X Modal
X Cultural
X Lingual
X Domain
X Site
Global OER Network

Grant Agreement Number:	761758
Project Acronym:	X5GON
Project title:	Cross Modal, Cross Cultural, Cross Lingual, Cross Domain, and Cross Site Global OER Network
Project Date:	2017-09-01 to 2020-12-31
Project Duration:	36 months
Deliverable Title:	D1.4 – Advanced Content Representations
Lead beneficiary:	UCL
Type:	Report
Dissemination level:	Public
Due Date (in months):	36 (August 2020)
Date:	31-December-2020
Status (Draft/Final):	Final
Authors:	Sahan Bulathwela, Maria Perez-Ortiz, E. S. V. Ranawaka, R. I. P. B. B. Siriwardana, G. A. K. Y. Ganepola, Thiruparan Ravikkumar, Shenal Pussegoda, Erik Novak, Emine Yilmaz and John Shawe-Taylor
Contact persons:	Sahan Bulathwela, Maria Perez-Ortiz, Emine Yilmaz and John Shawe-Taylor

Revision

Date	Lead author(s)	Commments
01/12/2020	Sahan Bulathwela Maria Perez-Ortiz, Emine Yilmaz and John Shawe-Taylor	Initial Draft
02/12/2020	Erik Novak	Added Contribution from JSI
04/12/2020	Thiruparan Ravikkumar and Shenal Pussegoda	Added Section on YouTube Tool
08/12/2020	E. S. V. Ranawaka R. I. P. B. B. Siriwardana and G. A. K. Y. Ganepola	Added Chapters on Language Detection and Duplicate Detection
14/12/2020	Alfons Juan	Internal Review
28/12/2020	Sahan Bulathwela	Final Version

Contents

List of Figures

List of Tables

Abstract

Availability of Open Educational Resources (OERs) has opened doors to many opportunities relating to providing timely, high quality education to the global. One of the key challenges X5GON project tries to address is aggregating OERs scattered around the world. Towards overcoming this challenge, content representation models for personalising educational materials (TrueLearn models) were proposed in deliverable D1.3 – Initial Content Representations [1]. Through this report, we advance and improve the content representation models by proposing multiple content representation models to (i) detect content language, (ii) detect duplicate materials and (iii) personalise educational materials by incorporating semantic relatedness between Knowledge Components (KCs). Multiple off-the-shelf language detection libraries are identified. Two content representations are identified for duplicate material detection and Semantic TrueLearn learner model is proposed. Furthermore, features that improve the user experience of the X5Learn learning interface are identified and developed.

[1] `https://www.x5gon.org/science/deliverables/`

1 Introduction

Recommendation systems have become one of the most popular forms of information retrieval in the recent years. With the vast amount of information available to the general public via the World Wide Web, the importance of information retrieval and recommendation systems grows by the day. In the context of applying Artificial Intelligence and Machine Learning in education, the research landscape has evolved from the first generation of intelligent tutoring systems [2] to fully fledged operational recommendation systems that use advanced statistical techniques and deep learning [3]. With the emergence of OERs as a cost-effective, scalable solution addressing Sustainable Development Goal 4 (SDG 4): *Ensuring inclusive and equitable quality education and promoting lifelong learning opportunities for all* [4], it is greatly imperative that Artificial Intelligence (AI) come to the help of managing the large collection of learning resources in an scalable, transparent and effective way.

Developing artificial intelligence systems that, mildly at least, understand the structure of knowledge is foundational to building an effective recommendation system for education[5, 3], as well as for many other applications [6, 7] related to knowledge management and tracing. From its inception, the intelligent tutoring and educational recommender communities have heavily relied on manually labelling the Knowledge Components (KCs) in a material or exercise [8]. However, this is not scalable in practice [2, 9].

1.1 Open Educational Resources (OER)

The global population grows in a rapid pace demanding more creative and innovative approaches to be devised in order to maintain providing high quality education to masses of learners. These learners can be diverse in many different ways including and not limited to dimensions such as cultural background, language, geographies, learning preferences and etc. Providing equal opportunities to such a diverse population can become very challenging. This has motivated the United Nations to include *Ensuring inclusive and equitable quality education and promoting lifelong learning opportunities for all*, in the sustainable development goals (SDGs) [10], the world's best plan to build a better world for people and our planet by 2030. As part of this movement, the concept *Open Educational Resources* was adopted and promoted by the UNESCO since 2002, encouraging a new family of educational resources that are geared towards democratising learning.

Figure 1: UNESCO Global Open Educational Resources Logo

1.2 AI in Education (AIEd)

In the recent years, Artificial Intelligence (AI) and Machine Learning (ML) have revolutionised how information is personalised to user needs while it shows potential in multiple domains. Among them, personalised education is very important due to the social impact it makes [11]. AI plays a significant role in maintaining quality of education provided to learners by identifying efficient learning pathways that works best for individual learners and their learning needs.

With the recent popularity of online learning [12], we can also observe that the creation of educational resources has also increased rapidly. Promotion of online learning in both commercial

(e.g. Udemy[2] and Udacity[3]) and non-commercial (e.g. Khan Academy[4] and MIT OpenCourseWare[5]) spheres have enabled the creation of an abundance of educational materials that are available in the Internet. Intuitively, large scale creation of educational material opens up opportunities for better personalisation as a wider spectrum of diverse learning resources are available.

This opportunity opens up potential to break from the traditional line of thought that heavily focuses on in-class learning to more ambitious use-cases such as Distance Learning, Massive Open Online Courses (MOOCs) and Lifelong Learning opportunities. Formal evaluations have shown that intelligent tutoring systems produce similar learning gains as one-on-one human tutoring, which has the potential to increase student performance to around the 98 percentile in a standard classroom [13, 14, 15]. Additionally, intelligent tutors could effectively reduce by one-third to one-half the time required for learning [13], increase effectiveness by 30% as compared to traditional instruction [13, 16, 17], reduce the need for training support personnel by about 70% and operating costs by about 92% and facilitate education in developing countries [18, 11]. Thus, the idea of building intelligent tutoring systems that provide online personalised education has gained a lot of traction in the recent years and will continue to do so.

Education will be impacted, and possibly transformed, by AI. AI is already changing the knowledge and skills needed in our global and innovation centred world. But AI is also enabling innovative methods of teaching and learning [19]. This report offers a summary and some examples of how AI could support the task of learners and teachers around the world, reflecting on some of the social aspects of AI and the technical and pedagogical challenges of the field. AI in Education (AIEd) includes everything from AI-driven, personalised and conversational educative systems, intelligent agents, automatic scoring and assessment and learner-support chatbots, to AI-facilitated matching of learner to learner/teacher and collaborative learning, putting learners in full control of their

1.3 Overview of Chapters

Through this work, our main contribution is proposing *Semantic* TrueLearn, a novel and transparent learner model that incorporates automatic entity linking and Wikipedia (a publicly available, humanly-intuitive, domain-agnostic and ever-evolving) knowledge graph, as a first step towards building an educational recommender that automatically labels materials and embeds the structure of universal knowledge. This algorithm is the sequel of TrueLearn Novel algorithm, our previous work outlined in deliverable *D1.3 – Initial Content Representations* [20]. Semantic TrueLearn maintains a symbolic representation of learners that allows explanations, rationalisations and scrutinising while leveraging the power of Bayesian learning. In addition, we also propose multiple models that can address language detection and duplicate detection that leads to higher quality content management within X5GON database. We also outline the improvements made to X5Learn learning platform which are subsequent improvements made on top of the X5Learn interface proposed in our previous work outlined in deliverable *D1.3 – Initial Content Representations* [20].

