Mining Videos from the Web for Electronic Textbooks

Rakesh Agrawal[1], Maria Christoforaki[2⋆], Sreenivas Gollapudi[1], Anitha Kannan[1], Krishnaram Kenthapadi[1], and Adith Swaminathan[3⋆]

[1] Microsoft Research
[2] Polytechnic Institute of New York University
[3] Cornell University

Abstract. We propose a system for mining videos from the web for supplementing the content of electronic textbooks in order to enhance their utility. Textbooks are generally organized into sections such that each section explains very few concepts and every concept is primarily explained in one section. Building upon these principles from the education literature and drawing upon the theory of *Formal Concept Analysis*, we define the *focus* of a section in terms of a few *indicia*, which themselves are combinations of concept phrases uniquely present in the section. We identify videos relevant for a section by ensuring that at least one of the indicia for the section is present in the video and measuring the extent to which the video contains the concept phrases occurring in different indicia for the section. Our user study employing two corpora of textbooks on different subjects from two countries demonstrate that our system is able to find useful videos, relevant to individual sections.

1 Introduction

It is inevitable that the traditional paper-based textbook will gradually evolve into electronic textbooks accessible from computing devices connected to the Internet. To enhance the utility of such electronic textbooks, we propose the problem of mining from the web a few selective videos related to a section in a textbook and present effective techniques for this purpose. Our techniques can be used to obtain a candidate set of relevant videos, which can then be used by different stakeholders: by teachers when preparing lectures on the material, by authors when creating pointers to supplementary video material for the textbook, and by students for reinforcing their learning from an alternative exposition.

The problem of finding suitable videos for a textbook section is quite different from that of finding videos relevant to a stand-alone piece of text. Textbooks are written following certain organizational principles in order to enable the reader to understand their content without incurring undue comprehension burden [13, 22]. Two properties of a well-written textbook of particular relevance to the present work are: (1) *focus* that says that each section explains a few concepts, and (2) *unity* that implies that for each concept there is a unique section that best explains the concept. In the presence of the unity property, the focus of a section can be viewed as the the unique contribution of

⋆ Work done at Microsoft Research.

the section to the textbook. The conventional information retrieval methods (*e.g.*, TF-IDF [32], LSA [18], and LDA [9]) are not adept at representing the focus of a textbook section (see [3]).

Hence, we take a departure from traditional retrieval methods and present an approach that first infers the focus of each section, taking into account the content of all other sections, and then finds videos relevant to that focus. Our representation of the focus is derived from the theory of *Formal Concept Analysis* [20]. We represent the focus using, what we call, *indicia*. An *indicium* for a section is a maximal combination of concept phrases that occurs frequently in that section but is not present in any other section. We also associate a score with each indicium based on the importance of the underlying concept phrases that captures the significance of the indicium to the section. We identify videos relevant for a section by ensuring that at least one of the indicia is present in the video and measuring the extent to which the video contains the concept phrases occurring in different indicia, after taking into account their significance. We study the efficacy of our system using textbooks on different subjects from two different countries. This extensive user study shows that our system is able to find useful videos relevant to the individual sections of a textbook.

1.1 Assumptions and Scope of the Paper

Before delving into details, we offer a few clarifications. We are assuming an evolutionary transformation of the current textbooks to their electronic versions. Undoubtedly, in the future, there will be textbooks written in a way to specifically exploit the functionalities provided by the electronic medium, but that will take time. Meanwhile, we are interested in taking the current books and enhancing the experience of studying from them. In the same vein, one can even question the continued need for textbooks. However, years of educational research have shown that the textbooks are the educational input most consistently associated with improvements in student learning [52]. They serve as the primary conduits for delivering content knowledge to the students and the teachers base their lesson plans mainly on the material given in textbooks [21]. Pragmatically, their importance in educational instruction is unlikely to diminish in the foreseeable future.

We should also clarify why enhancing electronic textbooks with videos has a high payoff. A number of pilot studies have established the importance of using multimedia content in educational instruction. In a recent work [36], Miller showed that the use of multimedia content is "particularly valuable in helping students acquire the initial mental imagery essential for conceptual understanding". Tantrarungroj [50] used a month-long longitudinal study to show that the students have much better content retention when they are presented with multimedia content along with textual material. The visual modality is particularly strong in many people because a child "sees and recognizes before speaking" [8]. The educational pedagogy informs us that any supplementary material is most effective when it is presented in close proximity to the main material [16]. We therefore augment videos at the section level.

We present the technology core for identifying relevant videos, but do not discuss the mechanisms for integrating them into the textbook. Issues such as implications for royalty sharing and intellectual property rights are outside the scope of the paper. It

is known that learning outcomes depend not only on the availability of educational materials, but also on how they are used by the teachers and students and how effectively they have been integrated with other interventions [21, 37]. While such deployment issues are critically important, they are beyond the scope of this paper.

