Data-driven Automated Induction of Prerequisite Structure Graphs

Devendra Singh Chaplot
School of Computer Science
Carnegie Mellon University
dchaplot@cs.cmu.edu

Yiming Yang
School of Computer Science
Carnegie Mellon University
yiming@cs.cmu.edu

Jaime Carbonell
School of Computer Science
Carnegie Mellon University
jgc@cs.cmu.edu

Kenneth R. Koedinger
School of Computer Science
Carnegie Mellon University
koedinger@cmu.edu

ABSTRACT

With the growing popularity of MOOCs and sharp trend of digitalizing education, there is a huge amount of free digital educational material on the web along with the activity logs of large number of participating students. However, this data is largely unstructured and there is hardly any information about the relationship between material from different sources. We propose a generic algorithm to use educational material and student activity data from heterogeneous sources to create a Prerequisite Structure Graph (PSG). A PSG is a directed acyclic graph, where the nodes are educational units and the edges specify the pairwise ordering of the units in effective teaching by instructors or for effective learning by students. We propose an unsupervised approach utilizing both text content and student data, which outperforms to supervised methods (utilizing only text content) on the task of estimating a PSG.

1. INTRODUCTION

Students need prior knowledge for thorough understanding of educational content. This need imparts an implicit order in learning educational concepts. Determining this order requires significant human time and effort. Furthermore, relying on expert knowledge to determine this order is subject to inconsistencies due to 'expert blind spot' [8]. We aim to leverage free educational material on the web, and huge amount of student activity logs associated with them, to create a universal Prerequisite Structure Graph (PSG). We define PSG as a directed acyclic graph, where the nodes are the universal concepts in an educational domain and the edges specify the pairwise ordering of concepts in effective teaching by instructors or for effective learning by students. The proposed unsupervised methods utilize both textual content and student performance data to perform better than supervised methods utilizing textual content. They can be generalized to find the learning order between any pair of educational elements from heterogeneous resources, at any level of granularity (courses, units, modules, skills, etc.).

The rest of the paper is divided as follows. The related work pertaining to the proposed methods is discussed in Section 2. Section 3 describes the dataset used for experiments. Performance-based and text-based unsupervised induction of a PSG are described in Sections 4 and 5, respectively. We describe the method of combining text-based and performance-based approaches in Section 6. Experiments and results are presented in Section 7. In Section 8, we analyze whether the concepts extracted by proposed methods are meaningful. Conclusions and future directions are covered in Section 9.

2. RELATED WORK

Currently, the construction of Concept Graphs majorly depends on manual work of domain experts. Recent work by [10] on Concept-Graph-Learning (CGL), focuses on determining the relationship between different University courses and MOOCs by inferring concepts from course descriptions. The proposed methods are completely unsupervised as compared to supervised CGL which requires partial instructor-specified links. One other recent work includes extracting a concept-hierarchy from textbooks [9], where the focus is only on extracting the hierarchies between concepts and the learning is only done at the concept level. We differentiate ourselves from this work with the fact that we learn the prerequisite relationships between educational concepts rather than hierarchies, and our method is generalizable to any granularity of educational elements.

Another indicator of prerequisite links between educational elements is student performance. An early approach to inferring prerequisite graphs from student performance data is knowledge spaces [2], which uses associations between student success on different classes of tasks to infer prerequisite relationships. The essential idea is that if students are highly likely to get tasks of type A correct (e.g., finding least common multiples) conditioned on getting tasks of type B correct (e.g., adding fractions with unlike denominators) but not the other way around (i.e., many students that can find common multiples fail at adding fractions), then A is a pre-

requisite of B. Subsequently, algorithms for inferring cognitive models of student learning from data have been developed and it is possible to infer prerequisite relationships from the results of these models [1]. The methods we propose are different as we utilize not only the student performance data, but also student activity data along with large amounts of text in course material. Also, previous approaches assume that there is no learning between attempts at different problems, which is suitable for standardized testing scenario but not true for student performance logs of complete courses.

