Constructing a Linked Concept Map from Textbooks Using DBpedia Spotlight

Patrick W. Miller
Georgia Institute of Technology
pmiller42@gatech.edu

ABSTRACT
This paper presents a preliminary study on constructing concept maps with prerequisite structure from textbooks. Topic recognition, extraction and disambiguation are performed with DBpedia Spotlight, and the prerequisite structure is identified through the location of the occurrence of each concept in the text. This study shows that the simple use of DBpedia is a viable method for automatically creating a concept map from a corpus of textbooks.

Author Keywords
Concept maps; textbooks; knowledge graphs; DBpedia Spotlight; prerequisite structure; text mining

Categories and Subject Descriptors
I.2.7 **[Natural Language Processing]**: Text analysis; I.2.6 **[Learning]**: Knowledge acquisition; Concept learning; I.7.5 **[Document and Text Processing]**: Document Capture—Document Analysis;

INTRODUCTION
Textbooks contain massive amounts of knowledge and have been foundational in guiding the learning process for many students. This knowledge has largely been untapped by data-driven approaches seeking to summarize educational concepts. This exploratory study examines how using natural language processing (NLP) and linking to DBpedia can produce a prerequisite concept map from textbooks. The directed knowledge graph that is created can be used in Intelligent Tutors to create prerequisite readings and adjust learning plans for different types of students. The final concept map provides a novel, automated view of prerequisite structure across educational topics and helps with understanding the topics a textbook covers along with an idea of its required expertise level.

There are a few research projects and papers aimed at using NLP to link textbooks together. Guerra et al. examines linking multiple textbooks using probabilistic topic models [6]. From their experiment they conclude that latent Dirichlet allocation (LDA) should be able to successfully link textbooks. Huang et al. extend on this approach through a variety of different models and combine student learning patterns to analyze personalized learning [7]. Meng et al. use a semantic-based approach to knowledge component extraction from textbooks instead of just relying on key terms [9]. Chen et al. discuss building concept maps from academic topics [3]. They take the keywords explicitly listed in the metadata of journal articles as concepts and define similarity of concepts through the co-occurrence of these keywords. Chaplot et al. build a Prerequisite Structure Graph using online educational material and student activity [2]. In Wang et al., the authors present the idea of using Wikipedia as a base of knowledge for extracting concepts from textbooks [11, 12]. They build a concept hierarchy by considering both “local relatedness” and “global coherence”, which examines both chapter by chapter concepts and the similarity of concepts throughout. Prerequisite relationships are defined sequentially by looking at concepts on a chapter-by-chapter basis. This study follows the ideas of Wang et al. in relating concepts together.

Instead of using traditional NLP methods to extract topics, this study utilizes DBpedia Spotlight, an API for annotating text with concepts defined in DBpedia [5]. There have been many examples of annotating text with Wikipedia or DBpedia topics [10, 13]; however, this methodology has not been extensively used in linking textbooks. Contractor et al. used DBpedia Spotlight to label education content in textbooks from India, but they did not also link together concepts in a prerequisite structure [4].

The contributions of this study are mostly in the successful utilization of DBpedia Spotlight as an efficient concept extraction tool. While the heuristics used here for linking topics together in a prerequisite structure are not complex, they manage to perform well at identifying major related concepts.

METHODOLOGY

Two subjects were chosen for this exploratory study: biology and psychology. For both of these subjects, a few textbooks were selected (Table 1) – some introductory and others more specialized. Biology and psychology have relatively well-defined terms, so it was expected that topic disambiguation would be easier. The workflow for developing the concept map is as follow:

1. For each textbook:
   i. Load the textbook and trim off the table of contents and appendices
   ii. Annotate the text using DBpedia Spotlight
   iii. Extract concept and location from each annotation
   iv. Divide the textbook into 50 parts
   v. For each part, count the number of times each concept occurs
   vi. Calculate average location heuristics for each concept
   vii. Calculate the distance pairwise between each concept
2. Combine each concept pair's distance, weighted by their co-occurrence in each textbook
3. Filter out concept pairs that don't co-occur at least 50 times
4. Filter out concept pairs that are on average more than 8% apart in a textbook
5. Construct directed graph using from the adjacency list: (concept, concept, distance)

