PEEK: A Large Dataset of Learner Engagement with Educational Videos

Sahan Bulathwela[1], María Pérez-Ortiz[1], Erik Novak[2], Emine Yilmaz[1] and John Shawe-Taylor[1]

m.bulathwela@ucl.ac.uk

[1] Centre for Artificial Intelligence, University College London (UK)

[2] Jožef Stefan Institute, Jožef Stefan International Postgraduate School (Slovenia)

ABSTRACT

In this work, we release a large and novel dataset of learners engaging with educational videos in-the-wild. The dataset, named Personalised Educational Engagement with Knowledge Topics (*PEEK*), is one of the first publicly available datasets that address personalised educational engagement. Educational recommenders have received much less attention in comparison to e-commerce and entertainment-related recommenders, even though efficient personalised learning systems could improve learning gains significantly. One of the main challenges in advancing this research direction is the scarcity of large, publicly available datasets. In the PEEK dataset, educational video lectures have been associated with Wikipedia concepts related to the material of the lecture, thus providing a humanly intuitive taxonomy. We believe that granular learner engagement signals, in unison with rich content representations, will pave the way to building powerful personalisation algorithms that will revolutionise educational and informational recommendation systems. Towards this goal, we 1) construct a novel dataset from a popular video lecture repository, 2) identify a set of benchmark algorithms to model engagement, and 3) run extensive experimentation on the PEEK dataset to demonstrate its value. Our experiments with the dataset show promise in building powerful informational recommender systems. The dataset and the support code is available at https://github.com/sahanbull/PEEK-Dataset.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; **Personalization**; **Recommender systems**; Information extraction; • **Applied computing → Interactive learning environments**.

Reference Format:
Sahan Bulathwela[1], María Pérez-Ortiz[1], Erik Novak[2], Emine Yilmaz[1] and John Shawe-Taylor[1]. 2021. PEEK: A Large Dataset of Learner Engagement with Educational Videos. In *4th Workshop on Online Recommender Systems and User Modeling (ORSUM 2021), in conjunction with the 15th ACM Conference on Recommender Systems, October 2nd, 2021, Amsterdam, The Netherlands (Remote)*.

1 INTRODUCTION

Developing artificial intelligence systems that, mildly at least, understand the structure of knowledge is foundational to building effective recommendation systems for lifelong education[13, 41, 62], as well as for many other applications related to knowledge management and tracing[35, 61]. In the context of personalised education,

the research communities have tirelessly worked on building Intelligent Tutoring Systems (ITS) which have been their focus from the early applications of AI in education[19]. ITS focus heavily on personalising testing opportunities that will allow learners to demonstrate their mastery of a Knowledge Component (KC), an atomic unit of knowledge that can be learned and mastered. These systems are designed for learning experiences with limited scope (such as learning about a specific topic, short course, etc.) which makes hand crafting KCs and encouraging repetitive test taking operationally feasible. More recent developments in online education has led to the boom of Open Educational Resources (OERs) and Massively Open Online Courses (MOOCs) moves from ITS to educational recommendation systems and personalised e-learning platforms that can cater a more informal learning setting where self motivated learners discover educational materials online to pursue their lifelong learning goals. To succeed in this landscape, a wider range of factors such as the diversity of learners, their personal drivers and how these drivers change over time has to be accounted for. One of the major barriers in building next generation educational recommendation systems is the scarcity of publicly available datasets.

1.1 Our Contribution

We publish *(P)ersonalised (E)ducational (E)ngagement Linked to (K)nowledge Topics* (PEEK), a novel dataset that comprises of watch time interactions of over 20,000 informal learners watching over 10,200 unique educational video lectures in an OER repository. Our contribution is two-fold, consisting of: i) constructing a novel dataset to predict learner engagement with educational videos and ii) formalising a prediction task with several baselines. The video lectures are partitioned into multiple fragments where the parts are transcribed and associated with Wikipedia topics using entity linking. PEEK dataset uses a humanly intuitive content representation where the topics are atomic Wikipedia pages. Having humanly intuitive content representations allows building explainable learner models that encourage learner meta-cognition and self-regulation, a valuable feature of a technology enhanced learning system[16]. The normalised watch time of individual users is discretised and provided as a target label defining a binary (engaged vs. not engaged) prediction task. We further identify several baselines from prior work and establish a formal task to benchmark predictive performance on the proposed dataset.

2 RELATED WORK

Publication of this dataset is inspired by advances in multiple research verticals. PEEK dataset is an educational recommendation dataset that contains learner interactions with fragments of video

lectures. We outline the relevant works from these different research verticals in this section.

2.1 Intelligent Tutoring Systems and Educational Recommenders

Knowledge Tracing (KT) and *Item Response Theory* (IRT), the foundational methods used in Intelligent Tutoring Systems, are evolving from traditional machine learning[19, 42, 43, 54, 63] to deep learning models[33, 46] that improve predictive performance by sacrificing interpretability and transparency, crucial requirements for many of these systems[2, 15]. It is also hard to overlook the data hunger of neural approaches considering the possible data scarcity in an educational setting where additional information should be exploited. While accompanying these additional drawbacks, deep learning models used in these tasks are still not guaranteed to outperform classical approaches[59].

Another major drawback of ITS research is its primary focus into predicting knowledge mastery via explicit learner feedback that comes through test taking. Although test taking is a reliable approach to verify learning, this approach comes with its fair share of limitations. Tests need to be carefully crafted by experts in a way that the skill mastery of learners on specific skills can be carefully verified with accompanying hints[18, 25, 26] costing substantial human capital. Scaling automatic question generation is an open challenge. The other stakeholder segment put into pressure by question solving is the learners themselves. Putting the learner in front of tests too frequently can significantly hinder the user experience driving them to abandon the system.

Beyond the use of costly explicit user interactions, there are various other implicit interactions learners execute within an interactive e-learning system that can be potentially exploited. Learner engagement with educational materials is also indicated by implicit interactions such as watching/pausing/rewinding/replaying video lecture, taking part in discussion fora etc.[48]. In the context of videos, although the potential of using watch time for engagement prediction[60] and personalisation[20] has led to positive results, this idea is under-explored in the education domain with only a handful of studies showing promise[9, 29] in population-based engagement prediction. Although there is a strong interest in building state-aware, personalised educational systems in the recent years[10, 11], an obvious reason for slow growth of personalised educational recommenders is the scarcity of publicly available datasets.

Unfortunately, there is very limited work in the Information Retrieval (IR) community that deals with a personalisation scenario similar to PEEK dataset as IR scenarios in education are significantly different from general click behaviours for which datasets are abundantly available. TrueLearn Novel model[14] is the only recent algorithmic contribution that deals with a similar dataset where topical contents automatically extracted from educational videos are used to predict learner engagement. Apart from sophisticated learner modelling techniques such as TrueLearn Novel, other basic recommendation strategies, such as content-based filtering and collaborative filtering, are also applicable to this dataset[34].