3. DATASET

We use the text content and student activity and quiz performance data from Georgia Tech's "Introduction to Psychology" MOOC which uses content from the Open Learning Initiative of Carnegie Mellon University [6]. The course spans over 12 weeks and a major topic of Psychology (like intelligence, personality, psychological disorders, etc.) was covered in each week of class. For each week, the text content from the corresponding unit(s) was extracted. The unit(s) covered in each week are shown in Table 1. On an average, each unit contained 12545 word instances with a standard deviation of 3730. For simplicity, we will use Unit i to denote the content covered in Week i, although the content covered in week i might include multiple units in the course. Besides the text inside course units, we also used text in the weekly quizzes separately to evaluate our text-based methods.

The course also contained ungraded practice activities within each unit. At the end of each week (from week 1 to week 11), students were assessed by a high stakes quiz containing questions from content covered in the corresponding week. The dataset includes the number of interactive activities and quiz scores of 1154 students for each week.

This dataset is ideal for our analysis since it has both the textual data of course material and the student activity and quiz performance data. We aim to predict prerequisite links between weeks using this data, which will imply prerequisite links between corresponding units. For example, a prerequisite link from Unit 9 to Unit 11 implies prerequisite link from Personality to Disorders, or in other words, a student who has learned Personality will be better able to learn Disorders.

For evaluation, the dataset was first annotated by three non-experts who determined whether a prerequisite link exists between content covered in any two units. If a prerequisite link exists from Unit i to Unit j, we call it a positive link, and conversely, if there is no link, it is called a negative link. The average percentage agreement for positive links between each pair of annotator was 29.6% while the percentage agreement for positive links among all the annotators was 18.7%. Since the inter-annotator agreement was very low, we got the dataset annotated by a domain expert. All the links marked positive by all three non-expert annotators were also marked positive by the domain expert, except one link. Finally, we took 15 links marked positive and domain expert, and 1 more link marked positive by all non-expert annotators as the set of positive links. Therefore, among 110 possible links, 16 links were labeled as positive and rest negative. Note that 55 out of 110 possible links are backward (i.e. from Unit i to Unit j such that $i > j$), which should be implicitly negative, but we will not use the information about the ordering of units in any of the proposed methods so that our methods are generalizable to any pair of educational elements: modules, chapters or whole courses.

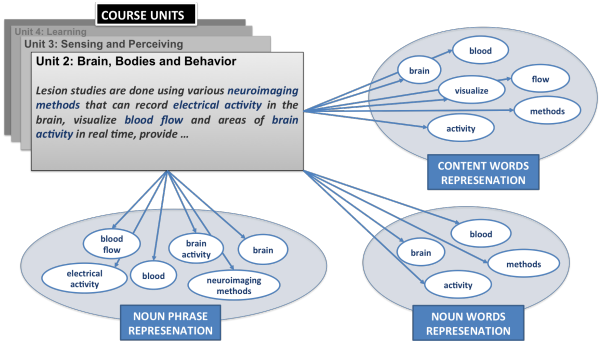


Figure 1: Example of three types of concept space representation schemes: Content Words, Noun Words and Noun Phrases.

Week	Unit(s) Covered
1	Introduction and Methods
2	Brains, Bodies, and Behavior
3	Sensing & Perceiving
4	Learning
5	Memory
6	Language and Intelligence
7	Lifespan development
8	Emotion and Motivation
9	Personality
10	Psychology in Our Social Lives
11	Disorders

Table 1: Unit(s) covered in each week of "Introduction to Psychology" course.

4. TEXT-BASED METHODS

Each educational unit consists of a set of canonical educational concepts. The text content in each educational unit can be used to find the concepts involved in it. The set of concepts in all units is defined as the universal concept space [10]. We define three concept space representation schemes as follows:

- **Content Words Representation (Word)**: The set of content words (Nouns, Verbs, Adjectives and Adverbs) occurring in the course content is used as the concept space. The words are lemmatized using MIT Java Wordnet Interface (JWI) [3].
- **Noun Words Representation (Noun)**: In this representation scheme, we only use set of nouns occurring in the course content as the concept space rather than all content words. These are again lemmatized using MIT JWI.
- **Noun-Phrase Representation (NP)**: In this representation scheme, the set of noun phrases (of depth less than 5) occurring in the course content is used as the concept space.