Title	Author	Subject
Molecular Biology: Principles and Practice	Cox, Michael	Biology
What Is Life? A Guide to Biology	Phelan, Jay	Biology
Biology of Plants	Raven, Peter	Biology
Life: The Science of Biology	Sadava, David	Biology
Abnormal Psychology	Comer, Ronald	Psychology
Psychology: A Concise Introduction	Griggs, Richard	Psychology
Psychology	Myers, David	Psychology
Exploring Psychology	Myers, David	Psychology
Introducing Psychology	Schacter, Daniel	Psychology

Table 1. The 9 textbooks from the fields of biology and psychology used in this study.

The workflow uses 0.6 as the confidence parameter for DBpedia Spotlight. This parameter setting was settled on after surveying 30 graduate students with knowledge of psychology or biology on the relevancy of concepts returned by different parameterizations of DBpedia Spotlight.

RESULTS

The final result of the study is a directed graph representing prerequisite structure between 480 total concepts with over 12,000 total relationships. The following sections summarize some of the interesting parts.

Key Concepts

Counting the number of occurrences for each concept provides a high level view of what the most important topics are in each textbook. Because only 9 textbooks were used in this study, the concepts are ranked according to their raw counts. With more textbooks, term frequency-inverse document frequency (tf-idf) could be used to more accurately represent the key concepts. Even with just the raw counts, the key concepts are almost exclusively relevant to the subject. Figure 1 contains the top 5 concepts by raw counts for each textbook.

MOLECULAR BIOLOGY: PRINCIPLES AND PRACTICE - DNA, Protein, Messenger RNA, RNA, Transcription (genetics)

WHAT IS LIFE? A GUIDE TO BIOLOGY - DNA, Allele, Bacteria, Protein, Ampere

BIOLOGY OF PLANTS - Meristem, Protein, Xylem, DNA, Micrometre

LIFE: THE SCIENCE OF BIOLOGY - Protein, DNA, Enzyme, Gene, Oxygen

ABNORMAL PSYCHOLOGY - Schizophrenia, Major depressive disorder, Anxiety, United States, Depression (mood)

PSYCHOLOGY: A CONCISE INTRODUCTION - Neuron, Classical conditioning, Schizophrenia, Operant conditioning, Long-term memory

PSYCHOLOGY - Psychology, Anxiety, Schizophrenia, Major depressive disorder, Depression (mood)

EXPLORING PSYCHOLOGY – Information technology, Psychology, Anxiety, Major depressive disorder, Depression (mood)

INTRODUCING PSYCHOLOGY - Neuron, Psychology, Anxiety, Psychotherapy, Classical conditioning

Figure 1: Top 5 concepts by count in each textbook. With a few exceptions, the top concepts that are extracted with DBpedia Spotlight are pertinent to the subject matter – even without performing something like tf-idf.