2.2 Video Fragments

Although majority of studies in the field focus on recommending content items that are relevant to a learner, the possibility of recommending parts of items (e.g. a part of a video in contrast to an entire video) has been investigated lately. Proximity-aware information retrieval that exploits the positional structure of tokens is not a new idea[50] in IR. Segmenting videos and building a table of contents using video segments has proven to be useful in efficiently summarising video content[28, 38]. Breaking informational videos into fragments has also shown promise in efficient previewing[8, 17, 44] and enabling non-linear consumption of videos[53]. Our prior proposal, TrueLearn Novel[14] model demonstrates the potential of using fragment-wise recommendation in educational recommenders. For these experiments, we created video parts of 5,000 characters (5 minutes) in the fragmentation process. This allows the e-learning system to have video fragments that contain a satisfactory amount of knowledge while keeping the video fragment length at a favourable value in terms of retaining viewer engagement[29]. While TrueLearn Novel demonstrates promise, we strongly believe that availability of such a dataset to the public is critical in pushing the frontiers of research in (fragment-based) educational recommendation. PEEK dataset addresses this need.

2.3 Scalable Feature Extraction

In addition to manual question generation, systems using KT and IRT often relies on expert labelling of the **Knowledge Components (KCs)**[51] (sometimes also for the hierarchy of knowledge[3] and defining a Q-matrix[52]), which is time consuming and not scalable to lifelong learning applications. There are various machine learning based unsupervised learning approaches that are used to extract latent topics from textual content. Approaches such as Latent Dirichlet Allocation (LDA)[5], Latent Semantic Analysis (LSA)[22] and other probabilistic approaches[31, 36] are potential candidates in this area. However, these unsupervised learning approaches suffer from complex hyper-parameter tuning processes and limited interpretability of the discovered *latent* KCs creating gaps in transparency. A step forward in this area has been using *explicit* representations based on taxonomies such as Wikipedia concepts[23, 27].

Wikification, a form of entity linking[6, 24] has shown great promise for automatically capturing the KCs covered in an educational resource. This technology is on a promising path towards providing *automatic*, *humanly-intuitive (symbolic)* representations from Wikipedia, representing at the same time *up-to-date knowledge* about *many domains*. Wikipedia, being one of the largest encyclopedias in the world, evolves with time due to the contributor population that constantly updates it. Due to these reasons, we use Wikification[6] to generate KCs that are included in the PEEK dataset.

2.4 Related Datasets

The majority of datasets that are related to this task come from the *knowledge tracing* domain[19, 46] which focuses on recovering knowledge/skill mastery of learners based on how they answer to specific exercises. This approach, as discussed in Section 2, is much more explicit and effort intense than inferring it using implicit

feedback (e.g. watched patterns of educational videos). Majority of the research in this domain has been based on ASSISTments data[51], that records a series of learners solving tests in an ITS platform. These datasets are heavily biased towards mathematics knowledge as the ASSISTments tool was used for mathematics education from the beginning[26, 39].

A few publicly available datasets to solve knowledge tracing problem exist. They include mathematics ASSISTments[1][32, 55], problem solving interactions[18], or multiple choice questions[2][56, 57]), all of which are noteworthy although they do not focus on implicit feedback and engagement. MOOCCube is a very recently released dataset that contains a spectrum of different statistics relating to learner-MOOC interactions including implicit and explicit test taking activity[62]. Although this dataset may contain data that can be used to predict learner engagement with implicit feedback, the dataset has only been used in prerequisite detection task which is very different. However, our prior experiments have consistently demonstrated that engagement prediction via online learning[14] can be achieved with a dataset that is similar to PEEK in structure.

Although previous work based on popular video repositories such as edX[29], Khan Academy[37], VideoLectures[13, 14] and YouTube[20, 60] has evidenced the existence of datasets that include implicit engagement signals, these datasets are never made public due to their proprietary nature. The scarcity of large scale publicly available datasets for predicting learner engagement with educational videos constrain the growth of the field. In this context, the Personalised Educational Engagement (*PEEK*) dataset, is the first and largest learner video engagement dataset that will be publicly released with humanly-interpretable Wikipedia concepts and the concept coverage associated with the video lecture fragments.

3 PEEK DATASET

In this section, we describe how the Personalised Educational Engagement linked to Knowledge topics (PEEK) dataset is constructed. Figure 2 outlines the overall process of creating the PEEK dataset.

3.1 Data Source

PEEK dataset is constructed using video metadata and learner activity data extracted from VideoLectures.Net[3] (VLN), a repository of scientific and educational video lectures. VLN repository records research talks and presentations from numerous academic venues accompanied by the lecture slides associated with the video. As the talks are recorded at peer-reviewed conferences and prestigious research venues, the lectures are reviewed and material is controlled for correctness of knowledge.

Figure 1 depicts the different components of the VLN user interface where a user engages with an educational video. Every lecture has a title, event details (e.g. conference venue, date, location etc.) and lecture meta data associated with it. Although most lectures consist of one video, some video lectures may contain more than one video as shown in Figure 1.

3.2 Fragmenting Videos

First, the videos in VLN repository are transcribed to its native language using the *TransLectures* project[4]. Then the non-English lecture videos are translated to English as we will use English Wikipedia for entity linking.

As described in Section 3.1, a lecture in VLN repository can contain one or more videos. To model user interactions in a much more granular level we further break each video into smaller parts. Once the transcription/translation is complete, we partition the transcript of each video into multiple *fragments* where each fragment covers approximately 5 minutes of lecture time. Having 5 minute fragments allow us to break the contents of a video into a more granular level while making sure that there is sufficient amounts of information in each video fragment.

3.3 Wikification of Transcripts

In order to identify the Knowledge Components (KCs) that are contained in different video fragments, we use Wikification[6]. This allows annotating learning materials with humanly interpretable KCs (Wikipedia concepts) at scale with minimum human-expert intervention. This setup will make sure that recommendation strategies build on this dataset will be technologically feasible for web-scale e-learning systems. Previous works have demonstrated the value of using Wikification in this task[14]. We restrict the dataset to English Wikipedia due to its richness in comparison to other languages.

3.4 Knowledge Component Ranking

As per [6], Wikification produces two statistical values per annotated KC, namely, *PageRank* and *Cosine Similarity* scores.

PageRank score is calculated by constructing a semantic graph where semantic relatedness ($SR(c, c')$) between each Wikipedia concept pair c and c' in the graph are calculated using equation 1.

$$SR(c, c') = \frac{\log(max(|L_c|, |L_{c'}|) - \log(|L_c \cap L_{c'}|)}{\log|W| - log(min(|L_c|, |L_{c'}|)} \quad (1)$$

where L_c represents the set of Wiki concepts with inwards links to Wikipedia concept c, $|\cdot|$ represents the cardinality of the set and W represents the set of all Wikipedia topics. This semantic relatedness graph is used for computing PageRank scores. PageRank algorithm[7] leads to heavily connected Wikipedia topics (i.e. more semantically related) within the lecture to get a higher score.

