Extracting Topics from Open Educational Resources

Mohammadreza Molavi[1], Mohammadreza Tavakoli[2], and Gábor Kismihók[2]

[1] Amirkabir University of Technology, Tehran, Iran
mr.molavi@aut.ac.ir
[2] German National Library of Science and Technology (TIB), Hannover, Germany
{reza.tavakoli,gabor.kismihok}@tib.eu

Abstract. In recent years, Open Educational Resources (OERs) were earmarked as critical when mitigating the increasing need for education globally. Obviously, OERs have high-potential to satisfy learners in many different circumstances, as they are available in a wide range of contexts. However, the low-quality of OER metadata, in general, is one of the main reasons behind the lack of personalised services such as search and recommendation. As a result, the applicability of OERs remains limited. Nevertheless, OER metadata about covered topics (subjects) is essentially required by learners to build effective learning pathways towards their individual learning objectives. Therefore, in this paper, we report on a work in progress project proposing an OER topic extraction approach, applying text mining techniques, to generate high-quality OER metadata about topic distribution. This is done by: 1) collecting 123 lectures from *Coursera* and *Khan Academy* in the area of data science related skills, 2) applying *Latent Dirichlet Allocation* (*LDA*) on the collected resources in order to extract existing topics related to these skills, and 3) defining topic distributions covered by a particular OER. To evaluate our model, we used the data-set of educational resources from *Youtube*, and compared our topic distribution results with their manually defined target topics with the help of *3* experts in the area of data science. As a result, our model extracted topics with **79%** of *F1-score*.

Keywords: Open Educational Resource · OER · Topic Extraction · Text Mining · Machine Learning.

1 Introduction

The world of education is changing rapidly due to the growing needs of personalized services. The quickly evolving labour market, its dynamically changing knowledge and skills demands [9], the global challenges for work and education due to the emergence of the COVID-19 pandemic, are all examples that highlight the increasing need for flexible and personalised education. Consequently, we have been facing with an exponential growth of distributed and heterogeneous educational materials (such as Open Educational Resources (OERs)) [6,2]. Enormous amount of OERs are provided and receiving more and more attention from

learners every day. However, OER authors often fail to provide metadata for their content, as they consider this as a time-consuming activity [3]. This notion leads to low-quality OER metadata, even though that high quality metadata is crucial for organizing OERs [3] and providing search and recommendation services [7]. Indeed, the lack of high quality metadata is one of the most important factors limiting the effectiveness of (OER based) personalised informal learning [10]. The *Covered Topics* (i.e. covered knowledge areas) is one of the essential (often ignored) metadata for learning resources, as it helps learners to find the most suitable educational resource for their learning objectives [6,2]. Having a clear picture about *Covered Topics* is especially important for OER users, who want to build their own learning trajectory [10]. As a potential solution, intelligent algorithms should be tailored to extract and identify topics from OERs automatically, which emerged as a key issue in e-learning in the recent decade [11]. To tackle this issue, researchers have elaborated a number of methods already. For instance, they have applied machine learning methods on educational materials [6,2,3], or analyzed and transformed educational content into appropriate data structures (e.g. trees)[10] to extract a particular topic from educational resources. However, the generalisability of these proposed approaches, so far, has remained limited [6].

In this paper, we address the above mentioned challenges and propose an automatic topic extraction model focusing on data science related OERs. This is done by 1) collecting OERs about data science related skills (e.g. machine learning, text mining, sql language) from two pioneer educational repositories (*Coursera, Khan Academy*), 2) identifying topics that should be ideally covered in each skill using topic modeling techniques (i.e. *LDA*), and 3) building topic models to extract the distribution of *Covered Topics* for a given OER. We evaluated our model by using a data-set of open educational videos from Youtube, assigning their topics (as labels) with the help of *3* data science experts, and calculating *F1-score* (harmonic mean of precision and recall) of our automatic topic distribution extractor, using manually assigned topics as ground truth.