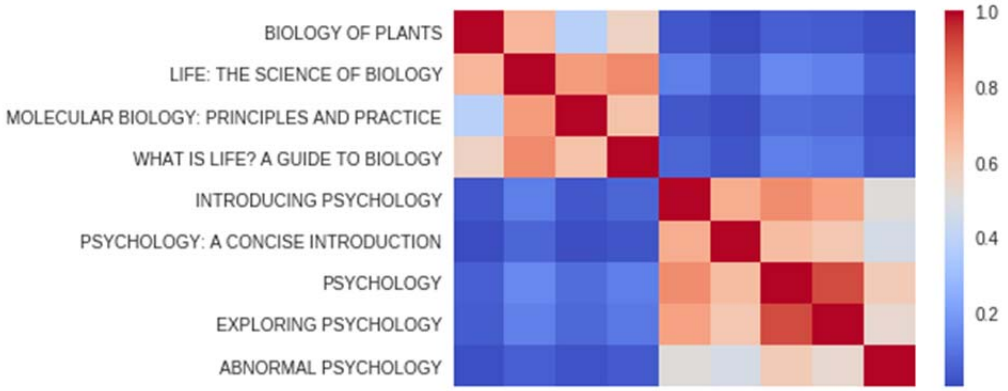

Figure 2: Correlation heat map showing relatedness of textbooks. There is clear clustering between the two subjects – biology books have more concepts in common with each other than with psychology textbooks.

Textbook Similarity

Another simple view on the effectiveness of DBpedia Spotlight with topic extraction is in how the concepts relate across textbooks. Figure 2 presents a correlation heat map of the concepts extracted from each textbook. The correlations clearly show a clustering of the two subjects. Biology textbooks have more concepts in common with each other than they do with psychology textbooks. *Molecular Biology* is the least similar to the other biology textbooks likely because it is an advanced level textbook and thus has more advanced concepts not shared in the other titles.

Concept Map

The goal of this study was to create a directed concept map with prerequisite structure. Because the end result is a directed graph with over 12,000 edges, the graph itself cannot be displayed here. Instead, figure 3 presents two example concepts and their most commonly related concepts. The majority of the prerequisite relationships are relevant and accurate; however, a few either do not fit or are too general. For example, "DNA" is a concept that is prerequisite to "Gene", but "Neurotransmitter" is not a prerequisite for "United States." This phenomenon could likely be solved with a larger set of textbooks and the utilization of tf-idf.

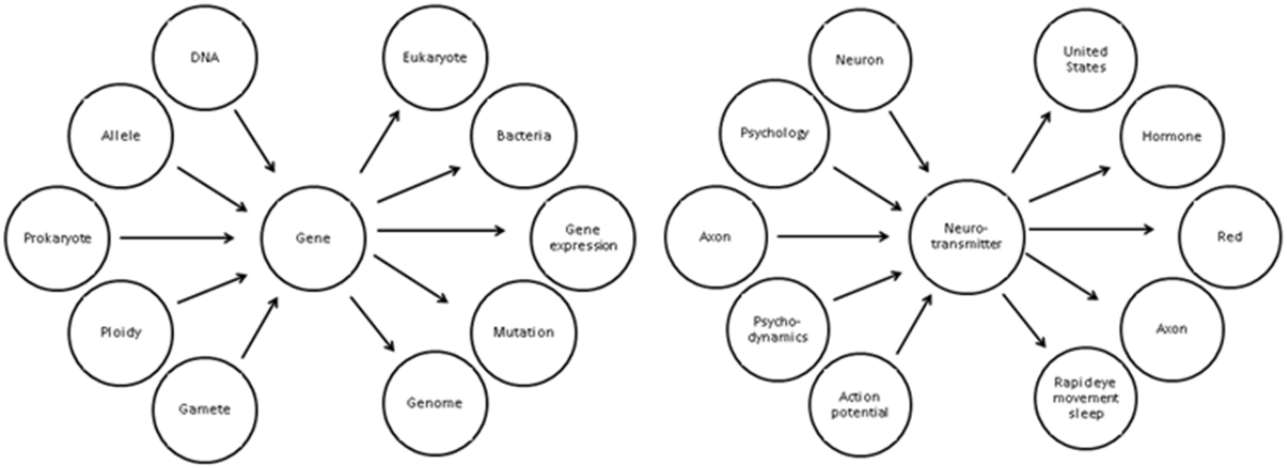

Figure 3: Most common related concepts for "Gene" (left) and "Neurotransmitter" (right). All of the relationships are within an average distance of 8% (measured as % of a textbook) of each other. "Axon" is listed both before and after neurotransmitter because they both occur on average in the same part of the textbook. The full concept map can be found at https://github.com/patrick-miller/textbook-concept-map.

