Web as a textbook: Curating Targeted Learning Paths through the Heterogeneous Learning Resources on the Web.

Igor Labutov
Cornell University
iil4@cornell.edu

Hod Lipson
Columbia University
hod.lipson@columbia.edu

ABSTRACT

A growing subset of the web today is aimed at *teaching* and *explaining* technical concepts with varying degrees of detail and to a broad range of target audiences. Content such as tutorials, blog articles and lecture notes is becoming more prevalent in many technical disciplines and provides up-to-date technical coverage with widely different levels of prerequisite assumptions on the part of the reader. We propose a task of organizing heterogeneous educational resources on the web into a structure akin to a textbook or a course, allowing the learner to navigate a sequence of webpages that take them from point A (their prior knowledge) to point B (material they want to learn). We approach this task by 1) performing a shallow term-level classification of what concepts are *explained* and *assumed* in any given text, and 2) using this representation to connect web resources that explain concepts to those web resources where the same concepts are assumed. The main contributions of this paper are 1) a supervised classification approach to identifying explained and assumed terms in a document and 2) an algorithm for finding optimal paths through the web resources given the constraints of the user's goal and prior knowledge.

Keywords

web resources; optimizing learning

1. INTRODUCTION

No scholar is born at the frontier of knowledge — early learning and lifelong learning both play a defining role in shaping the research vector of an academic [7]. More alarming, recent research [6] demonstrates that the pre-career idle time of an up-and-coming researcher has been on the steady rise during the last century, attributing to the "burden of knowledge" phenomenon — the inflation of the body of prerequisite prior knowledge to be mastered before being able to contribute to the field with original research. The hypothesis of [9, 10] is that facilitating effective early and lifelong learning practices is a viable way for easing the "burden of knowledge".

While physical textbooks and classrooms traditionally assumed the role of knowledge curators, they also present a bottleneck in today's rapidly growing web of up-to-date technical and academic content — peer-reviewed articles, lecture notes, tutorials, slides etc — from academics and "citizen scientists" alike. An automatic approach for "weaving" natural curricular progressions through the web of such heterogeneous academic/educational content, we believe, will catalyze early and lifelong learning by creating more efficient and goal-oriented curricula targeted to the level of the audience.

The web is the only collection of resources today where attempting this task becomes meaningful and promising. The reason for this is that the web contains an extensive amount of diversity in its content, i.e. content that explains the same concepts but in many different ways. Naturally this diversity reflects the diversity of the people who create this content, their backgrounds, styles of learning and ways of thinking about complex concepts, which would naturally match learners with similar characteristics. We believe that this diversity can be leveraged to create learning pathways that are not bound to the traditional curricula that are often constrained for no better than a historical reason. We propose instead to optimize a curriculum directly for *what you want to know* given *what you already know*.

We propose to tackle the problem of *curriculum mining* on the web, which broadly, involves linking technical resources on the web to other resources that explain a subset of concepts that are assumed in the original document. We propose to decompose the task into 1) understanding what is *explained* and *assumed* in a document on the part of the the reader and 2) use this document-level representation to sequence documents that guide the learner from their current state of knowledge towards their goal, for example, understanding a specific research paper or a set of lecture notes.

We propose a *term-centric* approach for inducing curricular relations between any pair of documents. Naturally, understanding a technical concept is more than being familiar with its surface term, and in this view an approach that operates at the level of individual terms may appear to be naïve. After all, to explain a new concept is to put together existing concepts in a novel way [13], and in the process introduce convenient nomenclature. However, we hypothesize, that by the virtue of seeking the shortest sequence of documents that "cover" (explain) multiple terms at once, the resulting bottle-

neck will implicitly "prefer" to link to prerequisite documents that introduce and explain whole concepts, i.e. groups of terms, as opposed to introducing terms one document at a time (an extreme example would be presenting a sequence of pages from a dictionary, each document defining a term independence; this is clearly undesirable). It will be our running assumption, that there exists a correlation between the knowledge of the terms and the understanding of the overarching concept.

Thus, to a first-order approximation, we model technical documents as "bags of terms", and in the interest of tractability set forth the following set of modeling assumptions:

- **Assumption 1** A document is a bag-of-technical-terms (multiset) that is further partitioned into two multisets: E (*Explained*), A (*Assumed*) — corresponding to the role (aspect) of the term within the document:

  Explained: The terms appear in the context that furthers the understanding of the concept corresponding to the term.

  Assumed: The concept corresponding to the term is assumed to be familiar, and is required for understanding the context in which it appears.

- **Assumption 2** The degree of reliance on the knowledge of a particular term in the document is proportional to the frequency of the term in the *Assume* multiset, i.e. which concepts are fundamental to the understanding of the document, and which are auxiliary is reflected in the number of occurences of the corresponding terms.

As an illustration, consider the following excerpt from Christopher Bishop's classic textbook *Machine Learning and Pattern Recognition* from the chapter that introduces the concept of *Expectation Maximization*:

> **Expectation Maximization**
> An elegant and powerful method for finding maximum likelihood solutions for models with latent variables is called the expectation maximization algorithm, or EM algorithm.