In chapter 2, we discuss the relevant related work that will pave the way to the models proposed in chapters 3 and 4. Chapter 3 of this report discusses about (i) language detection and (ii) duplicate detection models that will lead to improving the quality of the X5GON content database. Chapter 4 extends the TrueLearn Novel learner representation to introduce semantic relatedness information to the learner model. In chapter 5, we report the improvements done in the X5Learn learner interface and also introduces the YouTube tool that has been created to ingest more OERs. Finally, chapter 6 discusses the results and the achievements reported and concludes the report.

[2] `https://www.udemy.com`
[3] `https://eu.udacity.com`
[4] `https://www.khanacademy.org`
[5] `https://ocw.mit.edu`

2 Related Work

While, Open Educational Resources (OERs) have set themselves on a fast growth trajectory, gaining popularity. With innovative content creation models such as Content Explosion Model [21] and Open Educational Practice [22] boosting educational resource creation at scale, platforms such as X5GON [23] and X5Learn [24] progress towards making them easily accessible to learners.

However, trying to unify diverse repositories scattered across the world into one index pauses its own challenges. It is observed that most AI-powered enrichment models for tasks such as topic extraction, entity linking, cross-lingual translation etc. depend on prior knowledge such as language of the content. Another issue with OERs is that multiple repositories (sometimes the same repository) are likely to host multiple copies of the same educational resource. Due to effects as such, trivial components such as automatic language and duplicate detection can take a long way in terms of improving the quality of materials that are ingested and matched to the learners. Furthermore, identifying relatively less useful content (eg: parts of an examination, sets of references without any context to name a few) also enhances learner experience with indexes such as X5GON that ingest materials from multiple repositories.

The primary goal of this work is to leverage personalised recommendation of OERs to lifelong learners. This chapter discusses the prior art that lays the foundation towards building an educational recommendation system that improves learner engagement with educational materials. As scoped in section 1.3, the work that relates to capturing and improving contextual and context-agnostic aspects of learner-resource interaction is surveyed and discussed.

2.1 Automatic Language Detection

There are many examples for such instances such as web pages with extracts from other languages and European Union documents [25] that motivates understanding multi-language detection. Although multilingual document datasets such as ALTW 2010 [26], these datasets only contain multi-lingual documents which constrains them from carrying out evaluations on both mono and bi lingual performance analysis on the same dataset. Due to these reasons, the need arises for datasets that come from the same data source that can be used to evaluate monolingual and bilingual language detection. There is a whole community that is actively researching into this topic and proposing solutions that are already available in industry standard packages that can be reused with user-friendly licensing restrictions. A subset of these libraries are more suitable for X5GON project as they are implemented in JavaScript or Python programming languages.

2.1.1 Off-the-shelf Language Detection Libraries

- *textblob* detects language using Google Translate API. It doesn't provide probabilities for its predictions. This library only provides best match language meaning it doesn't support multi-lingual documents.
- *cld2* can detect multiple languages and provides confidence values. Although `cld3` exists, one reason to prefer `cld2` over `cld3` is because the `cld3` library doesn't do any preprocessing when it comes to emails, URLs etc. Since the OER materials are likely to contain these types of special text representations, `cld2` is preferable over `cld3`.
- *Polyglot* Depends on `cld2`. This library provides confidence. Multiple languages are detected while the library supports 196 languages.
- *langdetect* supports 55 detecting languages. However, this language detection algorithm is non-deterministic, which means that if the library is used on text that is either too short or too

ambiguous, different results will be obtained every time. This library can provide probabilities for top languages.

- *guess-language* supports over 60 languages. This library returns the language identifier, IANA language code for the top language detected.
- *langid* is pretrained on 97 languages that it can detect. This library is not sensitive to domain-specific features such as HTML, XML markup. It is also deployable as web service.
- *Fasttext* can be used to recognise 176 languages. Models were trained on data from Wikipedia, Tatoeba and SETimes datasets, used under CC-BY-SA.
- *nltkDetect* is integrated into the NLTK library and is based on TextCat algorithm [27]. The algorithm takes advantage of Zipf's law and uses n-gram frequencies to profile languages and text-yet to be identified-then compares using a distance measure.
- *whatthelang* is a "lightning fast" language prediction library as described in the GitHub page. It supports 176 languages. No confidence scores are provided for languages.
- *Langua* is a vectorised, improved version of `langdetect` library. It supports 55 languages.
- *Spacy* is a python library that provides industry strength Natural Language Processing (NLP) capabilities in Python. It includes a language detection module that supports 50+ languages. However, this feature only provides only the best matched language which makes it incompatible with multilingual documents.
- *franc* is a javascript library for language detection. It has multiple versions that support 82, 187 and 406 languages. This library also provides confidence scores on its predictions. The model is based on Universal Declaration of Human Rights (UDHR).

These libraries carry the accuracy realised by communities of researchers who are dedicated to improving the state-of-the-art performance and carries the technical stability of the developer communities that back these projects. This makes the above libraries suitable candidates for improving automatic language detection capabilities of X5GON.

2.2 Personalisation of Educational Resources

Recommendation systems are popular across multiple domains. Different approaches such as collaborative filtering [28], Bayesian match making [29] and extreme classification [30] are used to match resource with consumers. In the context of recommendation systems, state-aware machine learning systems have caught up alot of attention in the recent years [31, 32]. Contrary to conventional recommendation systems, a personalised learning system differs as its objectives are different. Rather than finding *similar* materials that is usually the case in a e-commerce or entertainment recommendation system, the primary objective is to discover sensible *learning trajectories* to a learner that will enable the learners to achieve their desired learning outcomes in the long run.

2.3 Personalised Learning Systems

An average student can achieve two standard deviations above an average control student taught under conventional group methods of instruction. The import line of research is to seek ways to accomplishing this under practical and realistic conditions than one-on-one tutoring [33]. When considering personalised learning systems, Intelligent Tutoring Systems is one of the most commonly researched application areas that attempts to apply AI in education. ITS community tries to address the above problem. An ITS usually contain i) The Domain Model, ii) The Pedagogy Model and iii) the learner model [34]. The domain model represents the subjects that are taught. The pedagogy model represents knowledge about effective approaches to teaching and learning including knowledge of instructional approaches [35], zone of proximal development [36] and etc. And the learner model is intended to build "a representation of the hypothesised knowledge state of the student" [37].

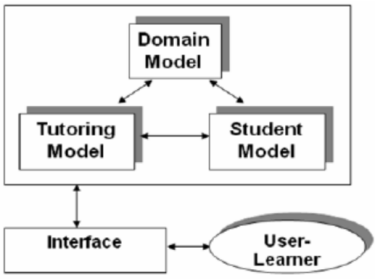

Figure 2: Structure of a Typical Intelligent Tutoring System

The primary goal of this work is to leverage personalised recommendation of OERs to lifelong learners through advanced content representations. This chapter discusses the prior art that can potentially contribute towards achieving this goal. As per section 1.3, the work that relates to capturing and improving resource quality (context-agnostic) and personalised recommendations (contextual) aspects of learner-resource interaction is surveyed and discussed.

