IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol. 001, LIIR036
DOI: https://doi.org/10.52731/liir.v001.036

Estimating the Difficulty of Courses from Syllabuses

Masako Furukawa [*], Yoshitomo Yaginuma [†]

Abstract

Syllabuses of many universities are published on the Web, and in some cases, past average scores are also published for the convenience of student registration. In this study, we investigate the relationship between the content of the syllabuses and the average scores, and clarify whether the difficulty level of the courses can be estimated from the syllabuses. In the proposed method, first, the topics that make up the syllabuses are extracted using topic analysis. Then, we find out how much each syllabus contains these topics. The difficulty of courses is estimated from these features using support vector machine. 10-fold cross-validation was performed for the evaluation, and it became clear that the difficulty level can be estimated with an accuracy of 62.3%.

Keywords: Syllabus, Topic modeling, Support vector machine.

1 Introduction

In face-to-face courses, the difficulty level can be adjusted while watching the student's reaction. On the other hand, in the case of online courses, teaching materials are often prepared in advance, so it is desirable to know the difficulty level when preparing the syllabuses, if possible.

There have been many studies on collecting and analyzing syllabuses on the Web. Moretti et al. proposed a system that analyzes syllabuses and data that evaluate subjects and teachers on the Web and used it for curriculum design [1]. Yasukawa et al. collected 6,493 syllabus documents from a national university in Japan and analyzed the data in order to clarify what kind of information must be included in a syllabus [2]. Sekiya et al. improved the topic analysis method called LDA (Latent Dirichlet Allocation), and analyzed the syllabuses of 10 American universities [3].

On the other hand, syllabuses of many universities are published on the Web, and in some cases, past average scores are also published for the convenience of student registration. Therefore, in this study, we investigate the relationship between the content of the syllabuses and the average scores, and clarify whether the difficulty level of the courses can be estimated from the syllabuses.

[*] National Institute of Informatics, Tokyo, Japan
[†] The Open University of Japan, Chiba, Japan

2 Estimating the Difficulty of Courses from Syllabuses

2.1 Proposed Method

In the proposed method, the difficulty level is estimated on the assumption that the content of the syllabus is related to the difficulty level, for example, the content related to mathematics has a low test score. Figure 1 shows the method for estimating the difficulty of courses from syllabuses. First, the topics that make up the syllabuses are extracted using topic analysis. Then, we find out how much each syllabus contains these topics. The difficulty of courses is estimated from these features using SVM (Support Vector Machine).

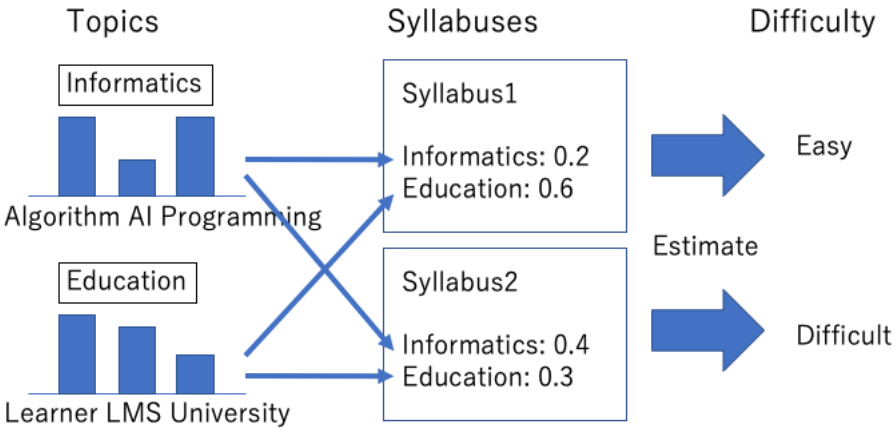

Figure 1: Proposed Method.

2.2 Extraction of Topics

We collected 183 syllabuses from a university that published syllabuses and average scores on the Web. These syllabuses contain information on average scores for the second semester of 2019 and the first semester of 2020. These 183 syllabuses were used to extract important keywords. Morphological analysis was performed, and only nouns that appeared more than once were extracted. As a result, 7286 independent nouns were obtained.

Next, TF-IDF was calculated for these nouns. TF is the number of occurrences of words, and its value becomes large when the word appears frequently. IDF is calculated by log of (total number of documents / number of documents including the word), and its value becomes large when the word appears only in a few specific documents. TF-IDF is the product of these values and is used as an index of important keywords. Since job titles and names of lecturers appear frequently and the value of TF-IDF increases, these words were excluded manually, and 100 words were extracted in descending order of TF-IDF. These words are Education, Nursing, Museum, Information, Politics, Psychology, Exhibition, Care, Support, Technology, and so on. Each syllabus can be represented by a 100-dimensional vector in which the number of occurrences of these words is arranged.

Then, topic analysis was performed using LDA, which is one of the algorithms used for topic extraction. Because the 183 syllabuses are not sufficient for topic analysis, 1938 syllabuses were additionally acquired from 4 universities, and a total of 2121 syllabuses were used for the extraction of topics.

Table 1 shows the topics extracted by the LDA when the number of topics is set to 10. Each extracted topic is a vector of 100 dimensions, and Table 1 shows the top 10 keywords that make up each topic. For example, Topic 4 is composed of keywords such as Data, Structure, Web, Ethics, Problems, and Issues. So, this topic is considered to be related to information and ethics. On the other hand, Topic 10 contains keywords such as Information, Research, Analysis, Society, Care, and Economy, and is considered to be related to the social use of information.