In the excerpt above, we solid–underline the terms that appear in the *Explained* aspect and dash–underline terms that appear in the *Assumed* aspect. Understanding the concept of *Maximum likelihood* is a prerequisite for understanding *Expectation Maximization*. It is no surprise that most resources that introduce the concept of *Expectation Maximization* implicitly assume that the reader is familiar with *Maximum Likelihood*. Academic and educational literature is fraught with such implicit assumptions that may be challenging to unravel for a learner especially new to the area. Note that on the surface it may seem that detecting instances of explained terms in the text is an equivalent task to finding instances of term definitions – a well studied task – but it is not so. Especially in technical disciplines, explaining a concept requires much more than giving a definition. A document defining a term, may or may not actually explain the concept behind it. For example, a document may define a term to refresh the reader's memory but otherwise assume the reader's familiarity with it. On the other hand, a document may explain a term without ever giving a one-sentence definition.

Finally, the proposed dichotomy may appear as a gross oversimplification, ignoring the entire continuum of pragmatics between the two extremes. We argue that while binary term-level classification alone may not capture the fine-grained aspect of any one term, combining it with the context of the entire document, will enable us to unravel the prerequisite relationships between documents.

2. RELATED WORK

Evidence of information overload in traditional textbooks Formal study of textbook organization conducted by [1] on a corpus of textbooks from India quantitatively addresses the issue known as the "mentioning problem" [12], where "concepts are encountered before they have been adequately explained and forces students to randomly 'knock around' the textbook". The work of [1] suggests that many traditional textbooks suffer from the resulting phenomenon of "information burden" and provide diagnostic metrics for evaluating it. A user study conducted by [2], though limited to electronic textbooks, demonstrated the utility of a navigational aid that links concepts and terms within a textbook and allows the user to navigate according to own preferences. This suggests the potential utility of tools that expand such "navigational ability" outside textbooks.

Attempts at manual curriculum curation There have been at least two efforts that we are aware of, that attempts to manually create "paths" between a selected set of resources on the web — two educational start-ups, Metacademy [5], and Knewton [4]. While motivated by the same goal, we believe that manual web-scale curriculum curation is akin to the manually-curated directory of the web (not too different from the original Yahoo directory from the 1990s), i.e. offering poor scaling capability in the dynamic, growing landscape of educational content on the web.

Attempts at automatic curriculum curation Most relevant to our task is the work of [11] that attempt to infer prerequisite relationships between a pair of Wikipedia articles. They frame the problem of prerequisite prediction as "link-prediction" between a pair of pages using primarily graph-derived (e.g. hyperlink structure) and some content-derived features (e.g. article titles). In contrast to their approach, we do not assume any existing structure connecting the web resources (e.g. within Wikipedia), as the majority of the educational content on the web is unstructured. Our approach also naturally facilitates a scalable assimilation of new content, as we require only a document-scoped term-level classification, without needing to explicitly construct or update a prerequisite graph. Furthermore, we develop an approach for optimizing curricular paths using the proposed representation. More recent work of [8] develop a method that does not rely on a manual annotation of the prerequisite relations as in [11], and instead uses the statistics of concept reference in a pair of pages to determine the prerequisite relation between them. Similar to [11], their focus is on the pairwise link prediction, in contrast to our goal of globally optimizing a learning curriculum.

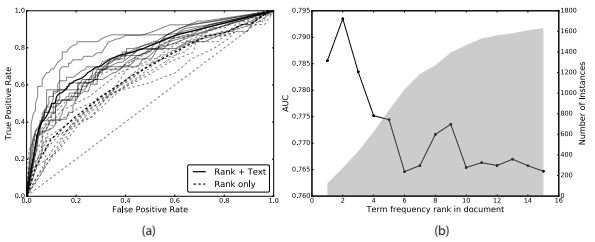

Figure 1: (a) ROC curves for the task of binary aspect classification. (b) AUC (left y-axis) of aspect classification for terms with a maximum document rank given on x-axis. Shaded region shows the number of terms up to the given maximum rank (right y-axis).

3. MODEL

3.1 Modeling explanations

We model the problem of identifying the explained and assumed terms in a document as a term-level binary classification task, i.e each term in the document is classified into one of the two categories. Although simple from an implementation perspective, this task is made difficult by the lack of annotated data in this domain. In this work, we rely on (i) manual annotation of the term aspects performed by us for one of the textbooks (Rice University's statistics text) and (ii) explicit annotations from the index of Bishop's Pattern Recognition and Machine Learning textbook that were made by the author of the text (the annotation is in the form of a location in the text where a particular concept is explained).

The Rice University's *Online Statistics Education: An Interactive Multimedia Course of Study* textbook, from hereon referred to as STATSBOOK consists of a total of 112 units, with a median of 12.5 unique technical terms per unit, for a total of 339 different technical terms in the book. We scrape the text content of the book from the web, replace all mathematical formulae and symbols with special tokens, and manually annotate each technical term mention with its representative form from the index, i.e. *normally distributed* with *normal distribution*. Manual term annotation obviates the need for introducing a word-sense disambiguation component and additional errors. We process the PRML dataset in an identical manner.

