.

3 Design of Our Approach

In order to generates module-topic hierarchies that are on par with human generated ones, we pay attention to the following factors:

- *Natural Language:* The topics that are generated by our model should be human-readable. This requires that topics are not just machine-readable mathematical representations, but grammatically-sound natural language phrases. We easily achieve this by using titles of Wikipedia articles as topic and module titles.

- *Maximum Coverage:* While generating the curriculum, we need to ensure that all key topics are included. While we filter out some topics on the basis that they are not noun-phrases, we ensure that all the remaining topics are included. As our topics correspond to Wikipedia article titles, we consider them as valid topics to be included in some module of the curriculum.

- *Relationship Sanity:* Understanding the relationship between modules and topics is paramount to the process of curriculum generation. While establishing links, we need to ensure that a module is paired with a topic if they are similar (using standard similarity measures of their word-probability distributions). It is also important to keep a module:topic pair disjoint if they are different. In our current approach, each topic is mapped to exactly one module. However, a topic's assignment to multiple modules may be permitted easily if desired. Existing models like TextRank and SingleRank employ ranks or thresholds to map topics to modules. In contrast, we use a clustering-based approach as we already have topics and their titles using

Wikipedia and only need to cluster them into modules.

- *Curriculum Ambiguity:* The keywords, topics and modules extracted are subjective, and there can be quite a bit of disagreement even among human-teams generating them. Variations in the topic distribution generated by different models are possible, and these can lead to different curricula. Thus, several different possible curricula generated can be considered valid since they each include keywords, topics and modules that describe the text. Hence, the validation of the model's performance cannot be restricted to one structure obtained from the document. We need to apply proper metrics which does not penalize the variations in the curriculum obtained.

For generating the curriculum, we need to generate topics (with human-readable titles) and then aggregate similar topics together to generate and name the modules.

4 Detailed Methodology

We propose an unsupervised, extractive model with a little abstraction offered from the external knowledge base to accomplish the task.

4.1 Candidate Generation

The initial step of the model is to extract keywords from the document. This is achieved by generating n-grams which will act as the candidates set for topics. During the exploration of the Wikipedia data dump; it was observed that 81.25% of the total (near 16 million) Wikipedia titles considered were made up of 1-3 words. The number of n-grams generated can be scaled with the size of processing text. The candidates set which occurs frequently with incorrect semantic structure does not add any importance, hence we eliminate the n-grams which are semantically or grammatically incorrect.
To accomplish the task of removing any semantically incorrect candidates, we consider the candidates which form a noun phrase. While exploring the data dump, it was also observed that more than 94% of the titles consisted of noun phrases. To incorporate this, we devised an approach to find candidate sets for different values of n. For unigrams/uni-grams, verify if the derived monogram is either a singular or plural noun. If the uni gram belongs to any other POS (parts of speech), discard it. The unigrams/uni-grams identified were also filtered based on occurrences for accurate prediction of titles. If the bigrams/bi-grams and trigrams/tri-grams are noun phrases with minor occurences of stopwords, they are added to the candidate set.

4.2 Using Wikipedia as external Knowledge Base

Wikipedia is the largest and most comprehensive knowledge source on the web with the latest information. It is well-structured with each Wiki page providing information on a particular topic and title serves as the main topic and references and links present show related topics. We have used close to 16 million titles in our task for generating titles based on the candidate set. As described previously, the model is developed with the focus to make it robust in its use. Our model can generate the titles from documents structured in different formats like articles, papers, transcribed speeches, scripts, comments etc. This model is also capable of segregating the modules belonging to different domains without compromising the module-topic relations. Wikipedia has information on various domains which expands our field of study into all those domains.

4.3 Search and Similarity Comparison

An efficient search engine was developed for our system for searching relevant titles from the Wikipedia title dump [1]. In the previous sections, we have discussed how the n-grams which constitute the candidate set are generated to find the topics. Each candidate can be considered as an entity adding significance to the document. We use these candidates to search for the appropriate titles from Wikipedia which can be used as the topics. For each candidate set, we retrieve an average of 15 titles which contain most of the keywords in the candidate set. However, all the titles that are retrieved will not be considered during the generation of the curriculum. These topics are used later for the hierarchical modelling which generates the curriculum.

