Assigning Educational Videos at Appropriate Locations in Textbooks*

Marios Kokkodis
NYU Stern
mkokkodi@stern.nyu.edu

Anitha Kannan
Microsoft Research
ankannan@microsoft.com

Krishnaram Kenthapadi
Microsoft Research
krisken@microsoft.com

ABSTRACT

The emergence of tablet devices, cloud computing, and abundant online multimedia content presents new opportunities to transform traditional paper-based textbooks into tablet-based electronic textbooks, and to further augment the educational experience by enriching them with relevant supplementary materials. The use of multimedia content such as educational videos along with textual content has been shown to improve learning outcomes. While such videos are becoming increasingly available, even a highly relevant video can be created at a granularity that may not mimic the organization of the textbook. We focus on the video assignment problem: *Given a candidate set of relevant educational videos for augmenting an electronic textbook, how do we assign the videos at appropriate locations in the textbook?* We propose a rigorous formulation of the video assignment problem and present an algorithm for assigning each video to the optimum subset of logical units. We also show that our objective function exhibits submodularity and hence admits an efficient greedy algorithm with provable quality guarantees, when the number of logical units is large. Our experimental evaluation using a diverse collection of educational videos relevant to multiple chapters in a textbook demonstrates the efficacy of the proposed techniques for inferring the granularity at which a relevant video should be assigned.

1. INTRODUCTION

Textbooks have been the primary teaching instrument since the 19th century. Education literature has extensively highlighted the central role played by textbooks in delivering content knowledge to the students, improving student learning, and in helping teachers prepare the lesson plans [14, 38]. The rapid proliferation of cloud-connected electronic devices has enabled the availability of textbooks in electronic format. However, many of these e-textbooks are predominantly digital versions of the printed books, and hence do not make use of the rich functionalities provided by the electronic medium (and/or the cloud-connectedness). Thus, we have the opportunity to enrich the reading experience by augmenting e-textbooks with supplementary materials appropriate to the learning style of the student, be it auditory, visual or kinesthetic style [7, 10, 11, 16, 32, 36].

Towards this goal, we focus on enriching the experience of reading from a textbook by interspersing rich video content at specific, appropriate locations in the textbook. In fact, the use of multimedia content along with textual material has been shown to result in better content retention [35] and improved concept understanding [28]. Our problem is further motivated by the rapid growth in online educational videos that are created and uploaded by self-appointed 'teachers'. YouTube Edu alone contains over 700,000 high quality educational videos from over 800 channels [27].

With the availability of abundant online video content, one can envision retrieval algorithms that can identify videos relevant to the textbook. In fact, we use one such existing algorithm to narrow the video collection to a relevant subset for the textbook [3]. However, this does not solve the problem entirely. Since the videos on the web are not created specifically for the textbook of interest, there are significant differences in the authoring style of a video creator versus that of a textbook author (we discuss other challenges in §1.1). The textbook author creates a logical hierarchy (chapter → sections → subsections, *etc.*) that is best suited for presentation of all the material that needs to be covered in the book. In contrast, the author of a video focuses only on the content to be presented in the video. This central difference makes it challenging to match videos to textbook units. While some videos may provide a high-level overview of the subject and hence may be appropriate at the granularity of the entire book, other videos may illustrate a specific concept or demonstrate an activity and hence may be appropriate at the level of a subsection or even a paragraph. Similarly, there may be videos that summarize a chapter or a section, and hence may be best placed at an intermediate granularity. For example, a video that contains material about different sections in a chapter can either be placed at the chapter beginning (if it provides an overview), or at the chapter end (if it helps to review the material in the chapter).

The focus of this paper is to recognize this mismatch and automatically determine the appropriate textbook locations for assigning the videos. More precisely [20, 21]:

*Microsoft Research, MSR-TR-2014-62, July 2014.

Given a textbook (or a chapter in a textbook) and a video relevant to the textbook (or the chapter), how do we identify the best subset of logical units (such as sections) that covers the material present in the video?

We propose a rigorous formulation of the video assignment problem and present an algorithm for assigning each video to the optimum subset of logical units. We also show that our objective function exhibits submodularity and hence admits an efficient greedy algorithm with provable quality guarantees, when the number of sections is large. As part of computing the objective function, we provide a novel representation for videos in terms of concept phrases present in the textbook, and their significance to the video. Our empirical study over a diverse collection of educational videos corresponding to multiple chapters in a textbook demonstrates the efficacy of the proposed techniques.

1.1 Other Considerations

In addition to identifying relevant videos and suggesting them at the appropriate granularity, there are other considerations that form important research questions beyond the scope of this paper. In particular, we need to consider aspects along three dimensions: the video, the viewer and the presenter. The relevancy of the video content to the textbook, the appropriate granularity in the textbook where the video should be placed, duration of the video, and the video quality [31] are examples of aspects related to the video. The appropriateness of the video to the viewer's prior subject knowledge and preference for the type of video such as lecture, demonstration, or animation are examples of aspects related to the viewer. The presentation style, accent, and diction are examples of the presenter aspects [25]. The rich diversity provided by the above dimensions also motivates the need for going beyond one source of videos (such as Khan Academy). For example, YouTube Edu alone has over 800 channels, and contains over 30 videos with nearly identical content but on different aspects of a single topic such as "the law of conservation of mass". Which of these 30 videos is the right augmentation for "the law of conservation of mass"? This is a function of all the above dimensions. In this paper, our focus is on the relevancy and the appropriate granularity: how do we automatically assign the candidate relevant videos at the appropriate granularity?