Each technical term in every unit of the book was annotated with the binary $\{explain, assume\}$ aspect, following the definitions outlined on the previous page. While for most terms, the application of these definitions is fairly unambiguous, for a significant number of term mentions, the aspects are not mutually exclusive, i.e. the term may be construed to belong to both aspects simultaneously. Often, in using (assuming) a term to explain a related concept, something about the assumed term is also explained as a side effect. The degree to which the explanation is distributed between the terms is difficult to judge objectively, and may vary between distinct mentions of the terms in different parts of the same document. We adopt a simple strategy for "breaking ties" in such cases: if we a judge a term as having been *intended* to be explained in the given context by the author, we mark it with the *explain* aspect, otherwise, the term is assumed to be *assumed*. In total across the entire STATSBOOK corpus, 1878 terms were annotated for their aspect (note that the same term appears in multiple documents with potentially different aspects), with a class ratio of 537 terms belonging to the *explain* and 1341 terms belonging to the *assume* aspect.

The PRML dataset contains a total of 3883 annotated terms, with 222 terms belonging to the *explain* and 3661 terms belonging to the *assume* aspect. The aspect of the term was determined from the index of the book, which explicitly specifies the pages where a term is explained.

A logistic regression model (LIBLINEAR [3] with default regularization parameter) was trained to predict a binary aspect of the terms and evaluated with 10-fold stratified cross-validation. A set of lexical and dependency features describing the context of each term (within a 1 sentence window), positional features describing the location of the term's mention within the document and sentences in which the term appeared, and the frequency rank of the term within the document were employed. We compare the performance of a classifier that uses all of these features with the one that uses only the rank. A classifier that is given rank as the only feature, will essentially learn a rank "threshold" that will decide the aspect of the term within the document, i.e. predict all terms above a certain rank as *explained*.

Figure 1(a) summarizes the performance of aspect prediction with the classifier trained using both linguistic and rank features (Rank+Text, AUC=0.76) versus a classifier trained using only the rank (Rank only, AUC=0.66) for the STATSBOOK corpus. As expected, rank is predictive of the aspect, but contextual linguistic cues provide a significant boost.

Keeping our end goal in mind, under Assumption 2 stated in the introduction, we hypothesize that the frequency rank of the term in a document correlates with the degree to which a term is either assumed or explained in that document. In the downstream task of linking documents to their prerequisites, getting the aspects of the more frequent terms correct is arguably more important than of the terms that only appear once or twice. We evaluate the performance of our aspect classifier as a function of the term's rank. Figure 1(b) illustrates predictive performance (AUC) on a subset of the data stratified by the term's frequency rank. We observe a favorable trend in increased predictive performance for higher ranked terms. An obvious explanation is that more frequent terms accumulate a larger set of features describing them (since each mention of the term contributes its context features), effectively decreasing variance in the predictions.

3.2 Optimal learning paths

Consider now that we have a large collection of documents (e.g. tutorials, papers, textbook chapters). Each such document explains some concepts but also assumes the reader's knowledge of other concepts (e.g. a tutorial may explain the concept of *normal distribution*, but may assume the knowledge of *probability* and *distribution*). We will now consider that we can reliably classify each term in each document into either the *Explained* or *Assumed* category. Consider that we also have a user who is interested in understanding a specific (target) document (or a set of target documents). The goal is to give a user a self-contained sequence of documents of

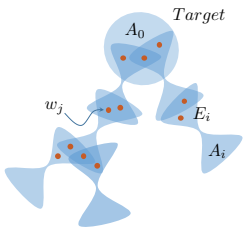

Figure 2: Each document is represented by a blue shaded region: the top part corresponds to the explained set E_i and the bottom part corresponds to the assumed set A_i. Red dots correspond to terms. This is an example of a feasible solution, where each document is *covered*.

minimal length that explains all of the concepts needed to understand the target document.

Formally each document d_i in our collection is a set of two sets of terms: the explained terms $E_i = E(d_i)$ and the assumed terms $A_i = A(d_i)$. A term in any document is either explained or assumed, but not both, i.e. $A_i \cap E_i = \emptyset$. We say that the document d_i is *covered* by a prerequisite set of documents P_i when:

$$A_i \subseteq \bigcup_{d_j \in P_i} E(d_j)$$

In other words the document is covered when every one of its assumed terms is explained by at least one document in the prerequisite set. For any prerequisite set that covers this document, the documents in the prerequisite set need to be covered as well, recursively until all documents have been covered. We assume the existence of documents with no prerequisites (leaves), i.e. those documents for which $A_\cdot = \emptyset$. The goal is to find a smallest *self-contained* set of documents P, i.e. a set of documents such that all the documents in P are covered and $d_0 \in P$, where $d_0 = \{A_0, E_0\}$ is the target document of interest to the user. Figure 2 illustrates a feasible solution to an example problem. Without additional restrictions, solutions to this problem can contain cyclical dependencies. Such cycles don't make sense in our setting. Thus an important restriction is that the the set of documents P can be ordered such that every document in the sequence is covered by the preceding documents in the sequence. Let $\mathbf{p}$ be a sequence of documents of length K, where $\mathbf{p}_k$ is the k^{th} document in the sequence, then we seek:

$$\begin{aligned} &\text{minimize } |\mathbf{p}| \\ &\text{s.t. } \forall k : A(\mathbf{p}_k) \subseteq \bigcup_{k'=0}^{k-1} E(\mathbf{p}_{k'}) \\ &d_0 \in \mathbf{p} \end{aligned} \tag{1}$$