The *Cosine Similarity score* is used as a proxy for topic coverage within the lecture fragment[14]. This score $cos(s_{tr}, c)$ between the *Term Frequency-Inverse Document Frequency (TF-IDF)* representations of the lecture transcript s_{tr} and the Wikipedia page c is calculated based on equation 2:

$$cos(s_{tr}, c) = \frac{\text{TFIDF}(s_{tr}) \cdot \text{TFIDF}(c)}{\|\text{TFIDF}(s_{tr})\| \times \|\text{TFIDF}(c)\|} \quad (2)$$

where $\text{TFIDF}(s)$ returns the TF-IDF vector of the string s while $\|\cdot\|$ represents the norm of the TF-IDF vector.

The authors of [6] comment that a linearly weighted sum between the PageRank and Cosine score can be used to rank the importance of Wikipedia concepts as per equation 3.

---

[1]http://www.assistmentstestbed.org/
[2]https://www.microsoft.com/en-us/research/event/diagnostic-questions-neurips2020/
[3]www.videolectures.net

[4]www.translectures.eu

Figure 1: Screen layout of VideoLecture.Net website from which the data for PEEK dataset was sourced. Every lecture can have one or more videos and is associated with metadata. The lecture transcripts are also available for processing.

$$\text{Rank}(c) = \alpha \cdot PageRank(c) + (1 - \alpha) \cdot cos(s_{tr}, c), \text{ where } \alpha \in [0, 1] \tag{3}$$

We experimented with different values for α in equation 3 and empirically validated suitable linear combinations of weights for PageRank score and Cosine Similarity score. We observed that a weight of 0.8 on PageRank and 0.2 on Cosine similarity leads to the most suitable ranking of KCs[12]. We use Rank to identify the five top-ranked KCs for each lecture fragment. The cosine similarity score as per equation 2 is included in the dataset as a proxy for coverage of that KC in the lecture fragment. We restrict the number of KCs to 5 as larger numbers (e.g. 10 KCs) have shown to degrade performance of learner models built with similar datasets[14]. Figure 4(ii) provides a word cloud of the most dominant KCs in PEEK dataset. It is evident that the majority of KCs (Wikipedia topics) associated to the lecture fragments in this dataset are related to artificial intelligence and machine learning.

3.5 Anonymity

We restrict the final dataset to lectures that have been viewed by at least 5 unique users to preserve k-anonymity[21] of users. Also, we report the timestamp of user view events in relation to the earliest event found in the dataset obfuscating the actual timestamp. That is we report the smallest timestamp in the dataset t_0 as 0s and any

timestamp t_i after that as $t_i - t_0$. This allows us to publish the true order and the real differences of duration between events without revealing the actual timestamps. Additionally, the lecture metadata such as title and authors etc. are not published to preserve the anonymity of the authors/lecturers. The motivation behind this decision is to avoid authors of the video lectures having unanticipated effects on their reputation by associating implicit learner engagement values to their content.

3.6 Labels

The user interface of VLN website also records the video watching behaviours of its users (Interaction logs in Figure 2).

The target label for learner engagement is a discrete variable based on *video watch time* which has been used as a proxy for video engagement in both non-educational[20, 60] and educational[14, 29] contexts. Normalised learner watch time $\bar{e}^t_{\ell,r}$ of learner ℓ with video fragment resource r_i at time point t is calculated as per equation 4.

$$\bar{e}^t_{\ell,r_i} = W(\ell, r_i)/D(r_i), \tag{4}$$

where $\bar{e}^t_{\ell,r_i} \in \{0, 1\}$, $W(\cdot)$ is a function that returns the *watch time* of learner ℓ for resource r_i and $D(\cdot)$ is a function that returns the duration of lecture fragment r_i. The ultimate label e^t_{ℓ,r_i} is derived by discretising $\bar{e}^t_{\ell,r_i}$ where $e^t_{\ell,r_i} = 1$ when $\bar{e}^t_{\ell,r_i} \geq .75$ and
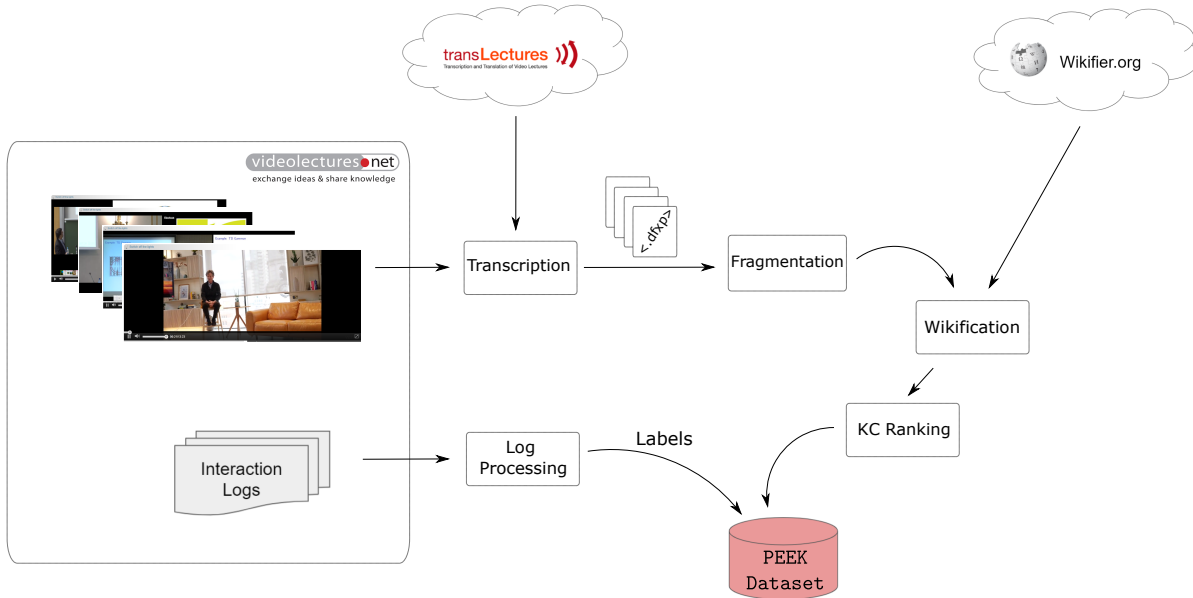
Figure 2: The video data and the learner interaction logs from VLN repository are processed separately to create the Wikipedia-based KCs and also the discrete engagement signals that are published in PEEK dataset.
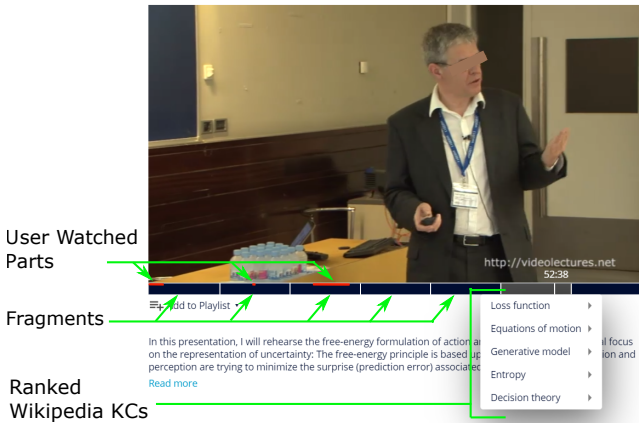

Figure 3: Visual representation of the different data items available in the PEEK dataset. Each video is broken into multiple, non-overlapping 5 minute fragments that are linked with ranked Wikipedia-based KCs. The watched parts of the video (in red) are used to create discrete engagement labels.