2 State of the Art

2.1 Semantic-based Methods

A number of studies use semantic methods and structured representation of data (such as taxonomies) to extract topics from educational resources [2]. For instance, [6] proposed a framework that combines semantic classification, taxonomies, and graph structures to extract topics and detect their relationships.

2.2 Text Mining Methods

Studies in this group analyse educational text and use text-related machine learning methods to detect topics in educational text [11]. For example, [10] created a system, which collects domain-specific content from online learning systems.

They extracted domain-specific terms by creating *Generalised Suffix Tree* (*GST*) from resources' text, and detected repeated sub-sequences as candidate terms to provide topic-specific recommendations for learners.

3 Method

3.1 Data Collection and Pre-processing

Target Skills To propose the first version of our approach, we extracted important skills for data science by mining relevant job vacancies [9, page_8]. In our online job vacancy data-set (from August 2019 to December 2019) the three most important data science skills were 1. *Machine Learning*, 2. *Text Mining*, and 3. *SQL Language*.

OER Resources In order to build our topic models, we collected 123 relevant online lectures (and their transcripts) from *Coursera*[3] and *Khan Academy*[4] related to our target skills (including 67 lectures for machine learning, 27 for text mining, and 29 for sql language). Moreover, to evaluate our proposed model, we used the data-set defined by [8, page_3] including 550 educational videos and their properties (e.g. rate, transcript, view-count) from *Youtube* in the area of data science. It should be mentioned that we applied the following pre-processing steps to prepare our collected OER transcripts for our analysis: 1) Removal of unimportant characters, punctuations, links, and stop words, and 2) Building TF-IDF representation.

3.2 Building Topic Models

To extract knowledge areas that are covered by particular educational resources in each of the target skills, we used *Latent Dirichlet Allocation* (*LDA*) [4]. *LDA* is a generative probabilistic topic model that considers each document as a distribution of different topics, each topic as a distribution of different words, and tries to extract existing topics together with their distribution of words for a corpus. To set the number of topics that *LDA* extracts, we calculated C_V *Coherence* [5] for different number of topics (between 2 to 10), and selected the topic amount with the highest coherence value. The following part of this section explains the detailed process of finding the most appropriate value of topic amounts, extracting topics, and assigning a name to topics (done manually, after executing *LDA* and based on the distribution of their words [1]) for *Text Mining* skill[5].

In order to build our topic models on *Text Mining* skill, we calculated C_V coherence for different number of topics on text mining-related educational resources, as shown on Figure 1. Based on the result, 7 topics provide us the best

[3] `https://www.coursera.org/`

[4] `https://www.khanacademy.org/`

[5] We report our result regarding topic amounts and evaluation for the other two skills: *Machine Learning* and *SQL Language*

coherence value[6]. Therefore, we set the parameter k of *LDA* to 7 and executed the analysis on our text mining corpus. The extracted topics, the assigned names, and the significant words of each topic are visible on Table 1.

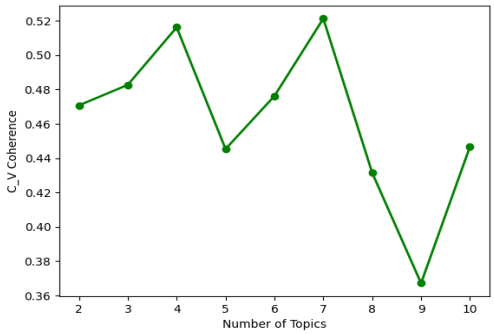

Fig. 1. C_V coherence for different number of topics in Text Mining corpus

Table 1. Output of LDA on Text Mining corpus

Topics	Assigned Name	Significant Words
$Topic_1$	Topic Modeling	lda plsa topic dirichlet parameters likelihood beta distribution alpha document
$Topic_2$	Sequence Models	prior string tag sequence markov hidden probabilities estimate position generate
$Topic_3$	Sentiment Analysis	features grams sentiment positive topic accuracy reviews negative idf tf
$Topic_4$	Matrix Factorization	matrix topic matrices squared diagonal svd factorization vectors approximation document
$Topic_5$	Text Classification	grams convolutional filter sentence corpus vec embeddings neural count modeling
$Topic_6$	Probabilistic Models	naive prior bayes given probability likelihood independent maximizes predicted significantly
$Topic_7$	Text Process & Feature Extraction	sentence document frequency features phrase vec grammar parse grams term

