Concept Focus

Semantic Meta-Data For Describing MOOC Content

Mesbah, Sepideh; Chen, Guanliang; Torre, Manuel Valle; Bozzon, Alessandro; Lofi, Christoph; Houben, Geert-Jan

DOI
10.1007/978-3-319-98572-5_36

Publication date
2018

Document Version
Final published version

Published in
Lifelong Technology-Enhanced Learning

Citation (APA)
Mesbah, S., Chen, G., Torre, M. V., Bozzon, A., Lofi, C., & Houben, G-J. (2018). Concept Focus: Semantic Meta-Data For Describing MOOC Content. In V. Pammer-Schindler, M. Pérez-Sanagustín, H. Drachsler, R. Elferink, & M. Scheffel (Eds.), *Lifelong Technology-Enhanced Learning: 13th European Conference on Technology Enhanced Learning, EC-TEL 2018 - Proceedings* (Vol. 11082, pp. 467-481). (Lecture Notes in Computer Science; Vol. 11082). Springer. https://doi.org/10.1007/978-3-319-98572-5_36

Concept Focus: Semantic Meta-Data for Describing MOOC Content

Sepideh Mesbah(✉), Guanliang Chen, Manuel Valle Torre, Alessandro Bozzon, Christoph Lofi, and Geert-Jan Houben

Delft University Of Technology, Delft, Netherlands
{s.mesbah,guanliang.chen,m.valletorre,a.bozzon,
c.lofi,g.j.p.m.houben}@tudelft.nl

Abstract. MOOCs promised to herald a new age of open education. However, efficient access to MOOC content is still hard, thus unnecessarily complicating many use cases like efficient re-use of material, or tailored access for life-long learning scenarios. One of the reasons for this lack of accessibility is the shortage of meaningful semantic meta-data describing MOOC content and the resulting learning experience. In this paper, we explore *Concept Focus*, a new type of meta-data for describing a perceptual facet of modern video-based MOOCs, capturing how focused a learning resource is topic-wise, which is often an indicator of clarity and understandability. We provide the theoretical foundations of *Concept Focus* and outline a methodical workflow of how to automatically compute it for MOOC lectures. Furthermore, we show that the learners' consumption behavior is correlated with a MOOC lecture's *Concept Focus*, thus underlining that this type of meta-data is indeed relevant for user-centric querying, personalizing or even designing the MOOC experience. For showing this, we performed an extensive study with real-life MOOCs and 12,849 learners over the duration of three months.

1 Introduction

Reusing and sharing teaching material is considered a central societal challenge by several policy makers. Despite continuously advancing open education policies [25], the vision of easy and personalizable access to open educational resources has still not been realized. To a large extent, this can be attributed to the lack of semantic capabilities of current courseware platforms: with access to only shallow *system-centric* meta-data (e.g. video length, authors names, publication date), these platforms are mostly degraded to be simplistic repositories for storing and serving learning resources. As a result, such platforms are often lacking in usability [29], and rarely take advantage of emerging technologies as for example intelligent digital assistants or conversational interfaces [27]. In this paper, we advocate for the availability of semantic meta-data for educational resources. In contrast to *system-centric* meta-data, *semantic* meta-data – e.g. didactic intent, perceived difficulty, required expertise, or educational quality – describes the

V. Pammer-Schindler et al. (Eds.): EC-TEL 2018, LNCS 11082, pp. 467–481, 2018.
https://doi.org/10.1007/978-3-319-98572-5_36

expected learning experience that a MOOC student might have with a given learning resource. This type of meta-data is generally hard to obtain as it either relies on subjective user-feedback, or needs to be indirectly approximated from the actual learning content. While some standards implicitly, introduce such meta-data types (e.g. LOM [6] – Learning Object Meta-data – covers "semantic density" or "difficulty"), it is usually not specified how such meta-data is defined, nor how it can be obtained from learning resources.

The main goal of this paper is to introduce the notion of **Concept Focus**, a measure of semantic relatedness of all concepts expressed in a learning resource. We set up a large-scale study on 3 MOOCs that engaged more than 12 K learners over the duration of three months. We show that *Concept Focus*, while describing an intrinsic property of the learning resource, is also closely related to learner behavior patterns that are usually associated with difficulty or obstacles in the learning process. This can allow future work to use *Concept Focus* as a lever for learning personalization, e.g. steering certain types of learners towards content with high or low focus based on their personalities and learning goals. In summary, our original contributions include:

- The theoretical foundations for *Concept Focus*, a novel meta-data type capturing a relevant aspect of the learning experience of a MOOC video.
- The design space for methods that automatically obtain *Concept Focus* scores of a given MOOC video in a unsupervised fashion.
- The analysis of 3 real-life MOOC courses featuring 67 videos and 12,849 enrolled learners. We show that *Concept Focus* is a characterizing property of video scripts, describing their topical depth or width. We also report the presence of a significant correlation between *Concept Focus* scores and behavioral patterns indicating learning difficulties, e.g. video watching behaviour, quiz scores, and number of forum questions.