We also do not discuss specific mechanisms for integrating the augmentations into the textbook, or their implications for royalty sharing and intellectual property rights. Further, issues and complementary approaches such as enhancing our results using collaboration and crowdsourcing [33] and integrating the augmentations with other interventions for improving the learning outcomes [14, 29] are very important, but are beyond the scope of this paper.

2. RELATED WORK

2.1 Enhancing Textbooks

There has been considerable work on augmenting textbook sections with relevant supplementary materials mined from the web [1, 3, 4]. In [4], the focus has been on finding textual content from the web that is relevant for a section. Somewhat related is the work proposed in [39] that augments textual documents such as news stories with other textual documents such as blogs. In [1], a method was proposed to identify the focus of the section, which was then used to obtain relevant web videos. However, it is not always possible to assign a video to a single section. A video may contain content that extends across sections, as the author of the video may have chosen a logical ordering different from that of the author of the textbook. In this paper, we present a technique that, given the videos relevant to the entire chapter, identifies the minimal combination of sections that best encapsulates the material covered in the video. Towards this goal, we infer a representation for a video as a byproduct of the COMITY algorithm [3] which we adapt to obtain relevant videos.

2.2 Re-ranking Search Results

There has been extensive work in the broad area of information retrieval for (a) identifying videos (or other content) relevant to a textual query, and (b) re-ranking these results based on a number of preferences and additional metadata. In this paper, we assume that the candidate set of relevant videos for each textbook chapter is provided by an oracle (described in §3). Therefore, we compare and contrast our work with literature on re-ranking along three dimensions: (a) diversification, (b) use of additional information, and (c) personalization based on preferences.

Diversification of results: The premise for diversification is as follows: As the user query can be ill-defined with respect to user needs, the retrieval system needs to trade-off between having relevant results of the 'dominant' intent and diverse results in the top positions. The Maximal Marginal Relevance (MMR) criterion was proposed in [8] to obtain a trade-off between redundancy and maintaining query relevance in re-ranking retrieved documents. Since then, many techniques have been proposed for diversifying retrieved results by making use of a number of additional signals such as click patterns [2, 9, 15, 37].

The focus of our work is to re-assign a video to the combination of sections that best describes the content of the video. Diversification techniques can be used to provide richer suggestions (when multiple videos are identified for the same set of sections). The diversification can be performed with respect to additional dimensions such as speaking styles (*e.g.*, liveliness) or presentation types (*e.g.*, demonstrations *vs.* lectures) used in the videos.

Leveraging video content for re-ranking: In order to refine rankings, a number of recent works has made use of visual, auditory and spoken words in the video. In [17], relevant results were further analyzed and re-ranked by identifying salient visual patterns of relevant and irrelevant shots. This work was further extended to incorporate recurrent shots using a random walk over the context graph where video stories are the nodes and the edges between them are weighted by contextual multimodal similarities based on visual cues and transcripts of the spoken words [18]. In a similar line of work, a graph based approach based on PageRank was proposed in [24] while a maximum weighted bipartite matching algorithm that uses video-clip similarities was proposed in [13].

In these works, the focus has been to enhance the retrieval

Algorithm 1 COMITY

Input: A textbook chapter j; Number of desired video results k; Number of desired video search results per query t; Number of desired concept phrases n.
Output: A list of top k video results from the web, along with relevance scores.

1: Obtain (up to) top n concept phrases from chapter j.
2: Form queries consisting of two concepts phrases each ($\binom{n}{2}$ queries in total).
3: Obtain (up to) top t video search results for each of the queries using a search engine.
4: Aggregate over $\binom{n}{2}$ video result lists to obtain relevance score, λ_{ij} associated with each video i for chapter j.
5: Return top k videos along with their λ_{ij} values.

scores by making use of rich signals provided in the visual and speech cues. Our work differs in multiple respects. First, given the relevant videos for a book chapter, we want to re-assign them to a subset of sections that best describes the content delivered in the video. Second, while one may resort to using automated speech recognition to extract the spoken words, this is not always possible: Educational videos from the wild (*e.g.*, from YouTube) have a wide range of recording conditions, speaker accents, and speaker age ranges making it quite challenging for automatic speech recognizers to extract transcripts reliably. Therefore, we propose a novel approach for inferring the underlying content in the video. In particular, we make use of the queries to a video search engine that led to the retrieval of the video. Using these queries as the basis, we design a representation that captures not only the salient concept phrases in the video, but also their relative importance to the video, and then perform assignment making use of this representation (§4).

Personalization of results: There is also a line of work on re-ranking the video results based on user preferences. For instance, an algorithm to provide personalized video suggestions by making use of the user-video graph was proposed in [6]. Such approaches can be used to further personalize the assignments, for instance, factoring in learner preference for a certain type of video.

3. CANDIDATE VIDEO SELECTION USING COMITY ALGORITHM

In this paper, we assume that we have access to the candidate set of videos relevant to a textbook chapter. In order to obtain that relevant set, we adapted one of the methods, namely, COMITY algorithm, proposed in [3] for augmenting textbook sections with images. We describe our adaptation of COMITY algorithm next as we also use the internals of this algorithm for other dependent tasks in our approach.

In [3], COMITY algorithm was used on a per-section basis to mine relevant images from the web. We found that when this technique was applied at the section level for retrieving videos, there was a huge redundancy in the retrieved videos, across multiple sections[1]. We highlight two key observations: First, the content of the same video could be shared

[1]Similar observation was also made in [3] with respect to the images retrieved.