When proposing a video, there are multiple considerations that must be taken into account. They can be broadly grouped into aspects related to the video, the viewer, and the presenter respectively. Video considerations include the relevancy of the content of the video to the textbook section, duration of the video, and the video quality [41]. Viewer considerations include the appropriateness of the video to the viewer's prior knowledge of the subject matter and preference for the type of video such as lecture, demonstration, animation, or enactment. Presenter considerations include the presenter speaking style [33], diction, and accent. In this paper, we address the problem of relevancy: how do we augment textbook sections with relevant videos available on the web?

1.2 Textbook Corpora

Our study uses publicly available school-level textbooks on different subjects from two different countries. The first corpus consists of books published by the CK-12 Foundation, U.S.A. that are available online from *ck12.org*. The second corpus comprises of books published by the National Council of Educational Research and Training (NCERT), India. These books are also available online from *ncert.nic.in* and they have been used in prior studies related to textbooks [4, 5, 6]. The language of these books is English. We generate augmentations for every section in every chapter of books in our corpora.

Respecting the space constraint, we present and discuss in depth the results for two books. From the CK-12 corpus, we provide results for the middle school Biology textbook. This book introduces various themes in Biology including Molecular Biology and Genetics, Cell Biology, Prokaryotes, Animals, Plants, and Human Biology. The book consists of 26 chapters, spanning 147 sections, and we consider the augmentations for all the sections in our performance evaluation.

From the NCERT corpus, we present results for the Grade XII Physics textbook. The broad theme of this book is electricity and magnetism. It covers electric charges and fields, electrostatic potential and capacitance, current electricity, moving charges and magnetism, magnetism and matter, electromagnetic induction, alternating current, and electromagnetic waves. This book consists of 15 chapters, spanning 200 sections, and again our evaluation considers the results for all the sections.

Hereafter, we refer to these books as Biology and Physics textbooks respectively.

1.3 Organization

The rest of the paper proceeds as follows. We start off by discussing the related work in §2. We then describe our model for representing the focus of a textbook section in §3. We describe how we use our representation of the focus of a textbook section for finding videos relevant to it in §4. We present the results of the user study in §5. We conclude with a summary and directions for future work in §6.

2 Related Work

Aboutness: The problem of formally defining the focus of a textbook section is related to the question of "what a document is about?". The latter has been been extensively investigated in the information retrieval literature from both theoretical (*e.g.*, [10, 24, 27]) as well as pragmatic perspectives (*e.g.*, [28, 32, 39]). However, the conventional information retrieval techniques are not adept at capturing the focus of a textbook section [3]. Our proposed representation of the focus is rooted in the theory of Formal Concept Analysis [20]. It also agrees with the properties of well-written textbooks enunciated in the education literature [13, 22]. We also validate its efficacy through the application of finding relevant videos for different textbook sections.

Content-based Video Retrieval: Quite innovative research has been reported in content-based video retrieval where the emphasis is on retrieving videos based on pre-specified physical object categories such as cars and people and their instances [40, 45]. There is also work on recognition and retrieval for certain classes of events for these objects (*e.g.*, human actions such as handshakes and answering phones [51], sporting events [57], or traffic patterns [25]). Retrieval is initiated by providing a textual query, or a representative image, or a region of the image depicting the object of interest. The TREC Video Retrieval Evaluation [38] has played a key role in the development of methods for content-based exploitation of digital videos. These methods have been designed to recognize objects that can be represented using visual pixels, and thus are inapplicable for recognizing abstract concepts such as 'kinetic energy' that are common and central in textbooks.

Video Search Engines: Popular search engines such as Google and Bing include support for video search. These search engines work by indexing the associated metadata and matching keyword queries with the stored metadata. The metadata may include textual description and tags, user comments and ratings, and queries that led to the video. One might be tempted to provide the text string of a textbook section as query to a video search engine and obtain the relevant videos. However, it is well known that the current search engines do not perform well with long queries [26, 29]. Indeed, when we ourselves experimented by querying the search engines using the first few lines of a section, we got none or meaningless results. In one major stream of research on information retrieval with long queries, the focus is on selecting a subset of the query, while in another it is on weighting the terms of the query [55]. This body of research however is not designed to work for queries consisting of arbitrarily long textbook sections.

Textbook Augmentation: It has been empirically observed that the linking of encyclopedic information to educational material can improve both the quality of the knowledge acquired and the time needed to obtain such knowledge [17]. Motivated by this finding, techniques for mining the web for augmenting textbooks with selective links to web articles and images have been presented in [4, 6]. We extend this line of research by investigating video augmentations. We also introduce new abstractions and techniques.

Massive Open Online Courses: Several institutions have made available the videos of the course lectures, and there are websites (*e.g.*, EducationalVideos.com, VideoLectures.net, WatchKnowLearn.org) that aggregate links to them. Massive open online courses (MOOCs) are a relatively new phenomenon to enable teachers to reach a global student population through video-based pedagogy. Coursera, edX, Khan Academy, and Udacity are examples of platforms that have sprung up to support such courses. We view these platforms as video sources for textbook augmentation, as well as potential consumers of our research.