2.4 Knowledge Tracing, Predictive Power and Interpretability

Lately, *Knowledge Tracing* (KT), one of foundational methods for building educational recommenders, is evolving from traditional machine learning [2, 9, 38, 39] to deep learning models [40, 3] that improve predictive power by sacrificing interpretability and transparency, crucial requirements for many of these systems. However, majority of these techniques still struggles to work around the requirement of expert labelling of the Knowledge Components (KCs) [41, 8] (and often the hierarchy of knowledge [5]), which is time costly and not scalable to lifelong learning applications. Incorporation of semantic relatedness in KT systems has been attempted in terms or prerequisite modelling [42, 43, 44], exercise similarity [45, 46, 47, 48, 49] and many other approaches [5, 50]. *Wikification*, a form of entity linking [51, 52], has shown substantial progress and great promise for automatically capturing the KCs covered in an educational resource. Although algorithms such as TrueLearn [53] have managed to use more scalable Wikipedia-based KC annotations techniques to overcome this problem, no attempts have been made to exploit semantic relatedness in Wikipedia. This work attempts to fill that gap.

Entity linking has thus been positioned now as a promising path towards providing *automatic*, *humanly-intuitive (symbolic)* representations from Wikipedia, representing at the same time *up-to-date knowledge* about *many domains*. For example, TrueLearn, a recent educational recommendation algorithm [53] has shown great promise building a recommender using these automatic annotations. However, wikification uses Wikipedia pages as KCs, which leads to a vast amount of topics and a non-structured and sparse representation of the material and consequently, the learners. This could also be specially detrimental to the recommender performance because often these systems treat KCs as independent, even though many KCs may be semantically related. Various usable semantic relatedness measures exist [54]. These comprise simple ones such as Jaccard similarity (of outward Wikipedia links), Language Model-based, Point-wise Mutual Information [55] to more advanced ones such as Milne and Witten [56] and entity embedding [57]. To the best of our knowledge, this is the first attempt at considering the semantic relatedness of concepts in a recommender system.

3 Refining Content Representations

Upon reviewing the actual content in the X5GON OER database, there were two evident issues that need to be addressed to enhance the quality of service that can be provided.

1. **Language Detection:** Detecting the language of materials beforehand is fundamental to AI-powered enrichment tasks that are carried out by X5GON processing pipeline.
2. **Duplicate Detection:** Detecting duplicates of the same educational resource is useful in compiling information retrieval results (search and recommendation results) that are pleasing to the learners in terms of user experience.

This chapter discusses two components that were developed in order to address the aforementioned issues.

3.1 Automatic Language Detection

Several python libraries were tested for their accuracy and prediction time in detecting the language of a sentence. Initial research to this subject pointed out that automatic language detection is an area that is extensively researched. This led to the finding that there are many state-of-the-art language detection models that are publicly available. This fact motivates to rather benchmark the existing work to build a suitable language detection tool for X5GON rather than inventing yet another language detector.

Using existing tools provides the solution with multiple advantages:

- Possibility to leverage industry standard language detection tools that already exists
- Low cost to production
- Large research community around the tools that constantly improve them
- Larger developer community around the tools leading to technical stability

It also poses drawbacks:

- Models are general purpose models
- Most publicly available tools focus on mono-lingual documents
- Many models get confused with mathematical expressions.

In light of these facts, this work is scoped to detect the language of text-based documents that are being indexed in X5GON. Majority of OERs that are indexed in the X5GON database are PDF documents and other documents that come in text modality.

3.1.1 Methodology

The objective of this work is to use existing language detection models to leverage language detection with X5GON database. One of the main elements that is of interest is the performance of these models in multi-lingual documents that are present in X5GON database.

Dataset In order to evaluate this, a novel dataset was constructed by programmatically mixing parts of texts from multiple languages to create *multilingual documents*. More information regarding the construction of this dataset can be found in D1.5 – Evaluation Methodologies for Content Representation Models and Release of Datasets for Measuring Quality of OERs.

Table 1: Potential language detection libraries with URLs for further reference.

Library	Reference URL
textblob	`https://textblob.readthedocs.io/en/dev/`
cld2	`https://pypi.org/project/cld2-cffi/`
Polyglot	`https://polyglot.readthedocs.io/en/latest/`
chardet	`https://chardet.readthedocs.io/en/latest/index.html`
langdetect	`https://pypi.org/project/langdetect/`
guess-language	`https://pypi.org/project/guess-language/`
langid	`https://github.com/saffsd/langid.py`
Fasttext	`https://fasttext.cc/`
nltkDetect	`https://www.nltk.org`
whatthelang	`https://github.com/indix/whatthelang`
Langua	`https://github.com/whiletruelearn/langua`
Spacy	`https://pypi.org/project/spacy/`
franc	`https://github.com/wooorm/franc`

Libraries The most popular publicly available language detection libraries were identified as potential candidates. The tools that were identified are listed in table 1

3.1.2 Summary of Results

8 European languages were selected for testing: German, Dutch, English, Slovene, Slovak, French, Italian and Spanish. These languages were selected as they made the majority of materials indexed in X5GON database [23].

Evaluation was done based on predictive performance and computational performance. Deliverable *D1.5 – Evaluation Methodologies for Content Representation Models and Release of Datasets for Measuring Quality of OERs* carries an extensive description of the evaluation methodology.

An ensemble of `fasttext` and `cld2` was selected as the final model to classify mono/multi-lingual documents with languages due to their reported accuracy of 97.5% and 95.53% on the language dataset. An extensive description of the results and the selected model is found in Deliverable *D1.6 – Report on Selected Models and Content Representations.*

3.2 Duplicate Detection

X5GON detects and processes OERs that are hosted in many repositories that are scattered around the world. Figure 3 depicts how hundreds of OER repositories are scattered all over different geographical regions.

Due to the openly licensed nature of OERs, there are tendencies for the same material to he published in multiple repositories. Unfortunately, X5GON processing engine sees these multiple duplicates as distinct OERs as the URLs are different from each other. This phenomenon leads to the same educational content being indexed as distinct OERs. In terms of ingestion, this has no significant effect on the pipeline. However, the problems occur when these OER records are matched to users. X5GON has multiple AI services that utilise content similarity. Due to this reason, duplicate resources will filter into learners and hinder their learning experience.

Figure 3: Multiple OER repositories scattered all over the world. The bubbled numbers refer to the number of OER repositories located in the respective regions

3.2.1 Different Types of Duplication

Duplicates, in the scope of self learning materials in X5GON database can be segregated as below:

1. Identical content - same word count, same words used in the same order of words
2. Similar content with small differences
    - Having minor differences of text (Less than 10%). Addition/ Omission of a few sentences.
    - Additional watermarks, footnotes, copyright declaimers etc.

As one might expect, the former type of duplication is easy to to as the content is identical. However, the latter type is more challenged due to the non-triviality of recognising duplicates.

3.2.2 Features for Detecting Duplicate OERs

Two aspects were identified to evaluate similarity between the OERs. Featuresets that focus on derived from the textual content of the data is used to analyse how duplicate detection can be done. The two feature representations used are as follows:

1. **Term-Frequency Feature Set**: In this representation, the words of the document are represented using the Bag-of-Words representation where the frequency of terms is used as the numeric representation for each token.
2. **Wikification Representation**: The document is represented as a Bag-of-Wikipedia Concepts. The Wikipedia can be extracted via entity linking [51]. The cosine similarity between the Wikpedia Topic and the textual content of the document is treated as a proxy for topic coverage [53].