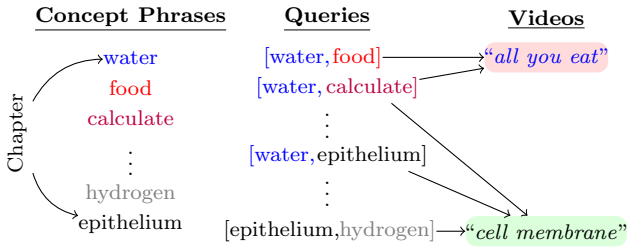

Figure 1: Query based video representation

across multiple sections, calling for an approach such as the one proposed in this paper to identify the combination of sections that best describes the video. Second, by applying the algorithm at the chapter level, we identify a richer set of videos, by exploiting dependencies between concepts elucidated in different sections.

Algorithm 1 provides a quick overview of our adaptation of COMITY algorithm. It uses top n concept phrases present in a chapter to query a commercial video search engine[2]. We will use *cphr* to denote a concept phrase present in a text. There are multiple approaches for identifying *cphrs* from a text. In this paper, we define the set of *cphrs* as the set of phrases that map to Wikipedia article titles [12, 26, 34]. This set is further refined using the techniques proposed in [4]. Since a *cphr* in isolation may not be representative of the text as the same text can discuss multiple concepts, COMITY forms $\binom{n}{2}$ video search queries by combining two *cphrs* each, in order to provide more context about the chapter. Figure 1 shows an example of how the queries are constructed from *cphrs* extracted from a textbook chapter on Biology. A relevant video for the chapter is likely to occur among the top results for many such queries. Thus, by aggregating the video result lists over all combinations of queries, we obtain the most relevant videos for the chapter. In particular, the score associated with a video i for a chapter j is given by $\lambda_{ij} := \sum_q (1/(p(i,q,R(q)) + \theta))$, where the summation is over $\binom{n}{2}$ queries issued and $p(i,q,R(q))$ denotes the position of video i in the result list $R(q)$ for query q if i is present in $R(q)$ and ∞ otherwise. θ is a smoothing parameter. This score captures the empirical observation that a video occurring among the top results for multiple queries was more relevant to the chapter than a video that occurred among the top results for only one query.

We slightly modified the algorithm presented in [3]: We used only a single search engine. We constructed queries by combining two *cphrs* while the original method used up to three *cphrs*. Since we considered the *cphrs* in the entire chapter, it was rather impossible to consider all $\binom{n}{3}$ queries.

4. APPROACH & ALGORITHMS

4.1 Representation of Textbook

Assume we have a textbook, consisting of K chapters. Each chapter is subdivided into sections. Let $\mathcal{C}_{book}$ denote the set

[2]A natural question is whether we could simply use the text string of a chapter to query a commercial video search engine and obtain the relevant videos. However, current search engines do not perform well with long queries [19, 22].

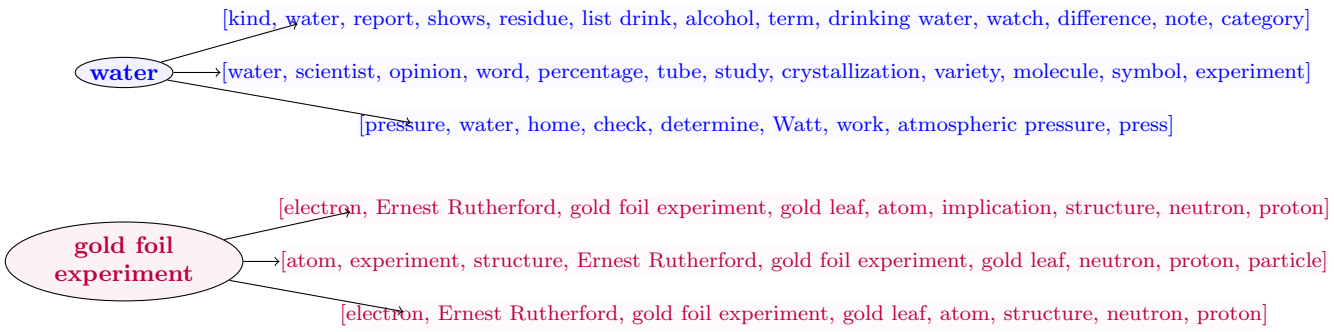

Figure 2: Illustration of important ('gold foil experiment') vs non-important ('water') concept phrases

of all *cphrs* (concept phrases) in the book. As discussed in §3, we define $\mathcal{C}_{book}$ to be the set of phrases in the book that map to Wikipedia article titles [12, 26, 34], further refined using the techniques proposed in [4].

Each *cphr* differs in its importance to the underlying text. For instance, a seemingly common *cphr* such as 'air' can be extremely important when we discuss 'air pollution'. Thus, we need to take into account the context in which the *cphr* occurs in order to compute its importance. Therefore, we introduce *context-dependent importance* score, $I(c)$ for a *cphr* c that is determined directly from the data.

How do we compute I(c)? We make the following observation: If a *cphr* is important for the context of the text, then the videos retrieved using it as *one of* the query terms will be related to each other. On the contrary, if the *cphr* is not, then the videos retrieved using it as *one of* the query terms will be very diverse and diffused. We operationalize this observation to infer the context-dependent score for each *cphr*.