Crowdsourcing: It was proposed in [1] to create an education network to harness the collective efforts of educators, parents, and students to collaboratively enhance the quality of educational material. Some websites (*e.g.*, Notemonk.com) allow students to download textbooks and annotate them. Such annotations can include links found interesting by the students, which can then be aggregated. Some allow teachers (*e.g.*, LessonPlanet.com) to find lesson plans, worksheets, and videos to assist them with their classroom presentations. Yet others (*e.g.*, Graphite.org) help educators to use and share apps, games, videos, and websites. One could view the techniques proposed in this paper as providing an initial consideration set of videos that gets refined using crowdsourcing and other manual approaches.

3 Focus of a Textbook Section

Our representation of the focus of a section in a textbook is derived from the *Formal Concept Analysis* (FCA) [20]. The theory of FCA has been shown to have connections to the philosophical logic of human thought [54]. We first provide a brief overview of FCA and then formally define the focus of a section in terms of *indicia*. Later, we evaluate the efficacy of our representation through the application of finding relevant videos for different textbook sections.

3.1 Formal Concept Analysis: An Overview

FCA postulates that we are given a context K consisting of a set of objects G, a set of attributes (properties) M, and a relation $I \subseteq G \times M$ specifying which objects have which attributes. A concept is then a pair (A, B) consisting of: i) its extent A, comprising all objects which belong to the concept, and ii) its intent B, comprising all attributes which apply to all objects of the extension. A formal concept is defined to be a pair of maximal subset of objects and maximal subset of attributes such that every object has every attribute.

The formal concepts are naturally ordered by the subconcept-superconcept relation as defined by: $(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow B_1 \supseteq B_2$. The set of all concepts together with the above partial order constitutes the *concept lattice* of the given context. For many applications, it is desirable to limit to the top-most part of a concept lattice since this region corresponds to concepts with a minimum support which are relatively stable to small perturbations (noise) in data, and also since the size of a concept lattice can be exponential in the size of the context in the worst case [30]. In [48], iceberg

concept lattices, based on frequent itemsets as known from data mining [7], were introduced to address this issue. Let $\mu \in [0, 1]$ be the minimum support. A concept (A, B) is said to be frequent if at least μ fraction of objects in G individually have every attribute in B. The set of all frequent concepts of a context K, together with the partial order between them, is called its *iceberg concept lattice*. See [42] for a comprehensive survey of recent advances in FCA and computational techniques. See [11, 14, 43] for overviews of several applications of FCA in information retrieval.

3.2 Using FCA to represent Focus

Assume we have a textbook, consisting of n sections, each of which is subdivided into paragraphs. The sections and paragraphs can be those specified by the author or they can be determined using techniques such as TextTiling [23]. We will use *cphr* to denote a concept phrase present in a text. Let $\mathcal{C}_{book}$ be the set of all *cphrs* in the book.[4]

Since the formal concepts are abstract, we can only observe their manifestations in the form of underlying *cphrs* appearing in various paragraphs. Given a textbook section s, treat different paragraphs of s as objects, different *cphrs* occurring in s as attributes, and define the relationship between objects and attributes based on occurrence of a *cphr* in a paragraph. Thus, a pair of maximal set of paragraphs P_C and maximal combination of *cphrs* C such that every *cphr* in C is present in every paragraph in P_C corresponds to a formal concept of the section.

Observe that the pair representation for a formal concept has redundancy built into it. Clearly, given a formal concept (A, B), the attribute set B completely determines the object set A, and vice versa. Thus, the iceberg concept lattice of section s can be thought of as corresponding to a partial order over sets of *cphrs* present in s. If $B_1 < B_2$ in this partial order then the set of *cphrs* corresponding to B_1 will be a superset of B_2. For compactness, therefore, we take the leaf nodes of the partial order since they correspond to the most specific sets of *cphrs* (or equivalently maximal combinations of *cphrs*) that are also frequent in the section.

Finally, since we are interested in concepts that are unique to each section, we add a uniqueness constraint to define the focus of the section. More precisely, we only include those leaf nodes that are rare in any other section [49].

Definition 1 (Indicium of a section). *A set of cphrs C present in a section s of the textbook constitutes an* indicium *of s if (1) C is frequent in s, (2) C uniquely occurs in s (i.e., there is no other section of the book in which C is frequent), and (3) C is maximal (i.e., there is no superset of C in s which is also frequent in s).*

[4] The identification of *cphrs* primarily involves detection based on rules or statistical and learning methods [28, 32]. In the former, the structural properties of phrases form the basis for rule generation, while the importance of a phrase is computed based on its statistical properties in the latter. Building upon [19, 34, 47], our implementation defines the initial set of *cphrs* to be the phrases that map to Wikipedia article titles. This set is refined by removing malformed as well as common phrases based on their probability of occurrence on the Web [53]. Our methodology is oblivious to the specific *cphr* identification technique used, though the performance of the system is dependent on it. Our implementation uses author provided sections and paragraphs.

pharynx, cellular respiration, transporting oxygen
cardiac muscle, connective tissue, gas transport
nasal cavity, connective tissue, gas transport

(a) Respiratory system

pharynx, respiratory system, epiglottis
emphysema, epiglottis, cigarette
bronchus, cigarette, respiratory system

(b) Respiratory diseases

pharynx, lipid digestion, pepsin
pharynx, large intestine, salivary gland
gall bladder, large intestine, pepsin

(c) Digestive system

Table 1: Indicia for consecutive sections in the Biology textbook.