3.2.3 Methods for Detecting Duplicate OERs

Exploratory analysis of the textual contents of OERs is the first step towards identifying the patterns that constitute duplication. Both feature sets proposed were experimented with the X5GON database to identify materials that are significantly similar to each other. The results showed that a hybrid method incorporating both word token and Wikication features was the most promising approach for detecting duplicates. An extensive report of the exploratory analyses and the chosen can be found in the Deliverable *D1.6 – Report on Selected Models and Content Representations.*

Once the model is identified, a subset of pairs of observations that were predicted to be duplicates were sampled randomly and labelled by humans annotators. This allowed us to calculate the precision of the unsupervised duplication detection mechanism that was identified to be most promising. This exercise also produced a gold standard dataset (derived from real life use-cases) that can be used to push the frontiers of duplication detection research field. More details about the produced dataset can be found in Deliverable *D1.5 – Evaluation Methodologies for Content Representation Models and Release of Datasets for Measuring Quality of OERs*

4 Semantic TrueLearn: Semantically-aware Educational Recommedations to Lifelong Learners

In this chapter, we propose *Semantic* TrueLearn, a novel and transparent learner model that incorporates automatic entity linking and Wikipedia (a publicly available, humanly-intuitive, domain-agnostic and ever-evolving) knowledge graph, as a first step towards building an educational recommender that automatically labels materials and embeds the structure of universal knowledge. Semantic TrueLearn maintains a symbolic representation of learners that allows explanations, rationalisations and scrutinising while leveraging the power of Bayesian learning. Towards achieving the goal of building an effective educational recommender, we i) propose different approaches for modelling semantic relatedness between the KCs, ii) propose a novel sub-symbolic Bayesian learner model that leverages universal knowledge while retaining transparency and iii) utilise an Open Education dataset to evaluate the performance of the proposed model.

4.1 User Modelling for Educational Recommendations using Learner Visits

The user modelling architecture [58, 1, 59] is designed to create a user profile based on the OERs the user has viewed. We developed the X5GON Connect [60], a small library that sends user activity data to the X5GON platform when it is integrated in an OER repository. With this data, the architecture is able to retrieve the OERs and use them to build or update the user models. Figure 4 shows the user modelling architecture.

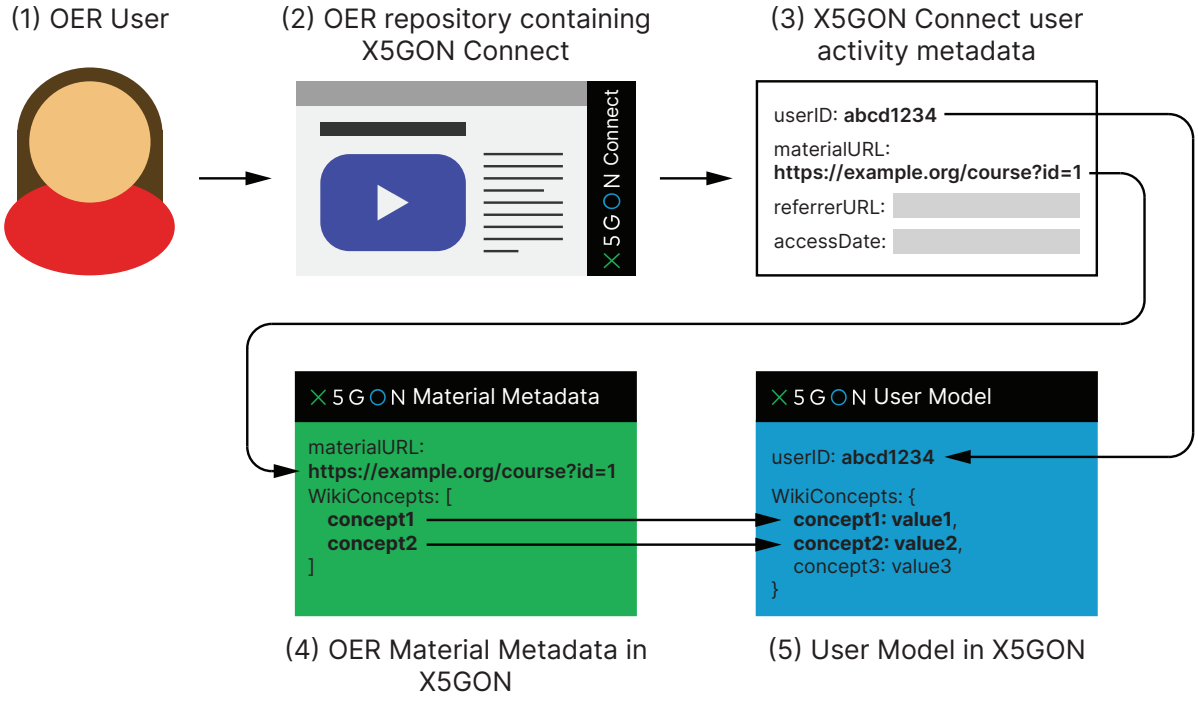

Figure 4: The user modelling architecture. The model is defined using the metadata of the OERs the user visited. Adapted from [1].

We present a short overview of the user model representation. A user model is generated using the viewed OER metadata. Each OER is presented as a dictionary of Wikipedia concepts found within

its content:

$$\text{material}(x) = \left\{ \begin{array}{c} \text{concept}_1 : \text{value}_{x,1} \\ \text{concept}_2 : \text{value}_{x,2} \\ \vdots \\ \text{concept}_N : \text{value}_{x,N} \end{array} \right\},$$

where x is the OER material, concept_i is the i-th Wikipedia concept, and $\text{value}_{x,i}$ is the number of times the Wikipedia concept is found within the material x. Afterwards, the representation is normalised such that $\sum_{i=1}^{N} \text{value}_{x,i} = 1$. Using the OER representations, we construct the user models as follows:

1. For each user u we initialize its corresponding user model with an empty dictionary: $\text{UM}(u)_0 = \{\}$.
2. When the user views a material x, we use its representation and update the model in the following way:
$$\text{UM}(u)_n = \frac{(n-1) \cdot \text{UM}(u)_{n-1} + \text{material}(x)}{n}, \quad (1)$$
where n is the the number of OERs user u viewed until this point.
3. The user model is then used to find relevant OERs by calculating the cosine distance between the user model and the material representations:
$$\text{cosine\_distance}(u, x) = 1 - \frac{\langle \text{UM}(u)_n, \text{material}(x) \rangle}{\|\text{UM}(u)_n\| \|\text{material}(x)\|},$$
where smaller distance means greater relevance.

The user models are stored in the database allowing us to update them in real-time. In addition, we developed the collaborative filtering method and integrated it into the recommender system.

4.1.1 Evaluation

We evaluated the user modelling architecture and the collaborative filtering method via the recommender plugin [61] integrated in the Videolectures.NET and UPV repositories. The recommender plugin provided a list of 20 OERs found relevant by one of the two approaches. The user selection from the recommended OERs was then stored in the X5GON database and analyzed [59, 62].

