Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 18(4): 4450–4460.
DOI: 10.3934/mbe.2021225
Received: 04 January 2021
Accepted: 16 April 2021
Published: 24 May 2021

Research article

Mathematical modeling and mining real-world Big education datasets with application to curriculum mapping

Kah Phooi Seng[1,*], Fenglu Ge[2] and Li-minn Ang[3]

[1] School of Engineering and Information Technology, UNSW Canberra, ACT 2612, Australia
[2] Pacific Telecom & Navigation Ltd., Hong Kong
[3] School of Science and Engineering, University of Sunshine Coast, Petrie, QLD 4502, Australia

* **Correspondence:** Email: kseng@unsw.edu.au.

Abstract: This paper proposes an approach for modeling and mining curriculum Big data from real-world education datasets crawled online from university websites in Australia. It addresses the scenario to give a student a study plan to complete a course by accumulating credits on top of subjects he or she has completed. One challenge to be addressed is that subjects with similar titles from different universities may put barriers for setting up a reasonable, time-saving learning path because the student may be unable to distinguish them before an intensive research on all subjects related to the degree from the universities. We used concept graph-based learning techniques and discuss data representations and techniques which are more suited for large datasets. We created ground truth of subjects relations and subject's description with Bag of Words representations based on natural language processing. The generated ground truth was used to train a model, which summarizes a subject network and a concepts graph, where the concepts are automatically extracted from the subject descriptions across all the universities. The practical challenges to collect and extract the data from the university websites are also discussed in the paper. The work was validated on nineteen real-world education datasets crawled online from university websites in Australia and showed good performance.

Keywords: graph learning; learning analytics; curriculum mapping; learning technologies; Big data

1. Introduction

Curriculum mapping is an important tool of modern educational design. A clearly mapped curriculum is able to show the links and relationships amongst the different courses in the curriculum to students and other stakeholders. An important part for the educational curriculum design of a course is to determine the prerequisites for the subjects contained in the course [1]. The prerequisite knowledge can be defined as the skills and information which are necessary for a student to succeed in a given instructional subject within a curriculum [2]. The subject prerequisites serve two purposes: 1) It ensures that a student has the necessary background knowledge before undertaking the subject to be studied; and 2) It forms a scaffolding for learning amongst the subjects so that students are led to take subjects with increasing levels of complexity (e.g., beginning from introductory subjects, to subjects with developing materials, and ending with subjects requiring more complex materials). This issue is particularly important for the rising popularity of online-based education [3] and the emergence of Massive Online Open Courses (MOOCs) [4]. One challenge to be addressed is that subjects with similar titles from different universities may put barriers for setting up a reasonable, time-saving learning path because the student may be unable to distinguish them before an intensive research on all subjects related to the degree from all universities. Therefore, it would be necessary to supply optimized relations amongst unlearned subjects based on the subjects previously learned.

Several authors have proposed mathematical models for linking the dependencies of courses within a curriculum to identify useful relationships and connections. The authors in [5] used a complex systems approach to propose a mathematical model termed as the Curriculum Prerequisite Network. In their work, the authors used a mathematical graph theory model (directed acyclic graph) where the nodes were used for representing the courses and the links between nodes represented the course prerequisites. Another approach was taken by authors in [6] which used KDD (Knowledge Discovery in Databases) methods to mine useful information from a medical and healthcare curriculum. Their methodology used the CRISP-DM (Cross-Industry Standard Process for Data Mining) reference model [7] and investigated the following queries: 1) Educational disciplines which do not conform to the curriculum data; 2) Identification of overlapping areas across the curriculum; 3) Identification of discipline clusters within the curriculum; and 4) Identification of courses which belong to important parts of the curriculum.

Some recent approaches and investigations for mining prerequisite information from education curriculum datasets have been proposed by authors in [8,9]. The authors in [8] proposed an approach termed as concept graph learning (CGL). In this work, the authors identified an issue where the course descriptions from different universities may contain descriptions and terms which vary amongst the universities. The authors proposed using an intermediate mapping termed as "universal concepts". The key idea is that courses from different universities would use the same terms in the universal concepts mapping and this allows courses from different universities to be mapped to the common space. The model can then be used to predict implicit prerequisites amongst the various courses. Their work was validated from educational and curriculum datasets retrieved from four universities in the United States. The authors in [9] presented a survey of analytics and techniques for Big education datasets. The work also reported on graph learning for real-world educational datasets from four universities in Australia. The works in [8,9] have demonstrated the validity of mining curriculum data from education datasets. However, the full potential and various challenges of the curriculum mapping has not been investigated as only the data from four universities have been considered in the earlier works.