CONCLUSION

This exploratory study has shown the effectiveness of using DBpedia Spotlight to extract concepts from textbooks and link them together in a prerequisite concept map. Using only 9 textbooks, the process identified key concepts that were relevant to the subject matter and consistent with the division between biology and psychology. Furthermore, the precedence relationship between concepts was largely relevant and accurate.

Future work

Given the promising results, I plan to expand the number of textbooks and subjects included. Having more textbooks will allow for better filtering of concepts and a more consistent concept map. On the other hand, including subjects that may not have well-defined terms may present some difficulties in topic identification and disambiguation.

In order to better prune the concepts that are included in the graph, I will explore using term frequency-inverse document frequency (tf-idf). Very general terms, such as "the United States", were included in the concept map even though they were not relevant to the topics covered in the textbooks. With a larger corpus of textbooks, I can better target concepts that occur more frequently together than not.

Finally, there needs to be more work on optimizing and pruning the edges in the graph. Relative distance between concepts is only a simple heuristic. Other methodologies for identifying a more sparse prerequisite structure—such as suggested in Wang et al. [11, 12]—would prove useful. Additionally, a more rigorous evaluation of the aptness of the final concept map would give more confidence that the methodology works.

ACKNOWLEDGMENTS

A big thank you to Macmillan Learning for allowing me to use some of their textbooks in this project. Also, to Ken Brooks and David Joyner for the guidance.

REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Kannan, and K. Kenthapadi. Data mining for improving textbooks. In *SIGKDD Explorations Newsletter*, 13(2):7-19, 2012.

[2] D. Chaplot, Y. Yang, J. Carbonell, and K. Koedinger. Data-driven Automated Induction of Prerequisite Structure Graph. In *EDM*, 2016.

[3] N. Chen, Kinshuk, C. Wei, and H. Chen. Mining e-Learning domain concept map from academic articles. In *Computers & Education*, 50:1009-1021, 2008.

[4] D. Contractor, K. Popat, S. Ikbal, S. Negi, B. Sengupta, and M. K. Mohania. Labeling Educational Content with Academic Learning Standards. In *SIAM International Conference on Data Mining*, pages 136-144. SIAM, 2015.

[5] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)* 2013.

[6] J. Guerra, S. Sosnovsky, and P. Brusilovsky. When one textbook is not enough: Linking multiple textbooks using probabilistic topic models. In *Scaling up Learning for Sustained Impact*, pages 125-138. Springer, 2013.

[7] Y. Huang, M. Yudelson, S. Han, and P. Brusilovsky. A Framework for Dynamic Knowledge Modeling in Textbook-Based Learning. In *UMAP*, pages 141-150. ACM, 2016.

[8] H. Liu, W. Ma, Y. Yang, and J. Carbonell. Learning Concept Graphs from Online Educational Data. In *Journal of Artificial Intelligence Research*, 55: 1059-1090, 2016.

[9] R. Meng, S. Han, Y. Huang, D. He, and P. Brusilovsky. Knowledge-Based Content Linking for Online Textbooks. In *IEEE*, pages 18-25. ACM, 2016.

[10] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM*, pages 233-242. ACM, 2007.

[11] S. Wang, C. Liang, Z. Wu, K. Williams, B. Pursel, B. Brautigam, S. Saul, H. Williams, K. Bowen, and C. L. Giles. Concept Hierarchy Extraction from Textbooks. In *DocEng*. ACM, 2015.

[12] S. Wang, A. G. Ororbia II, Z. Wu, K. Williams, C. Liang, B. Pursel, and C. L. Giles. Using Prerequisites to Extract Concept Maps from Textbooks. In *CIKM*, pages 317-326. ACM, 2016.

[13] G. Weikum and M. Theobald. From Information to Knowledge: Harvesting Entities and Relationships from Web Sources. In *PODS*. ACM 2010.