For each *cphr*, we consider top m most frequent videos retrieved when the *cphr* is used in conjunction with all other *cphrs* in the textbook. Figure 2 shows top *cphrs* associated with three most frequent videos for two *cphrs*, 'water' and 'gold foil experiment' (we describe the computation of *cphrs* in a video in §4.2). Consider the *cphr* 'water'. The intersection of the three sets of *cphrs* is only the *cphr*, 'water'. On the other hand, for the *cphr* 'gold foil experiment', the top three most frequent videos have a much larger set of common *cphrs*: {electron, Ernest Rutherford, gold foil experiment, foil, gold leaf, atom, structure, discovery, neutron, proton} (note that the intersection is computed over all the *cphrs* associated with the videos whereas only the top *cphrs* are shown). Thus, a specific phrase is likely to lead to videos that are more similar to each other than a generic phrase.

With this intuition, we measure I(c) as the average pair-wise inner product between top m videos retrieved in response to queries that contained c:

$$I(c) = \frac{\sum_{1 \leq i < j \leq m} < V_i, V_j >}{\binom{m}{2}}, \quad (1)$$

where V_i is the vector representation (in terms of *cphrs* and associated weights) for i^{th} top video for c.

While one can directly use $I(c)$, we found that these scores exhibited variances that are due to noise in the data, and not inherent to the *cphrs* themselves. We addressed this issue by clustering the scores into a small number of clusters (we used 4 clusters) and assigning cluster means as the importance scores. Table 1 shows examples of *cphrs* in the resulting four clusters for five chapters from a Science textbook: cluster 1 contains highly specific *cphrs* while cluster 4 has all the *cphrs* with very low *context-dependent importance* scores. We can see that the *cphr* 'cell' is inherently ambiguous (*e.g.*, cell can be used in the context of fuel cells, biological cells, *etc.*). However, since it is used specifically in the context of a Biology chapter and used in context with other *cphrs* with similar interpretation, this *cphr* is highly significant for the chapter. In contrast, the *cphrs* 'air' and 'place' are not.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
vacuole	magnesium chloride	myocyte	air
plastid	polyatomic ion	manure	ways
cotyledon	mole	hydrogen	hand
cell	monera	energy level	day
tissue	isotope	red blood cell	area
dicotyledon	j j thomson	cell (biology)	place
nervous tissue	physical property	diatomic molecule	water
plant cell	proportionality	cork	f.o.o.d

Table 1: *Cphrs* binned based on their *context-dependent importance* scores. Clusters 1 to 4 show decreasing importance. *Cphrs* are from a chapter in a Science textbook.

4.2 Representation of Candidate Videos

In order to match a video to a set of sections, we also need a representation of the video. While one obvious approach would be to use transcripts associated with videos, most videos in our corpus did not have high quality user-uploaded transcripts, and were too difficult for automatic speech recognizers to extract reliable speech signal (due to a wide range of recording conditions & speaker accents / age ranges).

Hence, we devise a representation motivated by the following observation: When a video is retrieved in a highly ranked position for a query, the corresponding query represents some aspects of the content of the video. As an example, consider Figure 1. The video "all you eat" describes dietary habits, and is retrieved as a top result for the queries "water, food" and "water, calculate". Thus, the *cphrs*, 'water', 'food', and 'calculate' can be associated with this video. Similarly, for the video "cell membrane", the relevant *cphrs* are 'epithelium', 'hydrogen', 'water', and 'calculate'. However, the relative importance between the *cphrs* that lead to retrieving

a video varies. In this example, the video on cell membrane should be related more to epithelium than to water. Therefore, we represent a video with not only the *cphrs* that led to the video, but also their importance to the video. For each *cphr* c and video v, we define the importance $w_{v,c}$ of c to v as the fraction of queries that contain c for which video v was retrieved as a top result:

$$w_{v,c} = \frac{\{q \in Q_c | (v \in TopResults(q)\}}{|Q_c|}, \quad (2)$$

where Q_c is the set of queries that contain *cphr* c. The intuition behind this definition is that the higher the fraction of queries that led to a specific video, the more related this phrase is with the video.

In our implementation, we restricted the possible *cphrs* that can lead to a video to be only those that are present in the textbook. However, one can extend this representation in many ways, *e.g.*., by using multiple books of the subject matter or identifying the *cphrs* in the transcript of the video, especially when the transcript is user-uploaded.

4.3 Section Subset Selection For Videos

For a given candidate video v, let $\mathcal{S}$ be the candidate collection of sections from the textbook chapter. From this large candidate set, we would like to select a *minimal subset* of top sections, $\mathcal{T} \subset \mathcal{S}$ that best covers the content in the video. We model this section subset selection problem as identifying a subset of sections that maximizes the following objective function:

$$\mathcal{F}(v, \mathcal{T}) = \text{cover}(v, \mathcal{T}) - \lambda |\mathcal{T}|, \quad (3)$$

where $\text{cover}(v, \mathcal{T})$ is a function that measures how well the set of sections $\mathcal{T}$ captures the content of the video v , which we describe momentarily. Our objective function incorporates a penalty for using more sections than required for explaining the video, by discounting for the number of sections $|\mathcal{T}|$. Thus, the objective function provides a trade-off between the extent to which the content of the video is captured and the number of sections used. Different trade-offs can be obtained through different choices of the non-negative parameter λ: A large value of λ corresponds to a greater penalty for having more sections.