Definition 2 (Focus of a section). *The set of indicia of a section s constitutes the* focus *of s, denoted by Ψ_s.*

We remark that our derivation of the definition of focus of a section agrees with the properties of well-written textbooks investigated in the education literature [13, 22]. For an author to have introduced a formal concept in a section, the *cphrs* underlying the formal concept must occur frequently across many paragraphs in the section. As a section contributes unique content to the book and introduces very few formal concepts, their underlying *cphr* combinations must appear uniquely in the section, and if not, then infrequently in other sections. We obtain concise representations as a side effect of the maximality constraint. Our implementation sets μ to require that an indicium must appear in at least two paragraphs in the section for it to be considered frequent.

We also remark that our notion of an indicium is related to the idea of a hypothesis for a class present in the FCA literature. Note that indicium C is a maximal frequent (and hence closed) itemset in class (text section) s, which is not frequent in another class (section). As defined in [31], a hypothesis for class s is a closed itemset occurring in s and not occurring in other classes. A minimal hypothesis is an inclusion-minimal hypothesis. An indicium is thus a "relaxation" of a minimal hypothesis, allowing it to occur in another class, but not frequently. Thus, the focus of a section consists of the set of relaxed minimal hypotheses for the section.

3.3 Illustrative Examples

Biology textbook: Table 1 shows top indicia for three consecutive sections in the Biology textbook (wherein the indicia are ordered by their significance score (see §4.1)). Table 1(a) gives indicia for the section on the anatomy of human respiratory system, Table 1(b) for the next section that discusses respiratory diseases, and Table 1(c) for the subsequent section that explains human digestive system.

We see that in each of these sections, there is at least one indicium that contains the *cphr* 'pharynx'. In human Biology, 'pharynx' refers to a part of the throat that participates in respiration and digestion. Hence, this phrase is discussed in all three sections and is present in the corresponding indicia. However, other *cphrs* occurring in these indicia provide the additional content (respiration or digestion) with which to disambiguate and represent the focus of the corresponding sections.

field line, magnetic field, monopole
field line, magnet, charged particle
electrostatics, field line, monopole

(a) Magnetism & Gauss' Laws

field line, magnetic field, earth
equator, meridian, southern hemisphere
earth, solar wind, poles

(b) Earth's Magnetism

Table 2: Indicia for adjacent sections in the Physics textbook.

The indicium ⟨pharynx, cellular respiration, transporting oxygen⟩ in the first section captures the working of the respiratory system in which oxygen enters through the mouth and nose and then travels through the pharynx to reach the lungs. In contrast, in the second section, the indicium ⟨pharynx, respiratory system, epiglottis⟩ captures how the valve, epiglottis, near the pharynx points upwards during respiration in order to enable breathing. In the third section on the human digestive system, the indicium ⟨pharynx, lipid digestion, pepsin⟩ differentiates the use of 'pharynx' by using digestion related concept phrases.

As another example, consider the indicium ⟨emphysema, epiglottis, cigarette⟩ in the second section. The *cphr* 'emphysema' refers to a progressive disease of the lungs caused mainly by smoking tobacco. Smoking tobacco also causes inflammation of epiglottis and hence can cause obstruction of oxygen through the 'pharynx'. Similarly, consider the indicium ⟨gall bladder, large intestine, pepsin⟩ in the third section. The *cphr* 'pepsin' refers to an enzyme that aids digestion of protein in the stomach and the *cphr* 'gall bladder' to the organ that stores bile and then secretes it to aid digestion.

Physics textbook: Table 2 shows top indicia for two adjacent sections in the Physics textbook. Table 2(a) shows indicia for the section on magnetism & Gauss' laws, while Table 2(b) shows them for the section on Earth's magnetism. The first section discusses the magnetic field and the physics behind their effects on moving particles. The second section discusses how Earth acts as a magnet. Consider the first rows of the two tables. They both contain *cphrs* 'field line' and 'magnetic field', but the *cphr* 'monopole' is unique to the indicium for the first section. The *cphr* 'monopole' appearing in the first section distinguishes this section on general magnetism from the section on Earth's magnetism: a magnetic monopole is a hypothetical particle in particle physics that is an isolated magnet with only one magnetic pole, and hence is not discussed in the context of Earth's magnetism as Earth has both poles. The *cphr* 'earth' is rather generic, but the indicium formed by combining it with 'field line' and 'magnetic field' is very pertinent to the section on Earth's magnetism.

4 Augmenting with Videos

A video might be associated with one or more of the following information: (a) images from the visual channel, (b) audio from the auditory channel, (c) video metadata consisting of title, description and any other video related properties such as duration and format, and (d) textual context (*e.g.*, webpage in which the video may have been embedded). One could attempt to match the textual content of a textbook section to the images from the visual channel of the video. However, today's video recognition systems can effectively recognize only the physical objects that are describable using vi-

sual pixels [38], whereas we need to be able to find videos relevant to textbook sections containing abstract concepts. Our system, therefore, employs transcript of the spoken words in the video to infer the relevance of the video to the textbook section. Many videos have such transcripts associated with them; otherwise, one can generate transcripts using speech recognition [46].