ILP formulation

We formulate an Integer Linear Program (ILP) that finds a minimum length self-contained sequence $\mathbf{p}$ of at most K documents such that it covers a user's document of interest d_0. Consider that we have a total of D documents. We define the following variables:

$$x_i^k \in \{0,1\} \quad \text{document } d_i \text{ is in } k^{th} \text{ position in the sequence}$$

We define the following constants:

$$\begin{aligned} e_{ij} &\in \{0,1\} \quad \text{Term } j \text{ is explained in document } i \\ a_{ij} &\in \{0,1\} \quad \text{Term } j \text{ is assumed in document } i \end{aligned}$$

Each assumed term in a document in position k must be explained by at least one document up to (but not including) the document in position k. This can be expressed via the following constraint:

$$\sum_{k'}^{k-1} \sum_{i}^{D} e_{ij} x_i^{k'} \geq \sum_{i}^{D} a_{ij} x_i^k \ \forall j \forall k$$

Each position in the sequence contains at most 1 document:

$$\sum_{i}^{D} x_i^k \leq 1 \ \forall k$$

User's preference of covering a document of interest d_0 is an additional constraint:

$$\sum_{k}^{K} x_0^k = 1 \ \forall k$$

Finally, the objective is to minimize the number of documents in the sequence:

$$\text{minimize } \sum_{k}^{K} \sum_{i}^{D} x_i^k$$

The above formulation also allows us to directly incorporate the user's prior knowledge into this optimization problem. If we represent a user as a set of *explained* terms, i.e. terms that the user is assumed to have mastered, then the constraints corresponding to these terms may simply be dropped from the formulation.

In the most general case, this formulation has D^2 variables and $O(D^2 \times V)$ constraints, where V is the number of terms in the vocabulary. In practice, however, we will often limit the maximum allowable sequence length to a fairly small constant (e.g. 10, as done in our experiments), reducing the order of the problem to $O(D)$ variables and $O(D \times V)$ constraints.

While in extremely large settings (hundreds of thousands of documents), even with a small K, solving this ILP directly is infeasible, in practice, we find that that we can can obtain exact solutions using LP relaxation and a vanilla Branch and Bound (using GLPK[1]) within several seconds, even with a many as 1,000 documents and hundreds of terms. Developing an approximation algorithm based on rounding the LP solution is our ongoing work.

[1] `https://www.gnu.org/software/glpk/`

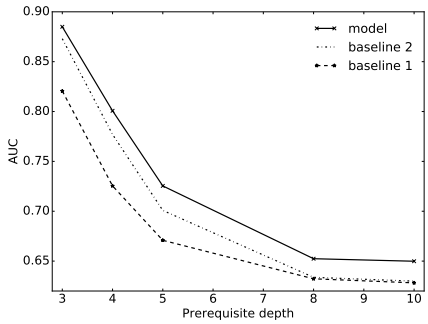

Figure 3: Term aspect classification is useful at the task of recovering prerequisites for units within a textbook. The y-axis is the average AUC at the task of predicting whether a particular unit is a prerequisite of another unit, based on three metrics. The metric that incorporates the *Explain/Assume* classifier performs best (solid line).

4. EVALUATION

4.1 Prerequisites

In order to evaluate the *Explain/Assume* classifier in an end-to-end setting, we employ the output of this classifier in the task of predicting prerequisites in a dataset where the prerequisites have been explicitly annotated. One such resource is *Rice University's Online Statistics Textbook*, which in addition to the text content, provides an explicit dependency graph annotating prerequisite relations between pairs of units (units are at the level of chapter sections). We propose a metric for scoring a pair of units according to their prerequisite relationship based only on the terminology of both units and the output of the *Explain/Assume* classifier. The proposed "prerequisite score" is defined as follows:

$$P(d_a \rightarrow d_b) = \frac{\sum_{t_i \in d_b} n_i^a \mathbb{1}[t_i \text{ assumed in } d_b \wedge \text{ explained in } d_a]}{\sum_{t_j \in d_a} n_j^a \mathbb{1}[t_j \text{ explained in } d_a]}$$

where n_i^a is the number of occurrences of term i in document d_a. Since the above score is guaranteed to be in the $[0, 1]$ range, we can interpret it as a probability $P(d_a \rightarrow d_b)$, a probability that document a is a prerequisite of document b. There is an intuitive interpretation to the above score: a document can be considered a strong prerequisite of a target document when it explains all of the assumed terms in the target document and nothing more. We can convince ourselves that in this case the score as defined above will be equal to 1. A document that explains too many unrelated concepts will suffer a penalty with respect to its prerequisite score to another document. Furthermore, we consider the relative frequency of the explained term in the prerequisite document as an additional signal of that term's importance. We find that this additional information increases the performance of prerequisite classification (discussed at the end of this section).