2 Concept Focus: Foundation and Implementation

Educational resources have been described by a multitude of different meta-data types, e.g. the IEEE LOM standard [7] includes a variety of different meta-data types, which can roughly be categorized into 9 groups. Most of these groups describe a learning object from a *system-centric* point of view: for example, general meta-data (e.g., id, title, language), technical aspects (e.g. length or size of videos), life-cycle (e.g., name of authors, version numbers), copyright, and usage restrictions. Only few types of meta-data actually cover the content itself: for instance, LOM group "classification" describes topic and keywords. Only one group of meta-data in LOM ("educational") is dedicated to *learners* and their actual *learning experience*, with information about interactivity, difficulty or semantic density. This is analogous to other educational meta-data standards, as for example Ariadne [9]. Additionally, also bottom up approaches employing folksonomy techniques have emerged [4], with *educational* meta-data related to topical depth and didactic purpose being of central importance there.

This *educational* meta-data has been shown to be very beneficial for personalization and querying (especially data on difficulty, interactivity and density [22]), and its effectiveness even increases when combined with content-related meta-data [1]. Despite this fact, educational meta-data is rarely used in real-life MOOC systems. This can be attributed to the fact that it is expensive to obtain, and usually either expert judgments or crowd-sourcing needs to be employed to this end [22]. Furthermore, in [8], it has been shown that for effective personalization, more semantically deeper types (like learning styles or content properties) are beneficial, as they would allow for more meaningful similarity measurements between learning resources [8] for recommendations and explorative queries. Also Concept Focus could be used to that end, allowing to distinguish broader lectures from topically narrower ones.

2.1 Intuition

We define *Concept Focus* as a measure of semantic relatedness of all concepts expressed in a learning resource (e.g. a recorded lecture, or a script). Intuitively, *Concept Focus* characterizes how strongly a learning resource focuses on a specific topic: Concept Focus is *high* when the concepts of a resource share topical affinity – e.g., a lecture on natural language processing, which discusses a technique like "word embeddings" is implemented, mentioning only related NLP techniques and mathematical concepts.

We will test in our evaluation the hypothesis that learning material covering different topics, possibly loosely related, lead to learning difficulties. Even in cases where low *Concept Focus* does not always lead to confusion and learning problems (as it might also characterize material giving summaries or overviews), we argue that it is in either case a valuable meta-data field to be considered by an educational personalizing information system, as we will show, it drives behaviours similar to the ones of meta-data that are harder to obtain, as for example clarity or difficulty. *Concept Focus* can be computed automatically by relying on a combination of NLP and information extraction techniques, thus overcoming the aforementioned limitation of prohibitively high costs of crowdsourcing or expert feedback. In short, *Concept Focus* can be realized as follows:

1. Extract all concepts (i.e., filtered named entities) from the textual representation of a given learning object.
2. Measure the *Semantic Relatedness* of a given concept in the learning resource, w.r.t. all other concepts in the same resource.
3. Calculate the *Concept Focus* of a resource, as a function of the semantic relatedness of all the concepts therein contained. Intuitively, if all concepts are semantically closely related, the *Concept Focus* focus of the resource is high; or low, otherwise.

2.2 Concept Extraction

In the following, we discuss how to extract concepts from videos, or more precisely the textual scripts of lecture videos. Arguably, the most important

educational material in MOOCs are the videos, as they are the principal mean for content delivery.

They are therefore our main object of analysis. Due to their interactivity, videos have the additional benefit of enabling in-video interaction analysis (i.e. users click actions such as pauses, replaying, etc.) to observe and assess the learning status of the students (e.g. difficulty in understanding the content) [15]. We exploit this fact in our evaluation.

Formally, a concept c can be defined as a k-gram that represents ideas and entities expressed in the video transcript text (e.g. "machine learning", "stock price index") [24]. Automatic concept extraction from text has received much attention in the past decade [3,18–20,26], and thus there exist a number of publicly available concept extractor tools, relying on techniques such as term-frequency analysis [26], co-occurrence graph [20], etc. Extracting concepts from MOOCs content is, however, a challenging task due to the low-frequency problem [23]: MOOCs videos are relatively short documents and due to the small number of words, statistical techniques (e.g. co-occurence) are not applicable. To cater for such limitations, we employ an ensemble approach, running a battery of concept extractor tools on a video's script, and extracting all the concepts contained in it. We adopt:

- **TF-IDF**[1]: A well-know Information Retrieval technique, used to rank candidate concepts based on their tf-idf (term frequency - inverse document frequency) in the corpus.
- **TextRank** [20]: A technique that extracts concepts by ranking them according to their co-occurrence graph.
- **TopicRank** [3]: An extension of Textrank. A graph-based concept extraction approach which relies on a topical representation of the text.
- **KPMiner** [10]: A simple technique, which employs a set of heuristic rules (e.g. length of the concept, position in the sentence) to extract concepts from the text.
- **Rake** [26]: Rapid Automatic Keyword Extraction is able to identify concepts by relying on the term frequency, term degree, and ratio of degree to frequency.
- **TextRazor**[2]: A text analysis API that returns detected entities, possibly decorated with links to the DBpedia or Freebase knowledge bases.

