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. Towards this goal, techniques have been proposed to automatically augment textbook sections with relevant web content such as online educational videos. However, 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. 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

Education literature has extensively highlighted the central role that textbooks play in delivering content knowledge to the students, improving student learning, and in helping teachers prepare lesson plans [19]. The rapid proliferation of cloud-connected electronic devices has enabled the availability of textbooks in electronic format. However, many of these e-textbooks are merely 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 [5, 6, 8, 15, 18]. In fact, studies show better content retention [17] and improved concept understanding [14] when educational multimedia content is shown along with textual material.

With the availability of abundant online video content [13], we can use retrieval algorithms [2] to narrow the video collection to a relevant subset for the textbook. 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. The textbook author creates a logical hierarchy (chapter → sections → subsections, *etc.*) that is suitable 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 [11]: *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. 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.

2. RELATED WORK

There has been considerable work on augmenting textbook sections with relevant supplementary materials mined from the web [1, 2, 3]. In [3], the focus has been on finding textual content from the web that is relevant for a section. Somewhat related is the work proposed in [20] 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 [2] which we adapt to obtain relevant videos.

3. CANDIDATE VIDEO SELECTION

We obtain the candidate set of videos relevant to a textbook chapter using an adaptation of COMITY algorithm [2] that was proposed in the context of augmenting textbook sections with images. We observed that when we applied this technique at the section level (§5), 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 can be shared 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 across sections.

Our adaptation of COMITY is presented in Algorithm 1. A chapter in a textbook is represented as a set of concept phrases (*cphrs*), obtained as the set of phrases that map to Wikipedia article titles [7, 16], and further refined using the techniques proposed in [3]. COMITY forms $\binom{n}{2}$ video search queries by combining two *cphrs* each, in order to provide more context about the chapter. Note that a *cphr* in isolation may not be representative of the text as the same text can discuss multiple concepts. At the same time, a single long query consisting of all concept phrases can lead to poor retrieval [9]. 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.

Algorithm 1 COMITY

```
Input: A textbook chapter; Number of desired video results $k$.
Output: Top $k$ video results from the web.
Obtain (up to) top $n$ concept phrases from the chapter.
Form $\binom{n}{2}$ queries consisting of two concept phrases each.
Obtain (up to) top $t$ video search results for each query.
Aggregate over $\binom{n}{2}$ video result lists, and return top $k$ videos.
```

4. APPROACH & ALGORITHMS

4.1 Representation of Textbook

Each section in a textbook represented by a set of *cphrs*, along with their *context-dependent importance* scores based on the importance of *cphrs* to the section. The computation of the score is based on 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. 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 the importance score, *I*(c) as the average pair-wise inner product between top m videos retrieved when c is used in conjunction with all other *cphrs* in the textbook.

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

where V_i is the vector representation (in terms of *cphrs* and associated weights) for i^{th} top video for c. We used $m = 3$ in our experiments. To account for variances in the scores due to sparsity, we also clustered the scores, and assigned the cluster means of the closest cluster to each of the *cphrs* [10].

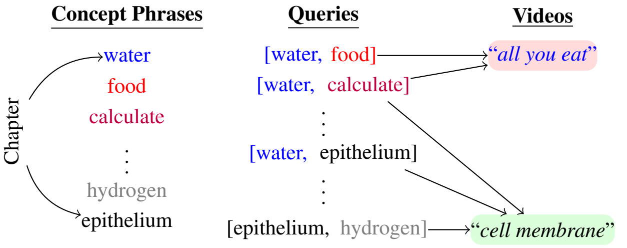

Figure 1: Query based video representation

4.2 Representation of Candidate Videos

We devise a representation for the videos 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:

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

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 by 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 and a large candidate set $\mathcal{S}$ of sections from the textbook chapter, our goal is 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 $\mathcal{T}^*$ that maximizes the objective function:

$$\mathcal{T}^* = \arg\max_{\mathcal{T} \in 2^{\mathcal{S}}} \left(\text{cover}(v, \mathcal{T}) - \lambda |\mathcal{T}| \right), \qquad (1)$$

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. 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. We estimated the value for the size penalty parameter λ using a cross validation set. This process resulted in $\lambda = 0.48$.

[1] Similar observation was made for image retrieval [2].

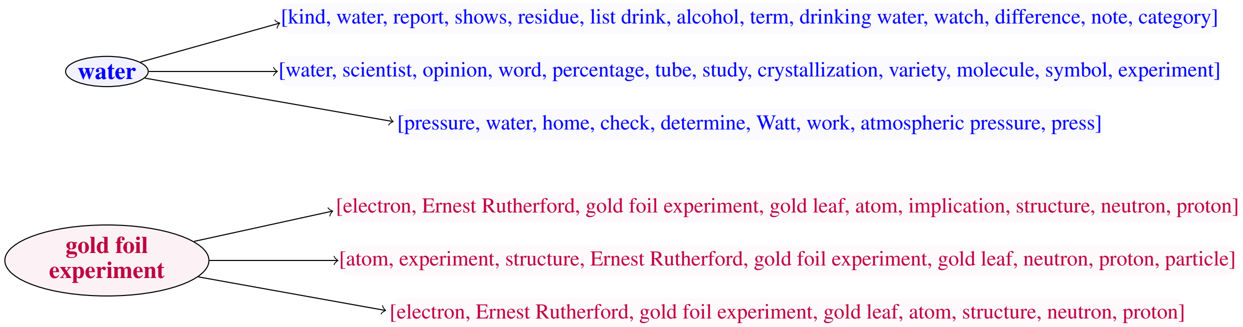
water
[kind, water, report, shows, residue, list drink, alcohol, term, drinking water, watch, difference, note, category]
[water, scientist, opinion, word, percentage, tube, study, crystallization, variety, molecule, symbol, experiment]
[pressure, water, home, check, determine, Watt, work, atmospheric pressure, press]
gold foil experiment
[electron, Ernest Rutherford, gold foil experiment, gold leaf, atom, implication, structure, neutron, proton]
[atom, experiment, structure, Ernest Rutherford, gold foil experiment, gold leaf, neutron, proton, particle]
[electron, Ernest Rutherford, gold foil experiment, gold leaf, atom, structure, neutron, proton]