4 Topic Model Extraction Evaluation

To evaluate our topic models, we used our Youtube data-set in which topics were assigned to videos manually. This manual assignment was done by 3 data science experts with at least 2 years of teaching experience and 5 years of industrial experience in data science related areas. It should be mentioned that each participant allocated at least 2 minutes for analysing each of the videos. Afterwards, we applied our topic extraction models on each video transcript (e.g. apply our machine learning topic model on the related educational videos on machine learning). Finally, we compared the manually assigned topics (by experts) and the output of our topic extraction models. As a result, we were able to determine the quality of our topic extraction models in relation to manual, expert topic assignments. We got *F1-score* of each topic extraction model as follows: *Text Mining*: *76%*, *Machine Learning*: *81%*, and *SQL Language*: *78%*. Therefore, our models were able to extract covered topics of educational resources with F1-score of **79%** in average.

[6] We did the same process and set C_V of other skills as follows: *Machine Learning*: 9, *SQL Language: 5*

5 Conclusion and Future Work

This study is one of the steps towards 1) dynamic definition of topics that should be covered in a particular knowledge areas, and 2) extracting the topic distribution for a given OER, as one of the most important metadata, to help learners to build their own learning path. We collected 123 educational lectures from two repositories related to 3 data science related skills. After that we applied *LDA* on the lectures' transcripts to extract the topic model for each skill. Finally, to evaluate the models, we used an educational Youtube data-set, assigned covered topics with the help of 3 data science experts. Subsequently, we applied our topic extraction models, and compared the output of our model with the manually assigned topics. This exercise revealed that our models can extract topics with *F1-score* of **79%**. As the next steps, we plan to add more educational resources to improve our models and also, apply our approach for other skills/knowledge.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3**(Jan), 993–1022 (2003)
2. de Carvalho Saraiva, M., Medeiros, C.B.: Finding out topics in educational materials using their components. In: 2017 IEEE Frontiers in Education Conference (FIE). pp. 1–7. IEEE (2017)
3. García-Floriano, A., Ferreira-Santiago, A., Yáñez-Márquez, C., Camacho-Nieto, O., Aldape-Pérez, M., Villuendas-Rey, Y.: Social web content enhancement in a distance learning environment: intelligent metadata generation for resources. International Review of Research in Open and Distributed Learning **18**(1), 161–176 (2017)
4. Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L.: Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. Multimedia Tools and Applications **78**(11), 15169–15211 (2019)
5. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on Web search and data mining. pp. 399–408 (2015)
6. Saraiva, M.d.C., et al.: Relationships among educational materials through the extraction of implicit topics: Relacionamentos entre materiais didáticos através da extração de tópicos implícitos (2019)
7. Tavakoli, M., Elias, M., Kismihók, G., Auer, S.: Quality prediction of open educational resources a metadata-based approach. arXiv preprint arXiv:2005.10542 (2020)
8. Tavakoli, M., Hakimov, S., Ewerth, R., Kismihók, G.: A recommender system for open educational videos based on skill requirements. arXiv preprint arXiv:2005.10595 (2020)
9. Tavakoli, M., Mol, S.T., Kismihók, G.: Labour market information driven, personalized, oer recommendation system for lifelong learners. arXiv preprint arXiv:2005.07465 (2020)
10. Wang, J., Xiang, J., Uchino, K.: Topic-specific recommendation for open education resources. In: International Conference on Web-Based Learning. pp. 71–81. Springer (2015)

11. Xie, M., Wu, C., Zhang, Y.: A new intelligent topic extraction model on web. JCP **6**(3), 466–473 (2011)