Inspecting the user models showed that when a user views a large number of OERs their representations become semi-stationary, e.g. does not update much with new OERs. This happens because of large n in equation 1, e.g. new OERs provide a small contribution to the user model. This could be improved by including the time component into equation 1, changing it into:

$$\text{UM}(u)_n = \left\{ \begin{array}{ll} \text{material}(x)_n, & n = 1, \\ \alpha \cdot \text{material}(x)_n + (1-\alpha) \cdot \text{UM}_{n-1}, & n > 1, \end{array} \right.$$

where the constant $\alpha \in (0, 1)$ represents the speed at which the user model forgets, and $\text{material}(x)_n$ is the representation of the material viewed at time n. This approach was inspired by the exponential moving average [63] in signal processing. Its development is a part of the future work.

4.2 User Knowledge Tracing using Video Watch Time

Consider a learning environment in which a learner ℓ interacts with a set of educational resources $S_\ell \subset \{r_1, \ldots, r_Q\}$ over a period of $T = (1, \ldots, t)$ time steps, Q being the total of resources in the system. Each resource r_i is characterised by the top KCs or topics covered $K_{r_i} \subset \{1, \ldots, N\}$ (N being the total of KCs considered by the system) and the depth of coverage d_{r_i} of those. We represent learner knowledge at time t as a multivariate Gaussian distribution $\boldsymbol{\theta}^t_\ell \sim \mathcal{N}(\boldsymbol{\mu}^t_\ell, \Sigma^t_\ell)$, $\boldsymbol{\mu}^t_\ell \in R^Q$ being the mean of knowledge and Σ^t_ℓ the covariance matrix. TrueLearn assumed that Σ is a diagonal covariance matrix in all cases and thus knowledge topics are completely independent from each other. The work in this paper builds towards considering a full covariance matrix, assuming that ρ_{ij} (estimated semantic relatedness) is a proxy for Σ_{ij} for topics i and j when $i \neq j$.

The key idea behind TrueLearn is to model the probability of engagement $e^t_{\ell,r_i} \in \{1, -1\}$ between learner ℓ and resource r_i at time t as a function of the learner skills/knowledge $\boldsymbol{\theta^t}_\ell$ and resource representation d_{r_i} for the top KCs covered K_{r_i}. When a new learner joins the recommender system, TrueLearn sets $\mu^0_\ell = 0$ and $\Sigma_{ii} = \beta$, where β is a hyperparameter of the system, and $\Sigma_{ij} = 0,\ i \neq j$. Then, when the learner consumes an educational fragment, TrueLearn updates the learner model/skills accordingly. Every skill that is not updated is set to the value from the last step, meaning at time t there might be many unobserved skills, specially given the amount of topics considered by the system (equal to the number of Wikipedia pages). Thus, TrueLearn assumes that the skill for topics in K_{r_i} can only be obtained through those topics and not semantically related ones.

4.3 Incorporating Semantic Relatedness Between Topics

The main assumption for Semantic TrueLearn, the algorithm proposed in this work, is that knowledge (or any other latent variable inferred by the system, such as interest, can be shared across semantically related topics. For this, we need to formally assume a relationship, which we set to:

$$\theta^t_{\ell,i} = \sum_{j \in \Omega_{\ell,i}} \frac{1}{|\Omega_{\ell,i}|} \cdot \gamma_{ij} \cdot \theta^t_{\ell,j}, \tag{2}$$

where Ω_i represents the set of topics indices used to infer the representation of topic i (e.g. most semantically related), where $i \neq j$. The mixing factors γ_{ij} can be set to semantic relatedness ρ_{ij} (and thus Σ_{ij}) or to a factor of the standard error of topic j (meaning most observed topics are used).

The tasks we define for Semantic TrueLearn are the following: i) Propagate information from observed/known topics to semantically related unobserved topics (which will be the vast majority, specially in the cold-start phase of the system). ii) Propagate information from observed/known topics to semantically related observed topics to make a more efficient use of every observation in the system. iii) Refine Σ_{ij} based on observed correlations between knowledge topics in the same material. In this paper we focus on i), as a first step to validate the usefulness of semantic relatedness to improve TrueLearn predictions. More specifically, we exploit the existing information from the observed topics for a learner, in order to estimate the *likely* value for unobserved topics and predict more accurately future learning events. We do this only at prediction time, which is computationally and memory efficient as the operation is restricted to the set of unobserved topics of the event and the semantically related observed topics of the learner.

4.4 Inferring the knowledge for unobserved topics at prediction (*Predict Step*)

We use Eq. 2 to infer $\theta_{\ell,i}$, the *new/unobserved* KC for learner ℓ by using the information from *observed* and semantically related KCs, $\theta_{\ell,j}$ where $j \in \Omega_{\ell,i}$. Specifically, we assume that the unobserved topic

is a linear combination of the Gaussian random variables in $\theta_{\ell,j}$, where we use Σ_{ij} (i.e. ρ_{ij}) as weights for the linear combination. We devise two strategies for propagating to semantically related topics:

$$\hat{\theta}^t_{\ell,i} \sim \mathcal{N}\left(\sum_{j\in\Omega_i} \frac{1}{|\Omega_i|}\cdot\rho_{ij}\cdot\mu^t_{\ell,j}, \sum_{j\in\Omega_i}\left(\frac{1}{|\Omega_i|}\cdot\rho_{ij}\right)^2\cdot\left(\sigma^t_{\ell,i}\right)^2\right),\ \ \forall\theta^t_{\ell,i}=\theta^0_{\ell,i}, \tag{3}$$

where we assume that the new unobserved variable also follows a Normal distribution and is a linear combination of (uncorrelated) Gaussian random variables. When we assume a correlation between the semantically related topics the equation above becomes more complex, so we use this as a first approach and will experiment with a linear combination of correlated variables in future work.

4.5 Increasing the unobserved skill variance for the update (*Update Step*)

TrueLearn is built on top of TrueSkill [64], which effectively changes the magnitude of the update based on the *precision* of the Gaussian variable. In our paper, we use a similar mechanism and add a factor to the precision of unobserved skills. The intuition behind this is that once we observe the user interacting with semantically related topics to an unknown topic, we can increase our uncertainty about the user not being familiar with the unknown topic (this is, we can no longer assume that the user does not know about the unknown topic). To increase the variance of the new skill, we use a factor of the frequency of the number of updates made to each observed topic. Our initial experiments showed that an increase of variance logarithmic proportionate to the frequency of updates is more stable than a linear proportion. We define this method as `Update`, which increases the variance of the new KC parameter before *updating* it using the outcome e^t_{ℓ,r_i}. The variance $\hat{\sigma}^2_{\ell,i}$ during update phase is $\hat{\sigma}^2_{\ell,i}=\prod_{j=1}^{|\Omega_i|}\prod_{n=1}^{\log(f_{\ell,j})}\sigma^2_{\ell,j}\cdot(1+\eta\rho_{ij})$ where $f_{\ell,j}$ is the *frequency* of past updates for user ℓ on skill j, ρ_{ij} is the semantic relatedness between the topics j and j, and η is a tuneable hyper-parameter between 0 and 1, that determines the update rate.

4.6 Semantic relatedness metrics (SR Metric)

As mentioned in chapter 2, different measures of semantic relatedness, both graph-based and neural-based, exist. We empirically evaluate if the predictive performance of an educational recommender can be improved by incorporating 7 different measures of semantic relatedness. We devise Milne and Witten (`M&W`), Entity Embeddings (`W2V`), Point-wise Mutual Information (`PMI`), Language Model based (`LM`), Jaccard Similarity (`Jaccard`), Conditional Probability (`CP`) and Barabasi and Albert (`BA`) methods, where semantic relatedness values are pre-computed and publicly available [55].