Because the output of the *Explain/Assume* classifier is a probability, rather than a class, we can relax the above score to directly incorporate the uncertainty in the classification:

$$P(d_a \rightarrow d_b) = \sum_{t_i \in d_b} \frac{n_i^a P(t_i \text{ explained in } d_a)}{\sum_{t_j \in d_a} n_j^a P(t_j \text{ explained in } d_a)} \quad (2)$$

Note that in addition to relaxing the requirement of an explicit *Explain* or *Assume* label, we also drop the requirement that only the assumed terms need to be explained to count towards the prerequisite score. This distinction is optional, but it encodes an important assumption on the kinds of "prerequisites" that this score will discover. This also brings up the importance of being precise about the definition of a prerequisite. A document a is a strict prerequisite of document b, if document a explains a subset of the assumptions in document b. However, we can relax this definition by *not* requiring that the terms explained in the prerequisite (a) are strictly assumed in the target (b). In other words, a document that explains a subset of the terms also explained in the target and *nothing else*, will have a score of 1 according to the above equation. In practice this corresponds to documents that explain the same concepts but in a simpler way (since they explain only a subset of the explained concepts in the target), and this is often a desired behavior in a learning sequence. For example, before reading a more advanced article on *Support Vector Machines*, the learner might want to read a more basic introduction to *Support Vector Machines*, although from the perspective of term classifications, both documents explain the same concept.

4.1.1 Reconstructing prerequisites

Rice University's Online Statistics Textbook provides a valuable resource for evaluating the effectiveness of the *Explain/Assume* classification at the task of predicting prerequisite relations between documents. The textbook consists of 112 units at the granularity of chapter sections, annotated as a directed graph, i.e. specifying a directed edge between a pair of units if one unit is considered a prerequisite of another unit. We process the raw HTML files of the textbook by removing markup, segmenting sentences and extracting terminology (obtained from the index) features as described in Section 3.1. We pose the problem of prerequisite relation prediction as a standard binary classification task, i.e. predicting for each pair of units in the book whether one unit is a prerequisite of another, where we consider a pair of units to be in a gold-standard prerequisite relation if there is a directed path between them in the graph. AUC is a convenient metric for evaluating performance in this prediction task, as the output of our scoring metric (Equation 2) is already scaled between 0 and 1. Note that the model trained only on the PRML corpus was used for term-aspect classification in this task. Figure 3 illustrates the results for three different models, as a function of the prerequisite depth, i.e. stratifying the classification results for a pair of units by the maximum distance between them in the graph. The three models evaluated are as follows:

- **Model** Prerequisite score is computed with Equation 2.
- **Baseline 1** Prerequisite score is computed with Equation 2, but with all n_i^a, n_j^a and $P(t_\cdot \text{ explained in } \cdot)$ set to 1. This baseline is equivalent to a ratio between the number of overlapping terms between a pair of documents and the number of terms in the prerequisite, i.e. $\frac{|d_a \cap d_b|}{|d_a|}$.

- **Baseline 2** Prerequisite score is computed with Equation 2, but with $P(t.\text{ explained in }\cdot)$ set to 1.

Each baseline illustrates the effect of *not* including a component of the scoring function in Equation 2. Our first conclusion from the results in Figure 3 is that the output of the *Explain/Assume* classifier provides an important signal in predicting the prerequisite relationship between documents. Furthermore, the relative frequency of the explained terms in the prerequisite document provides an additional gain in performance. This can be explained by Figure 1(b): the performance of the *Explain/Assume* classifier is greater in the higher term-frequency regime; discounting low-frequency terms (that are also likely less important to the content) reduces the classification noise and boosts the performance at the prerequisite prediction task. An additional observation is that the performance of the pairwise prerequisite classification improves for pairs of units that are closer in the graph, i.e. with less units in between. This is easily explained: units that are farther apart typically share less terminology, making the estimates based on terminology overlap noisier.

It is also interesting to note that the simplest baseline that considers only the ratio of overlapping terms between a pair of documents to the total number of terms in the prerequisite document does surprisingly well, especially well for pairs of documents closer together. This can be explained as follows: in a sequence of units like those in a textbook, units that are prerequisites tend to be less advanced, i.e. have less terminology, since less of it was introduced up to that point. Thus, units that are prerequisites, at least in a textbook, would be fairly predictable from the relative frequency of overlapped terms alone.