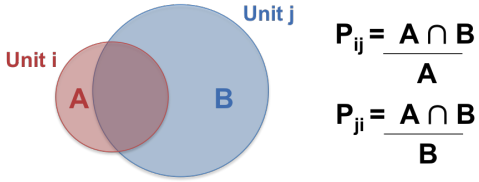


Figure 2: Overlap Method

An example of these three types of representation schemes is shown in Figure 1. The Concept space can be represented using other schemes such as Sparse Coding of Words and Distributed Word Embedding, but these produce latent concepts, which are not human understandable. Furthermore, previous results indicate word-based Representation scheme is more effective than latent concept based representation schemes [10].

Let the total number of the concepts in the concept space be p. Then the educational content in each unit can be represented by a p-dimensional vector, where each element is the frequency of corresponding concept (word, noun or noun-phrase) in the text content of the unit. The concept frequency can be normalized using the following quantities:

- **Collection Frequency (CF):** Total number of occurrences of the word in the collection or in our case, course. This normalizes concept frequencies such that all concepts are given equal weightage.
- **Document Frequency (DF):** Number of documents or in our case, units, that contain the concept. This gives less weightage to words occurring in most units such as module, learning objective, psychology, etc.
- **Wordnet Frequency (WF):** The frequency of word given in WordNet which represents the frequency of word in naturally occurring domain-independent text. This re-scales the frequencies such that domain-specific psychology terms have more weightage than generic terms.

We first describe an unsupervised method which determines prerequisite links based on only the text overlap between educational units. The key idea is that course unit u_i is a prerequisite of u_j to the extent that u_i is a probabilistic subset of u_j (i.e., most concepts involved in u_i are mostly involved in u_j) and u_j is not a probabilistic subset of u_i (i.e., most concepts involved in u_j are not involved in u_i). This idea of using asymmetry in computing the probabilistic subset is motivated by the theory of knowledge spaces [2], but we use text information rather than performance data.

Let x_i be a vector denoting the concept space representation of unit u_i. The length of this vector is the total number of the concepts. Each element of this vector is the frequency of the concept in the unit or one of the normalized versions of concept frequency (CF, DF or WF). The intuitive gloss on how we compute the probability that x_i is a probabilistic subset of x_j is by dividing the size of the intersection of x_i and x_j by the size of x_i (A is a subset of B if $A \cap B = A$ and less so to the extent that $A \cap B < A$, see Figure 2). Mathematically, we define P_{ij} as the ratio of sum of elements of pairwise minimum of x_i and x_j to the sum of elements of x_i:

$$P_{ij} = \frac{sum(min(x_i, x_j))}{sum(x_i)} \quad (1)$$

Then P_{ij} is the weight of the prerequisite link from unit i to unit j, which ranges from 0 to 1.

5. PERFORMANCE-BASED METHODS

Our particular approach for unsupervised induction of PSG based on student performance data grows out of recent analysis of student performance [6] which concludes that interactive activities are more indicative of learning gains than video watching or online text reading. In subsequent analysis, it was found that student learning within a course unit is more highly predicted by their activity within that unit than within other units [7]. However, there is an additional learning outcome boost associated with greater activities before a target unit, but not with greater activities after that unit. This result is consistent with there being prerequisite relationships between prior and later units and was the inspiration for new algorithm development on performance-based PSG inference.

The key idea behind the proposed performance-based methods is that more the activity in unit i predicts success in unit j, the more likely is unit i a prerequisite of unit j. This means that if students who do more activities in week i perform better in the week j quiz, as compared to students who do fewer activities in week i, then there is an evidence for a prerequisite link from content in week i to week j. Let

y_j be Quiz Scores in week j,

x_i be the number of interactive activities done in week i, and

w_{ij} be the parameters denoting the effect of activities in week i on quiz in week j, which we want to estimate. The value of the parameter is the strength of corresponding prerequisite relationship.