As a next step, we merge all the concepts individually extracted from each tool, filtering stopwords (e.g. something, anything, etc.) and concepts coming from "common" English language (e.g., "events", "data") that could be found in Wordnet. We retain only concepts that have been detected by the majority of the extractor tools (i.e. 4 out of 6) to filter out irrelevant concepts (e.g. "six months", "new stories"). Intuitively, a concept will be considered as a correct concept if it has been harvested by different combinations of concept extraction tools [5]. By merging all concepts extracted from a given video scripts v, we obtain a final list of Candidate Concepts $concepts(v) = \{c_1, ..., c_N\}$.

[1] http://www.hlt.utdallas.edu/~saidul/code.html.
[2] https://www.textrazor.com/.

2.3 Concept Focus

Concept Focus relies on measuring and aggregating the semantic relatedness of concepts contained in a lecture transcript: the higher the semantic relatedness between all concepts, the higher the focus of the lecture. While there can be many implementations for capturing semantic relatedness, previous studies [17] have shown that word embeddings [21] perform this task particularly well by e.g. measuring the cosine similarity of the word embedding vectors. We exploit Wikipedia to learn the word embedding representation of each concept. We first extract English articles from the latest publicly available Wikipedia dump[3]. Next, we built an embedding lexicon based on *fastText* [2]. *FastText* embeds each term (uni-gram and bi-gram) of a large document corpus into low-dimensional vector space (100 dimensions in our case) and overcomes the problem of out-of-vocabulary words by representing each word as a bag of character n-grams.

We adopt a typical measure of semantic relatedness $SR(c_1, c_2)$, that is computed between two specific concepts c_1 and c_2 by measuring the cosine similarity of their word embedding vectors [17].

In addition, we now also introduce the semantic relatedness $SR(c, v)$ between a concept c and all other concepts contained in a video transcript v. We also value the relatedness to the title of a video. For instance in a video v about "Propensity score matching"[4], concepts such as *propensity score*, *p-value* and *paired t-test* will get a higher semantic relatedness measure with respect to v, while a concept like *heart catheterization* is less related within v.

We define $SR(c, v)$ for a concept c and a MOOC video transcript v as follows:

$$SR(c, v) = \frac{\sum_{cv \in concepts(v)} SR(c, cv) * SR(c, titleOf(v))}{|concepts(v)|} \quad (1)$$

SR is a value in $[0, 1]$, where 1 represents the maximum relatedness that a concept can have in a video.

Consequently, the Concept Focus of a given lecture video v can be defined as the average concept relatedness of each concept in v within the context of v, i.e.:

$$CF(v) = \frac{\sum_{c \in concepts(v)} SR(c, v)}{|concepts(v)|} \quad (2)$$

CF is also in $[0, 1]$, where 1 is the highest *Concept Focus* value.

3 Evaluation

This section reports the results of an extensive study on real-life MOOCs, to showcase and discuss our new *Concept Focus* meta-data. We organize the study around the following research questions:

[3] https://dumps.wikimedia.org/enwiki/20180201/.

[4] https://www.coursera.org/learn/crash-course-in-causality/lecture/VtFdu/propensity-score-matching-in-r.

- **RQ1**: To what extent do properties of video scripts affect a course's *Concept Focus*? We investigate properties of learning material like video length, number of concepts, and position of the course in the MOOC.
- **RQ2**: To what extent does *Concept Focus* affect students' learning behaviour? We investigate the learners video watching behavior, quiz performance, and discussion behavior in relation to the Concept Focus of their consumed learning material.

3.1 Dataset Description

We analyze the log traces of learners collected from three MOOCs in edX[5]: itemize **DA** Data Analysis: Visualization and Dashboard Design, **IWC** Introduction to Water and Climate, **IWT** Introduction to Water Treatment.

We selected these 3 MOOCs for the following reasons: (1) they feature comparable amount of videos, and engaged students; (2) they cover a variety of topics; and (3) the scripts of their videos, and the interaction data for the engaged students are available. Table 1 summarizes the main properties of the selected MOOCs. We consider only *engaged* learners, i.e. learners that watched at least one video for more than 15 s. Interaction data is collected through click log traces. We analyzed in total 9,899,369 log trace records of 12,849 learners. Statistics of the MOOC and learners are summarized in Table 1.