This trade-off ensures that each additional section must explain a significant portion of the video. Suppose for example that the cover score for a video v and a set of sections $S' = \{S_2, S_3\}$ is 0.85, and for $S'' = \{S_2, S_3, S_4\}$ is 0.86. While S'' has a larger absolute cover score, it does not cover significantly more material than S' and hence we would prefer the assignment, S'. In other words, $\mathcal{F}(v, \mathcal{T})$ provides a trade-off between the extent of content match and the number of sections in the subset. Hence, our goal is to determine the subset $\mathcal{T}^*$ that maximizes the objective function:

$$\mathcal{T}^* = \arg\max_{\mathcal{T} \in 2^{\mathcal{S}}} \quad \mathcal{F}(v, \mathcal{T}). \quad (4)$$

Computing $\text{cover}(v, \mathcal{T})$**:** Let $C(v) \subseteq \mathcal{C}_{book}$ denote the set of *cphrs* present in our representation of video v and let $C(\mathcal{T}) \subseteq \mathcal{C}_{book}$ denote the set of *cphrs* present in the subset of sections $\mathcal{T}$. Then, $\text{cover}(v, \mathcal{T})$ captures how much of the video v is explained by the subset of sections $\mathcal{T}$. In our implementation, we define this function to be the weighted fraction of the *cphrs* in the video that is also covered by the subset of sections:

$$cover(v, \mathcal{T}) = \frac{\sum_{c \in (C(v) \cap C(\mathcal{T}))} w_{vc} I(c)}{\sum_{c \in C(v)} w_{vc} I(c)} . \quad (5)$$

The cover score takes values between 0 and 1, and the higher the value, the more video content is contained in the corresponding subset of sections.

Brute-force optimization: One approach to solve the above optimization is to exhaustively search over all possible subsets of $\mathcal{S}$ and pick the best subset as given by Eq. 4. This algorithm takes the set of sections in a textbook chapter and a candidate video as inputs. In addition to the size penalty parameter λ, we also include the coverage threshold θ that represents the minimum fraction of the video content that must be covered by including all sections in the chapter, in order to assign this video to any subset of sections. If the entire chapter covers less than θ fraction of the video content, then any subset of sections will also cover less than θ fraction (since $\text{cover}(v, .)$ is a monotonically increasing set function), and hence the algorithm chooses to return no assignment. Otherwise, the algorithm returns the subset of sections that maximizes $\mathcal{F}(v, \mathcal{T})$. We discuss the parameter choices in §5.3.

Greedy optimization: Clearly, brute force approach is prohibitive even for a reasonable size of $\mathcal{S}$. Hence, we exploit the fact that the functions, $\text{cover}(v, .)$ and $\mathcal{F}(v, .)$ are submodular (which we show below). While maximizing non-negative monotonically increasing submodular functions (subject to cardinality constraints) is intractable in general, one can design greedy solutions that are at most $(1 - \frac{1}{e})$ times worse than the optimal solution for this class of submodular functions [30]. While $cover(v, .)$ is a submodular function that satisfies the additional requirements, $\mathcal{F}(v, .)$ is not guaranteed to be monotonically increasing or even non-negative, and hence the results from [30] do not directly carry over. We next describe the greedy algorithm and show that we can still provide theoretical guarantees. Initialize with a section s that has the largest value of $\text{cover}(v, \{s\})$. Then, iteratively add sections, one at a time, such that the added section (along with previously chosen set of sections) gives the maximal incremental gain in the function, $\mathcal{F}(v, .)$. Note that the section that provides the maximal incremental gain in $\mathcal{F}(v, .)$ also gives the maximal incremental gain in $\text{cover}(v, .)$ (since the additional penalty for including any single section is λ, a fixed value irrespective of the section included). The algorithm continues until either all sections in the chapter have been included, or the maximal incremental gain in $\mathcal{F}(v, .)$ for any further section is negative (due to the penalty). This greedy approach is not guaranteed to result in an optimal solution, but is really efficient, which makes a potential extension from the chapter level to the book level fairly trivial and feasible.

Let k^* denote the number of sections included using the above greedy algorithm. Let $F_{k^*,greedy}$ and $C_{k^*,greedy}$ respectively denote the values of the function $\mathcal{F}(v, .)$ and the function $\text{cover}(v, .)$ evaluated at the corresponding greedy solution. Let $F_{k,opt}$ and $C_{k,opt}$ respectively denote the optimum values of the function $\mathcal{F}(v, .)$ and the function $\text{cover}(v, .)$

subject to the cardinality constraint that exactly k sections are present in the solution. Finally, let F_{opt} denote the optimum value of the function $\mathcal{F}(v,.)$ (that is, without any cardinality constraints). We provide a guarantee for the value of the greedy solution with respect to the optimum solution containing exactly the same number of sections, even though the guarantee is not with respect to the true optimum F_{opt}. We formally state the guarantees below.

THEOREM 4.1.

$$F_{k^*,greedy} \geq \left(1 - \frac{1}{e}\right) \cdot F_{k^*,opt} - \frac{\lambda \cdot k^*}{e}.$$

PROOF. The proof follows by making use of the result from [30] for the function, cover$(v,.)$ since this function satisfies the non-negativity and monotonically increasing requirements and further cover$(v,\phi) = 0$. Using this result, we can derive: $F_{k^*,greedy} = C_{k^*,greedy} - \lambda \cdot k^* \geq (1 - \frac{1}{e}) \cdot C_{k^*,opt} - \lambda \cdot k^* = (1 - \frac{1}{e}) \cdot F_{k^*,opt} - \frac{\lambda \cdot k^*}{e}$. □

Submodularity of $\mathcal{F}(v,.)$ and cover$(v,.)$: A submodular function g has the property of "diminishing returns": The difference in the value of the function that a single element makes when added to an input set decreases as the size of the input set increases. Formally, if $X \subset Y \subset \mathcal{S}$, then adding an element $z \in \mathcal{S} \setminus Y$ to both X and Y should satisfy:

$$g(X \cup z) - g(X) \geq g(Y \cup z) - g(Y) \tag{6}$$

Since the summation in the numerator in the function, cover$(v,.)$ is over *cphrs* that belong to the sections (intersected with those in video v), we can observe the following: With respect to the incremental gain in cover$(v,.)$ upon including z, the summation can include additional number of *cphrs* for set X compared to set Y. Hence, cover$(v,.)$ is a submodular function. Further, it is monotonically increasing, non-negative, and has zero value for empty set: cover$(v,\phi) = 0$. Since $\mathcal{F}(v,.)$ is obtained by adding a linear function to the function, cover$(v,.)$, it follows that $\mathcal{F}(v,.)$ is also submodular, although it does not satisfy these additional properties.