[1] Wikipedia Title Dump

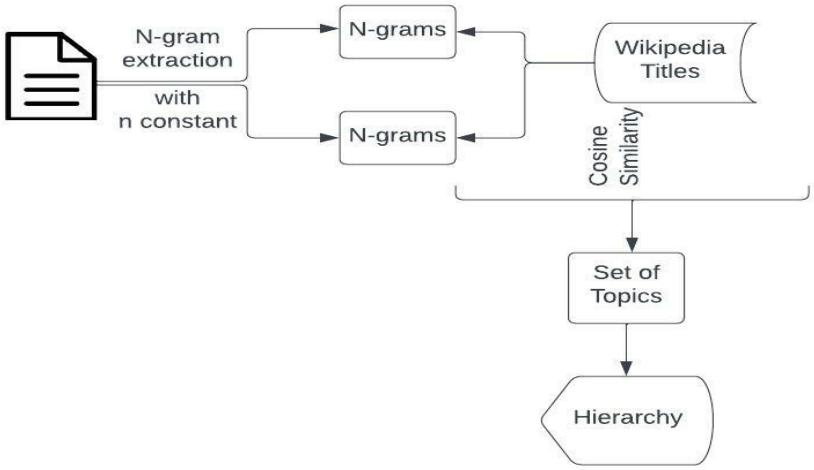

Figure 1: Methodology for leveraging Wikipedia Titles for Module Generation

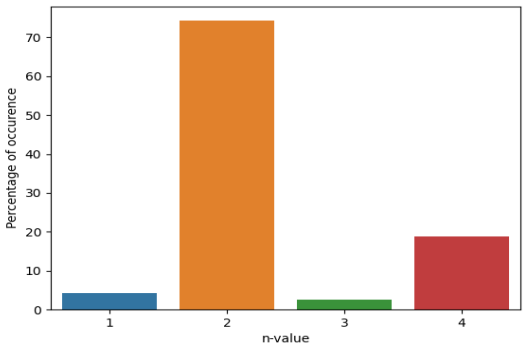

Figure 2: Percentage occurrence of n-grams

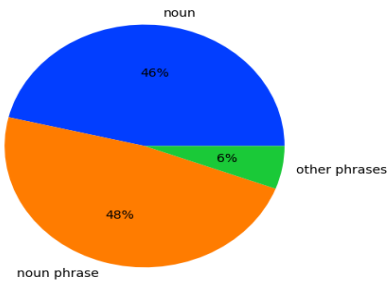

Figure 3: Percentage of occurence for noun phrases

The next step in topic generation is to remove any unwanted topics retrieved and segregate the remaining topics into modules to generate the curriculum. We have performed various experiments like distance metrics (L-norms), similarity metrics like levenshtein distance, cosine similarity etc to remove any unwanted titles. After experimentation, the best approach to get the titles was comparing cosine similarity of n-grams obtained and the Wikipedia titles derived. For any two vectors $v1$ and $v2$,

$$sim(v1, v2) = \frac{v1.v2}{||v1||.||v2||} \quad (1)$$

Before comparing similarity, we obtain a vector representation for the keyphrases and the titles which are to be compared. We achieve this using the unsupervised FastText representation over each keyphrase or title. Since the words can be from any domain, an unsupervised approach is recommended for vector embeddings of the words. Hence, we do not consider supervised representations like GLoVE, CoVE etc. The FastText model is trained on the wikipedia data dump before it is used for generating vector representations of each candidate generated. Cosine similarity is obtained between two sentence vectors obtained from the keyphrase and the title. The Wikipedia titles with high cosine similarity were considered to maintain accuracy in the titles. The result of this step is the topics and sub-topics for the given text document.

4.4 Hierarchical Modeling

A deep understanding of any module can occur if and only if the sub-topics can be clustered and put together to form a concept corresponding to that module. A 2-level hierarchy for a curriculum is the best way to portray the contents. Consider the matrix M where $M_{i,j}$ denotes the similarity between titles t_i, t_j. We use the Indicator function I_c defined as,

$$I_c(i,j) = \begin{cases} 1 & \text{if } M_{i,j} \geq \lambda \\ 0 & \text{if } otherwise \end{cases} \quad (2)$$

The proposed system derives a method wherein modules are formed by connecting all the Wikipedia titles with each other in a matrix based on similarity and classifying them into modules using the Indicator function mentioned above with a threshold(λ) as the clustering factor. Given a cluster of titles $m_i = \{t_1, t_2, t_3, ..., t_{N_i}\}$ where N_i denotes the number of titles in cluster m_i, the title of the module is given by,

$$title(m_i) = LCS(t_1, t_2, ..., t_{N_i}) \quad (3)$$

where, $LCS(.)$ is the longest common subsequence function. In our analysis, we have observed that the module titles are formed from the words that are common to two or more titles and form noun phrases. Hence, we consider this title function after verification using POS tagging.