Table 1. Overview of the three MOOC datasets analyzed. Legend: REG – Registered; Eng – Engaged; CR – Completion Rate

ID	Name	Start	End	Videos	# Learners		
					REG	ENG	CR
DA	*Data Analysis*	03/2016	06/2016	22	32,682	5,711	3.74%
IWC	*Introduction Water and Climate*	09/2014	11/2014	27	9,267	4,947	2.60%
IWT	*Introduction Water Treatment*	01/2016	03/2016	18	13,198	2,191	3.07%

Properties of MOOCs. To answer **RQ1**, we study the relation between the following features of videos in a MOOC, and their *Concept Focus*:

- `VD` – *Video Duration*: the length of a video, expressed in seconds.
- `VL` – *Average Video Length*: the average number of words in the video scripts of the given MOOC.
- `ANC` – *Average Number of Concepts*: the average number of concepts extracted from the video scripts of the given MOOC.
- `SC` – *Session of the Course*: the date the lecture was given (i.e. first session, second session, etc.)

Learners Behaviour. To address **RQ2**, we study the relationship between the measured behaviour of learners, and the *Concept Focus* score of videos. From

[5] https://www.edx.org/.

the log traces, we extracted the following 7 features. Each feature is calculated by aggregating all learner activities, including activities in the video player and in the course's forum, and their proficiency with the subject as assessed by the MOOC's grading system.

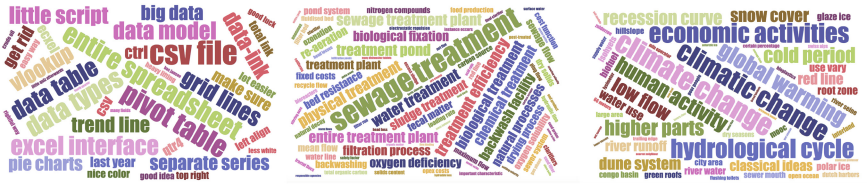
- `WT` – *Watching Time* of video material: the amount of time a learner has spent watching a video's material in the MOOC.
- `NWT` – *Normalized Watching Time* of video material: the total amount of time a learner has spent watching video material in the MOOC divided by the duration of the video.
- `FS` – *# Forward Seek*: the total number of times a learner seeks forward while watching a video.
- `BS` – *# Backward Seek*: the total number of times a learner seeks backward while watching a video.
- `SU` – *# Speed Up*: the total number of times a learner increases the play speed while watching a video.
- `SD` – *# Speed Down*: the total number of times a learner decrease the play speed while watching a video.
- `FG` – *Final Grade*: the percentage of quiz questions the learner. answered correctly after having interaction with a video.
- `NFP` – *# New Forum Posts*: the number of new forum posts (i.e., questions) created by the learner after having interaction with a video. Here we consider posts created within 15 min from the last interaction with a video.

3.2 RQ1: Video Properties Vs. Concept Focus

Table 2 summarizes the properties of the video scripts part of our analysis, including the number of unique concepts extracted from the MOOCs, the average, median and standard deviation number of concepts extracted from their videos, as well as the length of the videos in terms of the number of words. Here we consider extracted concepts that were also present in Wikipedia, and for which a vector representation exists. Notably, 98% of the candidate concepts extracted from the concept extraction phase have a vector representation in our corpus. Figure 1 shows samples of extracted concepts organized in word clouds, where the size of the concept is proportional to their Semantic Relatedness (SR) score.

`DA` videos, compared to `IWC` and `IWT`, feature on average 60% less concepts, and half the number of words per video. The standard deviation is proportionally higher, thus showing more variability within the course. Figure 2 shows the distribution of the *Concept Focus* for all the videos of the three MOOCs. The average *Concept Focus* for the courses are respectively 0.29 for `DA`, 0.26 for `IWT`, and 0.19 for `IWC`. An example of `IWC` video with low focus score ($CF = 0.16$) is the lecture "Urban Engineering"[6], which includes a rather diverse concepts such as "cloaca maxima", "city wall", or "permeable pavements". The lecture

[6] https://www.youtube.com/watch?v=nhMcB-bwSF0.



(a) DA (b) IWT (c) IWC

Fig. 1. Extracted concepts from video scripts of `IWC`, `IWT` and `DA`.

belongs to introductory course on Water Climate, a subject that is bound to embrace several topics. The "Solver" lecture in the `DA` course[7] is an example of very focused video ($CF = 0.36$), including concepts such as "data table", "excel sheet", or "spreadsheet". This is also expected, as the lecture is exclusively about an Excel plug-in program called "Solver".

Figure 3 shows the relation between the length of the video (in terms of words) and the *Concept Focus* for each MOOC. Intuitively, one would argue that the longer the text of the video script, the higher the number of concepts contained in it, thus the lower Concept Focus. Indeed, this is not necessarily the case. We can find a moderate significant positive correlation only for videos in the `IWC` course (Fig. 3c: $\rho = -0.59$, $p - value : 0.0069$). However, as shown in Fig. 4, videos with higher number of concepts do have lower concepts focus, but only for the `DA` course a moderate significant negative correlation could be found (Fig. 4a: $\rho = -0.60$, $p - value : 0.01$). These results show that *Concept Focus* is a lecture-specific property that is not biased by the length of a video or by the sheer number of concepts contained in it. Arguably, this is a desirable properties for a content-centric meta-data.