4.2 Scaling to the web

We collect and release two web corpora of educational content in the areas of Machine Learning and Statistics. Both corpora were collected using Bing Search API, by querying for short permutations of terms collected from the index of the *Pattern Recognition and Machine Learning* and *Rice University's Online Statistics Textbook*. The two corpora contain 42,000 and 1,000 documents respectively – a mixture of HTML and PDF files, pre-processed and converted to plain text. The difference in size of the two corpora is due to a smaller set of keywords used in the query set, and used primarily to rapidly validate the proposed model for path optimization. Consequently, because of a smaller term vocabulary, the smaller corpus is significantly less noisy (less irrelevant documents). The union of the terminology from the index of both textbooks was used as the vocabulary in processing each document. Additionally, terminology variations and abbreviations were consolidated using the link data from Wikipedia, e.g. terms *EM*, *E-M*, *Expectation-Maximization*, are all mapped to the same concept of *EM* in the terminology extraction stage.

Following the extraction of terminology from each webpage, each term is classified using the *Explain/Assume* classifier trained on the *Pattern Recognition and Machine Learning* textbook. We train this classifier in a fully supervised setting using all of the annotated data. In the next several sections, we present the analysis of the two web corpora and demonstrate the effectiveness of the proposed approach to connecting educational resources on the web.

4.3 Diversity of assumptions

The web is a unique setting, that unlike a traditional textbook or a course, offers a multitude of diverse explanations of the same concept. This diversity potentially enables the level of personalization that is not possible in traditional resources. We can analyze the diversity in the educational content on the web by looking at a slice of the web resources that share the same topic, but differ in their underlying assumptions and explanations. Figure 4 illustrates two articles that are both on the topic of *Expectation Maximization*. However, the two articles differ significantly in their assumptions on the background of the reader. Article 1 (left in Figure 4) is a very basic introduction to the topic and does not assume the knowledge of even the concept of *maximum likelihood*, which under most traditional curricula is assumed to be the prerequisite. Article 2 (right in Figure 4), however, assumes the knowledge of many more concepts such as *posterior probability*, *likelihood function* and *maximum likelihood*. This difference in the distribution of the underlying assumptions is explained by the fact the Article 1 s a very basic introduction to the topic, intended for an audience not in the area of statistics or machine learning. Article 2, however, is a significantly more thorough and a more technical introduction to the concept of the *Expectation Maximization* algorithm and thus assumes significantly more prerequisite background in the areas of statistics and machine learning. It's important to note that this distinction between the two documents cannot be easily made from their titles, or other surface cues: both documents are approximately the same length and their titles do not give away the level of technical detail. Their text content, however, provides the necessary cues to this information.

4.4 Fundamental prerequisites

Figure 5 illustrates the result of optimizing a learning path over the web corpus of 1,000 documents for the target webpage on the topic of "Maximum Likelihood Estimation". Sequences were optimized using the ILP formulation described in Section 3.2 using the GLPK Branch and Bound solver. Red rectangles correspond to terms for which the predicted label is *assumed* in the given document, and blue otherwise. In addition to the term-coverage diagram, we also illustrate the prerequisite dependencies extracted from the term coverage data: a directed edge is drawn to a document from the closest prerequisite in the sequence that covers at least one assumed term in the document. In the example in Figure 5, the target web-page is a fairly technical article on *Maximum Likelihood Estimation* that assumes the reader's understanding of the concepts such as the *likelihood function* which is pivotal for understanding the concept of *maximum likelihood*. As a consequence, the web-page that is placed immediately before in the optimal sequence are slides which consist of a more basic introduction to the *maximum likelihood*. Furthermore, the original target article assumes the reader's familiarity with *Generalized Linear Models* (which is in fact the previous section of the lecture notes of that series, indicating it as a prerequisite). The resulting sequence also contains an additional prerequisite on this topic. Finally, an interesting observation is that while the target article is fairly advanced in its assumptions about the reader's knowledge of

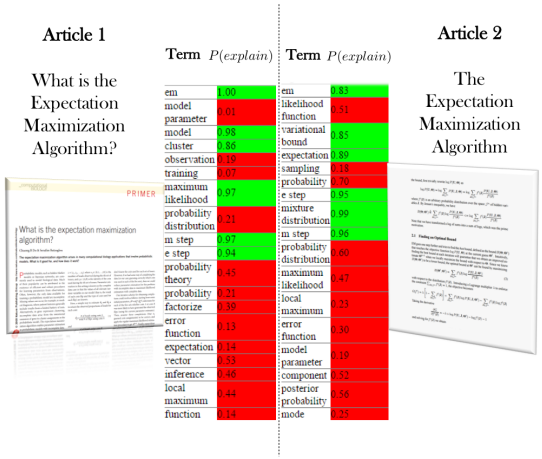

Figure 4: An example of two different web-pages about the same topic: *Expectation Maximization*, together with each page's terminology and its classification into either the *Explained* class (green) or the *Assumed* class (red). Observe that the two pages, while about the same topic, are different in what they assume about the reader. The article on the left is a very basic introduction to this topic, while the article on the right is written for experts.

probability, it actually goes into surprising depth in explaining the concept of a *derivative* and maximizing a function using derivatives from scratch, which is another important prerequisite to the concept of *maximum likelihood*. This is highly unconventional in traditional textbook and course curricula. This again underlines the advantage of working with the assumptions at document-level, allowing to leverage the diversity in explanations to find "shortcuts" through the learning paths.