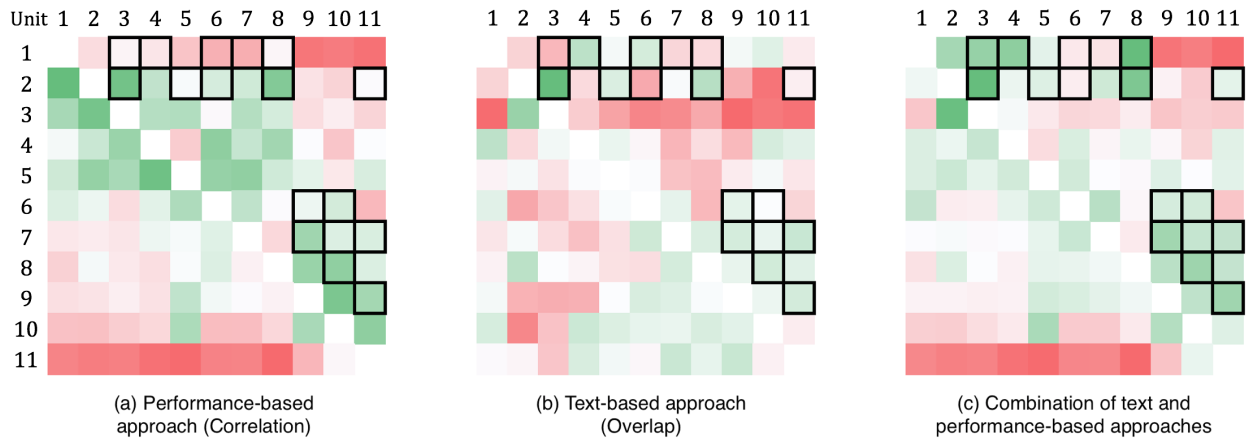
We define two methods for predicting prerequisite links using student performance data:

- **Correlation:** The effect of activities in week i on the performance in week j is estimated by the correlation between the number of activities by students in week i and the quiz scores of students in week j quiz. Let $\rho(X, Y)$ be the Pearson correlation coefficient between X and Y. Then,

$$w_{ij} = \quad \rho(x_i, y_j) = \frac{cov(x_i, y_j)}{\sigma_{x_i}\sigma_{y_j}}$$

- **Multiple Linear Regression:** We compute a linear regression for student quiz scores across the 11 units of the course where the dependent variable is student quiz score for the target unit and the independent variables are number of activities students do within each unit. Let $\boldsymbol{w_j} = [w_{1j}, w_{2j}, ..., w_{11j}]$, be a vector denoting the effects of activities in all weeks on quiz score of week j. We define multiple linear regression using lasso regularization as follows:

$$\boldsymbol{w_j^*} = \underset{\boldsymbol{w_j}}{\text{argmin}} \quad \sum_n (y_j - \boldsymbol{x^T w_j})^2 + \lambda \|\boldsymbol{w_j}\|$$



(a) Performance-based approach (Correlation) (b) Text-based approach (Overlap) (c) Combination of text and performance-based approaches

Figure 3: The heat map of strength of links from Unit i to Unit j for (a) Performance-based (Correlation) approach, (b) Text-based (Overlap) approach and (c) Combination of both. The black boxes represent the prerequisite links labeled by domain experts. Note that there is no link from Unit 7 to Unit 10, even though it appears to be surrounded by a black box.

Method Name	Method Type	Data Utilized	MAP	AUC
Regression	Unsupervised	Performance	0.562	0.571
Correlation	Unsupervised	Performance	0.604	0.720
Overlap	Unsupervised	*Quiz* Text	0.693	0.700
Overlap	Unsupervised	*Unit* Text	0.743	0.710
Overlap + Corr	Unsupervised	Performance & *Quiz* Text	0.798	0.820
Overlap + Corr	Unsupervised	Performance & *Unit* Text	**0.837**	**0.840**
CGL[10]	Supervised	*Unit* Text & Labeled links	0.747	0.820

Table 2: Comparison of all methods

6. COMBINING TEXT-BASED AND PERFORMANCE-BASED METHODS

We observed that most of the prediction errors in unsupervised text-based and performance-based methods were due to false-positives. This is because the dataset is imbalanced towards negative class with 85.45% negative labels. Unsupervised systems lacking this information predict positive and negative instance without any prior bias. In order to reduce the errors due to false positives, we propose to predict a positive link only when both methods indicate a positive link.