Considering the above challenges, in this paper, we discuss an approach for modeling and mining real-world Big education datasets with application to curriculum mapping. We use concept graph-based learning techniques, discuss data representations and techniques which are more suited for large datasets, and validate the work on nineteen real-world education datasets crawled online from university websites in Australia. The practical challenges to collect and extract the relevant data from the university websites are also discussed in the paper. The structure of this paper is as follows. Section 2 introduces the description of education graph learning and the data pre-processing. The proposed graph-based learning approach is discussed in Section 3. Section 4 gives discussions for the experimental results and validates the effectiveness of the proposed approach. Section 5 gives some concluding remarks.

2. Description of education graph learning and data pre-processing

2.1. Description of education graph learning

The overview of the education graph learning process is shown in Figure 1. The top part of the diagram shows examples of subjects in different universities. The second part of the diagram shows the Principal Concepts Pool (PCP), where the concepts of each subject are extracted by natural language processing (NLP) in the form of keywords. Given the ground truth data of prerequisite relations amongst subjects and the groups of concepts (keywords) of each subject, the relations of keywords between different group can be established, where every group of keywords represent one subject. After training based on all ground truth data, the relations of concepts will be achieved. Such learned relations amongst concepts are capable of inferring relations amongst subjects. For example, if many of the concepts of subject 1 are prerequisites of concepts of subject 2, it can be concluded that subject 1 should be a prerequisite of subject 2.

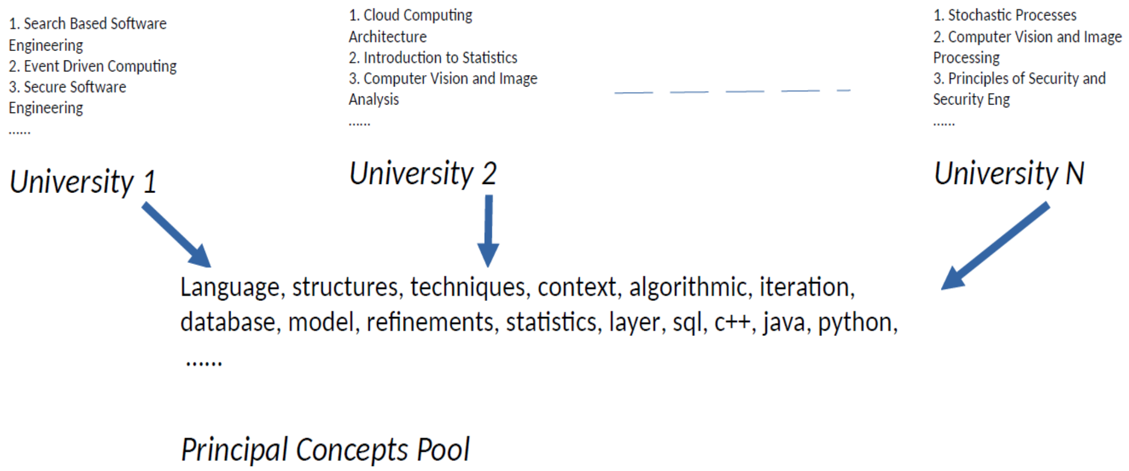
1. Search Based Software Engineering
2. Event Driven Computing
3. Secure Software Engineering
……
University 1
1. Cloud Computing Architecture
2. Introduction to Statistics
3. Computer Vision and Image Analysis
……
University 2
1. Stochastic Processes
2. Computer Vision and Image Processing
3. Principles of Security and Security Eng
……
University N
Language, structures, techniques, context, algorithmic, iteration, database, model, refinements, statistics, layer, sql, c++, java, python,
……
Principal Concepts Pool

Figure 1. Overview of education graph learning process.