Our problem now reduces to the following: given a textbook section (a query), search for related documents over the corpus of video transcripts. At a high level, this problem is similar to the query by document work [56] wherein given a news article (a query), techniques were proposed for identifying related documents from a corpus of blogs. However, our approach differs in two respects. We represent the textbook section using indicia which themselves are founded on formal concept analysis and properties of well-written textbooks, whereas their approach represents the given document by extracting key phrases. Our technique for using the representation to query the corpus (see below) also differs from their approach of issuing a conjunctive query of key phrases to a specialized blog search engine.

Given a section s and its set of indicia, Ψ_s, the videos relevant to the section are obtained using a two-step process. First, a candidate set of videos is selected by only including videos whose transcripts contain all *cphrs* from at least one indicium in Ψ_s. For each video in the candidate set, we assign a relevance score by measuring the combined significance of the indicia from the section that are present in the corresponding transcript. Let $\Psi_{s,v} \subseteq \Psi_s$ be the set of indicia of section s that are found in the transcript of video v. The relevance score for the video v is given by: $relevanceScore(v) := \sum_{C \in \Psi_{s,v}} f(C)$, where $f(C)$ is the significance score of indicium C. The videos are then ranked using this score, and the top k are chosen for augmenting the section.

4.1 Significance of an indicium

An indicium consists of a combination of *cphrs* that collectively represent the unique content a section, but many such combinations may exist for the same section. However, some indicium may offer a more significant representation than others. Hence, we associate a score denoting the significance of an indicium based on the *importance* of the underlying *cphrs*.[5] We first enunciate the desirable properties of significance score.

*Property 1 (*MONOTONICITY*).* The significance score of an indicium is a monotonically increasing function of the importance of its constituent *cphrs*.

This property is rooted in the intuitive notion that an indicium made up of more important *cphrs* is more significant. In particular, inclusion of an additional *cphr* to an indicium results in a more significant indicium (the uniqueness requirement is still preserved).

[5] Adopting the "keyphraseness" notion from [34, 35], our implementation defines the importance $\phi(c)$ of a *cphr* c in terms of the likelihood that the *cphr* is hyperlinked to the corresponding article in Wikipedia. The intuition is that more important *cphrs* are more likely to be hyperlinked in Wikipedia. Formally, $\phi(c) := n_{link}(c)/n_{all}(c)$, where $n_{link}(c)$ is the number of Wikipedia articles in which c occurs as a hyperlink and $n_{all}(c)$ is the total number of articles in which c appears. See [28, 32] for other possibilities.

*Property 2 (*CONCENTRATION*).* The significance score of an indicium increases as the importance of its constituent *cphrs* gets concentrated, that is, the importance is shifted from less important *cphrs* to more important *cphrs* retaining the same total importance.

This property stems from the observation that the more important *cphrs* tend to have a broader scope, for example, representing the entire chapter. By themselves, the less important *cphrs* may not represent a section and may even be ambiguous, but their combination with more important *cphrs* helps to narrow down to the focus of the section. The corresponding indicium can be thought of as anchoring to more important *cphrs*, and then refining their scope using less important *cphrs*.

For example, all three sections shown in Table 1 discuss the *cphr* 'pharynx'. The additional *cphrs* in the respective sections help to refine the scope of this *cphr* to either respiration or digestion as discussed in §3.3.

4.2 Characterization of Significance Score for an Indicium

We next show that the significance score of an indicium can be obtained using a broad category of simple functions that satisfy properties 1 and 2. Let $f(C)$ denote the significance score of indicium C. Let $c_1, c_2, \ldots, c_l$ be the *cphrs* present in C, listed in the decreasing order of their importance, that is, $\phi(c_1) \geq \ldots \geq \phi(c_l)$.

Claim. Suppose f is defined as the sum of a univariate function of the importance of constituent *cphrs*: $f(C) := \sum_{c \in C} g(\phi(c))$. Then, f satisfies properties 1 and 2 if $g(.)$ is a monotonically increasing non-negative convex function.

Proof. See [3].

Our implementation instantiates function g as $g(x) := e^x$. This function satisfies the requirements in Claim 4.2, and favors indicia for which the importance is concentrated in a few *cphrs*.

5 Performance

We now present the results of the user studies we conducted to quantify how well our approach is able to find videos relevant to the focus of each section. We first describe the video corpus, and then provide the results.

5.1 Video Corpus

The video corpus consists of education-related, short videos obtained from a focused web crawl [12, 44]. The crawler is seeded with educational videos from a few reputed sites. These videos span broad levels of education ranging from school to higher education to lifelong learning and originate from a variety of sources. Many of these videos had accompanying user uploaded transcripts of the video content. In order to remove variability arising out of the quality of speech recognition of the audio from the auditory channel of the videos, our experiments employed only those videos that contained author uploaded transcripts. There were nearly 50,000 such videos.

5.2 Experiments

We carried out two sets of experiments to assess how well our techniques are able to find relevant videos. The first experiment evaluates the proposed videos by measuring the precision of retrieval. The second experiment measures the congruence of the retrieval by computing agreement between the section and the retrieved video, in terms of overlap between concept phrases deemed important for the section and for the video by a panel of judges. We measure overlap using a number of similarity measures.