Figure 6 provides additional insightful examples of the generated sequences extracted from the term-coverage data of each sequence. Figure 6(d) is another example where the target document is a fairly advanced introduction to the topic (*Expectation Maximization*), which is preceded by a more gentle introduction to the same topic, as well as an additional prerequisite (*Maximum Likelihood*) which is a common prerequisite for this topic. Note, however, that while *Maximum Likelihood* is traditionally considered as a prerequisite for learning about *Expectation Maximization*, it is not the case for the more basic introduction to this topic (*What is the Expectation Maximization algorithm*), as that particular introduction aims to bring a very high-level understanding of the topic without burdening the reader with additional prerequisite requirements. Therefore, in that particular sequence, the reader is first given a gentle introduction to the topic, then the necessary prerequisite (*Maximum Likelihood*) for understanding the more advanced introduction.

4.4.1 Error analysis

The extracted sequences are not without errors. These errors stem from several potential sources, as a fairly involved pipeline lies between the raw document and the resulting optimal sequence, providing an opportunity for errors to propagate through the different stages. We break down these errors by their source to give a better understanding of how these problems need to be addressed in future work:

Terminology extraction: The greatest source of errors stems from errors in terminology extraction. There are two types of errors involved in terminology extraction: *false negatives* (missing terms) and *false positives* (term sense disambiguation errors). False negatives are more difficult to detect and often result in missing prerequisites; missing terms are especially difficult when relying on a finite vocabulary.

***Explain/Assume* classification:** The second greatest source of errors are the mistakes made by the aspect classifier. Classifying an explained term as an assumed term creates unnecessary prerequisites, while the reverse results in missing potentially important prerequisites.

Path optimization: because we solve the optimization problem exactly (i.e. find a global optimum), there are no errors stemming from the optimization itself (this will become a potential source of errors, however, when an approximation scheme, e.g. LP rounding, is used to obtain an approximate solution). However, the formulation of the optimization problem can be improved so as to introduce robustness to the errors in the earlier stages of the pipeline. As path optimization is the final stage that produces the final output, its sensitivity to the errors in terminology extraction and term aspect classification are directly reflected in the resulting output. Introducing robustness to these errors directly in the formulation of the optimization problem is potentially the most effective way to address the issues in the earlier stages of the pipeline. One issue with the current formulation is its inability to incorporate the relative frequency of the term into the optimization objective: ideally terms that appear less frequently in a document should have a lesser precedence for coverage than those that appear more frequently (Assumption 2 in the Introduction). The example in Figure 5 demonstrates the lack of robustness in the third document, where the appearance of the term *integral* creates an additional sequence of documents that cover this concept. From our earlier analysis is Section 3.1, we have shown that the errors in the *Explain/Assume* classifier are directly related to the relative frequency of the terms, and thus a way to incorporate these frequencies as weights into the optimization would potentially be the most effective way to deal with this noise.

5. CONCLUSION

We developed what we believe is the first end-to-end approach towards automatic *curriculum extraction* from the web, relying on the following pipeline: 1) extracting what is assumed vs. what is explained in a single document and then 2) connecting these documents into a sequence ensuring that the progression builds up the knowledge of the learner gradually towards their goal. We developed algorithms that addressed both of these components: 1) a semi-supervised approach for learning a term aspect classifier from a very small set of annotated examples and 2) an optimization problem for learning path recommendation based on the user's learning goals. To the best of our knowledge, we for the first time demonstrate and leverage the most unique characteristic of the web in the domain of learning: *diversity*, i.e.