Table 1: Extracted Topics

Topics	Top 10 Keywords
1	Buddhism, Modernity, Thought, Agenda, Society, Language, Museum, Nature, Environment, Problems
2	Education, Problems, Psychology, Society, Economy, Issues, Language, Research, Information, Examination
3	Activities, Companies, Issues, Support, Society, Life, Development, Health, Language, Management
4	Ethics, Problems, Issues, Data, Examination, Structure, Utilization, Web, Research, Universe
5	Activities, Problems, International, Society, Volunteers, Issues, Policies, Economy, Language, Companies
6	Activities, Communities, Society, Nature, Challenges, Economy, Arts, Research, Health, Elderly
7	Society, Disability, Politics, Issues, Language, Communication, Research, Economy, International, Problems
8	Health, Exercise, Technology, Language, Tasks, Commentary, Guests, Management, Communication, Life
9	Society, Development, Language, Issues, Research, Guests, Problems, Commentary, Activities, Organizations
10	Issues, Research, Information, Society, Care, Economy, Management, Technology, Sentences, Analysis

By performing topic analysis, we can also know the proportion of topics contained in each syllabus. For example, the course of "Introduction to Multimedia" was calculated to include

topic 4 at a ratio of 0.11 and topic 10 at a ratio of 0.88. In this way, each syllabus can be expressed as a 10-dimensional vector representing topic proportions, and these 10-dimensional vectors are used as features when estimating the difficulty level.

2.3 Estimation of Difficulty of Courses

There are various possible indicators of the difficulty of courses, but we simply used the average scores of the courses. The 183 courses have information on the average scores for the second semester of 2019 and the first semester of 2020. Regarding the average scores in the first semester of 2020, the test method was changed due to the influence of COVID-19, and the average scores became higher and the difference between courses became smaller. Therefore, the scores in the second semester of 2019 were used to determine the difficulty level.

Figure 2 shows the distribution of average scores of 183 courses in the second semester of 2019. The overall average was calculated to be 73.7. Here, if the average score of a course is equal to or higher than this score, it is labeled as "Easy", otherwise it is labeled as "Difficult".

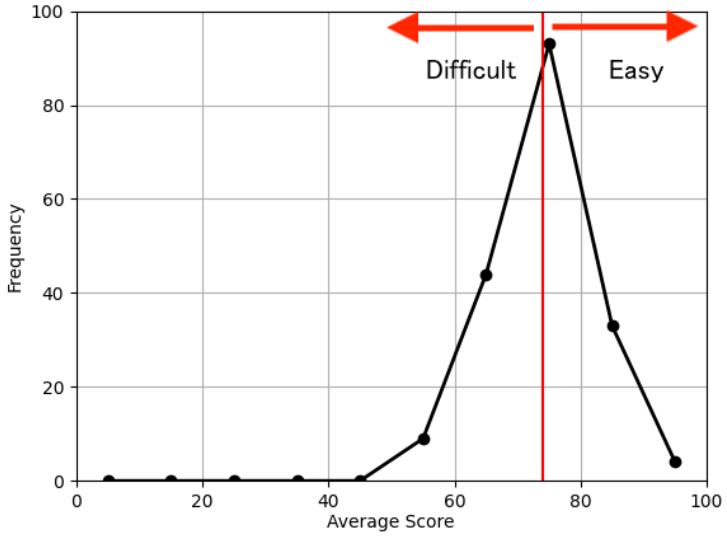

Figure 2: Distribution of Average Scores.

Then, we clarify how "Easy" and "Difficult" of the 183 syllabuses can be estimated from the 10-dimensional vectors representing topic proportions. The support vector machine was used for this estimation. Since the support vector machine converts the data into a high-dimensional space and then classifies the data, the accuracy of the estimation is generally higher than that of a simple linear discriminant analysis. When the training data and the evaluation data were the same, the accuracy was 69.4%.

In order to avoid the influence of overfitting, we performed 10-fold cross-validation, which is a method to obtain accuracy by separating training data and evaluation data. In this method, the data are divided into 10 subsets, and 9 subsets are used for learning, and 1 subset is used for evaluation. This process is repeated 10 times, and the average accuracy is taken as the final accuracy. The accuracy in this case was 62.3%. Although this accuracy is not high

enough, the difficulty level of the courses could be estimated from the syllabuses with a certain degree of accuracy.

3 Conclusions

In this study, we investigated the relationship between the content of syllabuses and the average scores, and estimated difficulty level of courses from the syllabuses. 10-fold cross-validation was performed for the evaluation, and it became clear that the difficulty level can be estimated with an accuracy of 62.3%. However, the proposed method does not consider student effort, different teachers' grading standards, syllabus writing principles, and so on. Improving the accuracy by using such additional information will be a future work.

References

[1] A. Moretti, J.P. González-Brenes, and K. McKnight, ”Data-Driven Curriculum Design: Mining the Web to Make Better Teaching Decisions, ” Proceedings of the 7th International Conference on Educational Data Mining, 2014, pp.421-422.

[2] M. Yasukawa, H. Yokouchi, and K. Yamazaki, ”Syllabus Mining for Analysis of Searchable Information, ” International Journal of Institutional Research and Management, Vol.4, No.1, 2020, pp.46–65.

[3] T. Sekiya, Y. Matsuda, and K. Yamaguchi, ”Curriculum analysis of CS departments based on CS2013 by simplified, supervised LDA, ” Proceedings of the 5th International Conference on Learning Analytics and Knowledge, 2015, pp. 330-339.