We get two square matrices of dimension equal to the number of units in the course, one each from text-based and performance-based methods. The $(i, j)^{th}$ element of these matrices represents the weight of the prerequisite link from unit i to unit j obtained from the corresponding method. We combine the two methods by first forcing diagonal entries (self-links) to be 0, then standardizing both the matrices such that both have zero mean and equal variance and then just applying a pairwise minimum over these standardized matrices. This approach predicts a link between any ordered pair of units only if both methods suggest that there should be a link between them. The combination of both methods using a pairwise minimum operation performed better than combination using pairwise summation, pairwise maximum and pairwise product. We also explored more complex models for combination, but found no evidence to justify model complexity.

7. EXPERIMENTS & RESULTS

We gathered and annotated the dataset for experiments as described in Section 3. For evaluation, we used macro-averaged Mean Average Precision (MAP) [5] and Area under ROC Curve (AUC) [4], which are popular metrics in ranked list retrieval and link detection evaluations [10].

The first two rows in Table 2 show the performance of two proposed performance-based methods: Multiple Linear Regression and Correlation. As the Correlation method performed better than Regression method (MAP 0.604 vs 0.562 and AUC 0.720 vs 0.571), we will use Correlation method for combining with text-based methods. The third and fourth column in Table 3 show the performance of text-based Overlap method over different concept space representation types and normalization types. The last two columns of this table show the performance of Overlap method combined with Correlation method. We compare this combined method to supervised Concept Graph Learning algorithm (CGL) [10]. The best results of all methods are summarized in Table 2, which shows that the unsupervised method which combines text-based and performance-based approaches outperforms supervised concept graph learning algorithm by a considerable margin (MAP 0.837 vs 0.747 and AUC 0.840 vs 0.820). As seen Table 3, the combined method performs better than CGL for most concept space representation and normalization types. Note that as compared to supervised CGL method, the proposed method ('Overlap+Corr') utilizes performance data in addition to the text content in educational material but doesn't require labeled links from experts. The results in Table 3 also suggest that on an average, Noun Phrase concept space representation works best for all text-based methods, although there is no clear winner among Normalization types.

Method Name		Overlap		CGL		Overlap+Corr	
Method Type		Text Unsupervised		Text Supervised		Perf+Text Unsupervised	
Rep Type	Norm Type	MAP	AUC	MAP	AUC	MAP	AUC
Word	None	0.656	0.640	0.685	0.789	0.686	0.750
	CF	0.667	0.680	0.742	0.805	0.717	0.800
	DF	0.638	0.660	0.638	0.766	0.836	0.830
	WF	0.693	0.660	0.676	0.781	0.730	0.800
NP	None	0.661	0.680	0.722	0.789	0.745	0.810
	CF	0.703	0.710	**0.747**	**0.820**	0.792	0.820
	DF	**0.743**	**0.710**	0.572	0.773	**0.837**	**0.840**
	WF	0.717	0.710	0.743	0.805	0.748	0.820
Nouns	None	0.734	0.670	0.751	0.805	0.746	0.820
	CF	0.681	0.680	0.687	0.797	0.821	0.810
	DF	0.721	0.680	0.535	0.766	0.755	0.830
	WF	0.738	0.680	0.696	0.797	0.748	0.820

Table 3: Comparison of different concept space representation schemes (Rep Type) and different Normalization schemes (Norm Type) over different text-based methods. CF, DF and WF refer to Collection Frequency, Document Frequency and WordNet Frequency, respectively, as described in Section 4. The best AUC and MAP scores for each method are marked in **bold**.