5. EVALUATION

We next perform empirical validation to demonstrate the efficacy of our approach in identifying the subset of sections that best covers the material presented in a video relevant to the chapter.

5.1 Dataset

We first construct a ground truth test set of videos for each textbook chapter. However, given the huge number of videos available online, it is infeasible to create such a set by inspecting all the videos. Therefore, we take a different approach: We consider the first five chapters of a 9^{th} grade science book. We chose this textbook for two reasons. First, these chapters span different sub-branches of science: Physics (Chapter 1: "Matter in our surroundings" and Chapter 2: "Is matter around us pure"), Chemistry (Chapter 3: "Atoms and molecules" and Chapter 4: "Structure of the atom"), and Biology (Chapter 5: "The fundamental unit of life"). There are about 5 sections (median value) in these chapters. Second, these chapters differ in the extent to which there is content overlap and commonality across sections. These differences help us to characterize when our approach is most beneficial. Although our approach uses COMITY algorithm at the *chapter level* to obtain the candidate set of relevant videos, for the purposes of comparative evaluation, we chose to apply COMITY algorithm at the *section level* (see §5.2 for details). That is, for each chapter, we run the COMITY algorithm, but by restricting to combinations of top n *cphrs* that are present in a section. We set $n = 20, t = 50$, and $k = 20$. This process resulted in 178 unique videos across all chapters. We assigned a human assessor to read all these five book chapters. After reading the chapters, the judge is asked to watch each video and manually identify all the sections that together capture the content of the video[3]. The judge can revisit the book to read multiple times. Note that the judge does not have access to the underlying algorithm that identified the video. The judge is also asked to remove videos that are irrelevant, or cover material beyond the scope of the book.

This judgment process resulted in 112 videos (denoted by $\mathcal{V}$) along with their best set of sections assignments that describe the content of each video. For each video v, we denote the set of sections that are assigned by this process by $\mathcal{S}_v^G$, where G stands for the ground truth set.

5.2 Algorithm used for comparison

Besides using the COMITY algorithm to identify the candidate set of videos for creating ground truth, we also used COMITY's resulting assignments as a baseline algorithm for comparison. Specifically, we associate each video with all the sections for which that video was retrieved by the algorithm. In particular, for video v, we denote $\mathcal{S}_v^C$ be the set of sections for which COMITY algorithm retrieved v as one of the top ranking videos. Since our goal is to compare the performance of our approach to this COMITY baseline, for the purposes of evaluation, we only included videos that are retrieved by running COMITY algorithm at the *section level* (that is, not at the chapter level).

In Figure 3, we show the overlap of the videos across the sections. We can see that only about 50% of the videos are assigned to a single section, 25% to two sections and the remaining to more than two sections. From this figure, we would like to highlight two main observations. First, COMITY can be used as a baseline since it also identified multiple sections for the same video (in nearly half the cases). Second and more importantly, since there is sufficient content that is shared across multiple sections, we need an assignment algorithm that identifies the correct granularity to which a video needs to be assigned.

5.3 Choice of Parameters

Our algorithm uses two parameters, θ and λ. The coverage threshold θ determines if the algorithm will match the video to any section. We set $\theta = 0.8$. We also varied θ in the range $[0.6, 0.9]$ obtaining very similar results indicating that the algorithm is not sensitive to this parameter. We estimated the

[3]Due to the volume of work needed per judge, this task was not a suitable candidate for use of workers from Amazon Mechanical Turk platform. In fact, our initial experimentation also confirmed this observation.

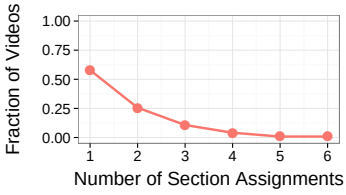

Figure 3: Number of sections that COMITY assigns for each of the 112 videos.

value for the size penalty parameter λ using a cross validation set. This process resulted in $\lambda = 0.48$. We used top three videos ($m = 3$) for computing the context-dependent importance score of each *cphr*.

5.4 Metrics

For each video v, let $\mathcal{S}_v^P$ be the set of sections that are identified by our proposed algorithm. We propose two metrics, 'Accuracy' and 'Relaxed Accuracy' to compare the performance.

Accuracy: This metric measures how accurately an algorithm can identify the entire set of sections that best captures the content in the video:

$$\text{Accuracy} = \frac{\sum_{v \in \mathcal{V}} I[\mathcal{S}_v^A = \mathcal{S}_v^G]}{|\mathcal{V}|}, \tag{7}$$

where $A \in \{C, P\}$ and $I[\mathcal{X} = \mathcal{Y}]$ evaluates to 1 if the sets $\mathcal{X}$ and $\mathcal{Y}$ have identical elements and 0 otherwise. $|\mathcal{V}|$ is the number of videos in the ground truth collection.