5 Experiments

5.1 Dataset

To show the results of our curriculum generation system, we used publicly available textbooks, where title structure has been removed, from the Central Board of Secondary Education (CBSE) [2] website from classes 8-12 and for different subjects but not limited to Biology, Physics and Social Sciences, available at National Council of Educational Research and Training (NCERT) [3] website. The curriculum within the books enabled us to compare our results with the curriculum generated with our model.

[2] CBSE official website
[3] NCERT Textbooks Link

5.2 Results

A quantifiable evaluation of the result is difficult due to lack of standard procedures for topic detection and curriculum generation tasks. However, we have showcased results obtained through LDA in Table 1, to compare as the baseline method. It is evident that we are extracting module titles which are monograms. LDA was developed with the intent to generate documents based on the keywords corresponding to them.

Topic Name
ACCELERATION
AXIS
BODY
CENTER
ENERGY
FORCE
LAW
MASS
MOMENTUM
MOTION
OBJECT
PARTICLE
POINT
SPEED
SYSTEM
TIME
VELOCITY

Table 1: LDA keyword extraction performed on 10th grade CBSE Physics Textbook

The results shown in Table 2, has 8 modules in contents whereas our model generated 12 learning objects with precise distinction. On evaluation from faculty and observation, it was noticed that our model has grouped sub-topics based on the right parameters and upon evaluation, it is noticed that all Wikipedia pages in sub-topics are related as references to the title Wikipedia page. The module names with no sub-topics are not grouped together because the model performs an extractive task and recognises words from the input text provided like the module *Kinematics* which would contain *average speed, average velocity, acceleration*.

The 12th grade Biology textbook considered for the model of Table 3 lists only topics in the curriculum page. Our system was able to generate sub-topics and depict a correlation between them. Similar results have been produced for several other textbooks and articles from the Internet. Apart from that, we were able to generate a inter-disciplinary

Module Name	Sub-Topics
GRAVITY	Center of Gravity Force of Gravity
UNITS	SI Units Base Units Derived Units
LAWS	Law of Gravitation Laws of Motion Laws of Nature
MOTION	Uniform Motion Translational Motion Rotational Motion
FRICTION	Static Friction Kinetic Friction Co-efficient of Friction Force of Friction
QUANTITIES	Base Quantities Physical Quantities
MOMENTUM	Total Angular Momentum Change in Momentum Linear Angular Momentum Angular Momentum
MOMENT	Moment of Inertia Moment of Force
AVERAGE SPEED	
KINETIC ENERGY	
ACCELERATION	
AVERAGE VELOCITY	

Table 2: Hierachy obtained on 10th grade CBSE Physics Textbook

curriculum for the given text with several modules formed for different subjects.

Module Name	Sub-Topics
REPRODUCTION	Human Reproduction Asexual Reproduction Sexual Reproduction Reproductive Health
GENETIC	Genetic Evolution Genetic Inheritance
HUMAN WELFARE	Human Biological Welfare
BIOTECHNOLOGY	Principles of Biotechnology Biotechnology Applications
ECOLOGY	

Table 3: Hierachy obtained on 12th grade CBSE Biology Textbook

Though there are no established metrics for quantifying the quality of the modules and subtopics generated, considering the unsupervised learning criterion, we try to quantify it assuming the modules as clusters.

Subject	Intracluster	Intercluster
Biology	0.04	0.3
Physics	0.04	0.2
Physics and Politics	0.04	0.2

Table 4: Similarity metrics for the modules generated

In Table 4, we see the average intercluster and intracluster distances between the modules and the topics within them. We expect the intercluster distances to be high and intracluster distances to be low. By this, we can say that the modules generated are distinct from each other, and the topics within the module are similar to the module they belong to. Upon observing the values in the table, we can see that though the values are very low, relatively, intercluster distances are greater than intracluster distances. This shows that the modules generated are properly structured.