4.7 Summary of Results

It is clear that the click-based user modelling approach outlined in section 4.1 only uses the click information which can be very restrictive in terms of the signal it can provide. However, this model is useful as X5GON works with materials that go beyond Videos. On the contrary, learner watch time provides a significantly more informative signal regarding what exact parts of videos a learner consumes in the context of video modality. The evaluation of *Semantic TrueLearn Novel* demonstrates that it can outperform the *TrueLearn Novel* algorithm, which does not account for semantic relatedness between the Wikipedia topics when predicting future learner video watching behaviours. We include a detailed description of the dataset used and evaluation methodology of Semantic TrueLearn in Deliverable *D1.5 – Evaluation Methodologies for Content Representation Models and Release of Datasets for Measuring Quality of OERs*. The detailed results and the discussion of the chose models can be found in Deliverable *D1.6 – Report on Selected Models and Content Representations*.

5 Advancements in X5Learn Learner Interface

Long documents, such as e-books, conference talks and lecture videos, constitute a substantial fraction among educational resources that circulate in the Internet [23, 65]. While many of these materials are of high quality and potential value to learners, research in online learner behaviour has shown that long formats are often considered overwhelming and unwieldy in practice, preventing learners from engaging with these resources [65]. Since engagement has been shown to be a prerequisite for achieving impactful learning outcomes [66, 67, 68], we propose a novel interface that aims to describe educational resources to the learner by devising *navigable engaging fragments*, that can serve as effective entry points (or alternatives - depending on the learner's information need at hand) to the use of entire documents/ videos. The goal is to increase transparency and put the learner in control of their educational choices.

Keeping in mind that the learner-content interaction plays a key role in providing fruitful learning experiences [69], we design a novel system that combines Artificial Intelligence (AI) and Human Computer Interaction (HCI) to enhance the learner experience and encourage them to engage effectively with learning resources. We open the resultant system to the public via `http://x5learn.org/`. The benefits of the proposed system go beyond video lectures and apply to a wide range of long document formats including audio and PDF documents, making it a very powerful tool. The initial developments of this interface is outlined in Chapter 5 X5Learn: A Learner Facing Dashboard for Learning in Deliverable D1.3 – Initial Content Representations.

5.1 Improving the X5Learn Playlist Tool with Iterative Design

X5Learn is an AI-powered Intelligent Learning Platform that is motivated to increase accessibility of high quality open educational resources to everyone in the world. Initially, the current remote teaching setting in the universities gave us the opportunity to test the playlist creation tool with a few university lecturers. Most of the Open Educational Resources (OERs) indexed in X5GON are university level materials. This gives us the chance to encourage lecturers to use the search function to discover materials and compile them in study packages that can be dispatched to their students.

Multiple issues were identified by interviewing university lecturers who attempted to use the system for teaching. These findings led to many improvements in the X5Learn interface. Some of the issues such as errors in Wikification were also identified (such as the case shown in figure 5)

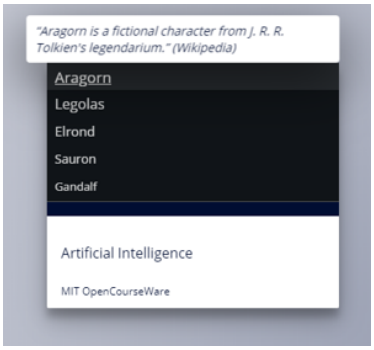

Figure 5: Search result for "Artificial Intelligence" in X5Learn initial interface.

A detailed account of the findings led improvements related to the X5Learn user interface is found in *D5.3 Final report on piloting.*

5.2 Discovering OERs in YouTube Video Repository

One of the key findings from speaking to X5Learn users is its scarcity of resources. University lecturers pointed out in their interviews that YouTube is one of the main sources for finding resources for their teaching. Many prior works also have demonstrated that YouTube contains thousands of useful educational resources [70, 65]. Researching deeper into YouTube educational materials also revealed that the platform gives the option for the authors to publish their videos under the Creative Commons BY 3.0 license [6] which makes the videos OERs [71].

[6] `https://creativecommons.org/licenses/by/3.0/legalcode`

Table 2: Identified YouTube channels which contain OERs with meta-information about the videos in them. The number of videos (Size) is rounded.

YouTube Channel	CC Licensed	Playlists available	Close Captioning	Language	Size
IIT Bombay	yes	yes	yes	en	3,400
Ted Ed	yes	yes	yes	en	1,700
Ted Ex	yes	yes	yes	en	160,000
Ted	yes	yes	yes	en	3,500
Ted Ex Student talks	yes	yes	yes	en	6,200
MIT Open Courseware	yes	yes	yes	en	6,200
METU Open Courseware	partially	yes	yes	multiple	800
Open Education and Culture	yes	yes	yes	en	200
Open Education Global	yes	yes	yes	en	200
Homework5: engageNY	yes	yes	yes	en	120
The Audiopedia	yes	no	yes	en	46,000
				TOTAL	**228,330**

`youtube_scraper` is a python package which can extract the relevant metadata from YouTube videos and transcripts of videos with timestamps. This python package includes several functions that allows you to extract the metadata provided by a YouTube channel based on the available playlists and able to extract all the videos metadata present in the channel. YouTube metadata data has been the primary data sources in many exciting research that look into popularity dynamics of videos [72], understanding engagement [70, 30] and many different NLP tasks . In addition to indexing OERs from YouTube in X5GON, extracting metadata

Following data can be extracted from each YouTube videos:

Channel Title	Title of the Video
Channel ID	Video URL
Published Time	Duration of Video
License under which it is published	View Count
Like Count	Dislike Count
Category ID	Description
Thumbnails	Transcripts with Timestamps

5.3 Results

Multiple improvements have been deplopyed to the X5Learn learning platform motivated by the findings from the interviews mentioned in section 5.1. Automatic thumbnail generation, Ability to change the title and description of playlist items and integration of PDF documents are few of these improvements. A detailed report of the above improvements can be found in D5.3 Final report on piloting.

At this point, we have identified several reputed channels in YouTube video repository that contain OERs that will enrich the X5GON index. Following table reports the identified channels with some background information about them. Although these videos have not been integrated to X5GON at this point, they can be easily extracted and ingested into X5GON platform which will magnify the potential of X5GON's usage significantly.

In order to make OER videos hosted in YouTube accessible via X5Learn platform, we have also integrated the YouTube Video player to X5Learn platform as well as depicted in figures 6 and 7

Figure 6: OER Videos coming from YouTube can appear in in X5Learn interface as search results. Notice that the source of these videos is "youtube.com"

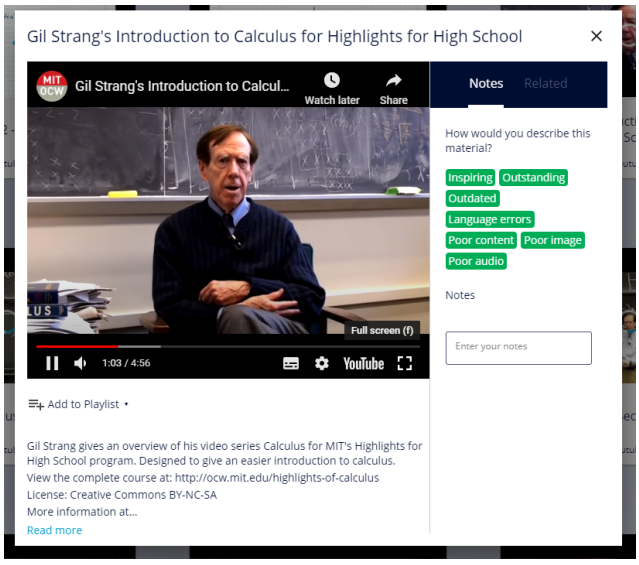

Figure 7: The YouTube video player replaces the conventional video player when the video resource comes from YouTube video repository.