$e^t_{\ell, r_i} = 0$ otherwise. The discretisation rule is motivated by the hypothesis that a learner should watch approximately 4 minutes of an educational video fragment that is approximately 5 minutes (duration of a video fragment that includes 5000 characters from the video transcript) in order to acquire knowledge from it[14].

3.7 Final Dataset

The final PEEK dataset consists of 290,535 interaction events from 20,019 distinct users with at least 5 events. These learner engage with 8,801 unique lecture videos that are partitioned into 36,408 fragments (4.14 fragments per video). The learner population in the dataset is divided into *Training* (14,050 learners) and *Test* (5,969 learners) datasets based on a 70:30 split. The label distribution in the dataset is also relatively balanced with only 56.35% of the labels being positive. As shown in Figure 4 (i), the majority of learners in the dataset have a relatively small number of events (under 80) making this dataset an excellent test bed for personalisation models designed to work in data scarce environments. VLN repository mainly publishes videos relating to Computer Science and Machine Learning leading to a learner audience who visit to learn about these subjects. This fact is confirmed by Figure 4 where it shows that the dataset is dominated by events with AI and ML related KCs.

3.8 Structure of the PEEK Dataset

The final dataset consists of 3 files.

(1) `train.csv`, used for hyperparameter validation and training parameters.
(2) `test.csv`, used as the held-out test set.
(3) `id_to_wiki_url_mapping.csv`, which contains the mapping between KC IDs and Wikipedia page URL.

`train.csv` and `test.csv` files contain the actual learner session data where their interaction with lecture fragments are recorded. The two files contains 70% and 30% of the learners respectively. Both files contain 15 columns that are described in Table 1. As

the name suggests `id_to_wiki_url_mapping.csv` file contains a mapping between the KC ID in PEEK dataset and the URL of the Wikipedia page of the concept associated with that KC.
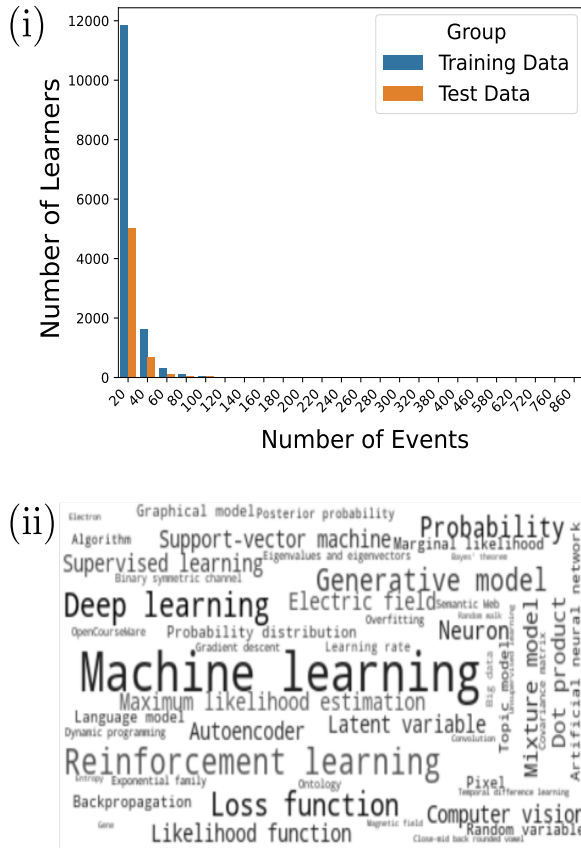

Figure 4: Characteristics of the PEEK dataset: (i) number of learners in the training/test dataset based on the number of events in the session for individual learner and (ii) word-clouds depicting the most frequently detected Wikipedia-based knowledge components.

4 MAIN PREDICTION TASK AND MODELS

We benchmark the predictive performance of this dataset with respect to learner engagement prediction by proposing a set of baseline models and measuring their performance with the PEEK dataset.

4.1 Data

Unlike previous studies, which relied on small, expert-curated collections of open educational materials, our dataset is built using in-the-wild watch patterns of multi-lingual open educational materials available online.

We construct two distinct datasets for our experiments. (i) `Active 20` is a smaller dataset that consists of 8,207 interaction events from the 20 most active users (410.35 events per learner) in the PEEK

dataset. This dataset contains a substantially small number of users who have relatively longer sessions. The smaller dataset will allow us to understand how the proposed learner models will behave with mature learners that have more interaction information in the system. On the contrary, (ii) PEEK is the full dataset that consists of 290,535 interaction events from 20,019 distinct users with at least 5 events. However, the majority of users in this dataset have relatively smaller sessions (average of 14.51 events per learner).

4.2 Prediction Task and Evaluation Metrics

As the dataset is already divided into train and test splits, all our experiments use a hold-out validation (train-test split) approach where hyperparameters and parameters are learned on 70% of the learners and the model is evaluated on the remaining 30% with the selected hyperparameter combination.

A sequential experimental design is employed, where $\hat{e}^t_{\ell, r_i}$, engagement with fragment t is predicted using fragments 1 to $t-1$. Since engagement is binary, predictions for each fragment can be assembled into a confusion matrix, from which we compute well-known binary classification metrics such as precision, recall and F1-measure. We focus on these measures as we are more interested in predictive lecture fragments that the learners are likely to engage with.

We average these metrics per learner and weight each learner according to their amount of activity in the system. We use F1-measure for model comparison as we are interested in improving both precision and recall.

The primary objective of the experiments is to evaluate how the different attributes of the dataset are useful in modelling engagement of learners. Towards this goal, we run two primary experiments, 1) investigate how different recommendation models perform with predicting engagement, and 2) how the number of KCs impact the prediction model. The benchmark algorithms used for the experiments are outlined in Section 4.3. The results of the experiments are outlined in Section 5.

4.3 Benchmark Models

PEEK is the first of its kind, a dataset that records in-the-wild engagement of informal learners with video lecture fragments. Due to the novelty of this dataset, we struggle to find already published baselines, except for the TrueLearn family of algorithms[14]. For the sake of comparing its predictive performance, we also propose a set of baselines that are based on content-based and collaborative filtering.