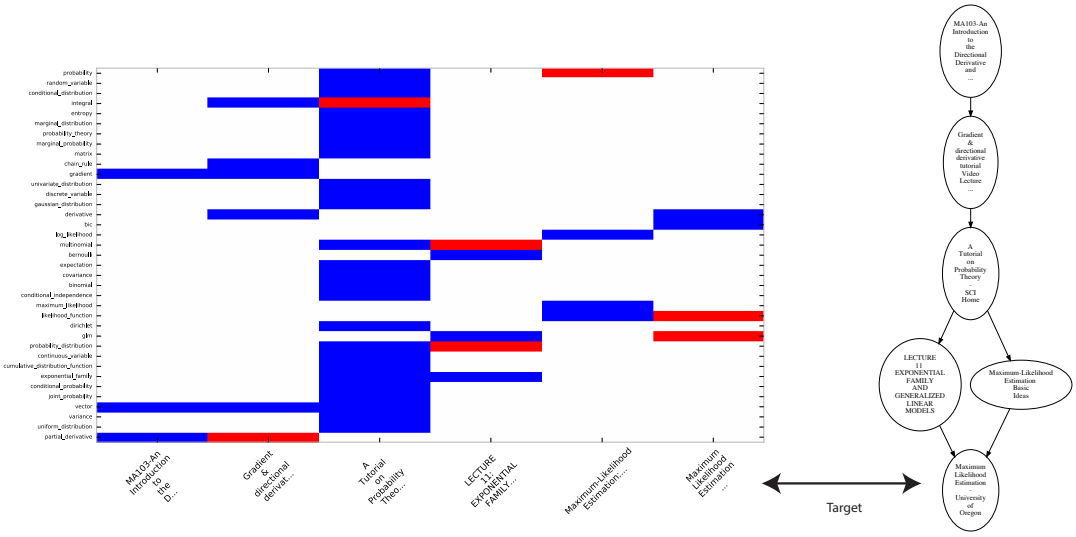

Figure 5: An example optimal sequence for the target document on *Maximum Likelihood Estimation*. Left: the term-coverage diagram. Each column represents a single web-page and each row a single term. Red rectangles correspond to terms that are classified as *assumed* in the corresponding document and *blue* corresponds to the *explained terms*. Right: the term-cover diagram is converted into a directed graph whereby an edge is drawn to a document from its closest prerequisite that explains at least one assumed term.

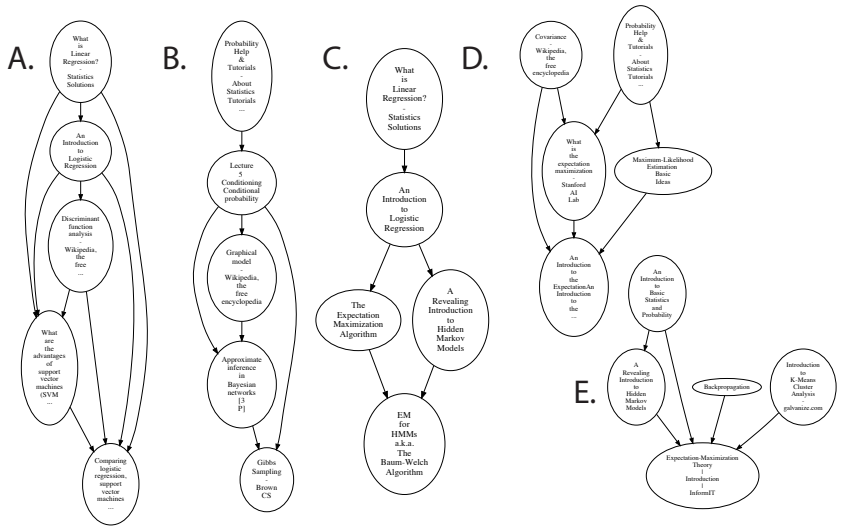

Figure 6: Additional examples of optimal paths generated from the 1,000-document web-page corpus for a select set of target web-pages. See text for details.

presence of content that explains the same concepts but in many different ways and from many different angles. This property of the web opens the doors to personalizing learning sequences that leverage the differences in explanations to find the most effective paths and shortcuts through the Internet. Finally, we outlined a set of important challenges that need to be addressed in order to make this task a practical reality at web-scale. We hope that this work, in addition to the datasets that we release, will serve to inspire interest from the community in what we believe is a challenging and an important task.

Acknowledgements

This research was funded by a grant from the John Templeton Foundation provided through the Metaknowledge Network at the University of Chicago. Computational resources were provided in part by grants from Amazon and Microsoft.

6. REFERENCES

[1] R. Agrawal, S. Chakraborty, S. Gollapudi, A. Kannan, and K. Kenthapadi. Empowering authors to diagnose comprehension burden in textbooks. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 967–975. ACM, 2012.

[2] R. Agrawal, S. Gollapudi, A. Kannan, and K. Kenthapadi. Study navigator: An algorithmically generated aid for learning from electronic textbooks. JEDM-Journal of Educational Data Mining, 6(1):53–75, 2014.

[3] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. JLMR, 9:1871–1874, 2008.

[4] J. Ferreira. Knewton. `http://http://www.knewton.org`.

[5] R. Grosse. Metacademy. `http://www.metacademy.org`.

[6] B. Jones, E. Reedy, and B. A. Weinberg. Age and scientific genius. Technical report, National Bureau of Economic Research, 2014.

[7] B. F. Jones. As science evolves, how can science policy? In Innovation Policy and the Economy, Volume 11, pages 103–131. University of Chicago Press, 2011.

[8] C. Liang, Z. Wu, W. Huang, and C. L. Giles. Measuring prerequisite relations among concepts.

[9] M. Peat, C. E. Taylor, and S. Franklin. Re-engineering of undergraduate science curricula to emphasise development of lifelong learning skills. Innovations in Education and Teaching International, 42(2):135–146, 2005.

[10] D. J. Rowley, H. D. Lujan, and M. G. Dolence. Strategic Choices for the Academy: How Demand for Lifelong Learning Will Re-Create Higher Education. The Jossey-Bass Higher and Adult Education Series. ERIC, 1998.

[11] P. P. Talukdar and W. W. Cohen. Crowdsourced comprehension: predicting prerequisite structure in wikipedia. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pages 307–315. Association for Computational Linguistics, 2012.

[12] H. Tyson-Bernstein. A conspiracy of good intentions. america's textbook fiasco. 1988.

[13] S. M. Wilson and P. L. Peterson. Theories of learning and teaching: what do they mean for educators? National Education Association Washington, DC, 2006.