6 Discussion and Conclusions

This section summarises the work that has been carried out in the direction of developing advanced content representations that are utilised in the X5GON learning platform. There are four major areas where content representation towards higher quality AI services has been embarked and advanced. They are:

- Automatic Language Detection of Text Documents
- Automatic Duplicate Detection with OERs
- Leveraging Semantic Relatedness between Knowledge Components to Improve Recommendations
- Improving X5Learn Interface to support more teaching/learning use-cases

Chapter 3 in the report presents language detection models and duplication detection strategies identified in order to improve the quality to content understanding and refining the results that are output from X5GON AI-based services (such as search and recommendation). It is observed that many open source language detection libraries with bleeding edge performance are available with open licences (such as Apache 2, MIT) off the shelf. This eliminates the need to reinvent the wheel by investing time and energy into creating yet, another language detection model. It is also observed that the X5GON processing pipeline already harvests content representations (Document content for TFIDF and Wikification output) that can be used to detect duplicates. Accurate language detection at early stages of the document processing improves the subsequent AI enrichment services such as Wikification and translation that depends on prior detection of the language. Duplication detection will allow X5GON to send out distinct documents as search and recommendation results that depend on content similarity.

Chapter 4 presents *Semantic TrueLearn* algorithm, a novel educational recommendation algorithm that leverages semantic relatedness information between knowledge components that are transparent and scalable. This algorithm builds its foundations on TrueLearn Novel algorithm described in *D1.3 – Initial Content Representations*. The model demonstrates how the initial content representations can be improved using semantic relatedness information that is publicly available. This model will act as the primary personalisation algorithm used in X5Learn learning platform which is described in chapter 5.

As chapter 5 describes, many improvements have been done to X5Learn platform driven by user feedback. Introduction of playlist generation tool, automatic thumbnail generation, customisable playlist items are some of the User Interface improvements that will support creating teaching materials using this intelligent user interface. The need to ingest more diverse learning materials that are more familiar to the teaching community is also identified and remedied though the development of the YouTube video ingesting tool.

6.1 Conclusion

Overall, various tasks relating to creating advanced content representations have been embarked by improving language detection, duplicate detection, personalisation model and the X5Learn user interface. The results also show that these improvements lead to better results in terms of predictive performance, higher quality document collection and ultimately, a superior learner experience with the X5GON services.

References

[1] N. Erik U. Jasna. Deliverable 4.2 - final prototype of user modelling architecture. https://www.x5gon.org/wp-content/uploads/2019/10/D4.2-Final-prototype-of-user-modelling-architecture-FINAL.pdf. Accessed in: 2020-12-02.

[2] Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

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

[4] UNESCO. Open educational resources (oer). https://en.unesco.org/themes/building-knowledge-societies/oer. Accessed: 2019-04-01.

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

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

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

[8] Douglas Selent, Thanaporn Patikorn, and Neil Heffernan. Assistments dataset from multiple randomized controlled experiments. New York, NY, USA, 2016. Association for Computing Machinery.

[9] Michael V. Yudelson, Kenneth R. Koedinger, and Geoffrey J. Gordon. Individualized bayesian knowledge tracing models. In H. Chad Lane, Kalina Yacef, Jack Mostow, and Philip Pavlik, editors, *Proc. of Artificial Intelligence in Education*, 2013.

[10] United Nations. Sustainable development goal 4. https://sustainabledevelopment.un.org/sdg4. Accessed: 2019-06-01.

[11] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, 11(1):1–10, 2020.

[12] I. Elaine Allen and Jeff Seaman. Online nation: Five years of growth in online learning. Technical report, 2007.

[13] Beverly Park Woolf. *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Morgan Kaufmann, 2010.

[14] Albert Corbett. Cognitive computer tutors: Solving the two-sigma problem. In Mathias Bauer, Piotr J. Gmytrasiewicz, and Julita Vassileva, editors, *User Modeling 2001*, pages 137–147. Springer Berlin Heidelberg, 2001.

[15] Benjamin S. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6):4–16, 1984.

[16] JD Fletcher. Does this stuff work? some findings from applications of technology to education and training. In *Proceedings of conference on teacher education and the use of technology based learning systems*. Society for Applied Learning Technology Warrenton, 1996.

[17] JD Fletcher. Intelligent training systems in the military. *Defense applications of artificial intelligence: Progress and prospects. Lexington, MA: Lexington Books*, 1988.

[18] Benjamin D Nye. Intelligent tutoring systems by and for the developing world: A review of trends and approaches for educational technology in a global context. *International Journal of Artificial Intelligence in Education*, 25(2):177–203, 2015.

[19] W. Holmes, M. Bialik, and C. Fadel. *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Independently Published, 2019.

[20] S. Bulathwela, M. Perez-Ortiz, E. Yilmaz, and J. Shawe-Taylor. D1.3 – initial content representations. `https://www.x5gon.org/wp-content/uploads/2019/10/X5GON_Deliverable_D1_3.pdf`. Accessed in: 2020-12-02.

[21] Jan M. Pawlowski, Volker Zimmermann, and Imc Ag. Open content: A concept for the future of e-learning and knowledge management?, 2007.

[22] Max Ehlers, Robert Schuwer, and Ben Janssen. Oer in tvet: Open educational resources for skills development, 2018.

[23] Erik Novak, Jasna Urbančič, and Miha Jenko. Preparing multi-modal data for natural language processing. In *Proc. of Slovenian KDD Conf. on Data Mining and Data Warehouses (SiKDD)*, 2018.

[24] Sahan Bulathwela, Stefan Kreitmayer, and María Pérez-Ortiz. What's in it for me? augmenting recommended learning resources with navigable annotations. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, IUI 20, 2020.

[25] Marco Lui, Jey Han Lau, and Timothy Baldwin. Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, 2:27–40, 2014.

[26] Timothy Baldwin and Marco Lui. Multilingual language identification: ALTW 2010 shared task data. In *Proceedings of the Australasian Language Technology Association Workshop 2010*, pages 4–7, Melbourne, Australia, December 2010.

[27] William B Cavnar, John M Trenkle, et al. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, volume 161175. Citeseer, 1994.

[28] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *Proc. of Int. Conf. on Data Mining*, volume 8, pages 263–272. Citeseer, 2008.

[29] Thore Graepel, Joaquin Quiñonero Candela, Thomas Borchert, and Ralf Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine.

[30] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proc. of ACM Conf. on Recommender Systems*, 2016.

[31] Sahan Bulathwela, María Pérez-Ortiz, Rishabh Mehrotra, Davor Orlic, Colin de la Higuera, John Shawe-Taylor, and Emine Yilmaz. Sum20: State-based user modelling. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM 20, pages 899–900, New York, NY, USA, 2020. Association for Computing Machinery.