Ideally, we would have liked to have as judges those students who had studied from the textbooks in our test corpus. In the absence of the access to this subject population to us, we carried out our user study on the Amazon Mechanical Turk platform, taking care to follow the best practices [2].

5.3 Precision

Setup: Taking cue from the relevance judgment literature [15, 38], we asked the turkers to read a section, watch a video, and then judge if the video was relevant to the section. The default choice in the HIT (Human Intelligence Task) was set to 'not-relevant' so that the judges needed to explicitly choose 'relevant' if they indeed found the video to be relevant. Each judge was required to spend a minimum of 30 minutes on a HIT. We rejected any HIT where the time spent was less than the minimum. Each HIT was judged by seven judges. In this manner, we computed the relevance of the top three videos proposed by our system over all sections in four randomly chosen chapters, for both the textbooks.

Metric: Our first metric is the commonly used $precision@k$ [32] which measures the fraction of retrieved videos in the top K positions that are judged to be relevant. For a section s, let $v_{s,j}$ be the retrieved video at position j. Let $rel(v_{s,j})$ be a binary variable that takes a value of 1 if the majority of judges voted $v_{s,j}$ to be relevant for s. Then,

$$precision@k = (\sum_{s \in S} \sum_{j=1}^{k} rel(v_{s,j})/k) \ / \ |S|,$$

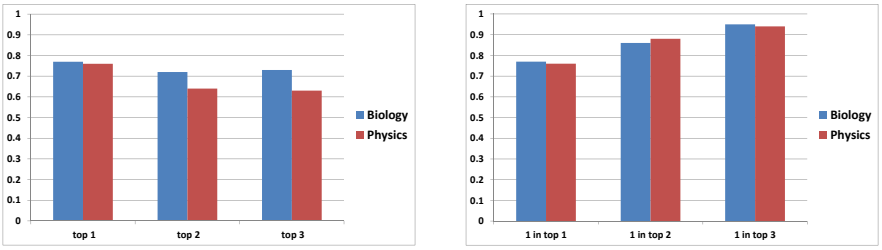
where k is the number of videos retrieved for each section and S is the set of sections.

We also measure whether the judges found at least i of the videos shown in top k positions for each section to be relevant, and compute the average across all sections:

$$precision@(i,k) = \sum_{s \in S} \delta[\sum_{j=1}^{k} rel(v_{s,j} \geq i]) \ / \ |S|,$$

where $\delta[x]$ is an indicator variable that evaluates to 1 if x is true, and to 0 otherwise. This metric is useful if the goal of video augmentation is to find a good candidate set of videos from which the final selection is made by an expert.

Results: Figure 1a shows the performance of our system under the first metric for $k = 1, 2, 3$. The results are quite encouraging. In 77% of the sections, the top video retrieved



(a) $precision@k$ (b) $precision@(1, k)$

Fig. 1: Retrieval precision.

by our system has been judged relevant. The performance is maintained at 73% even when both first and second videos are required to be judged relevant, and at 63% when all three videos are required to be judged relevant. We can also see that the performance is maintained across both the subjects.

Figure 1b shows the results under the second metric for $i = 1$ and $k = 1, 2, 3$. For 77% of the sections, judges agree with our top augmentation. This number goes up to 86% if we are willing to consider it a success if one of the first two videos is judged relevant. It shoots up to 95% if finding at least one out of three videos to be relevant is treated as success.

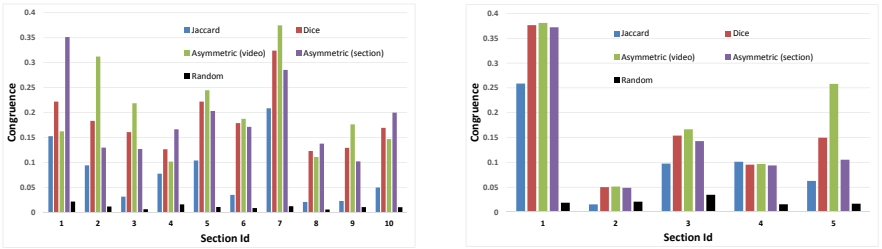
5.4 Congruence

This experiment measures the agreement between judges' collective understanding of the focus of a section and their collective understanding of the focus of the corresponding video. For this purpose, we designed two HITs, one for the section and the other for the video.

Setup: In *SectionHIT* (*VideoHIT*), the judge was asked to read the section (video) and provide top five phrases that best describe the section (video). We converted the phrases from all the judges into unigrams and removed stop words. Let Y_s be the set of unigrams obtained in this manner for section s, and $n_s[w]$ be the number of judges that included unigram w in one of the phrases for s. Similarly, Z_v and $n_v[w]$ for video v.

In this experiment also, judges were required to spend a minimum of 30 minutes on a HIT. The same section (and the corresponding video) was judged by five judges. We selected the judges who took part to be different from those who participated in the experiment reported in §5.3 to remove any biases.