2.2. Data pre-processing and representation

The experimental data was provided by real-world datasets crawled online from nineteen Australian university websites. The experimental data was obtained by regular web scraping techniques using Python on the respective university subject data websites which are available on the Internet. There were several challenges that had to be addressed for the data pre-processing. A practical challenge that was faced during the data collection process was to dynamically extract the generated subject data from the university websites. This process known as web-scraping was accomplished using Selenium [10]. A second challenge which was encountered was to clean the data to make it suitable for the experimental analytics. The raw data was cleaned by four methods including: 1) Conversion of data to lowercase—all characters were converted to lowercase; 2) Tokenization—sentences were broken down into tokens where each token represents a word, symbol or number; 3) Word stemming—removing the redundancy over words by conversion to their root words (e.g., conversion of "calculating", "calculated" to the root word "calculate"; and 4) Removal of stop words (e.g., "and", "the", "that"). The data representation used the Bag of Words (BoW) [11,12] approach. The BoW is a popular representation technique used for NLP to extract key attributes from large-scale text-based documents and objects where histograms of words are used as features to perform the classification.

The details of the experimental data are shown in Table 1. The first column shows the names of the universities in the experiment. The second column shows the number of subjects related to Computer Science (CS) and Information Technology (IT) in each university. The third column shows the total number of prerequisite relations relevant to CS and IT subjects in each university. The fourth column shows the number of key words retrieved from all CS and IT subjects in each university. The PCP contained a total of 30,840 concept keywords.

Table 1. The details of subject datasets used in the experimental work.

University	Subjects	Prerequisites	Key Words
University of Adelaide	75	110	1326
University of Canberra	110	29	1760
Central Queensland University	75	81	1256
Curtin University	73	56	1067
Deakin University	87	79	1723
James Cook University	61	52	1686
La Trobe University	71	78	966
RMIT and Melbourne Technical College	106	81	1897
University of Melbourne	49	88	1331
University of South Australia	67	39	1226
University of Western Australia	88	105	2324
University of Sydney	131	41	2518
Western Sydney University	96	93	1524
University of New South Wales	84	142	1603
University of Wollongong	84	72	1829
University of Queensland	97	101	1222
University of Tasmania	55	45	1075
University of Technology Sydney	80	86	1343
Victoria University	64	44	1949
Total number of concept keywords:		30840	

3. Methods

3.1. Notation and methodology

Table 2 shows a summary of the notations and terminology following the symbols used by the authors in [8]. Given a training set of courses with a bag-of-concepts representation per course as a row in matrix *X*, and a list of known prerequisite links per course as a row in matrix *Y*, the objective is to optimize the matrix *A* whose elements specify both the direction (sign) and the strength (magnitudes) of each link between concepts.

Table 2. Summary of notations and terminology.

Notation	Description
n	Number of subjects in a training set
c	Dimension of the universal concept space
$X = [x_1, x_2, ... x_n]$	$\in \mathbb{R}^{n\times c}$ is a set of *n* subjects, where $x_i \in \mathbb{R}^c$ is the BoW representation of the *i*-th subject
Y	*n*-by-*n* matrix where each entry represents relations of two subjects. $y_{ij} = 1$ if the *i*-th subject is a prerequisite of the *j*-th subject, otherwise $y_{ij} = -1$
A	$\in \mathbb{R}^{c\times c}$ where a_{ij} represents the strength of relations between the *i*-th concept/keyword and the *j*-th concept/keyword.

The relation between two subjects (subject *i* and subject *j*) can be expressed as:

$$F_{ij} = x_i^T A x_j \tag{1}$$

The objective is to find a matrix A by minimizing the cost function:

$$\sum_{ij} max((1 - y_{ij}F_{ij})^2) \ + \ \frac{\alpha}{2}\sqrt{Tr(AA^H)} \tag{2}$$

where, Tr represents the trace of a matrix and H is the conjugate transpose. We discuss two approaches to perform the optimization: 1) Rank approach; and 2) Learning a sparse concept graph.