Relaxed Accuracy: The above accuracy metric is stringent in that it requires all the sections identified by the algorithm to match with that of the ground truth. Consider the following example: Let the best assignment for a video be section 2. Consider two algorithms, A and B. Suppose that A identifies sections 2 and 3, while B identifies sections 4 and 5. Then, the above metric would penalize both algorithms equally. Clearly, algorithm A is better than algorithm B since the former identified section 2, that is present in the ground truth. We capture this intuitive notion using a distance function that takes into account how different is the inferred set from the ground truth set. In particular, inspired by the edit distance for string comparison [23], we compute "edit distance" (ED) between two sets as the number of sections that need to be inserted or deleted so that the predicted set matches the ground truth. For instance, ED between {2} and {4,5} is 3 while ED between {2} and {2,3} is 1. We define:

$$\text{Relaxed Accuracy} = \frac{\sum_{v \in \mathcal{V}} \left(1 - \frac{ED}{\max_D ED}\right)}{|\mathcal{V}|}, \tag{8}$$

where $\max_D ED$ is maximum edit distance that can be obtained in our dataset D. Note that *Relaxed Accuracy*$\in [0, 1]$, with 1 being perfect overlap and 0 being no overlap. We can also see that the edit distance is maximized when the sections identified correspond to all the sections in the chapter that are not part of the ground truth sections, so that the maximum edit distance equals the number of sections, $|\mathcal{S}_{all}|$ in the chapter. The edit distance between two sets equals their symmetric set difference. Thus, we can compute this metric as:

Figure 4: Breakdown of performance based on COMITY assignment

$$\text{Relaxed Accuracy} = \frac{\sum_{v \in \mathcal{V}} \left(1 - \frac{|\mathcal{S}_v^A \triangle \mathcal{S}_v^G|}{|\mathcal{S}_{all}|}\right)}{|\mathcal{V}|}, \tag{9}$$

where $A \in \{C, P\}$ and $\mathcal{S}_v^A \triangle \mathcal{S}_v^G$ denotes the symmetric set difference (edit distance) between the set of sections identified by an algorithm and the set of ground truth sections.

5.5 Results

We evaluated the algorithms based on two different ways of slicing the data: (A) grouping based on the number of sections assigned by COMITY to evaluate overall performance, and (B) chapter–wise results to understand performance based on chapter characteristics.

Performance based on COMITY assignments: Here, we would like to investigate the performance of our algorithm in comparison to COMITY on the basis of the number of sections that a video is assigned to by COMITY. We partitioned the videos into two groups: videos that are assigned to only one section by COMITY, and those that are not. From Figure 3, roughly 50% of the videos fall into either of these two groups.

Figure 4 shows the comparative results between the two methods for the two groups of videos under both metrics. We can see that when COMITY assigns a video to multiple sections, in many cases, it does so incorrectly, as shown by the achieved accuracy of 0.47. On the other hand, our approach is able to assign videos to the appropriate subset of sections with much higher accuracy (0.73). Under the relaxed accuracy metric, COMITY's performance is still lower than our approach (0.81 *v.s.* 0.90), indicating that even though the videos considered are relevant (recall our assumption that relevant videos are provided at the chapter level), COMITY either incorrectly assigns additional sections or finds only a subset of the ground truth sections. We further analyzed failure cases and found that our approach often fails to assign the right set of sections due to insufficient representation of the video, arising from the inherent restriction of issuing queries based on the section content.

For the group of videos where COMITY assigned to only one section, there is no significant difference in performance be-

tween the two methods. We investigated the reasons for this similar performance: For a video belonging to this group, the corresponding section often tends to be very focused on a particular topic (we discuss this next), and hence there is only a single logical section to which the video could be assigned. Consequently, the two methods result in similar performance for such videos.

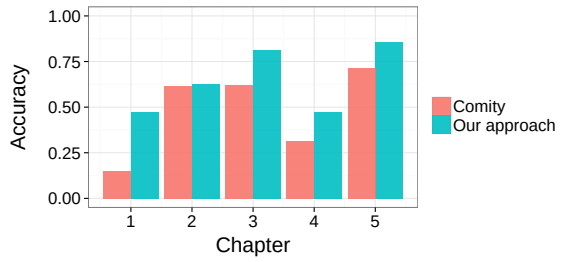

Figure 5: Chapter level comparison

Performance across chapters: We also investigated if there is difference in performance across chapters. Figure 5 shows the results. We further analyzed two chapters, one for which the two methods had similar performance and the other with huge difference in performance. For the former, we found that the corresponding sections in the chapter "Is matter around as pure" have unique focus: for instance, section 2 deals with different types of mixtures, while section 3 presents procedures for separating mixtures. These sections do not overlap much in terms of the concept phrases explained. As a result, videos assigned to each section are unique, and thus, the content of each video is not shared across sections in the chapter. In contrast, in chapter 1 titled "Matter in our surroundings", the first section explains the physical nature of matter, while the second one discusses the characteristics of particles of matter, leading to a huge overlap in the content of these sections. This commonality across sections results in videos that have similar content. Since our approach explicitly models these dependencies, it is able to assign the videos more accurately. In contrast, COMITY is myopic and hence is unable to tease out the interrelationships between sections in the chapter.

6. SUMMARY, DISCUSSION, AND FUTURE WORK

In this paper, we introduced the problem of identifying a set of logical units in a textbook (such as sections in a chapter) that best captures the content in an educational video that is relevant to the textbook. We provided a scalable solution that is effective across various subjects and for educational videos in the wild.

