International Journal of Institutional Research and Management
International Institute of Applied Informatics
2020, Vol. 4, No. 1, 46 – 65

Syllabus Mining for Analysis of Searchable Information

Michiko Yasukawa *, Hirofumi Yokouchi *, Koichi Yamazaki *

Abstract

Writing an effective syllabus is critically important for instructors to provide effective education at universities. However, little is known about how to create a well-written syllabus. It is necessary to elucidate what kind of information must be included in a syllabus. To achieve this goal, we focus on the searchable information in syllabi and analyze an actual syllabus collection that includes 6,493 syllabus documents of a national university in Japan. First, we investigate syllabus classification and syllabus search by using established text mining methods and an information retrieval method. The results of our experiments demonstrate that (i) knowledge discovery from syllabus documents is a challenging and non-trivial task, and (ii) just adding one particular word can already increase the searchability in syllabus search. Next, we investigate methods that provide word suggestions using deep learning approaches and large text corpora. In this experiment, we used a bibliographic database of university libraries in Japan, which contains 3,990,646 bibliographic entries, and a version of Japanese Wikipedia, which contains 2,351,545 articles. The results indicate that (iii) a vocabulary from a bibliographic database of university libraries is effective to ameliorate the efficacy measured by the mean reciprocal rank, and (iv) a wide range of vocabulary is essential in improving the recall in word suggestions. [1]

Keywords: database, faculty development, information retrieval, text mining

1 Introduction

Quality enhancement in education at universities is gaining importance in all respects. A well written syllabus is necessary to provide an effective description of an educational course. Davis suggested that a comprehensive syllabus is valuable for students[2]. Further, the information that can be provided in a syllabus was suggested by Walker[3], which included the main course goals, textbooks, and course schedule.

When every syllabus is informative and the number of syllabi is large, students face difficulties in finding useful information because of the quantity of data they must search. To address this problem, visualization of educational information to aid in retrieval was

* Graduate School of Science and Technology, Gunma University, Gunma, JAPAN

[1] An earlier version of this paper was presented at [1].

suggested by[4] and [5]. Faculty members should analyze syllabus documents when a curriculum is designed or evaluated at a university [6]. Based on these background studies, we elucidated the searchability of syllabi. Searchability indicates the ease of retrieving relevant information[7]. Specifically, we investigated syllabus searches assuming the following scenarios and research questions.

RQ1: When there are several thousands of syllabi, students need computer assistance because it is difficult to manually find relevant syllabi. Accessing such syllabi would require document processing, such as text mining. By applying machine learning methods to text mining, knowledge discovery from a large text is generally feasible. A syllabus collection is a sort of text data. Would knowledge discovery from a collection of syllabi be a trivial task? In other words, could we assume that simply utilizing established methods for text mining in the syllabi is automatically successful?

RQ2: If the question is "writing a syllabus" or "not writing a syllabus", writing is exemplary. If a syllabus entry has some text fields to fill in, all field should be filled in, rather than being left unfilled. Moreover, if a recent version of syllabus contains more searchable information than its previous version, the recent version would become more effectively utilized by students because they can easily find it. Under this scenario, what kind of words in a syllabus are effective to increase its seachability? If more words are added to a syllabus, can its searchability be increased accordingly?

RQ3-1: Let us assume that instructor X is not successful in writing an ideal syllabus voluntarily, but he or she can pick words to be added in the syllabus if relevant words were suggested by a system. Additionally, if instructor X is too busy and can read only a few word suggestions, what kind of a database would be effective for an external text resource for obtaining word suggestions?

RQ3-2: In another situation, instructor Y is not so busy and eager in reading many relevant words. What kind of word suggestions would be effective for him or her to obtain word suggestions? In other words, how can we increase the recall value of word suggestions for creating an ideal syllabus?

To our knowledge, this is the first study that thoroughly analyzed the searchable information in syllabi. The method and experiments of our study are explained in detail in the following sections. Particularly, RQ1 is examined in Section 3.2 and RQ2 is discussed in Section 3.3. Lastly, RQ3-1 and RQ3-2 are explored in Section 3.4.