3.2. Rank approach

It should be noticed that c or the length of x is large, and increases the computation cost of Eq (1). Therefore, it becomes necessary to solve the following problem.

$$\sum_{ij} max((1 - y_{ij}F_{ij})^2) \ + \ \frac{\alpha}{2}\sqrt{Tr(AA^H)} \ + \ < F - XAX^T, M > \tag{3}$$

$$F - XAX^T = 0 \tag{4}$$

The inner product in the angle brackets of Eq (3) is 0. The gradient of the Eq (3) over A is

$$\alpha A - X^T MX = 0 \tag{5}$$

An observation is that X^T is a $c \times n$ matrix, and the matrix M is a $n \times n$ matrix. Therefore, we only have to calculate the M matrix, then the matrix A can be projected back by Eq (5). The Rank approach algorithm based on Projected Gradient Decent (PGD) [8] is shown in Figure 2.

```
$K = XX^T ,M = 0, Q = 0$
$t = 1$
While not converge do
    $F = K\frac{M}{\alpha}K$
    $P = \frac{M}{\alpha} - \eta(\alpha F - 2K \triangle K)$
    $B = P + \frac{t-1}{t+2}(P - Q)/$
    $Q = P$
    $t = t + 1$
    $A = X\frac{M}{\alpha}X^T$
```

Figure 2. Rank approach algorithm (Algorithm 1).

3.3. Learning a sparse concept graph

A finding from the data collected from the 19 universities is that X is a sparse matrix. The matrix A is also a sparse matrix as many concepts have no relations between each other. Therefore, an optimization could be performed by calculating the cost function based on sparse matrices. The cost function is changed as shown in Eq (6).

$$\sum_{ij} max((1 - y_{ij}F_{ij})^2) \quad + \quad \alpha \sum_{ij} |a_{ij}| \tag{6}$$

The Learning a sparse concept graph algorithm based on Projected Gradient Decent (PGD) is shown in Eq (7) and Figure 3. For sparse techniques involving tree-structured data, the technique in [13] can be considered.

$$\mathrm{P^k = S_{\alpha t_k}(A^{k-1} - t_k)\nabla_g(A^{k-1})} \tag{7}$$

```
$A = 0_{c\times c}, P = 0_{c\times c}, Q = 0_{c\times c}$
$k = 1$
While not converge do
    $F = XAX^T$
    For j =1, 2,...,p do
        $P_j = S_{\alpha t_k}(A_j + 2t_k X^T \triangle X_j)$
    $A = P + \frac{k-1}{k+2}(P - Q)$
    $Q = P$
    $k = k + 1$
```

Figure 3. Learning a sparse concept graph algorithm (Algorithm 2).

4. Experimental results and discussion

4.1. Within-university prediction results and discussion

The crawled and pre-processed data from Section 2.2 were split into training sets and test sets. For an initial investigation, the models were used for within-university prediction. In this experiment, the models were trained from subject data from one university, and tested by the data from the same university. In this case, one third of the data set was used to train the model, another one third was used for validation, and the final one third was used to test the model. Validation was performed using three-fold cross validation. Two metrics MAP (mean average precision) and AUC (area under the curve) were used to evaluate the performance of the algorithms. MAP [14] is the average precision which is a popular metric used to measure the performance of models for the applications or tasks such as document/information retrieval, object detection. AUC [15] is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the receiver operating characteristic (ROC) curve, which is a graph showing the performance of a classification model at all classification thresholds. Table 3 shows the within-university performance for Algorithms 1 and 2. The following observations can be made from the comparisons in Table 3: 1) Algorithm 1 gave better performance than Algorithm 2 for the majority of the universities tested; and 2) In general, the AUC scores were higher than the MAP scores for the within-university performance. Based on the observations in 1) and 2), we will focus on using Algorithm 1 with the AUC metric for further investigations.

Table 3. Experimental results for within-university performance.