Through this work, we have only touched the tip of the iceberg for effective augmentation of textbooks with videos: In §1, we discussed multiple other considerations that need to be taken into account. Each of these dimensions is a promising direction for future work. Another important research direction is to design rigorous evaluation methodology factoring in these considerations and perform large scale user study in classroom settings [5]. More broadly, in a blended learning setting, a teacher may choose to combine course materials including multimedia presentations from multiple courses. Our work is a step towards addressing challenges that arise in such settings.

7. REFERENCES

[1] R. Agrawal, M. Christoforaki, S. Gollapudi, A. Kannan, K. Kenthapadi, and A. Swaminathan. Mining videos from the web for electronic textbooks. In *ICFCA*, 2014.

[2] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM*, 2009.

[3] R. Agrawal, S. Gollapudi, A. Kannan, and K. Kenthapadi. Enriching textbooks with images. In *CIKM*, 2011.

[4] R. Agrawal, S. Gollapudi, K. Kenthapadi, N. Srivastava, and R. Velu. Enriching textbooks through data mining. In *ACM DEV*, 2010.

[5] R. Agrawal, M. H. Jhaveri, and K. Kenthapadi. Evaluating educational interventions at scale. In *ACM L@S*, 2014.

[6] S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, and M. Aly. Video suggestion and discovery for YouTube: Taking random walks through the view graph. In *WWW*, 2008.

[7] W. Barbe, R. Swassing, and M. Milone. *Teaching through modality strengths: Concepts and practices.* Zaner-Bloser, 1981.

[8] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, 1998.

[9] H. Chen and D. R. Karger. Less is more: Probabilistic models for retrieving fewer relevant documents. In *SIGIR*, 2006.

[10] C. Claxton and P. Murrell. *Learning styles: Implications for improving educational practices. ASHE–ERIC Higher Education Report No. 4.* ASHE–ERIC, 1987.

[11] R. Dunn, J. S. Beaudry, and A. Klavas. Survey of research on learning styles. *Educational leadership*, 46(6), 1989.

[12] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *IJCAI*, 2007.

[13] Y. Gao and Q.-H. Dai. Clip based video summarization and ranking. In *CIVR*, 2008.

[14] J. Gillies and J. Quijada. Opportunity to learn: A high impact strategy for improving educational outcomes in developing countries. *USAID Educational Quality Improvement Program (EQUIP2)*, 2008.

[15] S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *WWW*, 2009.

[16] P. Honey and A. Mumford. *The manual of learning styles.* Maidenhead, 1992.

[17] W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking via information bottleneck principle. In *MULTIMEDIA*, 2006.

[18] W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking through random walk over document-level context graph. In *MULTIMEDIA*, 2007.

[19] S. Huston and W. B. Croft. Evaluating verbose query

processing techniques. In *SIGIR*, 2010.

[20] M. Kokkodis, A. Kannan, and K. Kenthapadi. Assigning educational videos at appropriate locations in textbooks. In *EDM*, 2014.

[21] M. Kokkodis, A. Kannan, and K. Kenthapadi. Assigning videos to textbooks at appropriate granularity. In *ACM L@S*, 2014.

[22] G. Kumaran and V. R. Carvalho. Reducing long queries using query quality predictors. In *SIGIR*, 2009.

[23] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, 1966.

[24] J. Liu, W. Lai, X.-S. Hua, Y. Huang, and S. Li. Video search re-ranking via multi-graph propagation. In *MULTIMEDIA*, 2007.

[25] S. Mariooryad, A. Kannan, D. Hakkani-Tur, and E. Shriberg. Automatic characterization of speaking styles in educational videos. In *ICASSP*, 2014.

[26] O. Medelyan, D. Milne, C. Legg, and I. Witten. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9), 2009.

[27] M. Meeker and L. Wu. Internet trends. Technical report, KPCB, 2013.

[28] M. Miller. Integrating online multimedia into college course and classroom: With application to the social sciences. *MERLOT Journal of Online Learning and Teaching*, 5(2), 2009.

[29] J. Moulton. How do teachers use textbooks and other print materials: A review of the literature. *The Improving Educational Quality Project*, 1994.

[30] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1), 1978.

[31] M. Pinson and S. Wolf. A new standardized method for objectively measuring video quality. *IEEE Transactions on Broadcasting*, 50(3), 2004.

[32] R. Schmeck. *Learning strategies and learning styles*. Plenum Press, 1988.

[33] B. W. Speck, T. R. Johnson, C. P. Dice, and L. B. Heaton. *Collaborative writing: An annotated bibliography*. Greenwood Press, 1999.

[34] M. Strube and S. Ponzetto. WikiRelate! Computing semantic relatedness using Wikipedia. In *AAAI*, 2006.

[35] P. Tantrarungroj. *Effect of embedded streaming video strategy in an online learning environment on the learning of neuroscience*. PhD thesis, Indiana State University, 2008.

[36] S. Tarver and M. Dawson. Modality preference and the teaching of reading: A review. *Journal of Learning Disabilities*, 11(1), 1978.

[37] E. Vee, U. Srivastava, J. Shanmugasundaram, P. Bhat, and S. A. Yahia. Efficient computation of diverse query results. In *ICDE*, 2008.

[38] A. Verspoor and K. B. Wu. Textbooks and educational development. Technical report, World Bank, 1990.

[39] Y. Yang, N. Bansal, W. Dakka, P. Ipeirotis, N. Koudas, and D. Papadias. Query by document. In *WSDM*, 2009.