We analyzed the weights of links predicted by different methods to understand how the combination of text and performance based methods affects our prediction. Figure 3 shows a heat map of strength of links between all pairs of Units. Each $(i, j)^{th}$ element in the matrix represents the strength of link from Unit i to Unit j, where green is denoting higher strength and red is denoting lower. Note that the heat of the colors is determined by relative value of the weights in one matrix and not absolute values across matrices. This is because AUC and MAP metrics evaluate relative value of predicted weights rather than absolute values. The black boxes represent the prerequisite links labeled by experts. The figure indicates that the estimates of performance and text-based approaches compliment each other to give better estimates when combined.

8. DISCUSSIONS

Figure 4 demonstrates a subset of prerequisite links identified by the proposed method and a subset of overlapping concepts occurring in them in the concept space. We would like to analyze whether the concepts identified by the proposed method are meaningful. Consider the relationship between Unit 11, 'Emotion and Motivation' and Unit 13, 'Psychology in Our Social Lives'. All the proposed methods estimate significant weights for link from Unit 11 to Unit 13. Figure 6 shows a part of the concept space representation using Content Words Representation scheme for these units. Overlap method indicates a strong prerequisite link from Unit 11 to Unit 13 due to significant overlap between the concepts in these units. Looking into the contents of these units, the Unit 11, 'Emotion and Motivation' consists of 'Human Motivation' module which involves understanding the motivation behind sexual behavior. It introduces concepts of 'attractiveness', 'proximity' and 'similarity' as motivating factors behind sexual interest. Unit 13, 'Psychology in Our Social Lives' requires the understanding of these concepts in order to understand 'Interpersonal Attraction' in 'Close Relationships' module. Since there are more concepts in Unit 13 like 'personality', 'aggression', 'stimulus', 'judgment', etc. which are not present in Unit 11, $P_{11,13}$ is greater than $P_{13,11}$. Thus, the concepts extracted by the proposed Concept Representation schemes appear to be interpretable and meaningful.

Similarly, we also try to interpret the performance-based results by inspecting the text of the interactive activities within the course. For example, the interactive activities in 'Human Motivation' module correspond to understanding concepts of 'attractiveness', 'proximity' and 'similarity'. The quiz at the end of unit on 'Psychology in Our Social Lives' also contains a question about role of proximity and similarity in interpersonal attraction. Therefore, the students who do more activities in week 8 (involving Unit 11 content) perform better on the week 10 (involving Unit 13) quiz (as compared to students who do fewer week 8 activities) and thus, performance-based approaches identify this relationship. Figure 5 shows the average number of activities of students in prior units as a function of their quiz scores in later unit for set of positive and negative links. The average number of activities in prerequisite units is greater than non-prerequisite units for all quiz scores which is a possible explanation of the effectiveness of performance-based methods. Also, the correlation between number of activities and quiz scores suggests that interactive activities are indicative of learning gains.

9. CONCLUSIONS & FUTURE WORK

We proposed completely unsupervised methods to leverage freely available textual content in educational resources and student performance & activity data for predicting prerequisite structure graph between arbitrary educational resources. Three different concept space representation schemes have been used for text-based methods with a variety of normalization methods for concept frequencies. We also show that when unsupervised text-based and performance-based methods are combined, they supplement each other to outper-

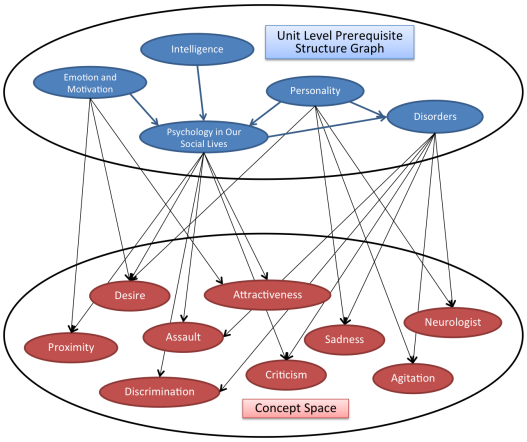


Figure 4: Demonstration of prerequisite links between different units in 'Introduction to Psychology' Course and a subset of overlapping concepts.