Table 2. Descriptive statistics for concepts C and number of words W of the analyzed MOOCs video scripts. Legend: UC, Unique Concepts; μ, average; m, median; σ, std.

MOOC ID	UC	μC	mC	σC	μW	mW	σW
DA	298	17	16	6	680	624	262
IWT	687	49	46	12	1268	1303	365
IWC	1095	43	43	7	1481	1398	366

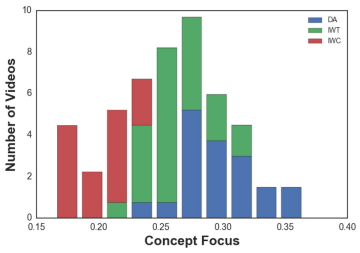
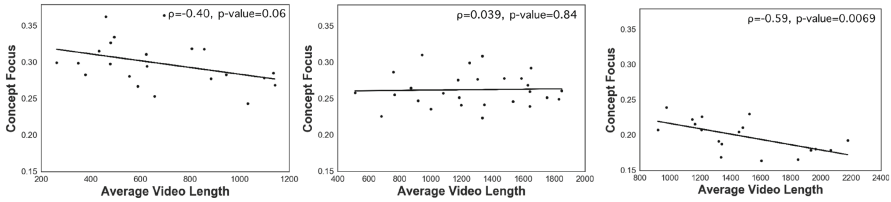


Fig. 2. Distribution of *Concept Focus* for the videos of *IWC*, *IWT* and *DA* in the shape of a stacked histogram

Finally, we study if the position of a video in a MOOC can be related to *Concept Focus*. Courses might feature different progression and organization of

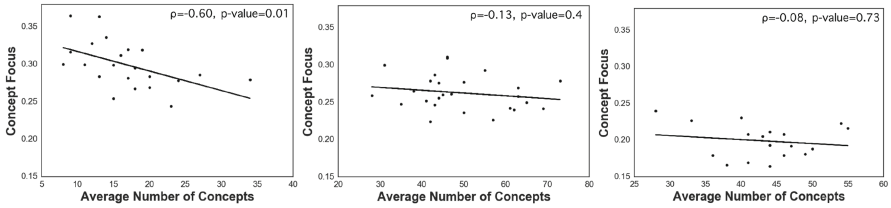
[7] https://www.youtube.com/watch?v=DgYmpmwBybQ.

subject, with introductory lecture in the beginning (low *Concept Focus*) and specialized lectures later on (high *Concept Focus*). As shown in Fig. 5, the three courses feature very different teaching profiles. Despite the lack of statistically significant relation with *Concept Focus*, we can see how `DA`, for instance, starts with two very focused videos while, over time, lectures show consistent variations of *Concept Focus* scores. In `IWT` and `IWC`, on the other hand, the first lecture has low *Concept Focus*, and there is less variations in score across lectures, roughly remaining the same.



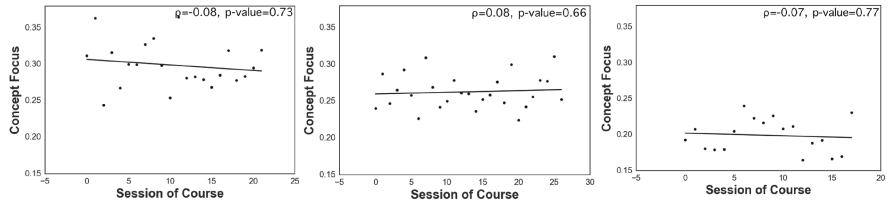
(a) DA $coeff = -4.645$, $intercept = 0.330$ (b) IWT $coeff = 2.112$, $intercept = 0.2604$ (c) IWC $coeff = -3.763$, $intercept = 0.2549$

Fig. 3. *Concept Focus* and the number of words in the video transcripts



(a) DA $coeff = -0.0026$, $intercept = 0.3430$ (b) WT $coeff = -0.0003$, $intercept = 0.28$ (c) IWC $coeff = -0.0005$, $intercept = 0.22$

Fig. 4. *Concept Focus* and the average number of concepts in video transcripts



(a) DA $coeff = -0.0007$, $intercept = 0.307$ (b) IWT $coeff = -0.00022$, $intercept = 0.2602$ (c) IWC $coeff = -0.0003$, $intercept = 0.2023$

Fig. 5. *Concept Focus* and the position of the related video in the MOOC.

3.3 RQ2: Learning Behaviour Vs. Concept Focus

We first study how the length of a video is related to the behaviour of learners, Fig. 6 summarizes the Spearman correlation between all measures as a heatmap. The *Video Duration* `VD` is obviously highly correlated with the learners *Watching Time* `VT`. The longer learners spends time watching videos, the higher the amount of video interactions such as `FS` (*# Forward Seek*), `SU` (*# Speed Up*) and `SD` (*# Speed Down*). We believe that the high `WT` is not associated with learning difficulty, as we observe a negative correlation between `WT` and `BS`, and positive correlation with `SD` which are indicators of higher level of difficulty [16].