[32] Sahan Bulathwela, María Pérez-Ortiz, Rishabh Mehrotra, Davor Orlic, Colin de la Higuera, John Shawe-Taylor, and Emine Yilmaz. Report on the wsdm 2020 workshop on state-based user modelling (sum'20). volume 54. ACM, 2020.

[33] Benjamin S. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6):4–16, 1984.

[34] Wayne Holmes, Maya Bialik, and Charles Fadel. Artificial intelligence in education: Promises and implications for teaching and learning, 2019.

[35] Carl Bereiter and Marlene Scardamalia. Intentional learning as a goal of instruction. In *Prepared on the basis of talks given at the Jul, 1985 symposium on Cognition and Instruction at the Learning Research and Development Center [LRDC], University of Pittsburgh, which was held in celebration of LRDC's 20th Anniversary and to honor Robert Glaser.* Lawrence Erlbaum Associates, Inc, 1989.

[36] Seth Chaiklin et al. The zone of proximal development in vygotsky's analysis of learning and instruction. *Vygotsky's educational theory in cultural context*, 1:39–64, 2003.

[37] John A Self. Student models in computer-aided instruction. *International Journal of Man-machine studies*, 6(2):261–276, 1974.

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

[39] Philip I. Pavlik, Hao Cen, and Kenneth R. Koedinger. Performance factors analysis –a new alternative to knowledge tracing. In *Proc. of Int. Conf. on Artificial Intelligence in Education*, 2009.

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

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

[42] C. Carmona, E. Millán, J. L. Pérez-de-la Cruz, M. Trella, and R. Conejo. Introducing prerequisite relations in a multi-layered bayesian student model. In *Proc. of the Int. Conf. on User Modeling*, UM'05, page 347–356, 2005.

[43] Tanja Käser, Severin Klingler, Alexander Gerhard Schwing, and Markus Gross. Beyond knowledge tracing: Modeling skill topologies with bayesian networks. In *Proc. of the Int. Conf. on intelligent tutoring systems*, pages 188–198. Springer, 2014.

[44] Penghe Chen, Yu Lu, Vincent W Zheng, and Yang Pian. Prerequisite-driven deep knowledge tracing. In *2018 IEEE Int. Conf. on Data Mining (ICDM)*, pages 39–48. IEEE, 2018.

[45] Andrew S Lan, Andrew E Waters, Christoph Studer, and Richard G Baraniuk. Sparse factor analysis for learning and content analytics. *The Journal of Machine Learning Research*, 15(1):1959–2008, 2014.

[46] Shalini Pandey and Jaideep Srivastava. Rkt: Relation-aware self-attention for knowledge tracing. *arXiv preprint arXiv:2008.12736*, 2020.

[47] Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, Guoping Hu, et al. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[48] Aritra Ghosh, Neil Heffernan, and Andrew S. Lan. Context-aware attentive knowledge tracing. In *Proc. of the 26th ACM SIGKDD Int. Conf. on Knowledge Discovery amp; Data Mining*, KDD '20, page 2330–2339, 2020.

[49] Hiromi Nakagawa, Yusuke Iwasawa, and Yutaka Matsuo. Graph-based knowledge tracing: Modeling student proficiency using graph neural network. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, 2019.

[50] Khushboo Thaker, Lei Zhang, Daqing He, and Peter Brusilovsky. Recommending remedial readings using student knowledge state. In *Proc. of Int. Conf. on Educational Data Mining*, pages 233–244, 2020.

[51] Janez Brank, Gregor Leban, and Marko Grobelnik. Annotating documents with relevant wikipedia concepts. In *Proc. of Slovenian KDD Conf. on Data Mining and Data Warehouses (SiKDD)*, 2017.

[52] Paolo Ferragina and Ugo Scaiella. 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, 2010.

[53] S. Bulathwela, M. Perez-Ortiz, E. Yilmaz, and J. Shawe-Taylor. Truelearn: A family of bayesian algorithms to match lifelong learners to open educational resources. In *AAAI Conference on Artificial Intelligence*, 2020.

[54] Marco Ponza, Paolo Ferragina, and Soumen Chakrabarti. On computing entity relatedness in wikipedia, with applications. *Knowledge-Based Systems*, 188, 2020.

[55] Francesco Piccinno. *Algorithms and data structures for big labeled graphs*. PhD thesis, Universitaḍe Pisa, 2017.

[56] David Milne and Ian H Witten. Learning to link with wikipedia. In *Proc. of the 17th ACM conference on Information and knowledge management*, pages 509–518, 2008.

[57] Stefan Zwicklbauer, Christin Seifert, and Michael Granitzer. Robust and collective entity disambiguation through semantic embeddings. In *Proc. of the Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR '16, 2016.

[58] N. Erik. Deliverable 4.1 - initial prototype of user modelling architecture. `https://www.x5gon.org/wp-content/uploads/2018/12/D4.1-Initial-Prototype-of-User-Modelling-Architecture.pdf`. Accessed in: 2020-12-02.

[59] N. Erik U. Jasna. D4.4 – final prototype of recommendation engine. `https://www.x5gon.org/wp-content/uploads/2019/10/D4.4-Final-prototype-of-recommendation-engine-FINAL.pdf`. Accessed in: 2020-12-02.

[60] X5GON. X5gon connect. `https://platform.x5gon.org/products/connect`. Accessed in: 2020-12-02.

[61] X5GON. X5gon connect recommender. `https://platform.x5gon.org/products/feed#connect-recommender`. Accessed in: 2020-12-02.

[62] N. Erik S. Ayşe Saliha. Deliverable 4.6 - final prototype of cross-language recommendation engine. `https://www.x5gon.org/wp-content/uploads/2020/04/D4.6-Final-prototype-of-cross-language-recommendation-engine.pdf`. Accessed in: 2020-12-02.

[63] Marcus Perry. *The Exponentially Weighted Moving Average*. June 2010.

[64] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill(tm): A bayesian skill rating system. In *Advances in Neural Information Processing Systems 20*, pages 569–576. MIT Press, January 2007.

[65] Philip J. Guo, Juho Kim, and Rob Rubin. How video production affects student engagement: An empirical study of mooc videos. In *Proc. of the First ACM Conf. on Learning @ Scale*, 2014.

[66] Andrew S Lan, Christopher G Brinton, Tsung-Yen Yang, and Mung Chiang. Behavior-based latent variable model for learner engagement. In *Proc. of Int. Conf. on Educational Data Mining*, 2017.

[67] Zachary A Pardos, Ryan SJD Baker, Maria OCZ San Pedro, Sujith M Gowda, and Supreeth M Gowda. 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.

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

[69] Andy Lane. Open information, open content, open source. pages 158–168.

[70] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. Beyond views: Measuring and predicting engagement in online videos. In *Proc. of the Twelfth Int. Conf. on Web and Social Media*, 2018.

[71] YouTube. Creative commons. `https://support.google.com/youtube/answer/2797468?hl=en`. Accessed: 2020-12-01.

[72] W. Hoiles, A. Aprem, and V. Krishnamurthy. Engagement and popularity dynamics of youtube videos and sensitivity to meta-data. *IEEE Transactions on Knowledge and Data Engineering*, 29(7):1426–1437, 2017.