University name	Algorithm 1 (Rank)		Algorithm 2 (Sparse)	
	AUC	MAP	AUC	MAP
University of Adelaide (UoA)	0.878	0.610	**0.908**	**0.654**
University of Canberra (UCanberra)	0.273	0.048	**0.907**	**0.355**
Central Queensland University (CQU)	**0.864**	**0.483**	0.706	0.304
Curtin University (Curtin)	**0.798**	**0.505**	0.488	0.399
Deakin University (Deakin)	**0.672**	**0.278**	0.543	0.261
James Cook University (JCU)	**0.741**	0.247	0.174	**0.405**
La Trobe University (LaTrobe)	**0.793**	**0.454**	0.267	0.137
Royal Melbourne Institute of Technology (RMIT)	0.758	**0.422**	**0.771**	0.380
University of Melbourne (UMelbourne)	**0.680**	**0.318**	0.368	0.290
University of South Australia (UNISA)	**0.663**	**0.228**	0.524	0.141
University of Western Australia (UWA)	**0.884**	**0.528**	0.612	0.211
University of Sydney (USydney)	0.579	**0.125**	**0.592**	0.089
University of New South Wales (UNSW)	**0.928**	**0.686**	0.766	0.419
Western Sydney University (WSU)	**0.907**	**0.558**	0.722	0.353
University of Wollongong (UWollongong)	**0.863**	**0.486**	0.826	0.434
University of Queensland (UQ)	**0.773**	**0.383**	0.393	0.262
University of Tasmania (UTasmania)	**0.918**	**0.440**	0.647	0.344
University of Technology Sydney (UTS)	**0.738**	**0.431**	0.451	0.273
Victoria University (VU)	**0.552**	0.118	0.467	**0.132**

Table 4. Experimental results for cross-university performance.

University	UoA	UCan	CQU	Curt	Deak	JCU	LaTr	RMIT	UMel	UNSA	UWA	USyd	UNS	WSU	UWol	UQ	UTas	UTS	VU
UoA	0.878	0.907	0.403	0.405	0.359	0.300	0.493	0.746	0.389	0.539	0.527	0.419	0.421	0.578	0.380	0.458	0.691	0.232	0.416
UCanberra	0.345	0.273	0.260	0.384	0.524	0.285	0.428	0.409	0.250	0.502	0.42	0.283	0.209	0.370	0.410	0.302	0.464	0.215	0.309
CQU	0.271	0.745	0.864	0.169	0.485	0.454	0.310	0.270	0.118	0.176	0.338	0.125	0.302	0.231	0.227	0.196	0.506	0.196	0.275
Curtin	0.556	0.614	0.348	0.798	0.295	0.176	0.335	0.282	0.260	0.095	0.261	0	0.238	0.396	0.372	0.024	0.375	0.242	0.232
Deakin	0.609	0.602	0.219	0.378	0.672	0.522	0.344	0.479	0.497	0.337	0.492	0.022	0.461	0.380	0.431	0.349	0.679	0.235	0.606
JCU	0.152	0.440	0.243	0.085	0.280	0.741	0.169	0	0.257	0.125	0.199	0	0.195	0.216	0.164	0.072	0.071	0.177	0.094
LaTrobe	0.260	0.703	0.253	0.087	0.260	0.227	0.793	0.134	0.155	0.223	0.201	0	0.098	0.151	0.212	0.065	0.191	0.127	0.045
RMIT	0.535	0.857	0.513	0.608	0.486	0.368	0.530	0.758	0.287	0.429	0.697	0.338	0.635	0.574	0.464	0.519	0.655	0.371	0.309
UMelbourne	0.321	0.425	0.263	0.281	0.292	0.446	0.307	0.195	0.680	0.213	0.465	0	0.287	0.288	0.262	0.210	0.452	0.246	0.057
UNISA	0.808	0	0.222	0.302	0.141	0.044	0.056	0.616	0.240	0.663	0.287	0.305	0.293	0.331	0.161	0.393	0.441	0.034	0.082
UWA	0.416	0.348	0.215	0.396	0.451	0.768	0.380	0.263	0.551	0.238	0.884	0.125	0.602	0.214	0.576	0.377	0.458	0.491	0.703
USydney	0.199	0.351	0.289	0.124	0.143	0.242	0.299	0.392	0.243	0.407	0.130	0.579	0.327	0.331	0.213	0.175	0.333	0.255	0.241
UNSW	0.658	0.819	0.599	0.608	0.514	0.398	0.468	0.653	0.510	0.600	0.693	0.489	0.928	0.689	0.620	0.500	0.702	0.432	0.547
WSU	0.792	0.795	0.499	0.771	0.586	0.483	0.609	0.735	0.439	0.539	0.639	0.338	0.654	0.907	0.610	0.568	0.685	0.454	0.516
UWollongong	0.367	0.378	0.362	0.307	0.383	0.224	0.324	0.574	0.405	0.586	0.554	0.092	0.623	0.531	0.863	0.371	0.655	0.362	0.198
UQ	0.367	0.637	0.573	0.304	0.417	0.459	0.369	0.392	0.439	0.546	0.484	0.206	0.311	0.468	0.350	0.773	0.679	0.309	0.303
UTasmania	0.611	0.355	0.154	0.220	0.220	0.029	0.324	0.295	0.351	0.267	0.491	0	0.326	0.135	0.271	0.093	0.918	0.236	0.212
UTS	0.456	0.502	0.170	0.515	0.456	0.278	0.400	0.269	0.500	0.418	0.542	0.335	0.479	0.389	0.569	0.217	0.530	0.738	0.337
VU	0.120	0.595	0.083	0.174	0.297	0.068	0	0.345	0.399	0.044	0.441	0.191	0.153	0.063	0.077	0.134	0.411	0.146	0.552