Table 3. Spearman correlation ρ between *Concept Focus* and learners behavioural features for all the videos in the dataset. $^{*}p - value < 0.05, ^{**}p - value < 0.001$

	ρ
`NWT` – # Normalized Watching Time	0.44**
`FS` – # Forward Seek	0.31*
`BS` – # Backward Seek	0.50**
`SU` – # Speed Up	−0.36**
`SD` – # Speed Down	−0.55**
`FG` –Final Grade	0.19
`NFP` – # New Forum Posts	−0.25*

Table 3 reports the measured Spearman correlation between the *learners behaviour* metrics and *Concept Focus* of the corresponding videos. *Concept Focus* is significantly correlated with `NWT`, `BS`, `SU`, `SD`, and `NFP`. We observe a moderate positive correlation between the amount of time learners spent watching video lectures and the number of times they seek backward - i.e., in the videos with higher Concept Focus, learners watch the video for a longer time and are more likely to re-watch parts of them. This observation aligns with the previous study [28] were the authors showed that difficulty correlates negatively with dwelling time (i.e. time students spend watching a video). We interpret this result as a sign of students disengaging with videos having lower focus i.e. that cover a wider range of concepts. A similar result can be found in [12] where it has been shown that many

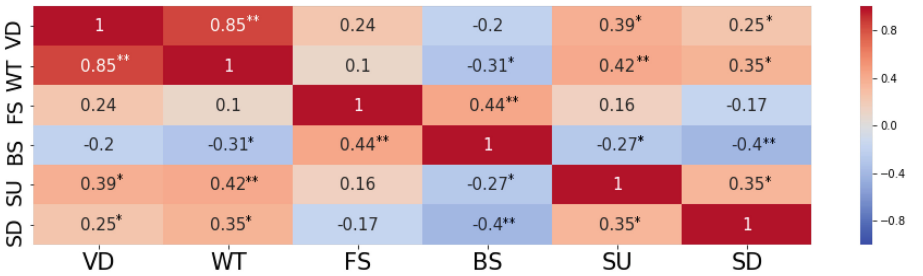


Fig. 6. Correlation heatmap of video interaction. Legend: *VD* – Video Duration; *WT* – Watching Time; *FS* – Forward Seeks; *BS* – Backward Seeks; *SU* – Speed Ups; *SD* – Speed Downs. $^{*}p - value < 0.05$, $^{**}p - value < 0.001$

students stop engaging with a courses (e.g. watching the videos) when they haven't enough knowledge to understand the context.

We also observe a weak negative correlation with the number of new forum post - i.e., after watching videos with lower Concept Focus, learners are more likely to post in the forum. This can be an indicator of having difficulty understanding the concepts in video scripts with low focus. The number of times the learner speed up and down the video have also a significant moderate negative correlation with the Concept Focus - i.e., in the videos with higher Concept Focus, learners continue watching the video without changing the speed of the video, possibly a sign of well-designed content progression. Finally we do not observe any statistically significant correlations between the final grade of the students and the Concept Focus.

The box plots in Fig. 7 depict the break down of the distribution of final grade, normalized watching time, # of new forum post, # of forward seek, # backward seek, # speed up and # speed down of three courses. In order to check if the samples are drawn from different population groups we performed the Kruskal-Wallis H Test (KWHT). In `DA`, where average *Concept Focus* is higher (0.29) than `IWT` (0.26) and `IWC` (0.19), the learners achieve a slightly higher grade (KWHT $statistic = 5.99$, $pvalue = 0.049$); a statistically significant higher normalized watching time (KWHT $statistic = 26.73$, $p - value = 1.56e - 06$), forward seek (KWHT $statistic = 10.49$, $pvalue = 0.005$) and back ward seek (KWHT $statistic = 17.31$, $pvalue = 0.0001$); and slightly lower number of speed up (KWHT $statistic = 9.94$, $pvalue = 0.006$) and speed down (KWHT $statistic = 1.35e - 05$, $pvalue = 22.42e - 05$). The difference in the distribution of number of new forum posts is not statistically significant (KWHT $statistic = 5.16$, $pvalue = 0.07$).

Altogether, these results show that *Concept Focus* is indeed a measure that relates to user-centric properties of videos, giving insights into potential engagement of learners, types of content, or potential learning problems.

4 Related Work

A growing body of literature has examined different attributes (e.g. video length [11], interface characteristics [13], video textual complexity [28], displaying the instructor's face to video instruction [14]) of MOOC videos and their effect on learners' dwelling time [15,16,28] or dropout [11].