Content-based Similarity. Content-based filtering can measure the similarity between two items. We compute a similarity value, $sim(r^{t-1}_{\ell, r_i}, r^t_{\ell, r_i})$ between two consecutive lecture fragments r^{t-1}_{ℓ, r_i} and r^t_{ℓ, r_i} in the learner ℓ's session. We use this similarity value to make an engagement prediction $\hat{e}^t_{\ell, r_i}$ based on equation 5.

$$\hat{e}^t_{\ell, r_i} = \begin{cases} 1 & \text{if } sim(r^{t-1}_{\ell, r_i}, r^t_{\ell, r_i}) \geq threshold \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

In this case, we investigate two similarity measures, namely 1) *Cosine*, 2) *Concept-based Jaccard* and *User-based Jaccard*. When computing cosine similarity, we represent each video fragment

with Educational Videos

ORSUM@ACM RecSys 2021, October 2nd, 2021, Amsterdam, The Netherlands (Remote)

Table 1: Detailed descriptions of the different columns of the `train.csv` and `test.csv` files included in the PEEK Dataset.

Column Number	Description	Details
1	Video Lecture ID	An integer ID associated with an individual video lecture
2	Video ID	An integer ID associated to every video belonging to the same Video Lecture ID (e.g. $1 \ldots v$ if the lecture has v videos)
3	Part ID	An integer ID associated with each video fragment (e.g $1 \ldots f$ for a video with f fragments)
4	Timestamp	Timestamp (to the nearest second) when the play event was initiated.
5	user ID	An integer ID associated with each unique learner in the dataset PEEK dataset (IDs in `train.csv` and `test.csv` files are mutually exclusive).
6,8,10,12,14	KC IDs	An integer ID associated with each unique Knowledge Component. This ID can be linked to the human readable Wikipedia concept names found in `id_to_wiki_url_mapping.csv` file
7,9,11,13,15	Topic Coverage	Proxy for coverage of the relevant KC in the fragment of interest. KC coverage is the cosine similarity as per equation 2.
16	Label	The binary label e_{ℓ,r_i}^t, 1 if the learner watched $\geq .75$ of the video fragment, 0 otherwise.

using the bag of concepts representation where the concepts are the super set of Wikipedia concepts mentioned in the dataset. The values in this sparse vector are the cosine similarities between respective Wikipedia concept and the lecture fragment transcript as per equation 2.

An alternative approach to finding concept-wise similarity is Jaccard similarity. Concept-based Jaccard similarity $\mathtt{Jaccard}_C(r_{\ell,r_i}^{t-1}, r_{\ell,r_i}^t)$, between lecture fragments r_{ℓ,r_i}^{t-1} and r_{ℓ,r_i}^t is computed based on equation 6.

$$\mathtt{Jaccard}_C(r_{\ell,r_i}^{t-1}, r_{\ell,r_i}^t) = \frac{C(r_{\ell,r_i}^{t-1}) \cap C(r_{\ell,r_i}^t)}{C(r_{\ell,r_i}^{t-1}) \cup C(r_{\ell,r_i}^t)} \qquad (6)$$

where $C(\cdot)$ is a function that returns the set of Wikipedia concepts in resource r_i

Similarly, one can also measure the similarity between two lecture fragments based on how many learners interact with both the lecture fragments. The user interactions in the training dataset is used exclusively to learn the similarity matrix in order to avoid data leakage. In this approach, we can calculate the user-wise jaccard similarity $\mathtt{Jaccard}_{\mathcal{U}}(r_{\ell,r_i}^{t-1}, r_{\ell,r_i}^t)$, as per equation 7.

$$\mathtt{Jaccard}_{\mathcal{U}}(r_{\ell,r_i}^{t-1}, r_{\ell,r_i}^t) = \frac{\mathcal{U}(r_{\ell,r_i}^{t-1}) \cap \mathcal{U}(r_{\ell,r_i}^t)}{\mathcal{U}(r_{\ell,r_i}^{t-1}) \cup \mathcal{U}(r_{\ell,r_i}^t)} \qquad (7)$$

where $\mathcal{U}(\cdot)$ is a function that returns the set of learners that interacted with resource r_i

Knowledge Tracing (KT). KT builds a learner representation of knowledge of the learner[63]. This learning model is then used in predicting engagement of learner ℓ with lecture fragment resource r_i at time t. As the PEEK dataset has a temporal dimension, we reformulate the KT algorithm into an online learning graphical model inspired by the reformulation in[4]. The skill variables in the KT model are Bernoulli variables ($\theta_{\ell,c}^t \sim \mathtt{Bernoulli}(\pi_{\ell,c}^t)$), assuming that a learner ℓ would have either mastered a skill/ concept c or not (represented by probability $\pi_{\ell,c}^t$). Skills are initialised ($\theta_{\ell,c}^0$) using a $\mathtt{Bernoulli}(.0)$ prior, assuming that the latent skill is not mastered in the beginning. A noise factor similar to what is found

in the conventional KT model[19] is added to this model and is tuned using a grid search.

TrueLearn Novel Model. Similar to KT model, TrueLearn model is also an online, graphical model that develops a learner model of the learner. This model is inspired by the TrueSkill model[30] which is an evolution of IRT[49]. Contrary to the KT model, TrueLearn model models skills as Gaussian variables ($\theta_{\ell,c}^t \sim \mathcal{N}(\mu, \sigma^2)$). In addition to modelling knowledge, TrueLearn Novel model also models novelty of content which is a key aspect of educational recommendation[14]. Enforcing the same assumptions of KT model, the TrueLearn Novel skill parameters are initialised using a $\mathcal{N}(.0, \sigma_0^2)$ prior where σ_0^2, initial variance is a hyperparameter tuned using a grid search.

5 PREDICTION TASK PERFORMANCE AND DISCUSSION

The predictive performance of multiple benchmark model (outlined in Section 4.3) with the PEEK dataset is evaluated. The hyperparameters of each model (including number of KCs) are tuned using a grid search. The model performance with the best hyperparameter combinations is presented in Table 2.

Good predictive power of the similarity based approaches validate the expressiveness of the content representations based on Wikipedia concepts and cosine similarity. The poor performance of $Jaccard_{\mathcal{U}}$ Similarity model with the top 20 user dataset is expected as this model relies on learning similarities based on training data. Although 20 user dataset contains active users with more events per session, there is less data in total due to the small number of users which leads to learning a mostly incomplete similarity matrix between lecture fragments. It is seen that this condition changes with the full PEEK dataset where more data is available. The results in table 2 shows that TrueLearn Novel model performs best among the benchmarks investigated. This result further reinforces the superiority of TrueLearn Novel model investigated with a similar dataset[14].

Table 2: Predictive performance presented using Precision, Recall and F1-Score for the baseline models. The most performant value and the next best value are highlighted in bold and *italic* faces respectively.