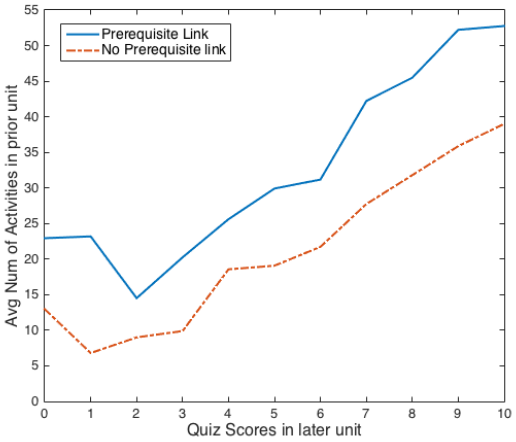


Figure 5: The average number of activities of students in prerequisite units as a function of their quiz scores in post-requisite unit.

form sophisticated supervised methods. Concepts extracted using the proposed representation schemes seem to be interpretable and meaningful from educational perspective.

While the results are encouraging, a limitation of the current work is the size of the dataset. Although the text content in the course and student activity and performance data is rich, the number of positive prerequisite relations in the dataset is low. Validation of proposed methods on diverse educational data from different courses is required to test their generalizability and scalability. Furthermore, conducting a long-term user-study involving students to verify if the predicted prerequisites help them improve their performance over a course, would be useful.

10. REFERENCES

[1] M. C. Desmarais, A. Maluf, and J. Liu. User-expertise modeling with empirically derived probabilistic implication networks. *User modeling and user-adapted interaction*, 5(3-4):283–315, 1995.

[2] J.-P. Doignon and J.-C. Falmagne. Spaces for the assessment of knowledge. *International journal of man-machine studies*, 23(2):175–196, 1985.

[3] M. A. Finlayson. Java libraries for accessing the princeton wordnet: Comparison and evaluation. In *Proceedings of the 7th Global Wordnet Conference, Tartu, Estonia*, 2014.

[4] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[5] K. Kishida. *Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments*. National Institute of Informatics Tokyo, Japan, 2005.

[6] K. R. Koedinger, J. Kim, J. Z. Jia, E. A. McLaughlin, and N. L. Bier. Learning is not a spectator sport: Doing is better than watching for learning from a mooc. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, pages 111–120, 2015.

[7] K. R. Koedinger, E. A. McLaughlin, J. Z. Jia, and N. L. Bier. Is the doer effect a causal relationship? how can we tell and why itâĂŹs important. In *Proceedings of the Sixth International Learning Analytics & Knowledge Conference*, 2016.

[8] M. J. Nathan, K. R. Koedinger, and M. W. Alibali. Expert blind spot: When content knowledge eclipses pedagogical content knowledge. In *Proceedings of the Third International Conference on Cognitive Science*, pages 644–648. Citeseer, 2001.

[9] 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 *Proceedings of the 2015 ACM Symposium on Document Engineering*, DocEng '15, pages 147–156, New York, NY, USA, 2015. ACM.

[10] Y. Yang, H. Liu, J. Carbonell, and W. Ma. Concept graph learning from educational data. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 159–168. ACM, 2015.

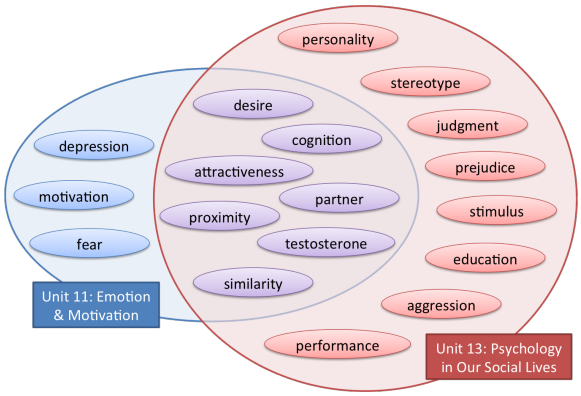


Figure 6: Demonstration of overlap of concepts between units on 'Emotion and Motivation' and 'Psychology in Our Social Lives' and prediction of prerequisite link using Overlap method.