Recently, several studies focused on the in-video interactions analysis (e.g. measuring the number of pauses, skipping, re-watching) to measure the level of the perceived video difficulty [15,16] and to model students learning behaviour [28]. The existing research capitalize on the relationship between the user and the content to measure the perceived video difficulty. We still have a limited understanding about the intrinsic properties of the text (i.e. without the interpretation of the users) that make a MOOC video clear for the students. Our work is inspired by [28], where the researchers focused on the textual analysis (e.g. word and sentence length, frequency of words, etc.) of the video scripts

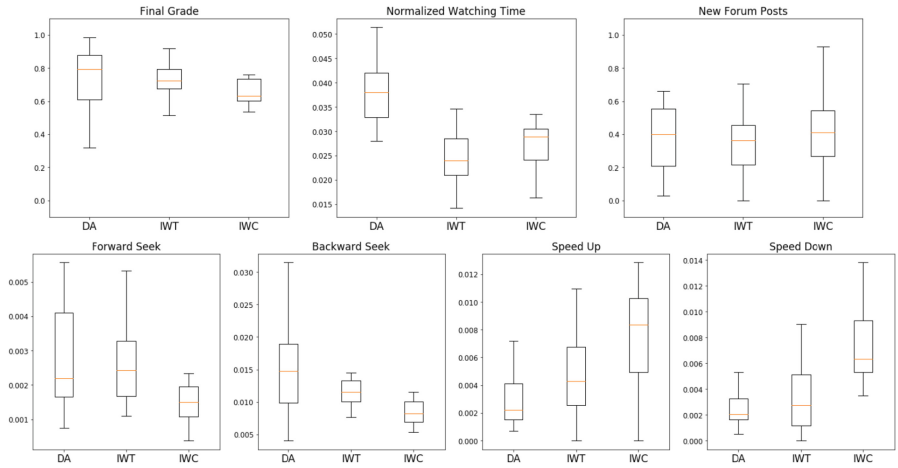


Fig. 7. Distribution of Final Grade (`FG`), Normalized Watching Time (`NWT`), # New Forum Posts (`NFP`), # Forward Seek (`FS`), # Backward Seek (`BS`), # Speed Up (`SU`) and # Speed Down (`SD`) for the three courses.

and showed the effect of video complexity on the users video interaction (i.e. dwelling time and rate of the learners). However, the properties of the concepts (i.e. k-grams that represent ideas and entities expressed in the text such as: machine learning, stock price index, etc.) used in the text and the semantic relation between them are not well understood to characterize the lecture clarity and understandability. Thus, in this paper we focus on analyzing the content of MOOC videos to obtain their concept focus topic-wise, which is often an indicator of clarity and understandability of a lecture.

5 Conclusion

In this paper, we introduced *Concept Focus*, a novel type of meta-data capturing an aspect of a user's learning experience when interacting with learning content in an online MOOC platform. *Concept Focus* describes how focused a learning resource is w.r.t. a restricted set of topics. It can be used to semantically characterize a learning resource (as for example an in-depth explanations vs. a general overview), but might also be an indicator for potential learning challenges. In contrast to other meta-data types, we show that *Concept Focus* can be computed fully automatically by relying on a combination of natural language processing and information extraction techniques, thus avoiding the common detriment of having to rely on costly crowd-sourcing or experts. We believe Concept Focus can play a role as part of the feature set of more elaborate methods for automatically deriving meta-data on teaching methods or learning styles.

We conducted an extensive study covering three real-life MOOCs with 67 videos on the edX MOOC platform. We show that *Concept Focus* is a property that does not depend on video length, it is lecture-specific, and it characterizes the organization of a MOOC. By analyzing the activity logs of 12,849 learners, we investigated their video watching behavior, quiz performance, and discussion behavior in relation to the concept focus of their consumed learning material. Furthermore, we investigated properties of learning material like video length or number of contained concepts. The analysis indicates a correlation between low *Concept Focus*, and behaviors which are associated with learning difficulties.

While these results are supported by general intuition and previous findings, our study is limited to three MOOCs. Additional studies are therefore necessary to better understand the relationship between this novel meta-data, and behavioural properties of learners.