Dataset Algorithm	Most Active 20 Users			PEEK Dataset		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Cosine Similarity	*67.799*	*87.194*	*73.635*	*57.859*	58.451	54.056
Jaccard$_C$ Similarity	67.157	80.901	70.818	57.810	60.362	55.026
Jaccard$_U$ Similarity	46.499	11.606	17.761	57.854	*72.755*	*61.219*
Knowledge Tracing	65.050	65.828	62.077	53.254	28.564	34.513
TrueLearn Novel	**78.790**	**89.390**	**83.090**	**58.290**	**79.240**	**64.710**

5.1 Number of Knowledge Components

The impact of the number of KCs for each model is also investigated and the results are reported in Figure 5.

The plots indicate that the similarity based approaches tend to perform better with more KCs in the representation. This is natural as Similarity based approaches need more information from the exact pair of contents that are being compared in order to make a fine grained prediction. On the contrary, Knowledge Tracing and TrueLearn novel Models that develop a learner model perform better with less number of KCs. This observation is also expected as the learner model has the capacity to store information about various different KCs that the learner encountered over the past. This allows the learner model to make a good prediction with little information from the current resource. In the TrueLearn model, the effect of adding extra KCs to the representation makes a very small difference in performance.

5.2 Discussion

Knowledge (skill mastery to be more specific) is usually measured from explicit test performance[19, 40, 46]. Contrary to narrowly scoped learning scenarios common to ITS, seeking explicit feedback is comparatively less realistic in lifelong learning settings where informal learners attempt to acquire knowledge without necessarily having the need to go through rigorous testing. Forcing learners into aggressive testing can hinder their learning experience. Additionally, technical challenges too exist in areas such as scalable test/ question generation and automatic test marking. However,[14] has demonstrated that modelling knowledge from educational watch patterns is a promising direction towards predicting future engagement. The opportunity that PEEK dataset brings is the capability to advance this research avenue by making the first large dataset publicly available. This dataset will allow researchers to improve knowledge representation and also include other crucial factors such as learner interest[13] in the spirit of building integrative educational recommenders using implicit feedback. As Figure 4(i) shows, PEEK dataset also contains a large number of learners with limited activity making this dataset an excellent candidate to evaluate low resource, data efficient personalisation algorithms.

PEEK dataset also acts as the bridge between multiple rich data sources. The Wikipedia concepts used in representing the KCs also enables connecting this dataset to the variety of auxiliary information available around Wikipedia, the world's largest encyclopedia. Some examples of possibilities are the ability to leveraging additional data structures such as Wikipedia page contents,

the hyperlink graph and the category tree. Other enriched derivatives of Wikipedia such as semantic relatedness[47] and knowledge bases[1, 58] are also directly fused with this dataset to build more powerful content representations. In addition, the Wikipedia taxonomy also provides humanly intuitive representations (grounded on Wikipedia concepts) that enhances interpretability. Figure 4 (ii) is a prime example of how humanly intuitive KCs improves the expressiveness of content representations. A humanly intuitive representation also paves way to much needed interpretable learner models that are essential to triggering meta-cognition within learners[16].

5.3 Relevance to Online Learning

PEEK dataset consists of a collection of learners making engagement choices across educational materials over time. This dataset clearly captures the temporal dynamics of the learners whose knowledge state and preferences change over time. The superiority of TrueLearn Novel in Table 2 also gives a strong indication on how a state-aware learner model that changes its learner representation over time, best captures the engagement dynamics of the dataset in comparison to naïve similarity-based approaches. This result reaffirms the relevance and usefulness of this dataset to invent online learning models that can identify personal and temporal dynamics of users.

5.4 Supported Tasks

This section introduces the reader to the tasks that the PEEK dataset could be used for. The main application area of the PEEK dataset is engagement prediction with video lecture fragments in an web-based learning setting as demonstrated in Section 4. However, this task can be further extended by fusing the engagement labels from all the fragments of a video to carry out video recommendation which is more common[20] (contrary to video fragment recommendation).

Moving away from engagement prediction, PEEK dataset can also be used to understand the structure of knowledge and learning pathways through it. Deducing the structure of knowledge using the co-occurrence patterns of KCs within the video lectures provide opportunities to understand inter-topic relationships and how knowledge is structured. Work in this direction can be used in identifying related materials and accounting for novelty in educational recommendation[14]. The dataset can also be used to identify different clusters/groups of learners and learning resources which will allow understanding the education landscape better. Studying the
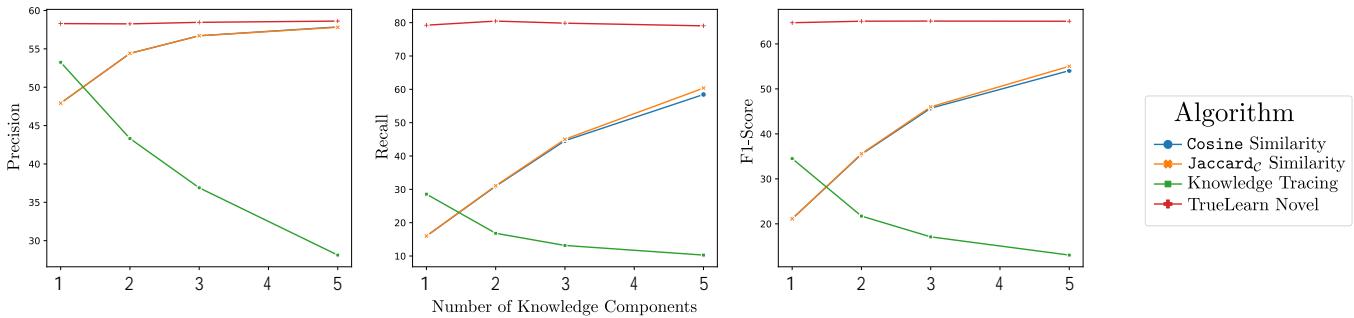
Figure 5: Predictive performance of PEEK dataset in terms of Precision (left), Recall (middle) and F1-Score (right) for the benchmark models when varying numbers of Knowledge Components (KCs) are used as the content representation.

evolution of KCs within learner journeys over time can also unlock opportunities to understand prerequisites[62].

6 CONCLUSIONS

This work releases *PEEK*, the first and largest publicly-available, humanly-intuitive dataset of learner interaction with educational videos opening up various avenues to the research community to push the frontiers of this line of research. It publishes learner interactions of over 20,000 learners with fragments of lectures over time. The predictive performance of several benchmark models is evaluated with this dataset. The power of this dataset is already portrayed by the promising experimental results presented.

As future directions, we see the importance of modelling semantic relatedness between KCs[47] that are likely to boost performance of learner models such as Knowledge Tracing and TrueLearn. Exploiting other side information from other Wikipedia such as the category tree is also an interesting direction to explore. The potential of detecting other user signals such as learner interest from PEEK dataset can also lead to insightful findings. PEEK dataset also allows experimenting with temporal dynamics such as interest decay[45].

ACKNOWLEDGMENTS

REFERENCES

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer, 722–735.