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

Computing $\text{cover}(v, \mathcal{T})$**:** Let $\mathcal{C}_{book}$ denote the set of all *cphrs* (concept phrases) in the book. 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 $\mathcal{T}$. We define $\text{cover}(v, \mathcal{T})$ to be the weighted fraction of the *cphrs* in the video that are also covered by the subset of sections:

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

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: Given the set of sections in a textbook chapter and a candidate video as inputs, our algorithm first checks whether a certain minimum fraction, θ of the video content can be covered by including all sections in the chapter, and if so, returns the optimal subset of sections (by exhaustively searching over all possible subsets). Upon performing sensitivity analysis, we observed that the algorithm is not sensitive to θ in the range $[0.6, 0.9]$, and hence we set $\theta = 0.8$ in our experiments.

Greedy optimization: In [10], we show that our objective function (Eq. 1) exhibits submodularity and hence admits an efficient greedy algorithm with provable quality guarantees, when the number of sections is large. Let k^* denote the number of sections included using this greedy algorithm, and $F_{k^*,greedy}$ denote the corresponding value of the objective function. Let $F_{k^*,opt}$ denote the optimum value of the objective function subject to the cardinality constraint that exactly k sections are present in the solution. We formally state the theorem below (see [10] for the proof).

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

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.

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* (further explained in the next subsection). 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[2]. 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 sections assignments. In particular, for each video v, $\mathcal{S}_v^G$ is the set of ground truth sections assigned.

Baseline algorithm: We also used COMITY algorithm's assignments as the baseline for comparison. Specifically, for each video v, we associate all the sections for which it was retrieved as a top ranking video, and we denote this set as $\mathcal{S}_v^C$. In fact, only about 50% of the videos are assigned to a single section, 25% to two sections and the remaining to more than two sections. Thus, COMITY can be used as a baseline since it also identified multiple sections for the same video (in nearly half the cases).

Metrics: For each video v, let $\mathcal{S}_v^P$ be the set of sections identified by our proposed algorithm.

Accuracy: This metric measures how accurately an algorithm can identify the entire set of sections that best captures the content in the video: Accuracy $= \frac{\sum_{v \in \mathcal{V}} I[\mathcal{S}_v^A = \mathcal{S}_v^G]}{|\mathcal{V}|}$, 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.

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. We define a relaxed version that takes into account how different the inferred set is from the ground truth set: Relaxed Accuracy $= \frac{\sum_{v \in \mathcal{V}} \left(1 - \frac{|\mathcal{S}_v^A \triangle \mathcal{S}_v^G|}{|\mathcal{S}_{all}|}\right)}{|\mathcal{V}|}$, where $A \in \{C, P\}$, $|\mathcal{S}_{all}|$ denotes the number of sections in the chapter, and $\mathcal{S}_v^A \triangle \mathcal{S}_v^G$ denotes the symmetric set difference between the set of sections

[2] Our initial experiments confirmed that this task was not suited for Amazon Mechanical Turk (due to the volume of work per judge).

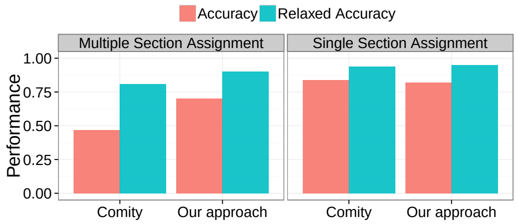

Figure 3: Performance based on COMITY assignment

identified by an algorithm and the set of ground truth sections.

5.1 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 compare the two algorithms based on the number of sections to which a video is assigned to by COMITY. To this effect, we partitioned the videos into two groups: videos that are assigned to only one section by COMITY, and those that are not. Roughly 50% of the videos fall into either of these two groups.

Figure 3 shows the results. 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 between 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.

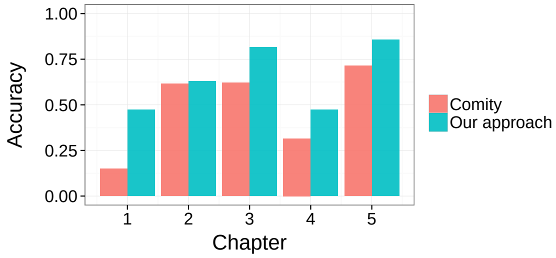

Figure 4: Performance across chapters

Performance across chapters: We also investigated if there is difference in performance across chapters. Figure 4 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 relationships between sections in the chapter.

6. SUMMARY AND FUTURE WORK

In this paper, we introduced the problem of identifying a set of logical units in a textbook that best captures the content in a relevant educational video. 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. There are multiple other considerations such as presenter [12] or presentation styles that need to be taken into account. We also need to design rigorous evaluation methodology factoring in these considerations and perform large scale user study in classroom settings [4]. 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. Kannan, and K. Kenthapadi. Enriching textbooks with images. In *CIKM*, 2011.

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

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

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

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

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

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

[9] S. Huston and W. B. Croft. Evaluating verbose query processing techniques. In *SIGIR*, 2010.

[10] M. Kokkodis, A. Kannan, and K. Kenthapadi. Assigning educational videos at appropriate locations in textbooks. Technical Report MSR-TR-2014-62, Microsoft Research, 2014.

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

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

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

[14] 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.

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

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

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

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

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

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