Subject	Min	Max	Avg
Physics	0.238	0.937	0.476
Biology	0.416	0.973	0.742
Physics and Politics	0.377	0.937	0.601

Table 5: METEOR Scores for the modules generated

In Table 5, we see the minimum,maximum and average METEOR[Lavie and Agarwal(2007)] scores for each textbook. We chose this metric over other machine translation outputs metrics because of it's additional feature of stemming and synonymy matching, along with greater co-relation with human judgment than the other metrics like ROUGE, BLEU etc. We have mapped our system-generated topic and module names with ones in our dataset and calculated the metric. As we can observe, the maximum METEOR score for all textbooks is 0.937, almost equal to 1, which demonstrates that generated modules are very close to the original textbook modules. The average score is almost 0.6, which shows that our system-generated topics and modules are analogous to textbook modules and topics.

The results in Table 6 depict the performance and distinguishability of our model when is the input is from two different disciplines but distinct modules with an inter-disciplinary hierarchy has been formed.

Module Name	Sub-Topics
HEAT	Heat and Electricity Heat and Light Conductors of Heat
CELL	Cell Structure Cell Membrane Cell Wall Plant cells Animal cells
FRICTION	Force of Friction Static Friction Sliding Friction Rolling Friction
REFLECTION	Laws of Reflection Angle of Reflection Diffused Reflection
POLLUTION	Noise Pollution Air Pollution
SOLAR SYSTEM	
FORCE	Applied Force Muscular Force Frictional Force
PRESSURE	Atmospheric Pressure
COMBUSTION	
COAL	
PETROLEUM	
DEMOCRACY	Democracy and Equality Development of Democarcy
HEALTH CARE FACILITIES	
GENDER	Gender Equality
MEDIA	
MARKETS	Putting-Out-System
WOMEN	Women Harassment Women Equality Women Empowerment

Table 6: Hierarchy obtained on 8th grade Science and 7th grade Social textbook

6 Conclusion

In this paper, we presented a pipeline of standard components and using Wikipedia as the external Knowledge Base to generate human interpretable 2-level hierarchies.

Based on the concept of candidate item set generation, we are able to create a set of unigrams/unigrams, bigrams/bi-grams and trigrams/tri-grams which are the learning objects and can be mapped to Wikipedia titles. The proposed model is evaluated with the help of general observations and experienced faculty on publicly available data sets. The input is not limited to a single subject textbook and can contain text from the web such as web content, news articles, blogs, etc.

The task of Curriculum Generation is carried out by an extractive model and therefore, titles which do not occur in text cannot be grouped under module names.

We believe that our model can be extended to developing deeper hierarchies beyond 2 levels. For future work, we will further improve our candidate item set generation techniques, taking into context the data they are present in. Moreover, we will utilize the linking structure between Wikipedia pages to develop a deeper hierarchy with better co-relations. Aside from the drawbacks of extractive models, we can also try to pursue the problem using abstractive approaches.

References

[David M. Blei and Jordan(2003)] Andrew Y. Ng David M. Blei and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research 3*, pages 993–1022. Version 3.

[J. Boyd-Graber and Mimno(2017)] Y. Hu J. Boyd-Graber and D. Mimno. 2017. *Applications of Topic Models. Foundations and Trends 244 in Information Retrieval*, volume 11.

[Jacobi(2014)] Steffen Krawatzeck Robert Dinter Barbara Lorenz Anja. Jacobi, Frieder Jahn. 2014. Towards a design model for interdisciplinary information systems curriculum development, as exemplified by big data analytics education. In *22nd European Conference on Information Systems*.

[Katuk and Ryu(2010)] Norliza Katuk and Hokyoung Ryu. 2010. Finding an optimal learning path in dynamic curriculum sequencing with flow experience.

[Lavie and Agarwal(2007)] Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. pages 228–231.

[Mihalcea and Tarau(2004)] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona,

Spain. Association for Computational Linguistics.

[P. Liu and Wang(2012)] W. Heng P. Liu, L. Li and B. Wang. 2012. Hlda based text clustering. *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems*, pages 1465–1469.

[Page et al.(1998)Page, Brin, Motwani, and Winograd] Larry Page, Sergey Brin, R. Motwani, and T. Winograd. 1998. The pagerank citation ranking: Bringing order to the web.

[Wan and Xiao(2008)] Xiaojun Wan and Jianguo Xiao. 2008. Single document keyphrase extraction using neighborhood knowledge.