Metric: We compute congruence using several similarity measures [32]. For a video v for section s, the congruence is computed on the sets Z_v and Y_s of unigrams provided by the judges for video v and section s, respectively. We used two symmetric measures: the weighted Jaccard $\left(\frac{\sum_{w \in Z_v \cap Y_s} \min(c_v[w], c_s[w])}{\sum_{w \in Z_v \cup Y_s} \max(c_v[w], c_s[w])}\right)$ and Dice $\left(\frac{2|Z_v \cap Y_s|}{|Z_v|+|Y_s|}\right)$. We also computed asymmetric measures with respect to the section and the video: $|Z_v \cap Y_s|/|Z_v|$ and $|Z_v \cap Y_s|/|Y_s|$ respectively.



(a) Biology textbook

(b) Physics textbook

Fig. 2: Congruence between section focus and retrieved video.

Results: Figure 2 shows the results. For each section (shown in X-axis), we selected the top video identified by our approach and computed congruence (shown in Y-axis) between the section and the corresponding top video. For comparison, we also did the following computation. For each section, we randomly sampled as many unigrams as provided by the judges. Similarly, we also randomly sampled unigrams from the matching videos. We used these two sets to compute average congruence over 100 random runs for each ⟨section, video⟩ pair. We can see that the congruence obtained using the unigrams provided by the judges is significantly higher than that of the randomly sampled unigrams under all the measures.

6 Conclusions

Motivated by the importance of textbooks in learning, we studied the feasibility of enhancing the predominantly text-oriented textbooks with a few selective videos mined from the web at the level of individual sections. We took an approach that does not view textbook sections as stand-alone pieces of text, but rather part of a logically organized work based on well-founded educational principles in which each textbook section contributes uniquely to the pedagogical objective of the book. Our main contributions are as follows:

- Inspired by the theory of Formal Concept Analysis, we propose that the focus of textbook sections can be defined and identified in terms of a small number of indicia, each of which consists of a combination of concept phrases appearing in the section. Indicia of a textbook section are unique relative to all other sections of the book and can be computed by considering all the sections jointly.
- On the video side, we propose making use of the transcript of the spoken words in the audio from the auditory track of the video. However, videos found on the web are independently produced and without necessarily following the organizational logic of textbooks. We therefore use indicia from a section to identify candidate videos and then score them based on the concept phrases present and their importance.
- We evaluated our video augmentation algorithm through extensive user studies of its performance. The video corpus used in the study consisted of nearly 50,000

videos crawled from the web. The textbook corpora consisted of publicly available school textbooks from two different sources, one from U.S.A. and the other from India. This empirical evaluation confirmed the effectiveness of our algorithm in finding relevant videos even at the fine granularity of individual sections of a textbook.

In developing our solution, we built upon work in various disciplines, including educational sciences, natural language and speech processing, knowledge representation and formal concept analysis, information retrieval and extraction, web and data mining, and crowdsourcing. As such, this work might serve as a bridge for researchers belonging to these communities.

For future, we would like to integrate considerations beyond relevance in our video mining system. We expect incorporating viewer aspects, especially appropriateness to viewer's background and prior knowledge, to be particularly valuable and challenging. It is possible for a video to contain not only content relevant for a particular textbook section, but also additional material. In such cases, we would like to be able to pinpoint the subset of the proposed video. The reader would have noticed that the ideas and techniques we have proposed are quite general and have broader applicability. We would like to explore their effectiveness in augmenting textbooks with other types of content that have been investigated in the past [4, 6].

Acknowledgments We wish to thank Sergei Kuznetsov for introducing us to FCA and and providing insightful feedback.

References

[1] *Improving India's Education System through Information Technology*. IBM, 2005.

[2] *Amazon Mechanical Turk, Requester Best Practices Guide*. Amazon Web Services, June 2011.

[3] R. Agrawal, M. Christoforaki, S. Gollapudi, A. Kannan, K. Kenthapadi, and A. Swaminathan. Mining videos from the web for electronic textbooks. Technical Report MSR-TR-2014-5, Microsoft Research, 2014.

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

[5] R. Agrawal, S. Gollapudi, A. Kannan, and K. Kenthapadi. Identifying enrichment candidates in textbooks. In *WWW*, 2011.

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

[7] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast discovery of association rules. In *Advances in knowledge discovery and data mining*, chapter 12. AAAI/MIT Press, 1996.

[8] J. Berger. *Ways of seeing*. Penguin, 2008.

[9] D. Blei, A. Y. Ng, and M. Jordani. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.

[10] P. D. Bruza, D. W. Song, and K.-F. Wong. Aboutness from a commonsense perspective. *Journal of the American Society for Information Science*, 51(12), 2000.

[11] C. Carpineto and G. Romano. *Concept data analysis: Theory and applications*. John Wiley & Sons, 2004.