[2] Jordan Barria Pineda and Peter Brusilovsky. 2019. Making educational recommendations transparent through a fine-grained open learner model. In *Proceedings of Workshop on Intelligent User Interfaces for Algorithmic Transparency in Emerging Technologies at the 24th ACM Conference on Intelligent User Interfaces, IUI 2019, Los Angeles, USA, March 20, 2019*, Vol. 2327.

[3] Konstantin Bauman and Alexander Tuzhilin. 2018. Recommending remedial learning materials to students by filling their knowledge gaps. *MIS Quarterly* 42, 1 (2018), 313–332.

[4] Christopher Bishop, John Winn, and Tom Diethe. 2015. *Model-Based Machine Learning*. Early access version (http://www.mbmlbook.com/). Accessed: 2019-05-23.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003).

[6] Janez Brank, Gregor Leban, and Marko Grobelnik. 2017. Annotating Documents with Relevant Wikipedia Concepts. In *Proc. of Slovenian KDD Conf. on Data Mining and Data Warehouses (SiKDD)* (Ljubljana, Slovenia).

[7] Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. In *Proc. of Int. Conf. on World Wide Web*.

[8] Sahan Bulathwela, Stefan Kreitmayer, and María Pérez-Ortiz. 2020. What's in It for Me? Augmenting Recommended Learning Resources with Navigable Annotations. In *Proc. of Int. Conf. on Intelligent User Interfaces Companion (IUI '20)*. 114âĂŞ115.

[9] Sahan Bulathwela, María Pérez-Ortiz, Aldo Lipani, Emine Yilmaz, and John Shawe-Taylor. 2020. Predicting Engagement in Video Lectures. In *Proc. of Int. Conf. on Educational Data Mining (EDM âĂŹ20)*. https://arxiv.org/pdf/2006.00592.pdf

[10] Sahan Bulathwela, María Pérez-Ortiz, Rishabh Mehrotra, Davor Orlic, Colin de la Higuera, John Shawe-Taylor, and Emine Yilmaz. 2020. SUM'20: State-Based User Modelling. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM '20)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3336191.3371883

[11] Sahan Bulathwela, María Pérez-Ortiz, Rishabh Mehrotra, Davor Orlic, Colin de la Higuera, John Shawe-Taylor, and Emine Yilmaz. 2021. Report on the WSDM 2020 Workshop on State-Based User Modelling (SUM'20). *SIGIR Forum* 54, 1, Article 5 (Feb. 2021). https://doi.org/10.1145/3451964.3451969

[12] Sahan Bulathwela, María Pérez-Ortiz, E. S. V. Ranawaka, R. I. P. B. B. Siriwardana, G. A. K. Y. Ganepola, Emine Yilmaz, and John Shawe-Taylor. [n.d.]. D1.6 âĂŞ Report on Selected Models and Content Representations. https://www.x5gon.org/wp-content/uploads/2021/01/D1.6-1.pdf. Accessed in: 2021-04-02.

[13] Sahan Bulathwela, María Pérez-Ortiz, Emine Yilmaz, and John Shawe-Taylor. 2020. Towards an Integrative Educational Recommender for Lifelong Learners. In *AAAI Conference on Artificial Intelligence*.

[14] Sahan Bulathwela, María Pérez-Ortiz, Emine Yilmaz, and John Shawe-Taylor. 2020. TrueLearn: A Family of Bayesian Algorithms to Match Lifelong Learners to Open Educational Resources. In *AAAI Conference on Artificial Intelligence*.

[15] Susan Bull. 2020. There are open learner models about! *IEEE Transactions on Learning Technologies* 13, 2 (2020), 425–448.

[16] Susan Bull and Judy Kay. 2008. Metacognition and open learner models. In *The 3rd workshop on meta-cognition and self-regulated learning in educational technologies, at ITS2008*. 7–20.

[17] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing*. 162–171.

[18] Youngduck Choi, Youngnam Lee, Dongmin Shin, Junghyun Cho, Seoyon Park, Seewoo Lee, Jineon Baek, Chan Bae, Byungsoo Kim, and Jaewe Heo. 2020. Ednet: A large-scale hierarchical dataset in education. In *International Conference on Artificial Intelligence in Education*. Springer, 69–73.

[19] Albert T. Corbett and John R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction* 4, 4 (1994).

[20] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *Proc. of ACM Conf. on Recommender Systems*.

[21] Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. ORCAS: 20 Million Clicked Query-Document Pairs for Analyzing Search.

In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 2983âĂŞ2989. https://doi.org/10.1145/3340531.3412779

[22] Susan T Dumais. 2004. Latent semantic analysis. *Annual review of information science and technology* 38, 1 (2004), 188–230.

[23] Ofer Egozi, Shaul Markovitch, and Evgeniy Gabrilovich. 2011. Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)* 29, 2 (2011), 1–34.

[24] Paolo Ferragina and Ugo Scaiella. 2010. TAGME: On-the-Fly Annotation of Short Text Fragments (by Wikipedia Entities). In *Proc. of ACM Int. Conf. on Information and Knowledge Management (CIKM '10)*.

[25] Emily R Fyfe. 2016. Providing feedback on computer-based algebra homework in middle-school classrooms. *Computers in Human Behavior* 63 (2016), 568–574.

[26] Emily R. Fyfe. 2016. Providing Feedback on Computer-Based Algebra Homework in Middle-School Classrooms. *Comput. Hum. Behav.* 63, C (Oct. 2016), 568âĂŞ574. https://doi.org/10.1016/j.chb.2016.05.082

[27] Evgeniy Gabrilovich, Shaul Markovitch, et al. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis.. In *IJcAI*, Vol. 7. 1606–1611.

[28] Google. 2021. YouTube Help: Video chapters. https://support.google.com/youtube/answer/9884579?hl=en-GB

[29] Philip J. Guo, Juho Kim, and Rob Rubin. 2014. How Video Production Affects Student Engagement: An Empirical Study of MOOC Videos. In *Proc. of the First ACM Conf. on Learning @ Scale*.

[30] Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. TrueSkill(TM): A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems 20*. 569–576. https://www.microsoft.com/en-us/research/publication/trueskilltm-a-bayesian-skill-rating-system/

[31] Thomas Hofmann. 2013. Probabilistic latent semantic analysis. *arXiv preprint arXiv:1301.6705* (2013).

[32] Michelle Hurst and Sara Cordes. 2018. Labeling Common and Uncommon Fractions Across Notation and Education. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society, CogSci 2018, Madison, WI, USA, July 25-28, 2018*. cognitivesciencesociety.org. https://mindmodeling.org/cogsci2018/papers/0354/index.html

[33] Weijie Jiang, Zachary A. Pardos, and Qiang Wei. 2019. Goal-based Course Recommendation. In *Proceedings of International Conference on Learning Analytics & Knowledge*.

[34] Gilad Katz, Nir Ofek, Bracha Shapira, Lior Rokach, and Guy Shani. 2011. Using Wikipedia to boost collaborative filtering techniques. In *Proceedings of the fifth ACM conference on Recommender systems*. 285–288.