References

1. Abdelali, S., et al.: Education data mining: Mining MOOCS videos using metadata based approach. In: Information Science and Technology (CiSt), pp. 531–534. IEEE (2016)
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
3. Bougouin, A., Boudin, F., Daille, B.: Topicrank: Graph-based topic ranking for keyphrase extraction. In: International Joint Conference on Natural Language Processing (IJCNLP), pp. 543–551 (2013)
4. Catarino, M.E., Baptista, A.A.: Relating folksonomies with dublin core. In: Dublin Core Conference, pp. 14–22 (2008)
5. Chen, L., Ortona, S., Orsi, G., Benedikt, M.: Aggregating semantic annotators. Proc. VLDB Endow. **6**(13), 1486–1497 (2013)
6. Learning Technology Standards Committee. IEEE Standard for learning object metadata. IEEE Standard, 1484(1), 2007-04 (2002)
7. Consortium, I.G.L.: Learning resource meta-data specification (2002). https://www.imsglobal.org/metadata/index.html. Accessed 26 Feb 2018
8. Dorça, F.A., Carvalho, V.C., Mendes, M.M., Araújo, R.D., Ferreira, H.N., Cattelan, R.G.: An approach for automatic and dynamic analysis of learning objects repositories through ontologies and data mining techniques for supporting personalized recommendation of content in adaptive and intelligent educational systems. In: Advanced Learning Technologies (ICALT), pp. 514–516. IEEE (2017)
9. Duval, E., Vervaet, E., Verhoeven, B., Hendrikx, K., Cardinaels, K., Olivié, H., Forte, E., Haenni, F., Warkentyne, K., Forte, M.W., et al.: Managing digital educational resources with the ariadne metadata system. J. Internet Cat. **3**(2–3), 145–171 (2000)
10. El-Beltagy, S.R., Rafea, A.: KP-miner: A keyphrase extraction system for english and arabic documents. Inf. Syst. **34**(1), 132–144 (2009)
11. Guo, P.J., Kim, J., Rubin, R.: How video production affects student engagement: An empirical study of mooc videos. In: ACM Conference on Learning @ Scale Conference, L@S 2014. pp. 41–50. ACM, New York, NY, USA (2014)

12. Khalil, H., Ebner, M.: MOOCS completion rates and possible methods to improve retention-a literature review. In: EdMedia: World Conference on Educational Media and Technology, pp. 1305–1313. Association for the Advancement of Computing in Education (AACE) (2014)
13. Kim, J., Guo, P.J., Seaton, D.T., Mitros, P., Gajos, K.Z., Miller, R.C.: Understanding in-video dropouts and interaction peaks inonline lecture videos. In: Conference on Learning@ Scale Conference, pp. 31–40. ACM (2014)
14. Kizilcec, R.F., Papadopoulos, K., Sritanyaratana, L.: Showing face in video instruction: effects on information retention, visual attention, and affect. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 2095–2102. ACM (2014)
15. Li, N., Kidzinski, L., Jermann, P., Dillenbourg, P.: How do in-video interactions reflect perceived video difficulty? In: European MOOCs Stakeholder Summit, No. EPFL-CONF-207968, pp. 112–121. PAU Education (2015)
16. Li, N., Kidziński, Ł., Jermann, P., Dillenbourg, P.: MOOC video interaction patterns: what do they tell us? In: Conole, G., Klobučar, T., Rensing, C., Konert, J., Lavoué, É. (eds.) EC-TEL 2015. LNCS, vol. 9307, pp. 197–210. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24258-3_15
17. Lofi, C.: Measuring semantic similarity and relatedness with distributional and knowledge-based approaches. Database Soc. Japan **14**(3), 1–9 (2016)
18. Mesbah, S., Fragkeskos, K., Lofi, C., Bozzon, A., Houben, G.-J.: Semantic annotation of data processing pipelines in scientific publications. In: Blomqvist, E., Maynard, D., Gangemi, A., Hoekstra, R., Hitzler, P., Hartig, O. (eds.) ESWC 2017, Part I. LNCS, vol. 10249, pp. 321–336. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58068-5_20
19. Mesbah, S., Lofi, C., Valle Torre, M., Bozzon, A., Houben, G.J.: TSE-NER: an iterative approach for long-tailentity extraction in scientific publications. In: International Semantic Web Conference. Springer (2018)
20. Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: Conference on Empirical Methods in Natural Language Processing (2004)
21. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
22. Miranda, S., Ritrovato, P.: Automatic extraction of metadata from learning objects. In: Intelligent Networking and Collaborative Systems (INCoS), pp. 704–709. IEEE (2014)
23. Pan, L., Wang, X., Li, C., Li, J., Tang, J.: Course concept extraction in MOOCS via embedding-based graph propagation. In: International Joint Conference on Natural Language Processing, vol. 1, pp. 875–884 (2017)
24. Parameswaran, A., Garcia-Molina, H., Rajaraman, A.: Towards the web of concepts: extracting concepts from large datasets. VLDB Endow. **3**(1–2), 566–577 (2010)
25. Parliament, E.: Opening up education: Innovative teaching and learning for all through new technologies and open educational resources. Communication from the commission to the European Parliament (2013)
26. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. In: Berry, M.W., Kogan, J. (eds.) Text Mining: Applications and Theory, pp. 1–20. Wiley, Hoboken (2010)
27. Sarikaya, R.: The technology behind personal digital assistants: an overview of the system architecture and key components. IEEE Signal Process. Mag. **34**(1), 67–81 (2017)

28. Van der Sluis, F., Ginn, J., Van der Zee, T.: Explaining student behavior at scale: the influence of video complexity on student dwelling time. In: ACM Conference on Learning @ Scale, L@S 2016, pp. 51–60. ACM, New York, NY, USA (2016)
29. Tsironis, A., Katsanos, C., Xenos, M.: Comparative usability evaluation of three popular MOOC platforms. In: 2016 IEEE Global Engineering Education Conference (EDUCON), pp. 608–612. IEEE (2016)