[12] S. Chakrabarti, M. Van den Berg, and B. Dom. Focused crawling: A new approach to topic-specific web resource discovery. *Computer Networks*, 31(11), 1999.
[13] M. Chambliss and R. Calfee. *Textbooks for Learning: Nurturing Children's Minds*. Wiley-Blackwell, 1998.
[14] J. M. Cigarrán, A. Peñas, J. Gonzalo, and F. Verdejo. Automatic selection of noun phrases as document descriptors in an FCA-based information retrieval system. In *ICFCA*, 2005.
[15] C. L. A. Clarke, N. Craswell, I. Soboroff, and E. M. Voorhees. Overview of the TREC 2011 web track. Technical report, NIST, 2011.
[16] J. Coiro, M. Knobel, C. Lankshear, and D. Leu, editors. *Handbook of research on new literacies*. Lawrence Erlbaum, 2008.
[17] A. Csomai and R. Mihalcea. Linking educational materials to encyclopedic knowledge. In *AIED*, 2007.
[18] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6), 1990.
[19] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *IJCAI*, 2007.
[20] B. Ganter and R. Wille. *Formal concept analysis: Mathematical foundations*. Springer, 1999.
[21] 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.
[22] W. Gray and B. Leary. *What makes a book readable*. University of Chicago Press, 1935.
[23] M. A. Hearst. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1), 1997.
[24] B. Hjørland. Towards a theory of aboutness, subject, topicality, theme, domain, field, content ... and relevance. *Journal of the American Society for Information Science and Technology*, 52(9), 2001.
[25] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank. Semantic-based surveillance video retrieval. *IEEE Transactions on Image Processing*, 16(4), 2007.
[26] S. Huston and W. B. Croft. Evaluating verbose query processing techniques. In *SIGIR*, 2010.
[27] W. J. Hutchins. On the problem of aboutness in document analysis. *Journal of Informatics*, 1(1), 1977.
[28] D. Jurafsky and J. Martin. *Speech and language processing*. Prentice Hall, 2008.
[29] G. Kumaran and V. R. Carvalho. Reducing long queries using query quality predictors. In *SIGIR*, 2009.
[30] S. O. Kuznetsov. On computing the size of a lattice and related decision problems. *Order*, 18(4), 2001.
[31] S. O. Kuznetsov. Complexity of learning in concept lattices from positive and negative examples. *Discrete Applied Mathematics*, 142(1), 2004.
[32] C. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.
[33] S. Mariooryad, A. Kannan, D. Hakkani-Tur, and E. Shriberg. Automatic characterization of speaking styles in educational videos. In *ICASSP*, 2014.
[34] O. Medelyan, D. Milne, C. Legg, and I. Witten. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9), 2009.
[35] R. Mihalcea and A. Csomai. Wikify!: Linking documents to encyclopedic knowledge. In *CIKM*, 2007.
[36] 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.

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

[38] P. Over, G. Awad, J. Fiscus, B. Antonishek, M. Michel, A. Smeaton, W. Kraaij, and G. Qunot. TRECVID 2011 – Goals, tasks, data, evaluation mechanisms and metrics. Technical report, NIST, 2011.

[39] D. Paranjpe. Learning document aboutness from implicit user feedback and document structure. In *CIKM*, 2009.

[40] B. Patel and B. Meshram. Content based video retrieval. *The International Journal of Multimedia & Its Applications (IJMA)*, 4(5), 2012.

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

[42] J. Poelmans, D. I. Ignatov, S. O. Kuznetsov, and G. Dedene. Formal concept analysis in knowledge processing: A survey on models and techniques. *Expert Systems with Applications*, 40(16), 2013.

[43] U. Priss. Formal concept analysis in information science. *Annual Review of Information Science and Technology*, 40, 2006.

[44] C. Shah. TubeKit: A query-based YouTube crawling toolkit. In *JCDL*, 2008.

[45] S. W. Smoliar and H. Zhang. Content based video indexing and retrieval. *IEEE MultiMedia*, 1(2), 1994.

[46] A. Stolcke, B. Chen, H. Franco, V. Gadde, M. Graciarena, M. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lei, et al. Recent innovations in speech-to-text transcription at SRI-ICSI-UW. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5), 2006.

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

[48] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal. Computing iceberg concept lattices with TITANIC. *Data and Knowledge Engineering*, 42(2), 2002.

[49] L. Szathmary, A. Napoli, and P. Valtchev. Towards rare itemset mining. In *ICTAI*, 2007.

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

[51] Y. Tian, L. Cao, Z. Liu, and Z. Zhang. Hierarchical filtered motion for action recognition in crowded videos. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 42(3), 2012.

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

[53] K. Wang, C. Thrasher, E. Viegas, X. Li, and P. Hsu. An overview of Microsoft Web N-gram corpus and applications. In *NAACL–HLT*, 2010.

[54] R. Wille. Formal concept analysis as mathematical theory of concepts and concept hierarchies. In B. Ganter, G. Stumme, and R. Wille, editors, *Formal concept analysis: Foundations and applications*. LNAI 3626, Springer, 2005.

[55] X. Xue, S. Huston, and W. B. Croft. Improving verbose queries using subset distribution. In *CIKM*, 2010.

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

[57] N. Zhang, L.-Y. Duan, L. Li, Q. Huang, J. Du, W. Gao, and L. Guan. A generic approach for systematic analysis of sports videos. *ACM Transactions on Intelligent Systems and Technology*, 3(3), 2012.