[35] Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401* (2020).

[36] Shangsong Liang. 2019. Collaborative, dynamic and diversified user profiling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4269–4276.

[37] Zachary MacHardy and Zachary A. Pardos. 2015. Evaluating The Relevance of Educational Videos using BKT and Big Data. In *Proc. of Int. Conf. on Educational Data Mining*.

[38] Debabrata Mahapatra, Ragunathan Mariappan, Vaibhav Rajan, Kuldeep Yadav, Seby A., and Sudeshna Roy. 2018. VideoKen: Automatic Video Summarization and Course Curation to Support Learning. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 239âĂŞ242. https://doi.org/10.1145/3184558.3186988

[39] Patrick McGuire, Shihfen Tu, Mary Ellin Logue, Craig A Mason, and Korinn Ostrow. 2017. Counterintuitive effects of online feedback in middle school math: results from a randomized controlled trial in ASSISTments. *Educational Media International* 54, 3 (2017), 231–244.

[40] Hiromi Nakagawa, Yusuke Iwasawa, and Yutaka Matsuo. 2019. Graph-Based Knowledge Tracing: Modeling Student Proficiency Using Graph Neural Network. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI '19)*.

[41] Zachary A Pardos, Ryan SJD Baker, Maria OCZ San Pedro, Sujith M Gowda, and Supreeth M Gowda. 2014. Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. *Journal of Learning Analytics* 1, 1 (2014).

[42] Philip I. Pavlik, Hao Cen, and Kenneth R. Koedinger. 2009. Performance Factors Analysis –A New Alternative to Knowledge Tracing. In *Proc. of Int. Conf. on Artificial Intelligence in Education*.

[43] Radek Pelánek, Jan Papoušek, Jiří Řihák, Vít Stanislav, and Juraj Nižnan. 2017. Elo-based learner modeling for the adaptive practice of facts. *User Modeling and User-Adapted Interaction* 27, 1 (01 Mar 2017), 89–118.

[44] María Pérez-Ortiz, Claire Dormann, Yvonne Rogers, Sahan Bulathwela, Stefan Kreitmayer, Emine Yilmaz, Richard Noss, and John Shawe-Taylor. 2021. X5Learn: A Personalised Learning Companion at the Intersection of AI and HCI *(IUI '21)*.

Association for Computing Machinery, New York, NY, USA, 70âĂŞ74. https://doi.org/10.1145/3397482.3450721

[45] Guangyuan Piao and John G. Breslin. 2016. Exploring Dynamics and Semantics of User Interests for User Modeling on Twitter for Link Recommendations. In *Proceedings of the 12th International Conference on Semantic Systems* (Leipzig, Germany) *(SEMANTiCS 2016)*. Association for Computing Machinery, New York, NY, USA, 81âĂŞ88. https://doi.org/10.1145/2993318.2993332

[46] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep Knowledge Tracing. In *Advances in Neural Information Processing Systems*.

[47] Marco Ponza, Paolo Ferragina, and Soumen Chakrabarti. 2020. On Computing Entity Relatedness in Wikipedia, with Applications. *Knowledge-Based Systems* 188 (2020). https://doi.org/10.1016/j.knosys.2019.105051

[48] Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume III, and Lise Getoor. 2014. Learning latent engagement patterns of students in online courses. In *Proc. of AAAI Conference on Artificial Intelligence*.

[49] G.E. Rasch. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests.* Vol. 1.

[50] Ralf Schenkel, Andreas Broschart, Seungwon Hwang, Martin Theobald, and Gerhard Weikum. 2007. Efficient text proximity search. In *International Symposium on String Processing and Information Retrieval*. Springer, 287–299.

[51] Douglas Selent, Thanaporn Patikorn, and Neil Heffernan. 2016. ASSISTments Dataset from Multiple Randomized Controlled Experiments. In *Proc. of the Third (2016) ACM Conf. on Learning @ Scale (L@S '16)*. https://doi.org/10.1145/2876034.2893409

[52] Kikumi K Tatsuoka. 1983. Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of educational measurement* (1983), 345–354.

[53] Gaurav Verma, Trikay Nalamada, Keerti Harpavat, Pranav Goel, Aman Mishra, and Balaji Vasan Srinivasan. 2021. Non-Linear Consumption of Videos Using a Sequence of Personalized Multimodal Fragments. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 249âĂŞ259. https://doi.org/10.1145/3397481.3450672

[54] Jill-Jênn Vie and Hisashi Kashima. 2019. Knowledge tracing machines: Factorization machines for knowledge tracing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 750–757.

[55] Candace Walkington, Virginia Clinton, and Anthony Sparks. 2019. The effect of language modification of mathematics story problems on problem-solving in online homework. *Instructional Science* 47, 5 (2019), 499–529.

[56] Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Yordan Zaykov, Jose Miguel Hernandez Lobato, Richard Turner, Richard Baraniuk, Craig Barton, Simon Peyton Jones, Simon Woodhead, and Cheng Zhang. 2020. Diagnostic Questions: The NeurIPS 2020 Education Challenge. https://www.microsoft.com/en-us/research/publication/diagnostic-questions-the-neurips-2020-education-challenge/

[57] Zichao Wang, Sebastian Tschiatschek, Simon Woodhead, JosĂľ Miguel HernĂ̧andez-Lobato, Jones Simon Peyton, Richard G. Baraniuk, and Cheng Zhang. 2021. Educational Question Mining At Scale: Prediction, Analysis and Personalization. In *Symposium on Educational Advances in Artificial Intelligence (AAAI-EAAI)*. https://www.microsoft.com/en-us/research/publication/educational-question-mining-at-scale-prediction-analysis-and-personalization/

[58] Wikimedia. 2018. *Wikimedia Downloads*.

[59] Kevin H Wilson, Yan Karklin, Bojian Han, and Chaitanya Ekanadham. 2016. Back to the basics: Bayesian extensions of IRT outperform neural networks for proficiency estimation. (2016).

[60] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. 2018. Beyond Views: Measuring and Predicting Engagement in Online Videos. In *Proc. of the Twelfth Int. Conf. on Web and Social Media*.

[61] Tae Yano and Moonyoung Kang. 2016. *Taking advantage of wikipedia in natural language processing*. Technical Report. Technical report, Carnegie Mellon University Language Technologies Institute.

[62] Jifan Yu, Gan Luo, Tong Xiao, Qingyang Zhong, Yuquan Wang, Wenzheng Feng, Junyi Luo, Chenyu Wang, Lei Hou, Juanzi Li, et al. 2020. MOOCCube: a large-scale data repository for NLP applications in MOOCs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3135–3142.

[63] Michael V. Yudelson, Kenneth R. Koedinger, and Geoffrey J. Gordon. 2013. Individualized Bayesian Knowledge Tracing Models. In *Proc. of Artificial Intelligence in Education*, H. Chad Lane, Kalina Yacef, Jack Mostow, and Philip Pavlik (Eds.).