4.2. Cross-university prediction results and discussion

For a next investigation, the models were used for cross-university prediction. In this experiment, the models were trained from subject data from one university, and tested by the data from a different university. The keywords from the various universities were combined to create a dictionary which was then used for coding concepts for all universities. Table 4 shows the cross-university performance using Algorithm 1 and the AUC metric. The following observations can be made from the comparisons in Table 4: 1) For most universities, the inter-university performance (shown by the diagonals of the matrix) gave better prediction performance that the cross-university performance. However, in some cases, the cross-university performance gave better performance. This shows that having universal concepts and keywords amongst different universities could allow for significant cross-university prediction (i.e., transfer learning amongst universities, e.g., UoA and UCan which gave a performance of 0.907); and 2) There were some cross-university AUC results which gave zero values. This shows the converse situation where the different universities do not have any keywords or have very few keywords in common. In this case, the subject data learnt from one university is not able to help in the prediction performance for the other university.

For a final investigation, we performed a comparison of the number of keywords versus the sum of the predictive performance for all universities. Figure 4 shows a summary of the results for the various universities. Although, there were universities with many keywords which gave high predictive performance, there were also universities with fewer keywords which gave comparable performance (e.g., USydney with 2518 keywords gave a performance of 4.694, however UTasmania with 1075 keywords gave a comparable performance of 4.590). This shows that the keywords which are used are more important than the total number of keywords. It is important to note that the experimental data was automatically provided by real-world datasets crawled online from Australian university websites. No further pre-processing or keyword selection was performed other than that discussed in Section 2.2.

