Assigning Videos to Textbooks at Appropriate Granularity

Marios Kokkodis
NYU Stern

Anitha Kannan
Microsoft Research

Krishnaram Kenthapadi
Microsoft Research

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. *Given a candidate set of relevant educational videos for augmenting an electronic textbook, how do we assign the videos at the appropriate granularity (a collection of logical units in the book)?* 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.

REPRESENTATION OF TEXTBOOK

Consider a textbook, consisting of K chapters, each subdivided into sections. We define $\mathcal{C}_{book}$ to be the set of concept phrases (*cphrs*) in the book that map to Wikipedia article titles, further refined as in [1]. We define *context-dependent importance* score, $I(c)$ for a *cphr* c as follows. 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. We measure *I*(c) as the average pair-wise inner product between top m videos retrieved in response to queries that contained c: $I(c) = \frac{\sum_{1 \leq i < j \leq m} < V_i, V_j >}{\binom{m}{2}}$, where V_i is the vector representation (in terms of *cphrs* and associated weights) for i^{th} top video for c (explained in the next section).

VIDEO CANDIDATE SELECTION AND REPRESENTATION

We first obtain the candidate set of videos relevant to a textbook chapter using an adaptation of COMITY algorithm [1]. Given top n concept phrases (*cphrs*) present in a chapter, $\binom{n}{2}$ queries are formed by combining two *cphrs* each, and issued to a commercial video search engine. The most relevant videos for the chapter are obtained by aggregating the video result lists over all combinations of queries.

In order to match a video to a set of sections, we also need a representation of the video. While, in principle, one can use transcripts associated with videos and identify the *cphrs* in them (similar to identification in textbooks), most videos in our corpus did not have high quality user-uploaded transcripts, and further, we found the transcripts extracted by automatic speech recognizers to be of poor quality. Instead, we use a different approach for video representation based on the queries from the textbook that led to the videos. For each *cphr* c and video v, 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.

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:

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

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.

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 $\mathcal{T}$. 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:

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

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). In our experiments, we used $\theta = 0.8$, and estimated λ to be 0.48 through cross validation.

EVALUATION

We evaluate our approach over the first five chapters of a 9^{th} grade science book, spanning different sub-branches of science. We obtained an initial set of 178 videos by running

L@S'14, March 4–5, 2014, Atlanta, Georgia, USA.
ACM 978-1-4503-2669-8/14/03.
http://dx.doi.org/10.1145/2556325.2567880

Algorithm 1 Section Subset Selection For Videos

Input: Set of sections $\mathcal{S}$ in a given textbook chapter; A candidate video v; Coverage threshold θ.
Output: The optimal subset of sections $\mathcal{T}^* \subseteq \mathcal{S}$ (or $null$ depending on the coverage threshold).

1: **if** $cover(v, \mathcal{S}) < \theta$ **then return** $null$.
2: **return** $\arg\max_{\mathcal{T} \in 2^{\mathcal{S}}} \ (\text{cover}(v, \mathcal{T}) - \lambda|\mathcal{T}|)$.

COMITY algorithm (at the *section level*) across sections in all chapters. A human assessor was asked to read all five chapters, and then watch each video and manually identify all the sections that together capture the content of 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 every video v, denote the set of sections that are assigned by this process by $\mathcal{S}_v^G$.

Baseline algorithm: For video v, we obtained the baseline as the set $\mathcal{S}_v^C$ 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).

We empirically validated that COMITY can be used as a baseline since (a) it also identified multiple sections for the same video (in nearly half the cases), and (b) there is sufficient content that is shared across multiple sections.

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: $\text{Accuracy} = \frac{\sum_{v \in \mathcal{V}} I[\mathcal{S}_v^A = \mathcal{S}_v^G]}{|\mathcal{V}|}$, 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 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. We define a relaxed version that takes into account how different the inferred set is from the ground truth set: $\text{Relaxed Accuracy} = \frac{\sum_{v \in \mathcal{V}} \left(1 - \frac{\mathcal{S}_v^A \triangle \mathcal{S}_v^G}{|\mathcal{S}_{all}|}\right)}{|\mathcal{V}|}$, 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 (edit distance) between the set of sections identified by an algorithm and the set of ground truth sections.

Results:

Methods	Accuracy	Relaxed accuracy
COMITY	0.513	0.877
Proposed approach	0.649	0.908

We can see that under the stringent metric of **Accuracy**, our approach performs significantly better than the baseline (COMITY). With the **Relaxed Accuracy** metric, our approach performs slightly better than the baseline.

DISCUSSION

The recent upsurge in new models of learning such as blended learning, massive open online courses, and flipped classrooms [2, 3, 4, 6, 7, 8] emphasizes the importance of audio-visual learning. This trend begs the question: Can we eliminate textbooks, altogether? Our answer is a qualified no. We believe that textbooks will continue to play a central role in educational instruction, with videos enabling the additional modality to learn from (*e.g.*, [9]). In fact, education literature has extensively highlighted the importance of textbooks in delivering content knowledge to the students, improving student learning, and in helping teachers prepare the lesson plans [5, 10]. However, with the emergence of abundant educational content available online, cloud-connected electronic devices and electronic textbooks, we are now well positioned to integrate multimedia content to personalize textbooks based on the learning style of the user. In this work, we took a step towards addressing associated challenges: how do we effectively match huge educational content available online to the textbook of interest *at appropriate granularity*? An important subsequent work is to design rigorous evaluation methodology and perform large scale user study among students to quantify the effectiveness of using such an enriched textbook.

While our current approach focused on the relevancy and the appropriate granularity of the video, several dimensions pertaining to the video, the viewer and the presenter need to be taken into account for effective augmentation of textbooks with videos. Each of these dimensions is a promising direction for future work.

REFERENCES

1. Agrawal, R., Gollapudi, S., Kannan, A., and Kenthapadi, K. Data mining for improving textbooks. *ACM SIGKDD Explorations Newsletter 13*, 2 (2011).
2. Bergmann, J., and Sams, A. *Flip your classroom: Reach every student in every class every day*. International Society for Technology in Education, 2012.
3. Dellarocas, C., and Alstyne, M. V. Money models for MOOCs. *Communications of the ACM 56*, 8 (2013).
4. Garrison, D. R., and Kanuka, H. Blended learning: Uncovering its transformative potential in higher education. *The internet and higher education 7*, 2 (2004).
5. Gillies, J., and Quijada, J. Opportunity to learn: A high impact strategy for improving educational outcomes in developing countries. *USAID Educational Quality Improvement Program (EQUIP2)* (2008).
6. Martin, F. G. Will massive open online courses change how we teach? *Communications of the ACM 55*, 8 (2012).
7. Staker, H., and Horn, M. B. Classifying K–12 blended learning. *Innosight Institute* (2012).
8. Strayer, J. *The effects of the classroom flip on the learning environment: A comparison of learning activity in a traditional classroom and a flip classroom that used an intelligent tutoring system*. PhD thesis, Ohio State University, 2007.
9. Tisdell, C. C. *Engineering mathematics: YouTube workbook*. Bookboon, 2013.
10. Verspoor, A., and Wu, K. B. Textbooks and educational development. Tech. rep., World Bank, 1990.