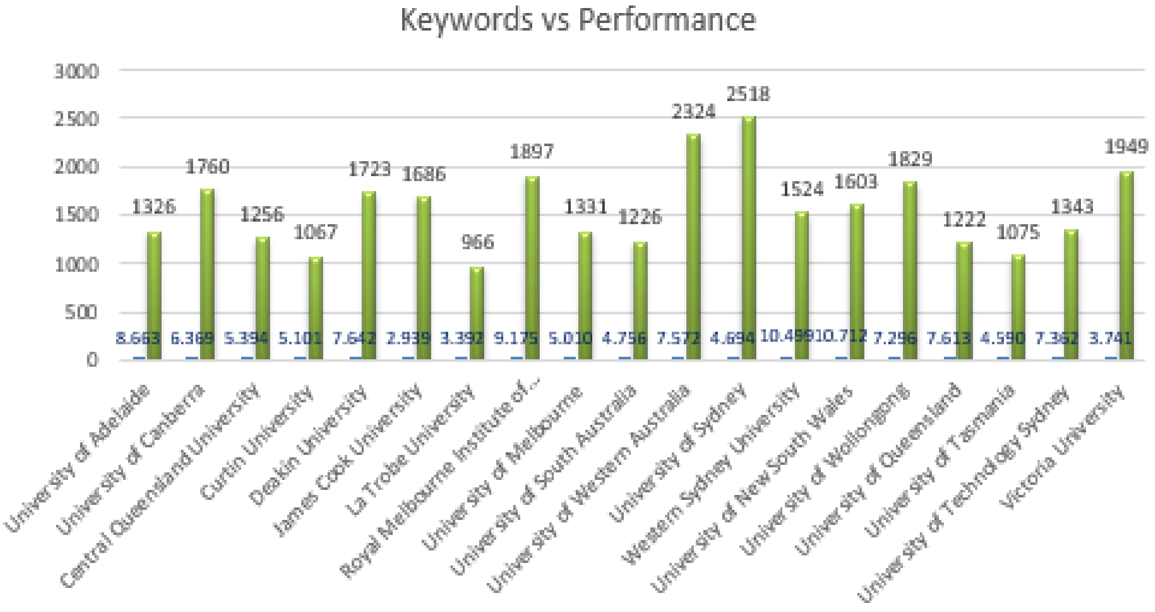

Figure 4. Comparison of number of keywords vs performance.

5. Conclusions

This paper has proposed an approach for modeling and mining curriculum Big data from real-world education datasets crawled online from university websites. The practical scenario is to enable students to have a study plan to complete a course by accumulating credits on top of subjects he or she has already completed from different universities. Other than the practical use for students, the work also demonstrates the value of transfer learning amongst different universities for Big education data where the subject data from universities can be used for predicting pre-requisite links amongst subjects from other universities. The work has been validated on nineteen (50% of the total number of public universities in Australia) real-world education datasets from university websites in Australia and showed good performance. Our future work aims to extend the investigation to all 38 public universities in Australia, and also investigate the transfer learning for universities in different countries.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. G. O'Neill, *Curriculum design in higher education: theory to practice*, Research Repository UCD, University College Dublin, Teaching and Learning, 2015.
2. A. Vuong, T. Nixon, Brendon Towle, A method for finding prerequisites within a curriculum, in *Proceedings of the 4th International Conference on Educational Data Mining*, (2011), 211–216.
3. P. Hill, Online educational delivery models: a descriptive view, *Educause Rev.*, **47** (2012), 84–86.
4. L. Yuan, S. J. Powell, *MOOCs and open education: Implications for higher education*, Cetis, 2013.
5. P. R. Aldrich, The curriculum prerequisite network: modeling the curriculum as a complex system, *Biochem. Mol. Biol. Edu.*, **43** (2015), 168–180.
6. M. Komenda, M. Víta, C. Vaitsis, D. Schwarz, A. Pokorná, N. Zary, et al., Curriculum mapping with academic analytics in medical and healthcare education, *Plos One*, **10** (2015), e0143748.
7. R. Wirth, J. Hipp, CRISP-DM: Towards a standard process model for data mining, in *Proceedings of the 4th International Conference Practical Applications of Knowledge Discovery and Data Mining,* (2000), 29–39.
8. H. Liu, W. Ma, Y. Yang, J. Carbonell, Learning concept graphs from online educational data, *J. Artif. Intell. Res.*, **55** (2016), 1059–1090.
9. K. L. M. Ang, F. L. Ge, K. P. Seng, Big educational data & analytics: Survey, architecture and challenges, *IEEE Access*, **8** (2020), 116392–116414.
10. S. Avasarala, *Selenium WebDriver Practical Guide*, Packt Publishing Ltd, 2014.
11. Y. Zhang, R. Jin, Z. H. Zhou, Understanding bag-of-words model: a statistical framework, *Int. J. Mach. Learn. Cybern.*, **1** (2020), 43–52.
12. Y. HaCohen-Kerner, D. Miller, Y. Yigal, The influence of preprocessing on text classification using a bag-of-words representation, *Plos One*, **15** (2020), e0232525.

13. H. Zhang, S. Wang, X. Xu, T. W. S. Chow, Q. M. J. Wu, Tree2Vector: learning a vectorial representation for tree-structured data, *IEEE Trans. Neural Networks Learn. Syst.*, **29** (2018), 5304–5318.
14. Breaking Down Mean Average Precision (MAP). Available from: https://towardsdatascience.com/breaking-down-mean-average-precision-map-ae462f623a52.
15. J. Huang, C. X. Ling, Using AUC and accuracy in evaluating learning algorithms, *IEEE Trans. Knowl. Data Eng.*, **17** (2005), 299–310.