2 Method

If the number of syllabus documents is significantly large, it is challenging to satisfy the information need of students. The search process requires more time and effort because word usage may be marginally different in different documents. It is possible that a word is a polyseme (has the same spelling with a different meaning) or a synonym (has the same meaning with a different spelling). If an initial search query is not effective, adding more words can result in an improved set of results. This is a well-known approach in information retrieval and is called "query expansion." [8]

In information retrieval, the relationship between documents and index terms in a document corpus is generally represented by a vector space model [9]. Consider a document-term matrix, $D_c \in \mathbb{R}^{n \times m}$ for a given document corpus as

$$D_c = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1m} \\ t_{21} & t_{22} & \cdots & t_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \cdots & t_{nm} \end{pmatrix}.$$

In the matrix D_c, a row represents a document in the corpus and columns represent the m-dimensional index space, where each document contains up to m distinct index terms. The index terms are weighted according to their importance, which is based on the frequency of occurrence of words in the document corpus. Similar to the documents, a search query, D_q, is represented in the m-dimensional index space and is defined as

$$D_q = \begin{pmatrix} t_1 & t_2 & \cdots & t_m \end{pmatrix}.$$

For the same document corpus, we can also define a term-document matrix, $T_c \in \mathbb{R}^{m \times n}$, which is the transposed matrix of a matrix D_c as

$$T_c = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{pmatrix}.$$

In the matrix T_c, a row represents an index term in the corpus and columns represent the n-dimensional document space, where each index term is contained by up to n distinct documents. For this matrix, search query T_q is represented in the n-dimensional document space and is defined as

$$T_q = \begin{pmatrix} d_1 & d_2 & \cdots & d_n \end{pmatrix}.$$

For a better understanding of the above definitions, we present a toy example as follows. Document-term matrix D_{toy} represents a three-dimensional index space, where four documents contain three distinct terms (e.g., "aeronautical," "biological," and "computational") as

$$D_{toy} = \left(\begin{array}{c|ccc} & aero. & bio. & comput. \\ \hline syll_1 & 0.1 & 0.0 & 0.9 \\ syll_2 & 0.0 & 0.5 & 0.5 \\ syll_3 & 0.2 & 0.2 & 0.6 \\ syll_4 & 0.0 & 0.0 & 1.0 \end{array}\right).$$

We can define term-document matrix, T_{toy}, that represents the same document corpus as

$$T_{toy} = \left(\begin{array}{c|cccc} & syll_1 & syll_2 & syll_3 & syll_4 \\ \hline aero. & 0.1 & 0.0 & 0.2 & 0.0 \\ bio. & 0.0 & 0.5 & 0.2 & 0.0 \\ comput. & 0.9 & 0.5 & 0.6 & 1.0 \end{array}\right).$$

By observing the relationship between documents and index terms, we can identify that the most important term in document $syll_1$ is "computational," and that this index term has the largest importance in document $syll_4$ among the corpus. An information retrieval model using the associative relationship between documents and index terms (i.e., from documents

to index terms, and from index terms to documents) is called an "Associative Search" and was applied to an interactive search system by Takano et al. [10]

In this study, for the analysis of syllabus documents, we use the relationship between documents and index terms to compare a baseline method (a search conducted by a short ambiguous query) and an ameliorated method (a search conducted based on a reformed query with additional words). Specifically, the reformed query, D_{qe}, of the baseline query, D_{qb}, used to obtain the ameliorated results is calculated by adding a new query vector, D_{qx}, and rewriting the term weights as

$$D_{qe} = D_{qb} + D_{qx}.$$

The similarity between query q and document d is computed using a *pivoted document length normalization* (PDLN) that was proposed by Singhal et al.[11]. It should be noted that q may be either D_q or T_q, and d may be a row vector in D_c or T_c, depending on the purpose of the search.

Word search by word is performed in the same way as a document search by document because the document-term matrix and the term-document matrix are transposed to each other. The definition of the similarity equation is as

$$\mathrm{sim(d|q)} = \frac{1}{\mathrm{PDLN(d)}} \times \sum_{t \in q,d}^{n} \mathrm{wq(t|q)} \times \mathrm{wd(t|d)}.$$

Symbols in the above equation are explained as follows.

- t, d, q, and n respectively denote words, documents, queries, and the number of words in the queries and documents.
- $\mathrm{PDLN(d)}$ represents the normalization value for document d.
- $\mathrm{wq(t|q)}$ and $\mathrm{wq(t|d)}$ represent the weight values for word t in query q and document d, respectively.

The definitions of $\mathrm{wq(t|q)}$, $\mathrm{wq(t|d)}$, and $\mathrm{PDLN(d)}$ are as

$$\mathrm{wq(t|q)} = \frac{1 + \log(\mathrm{TF(t|q)})}{1 + \log(\mathrm{aveTFq}),}$$

$$\mathrm{wq(t|d)} = \frac{1 + \log(\mathrm{TF(t|d)})}{1 + \log(\mathrm{aveTFd}),}$$

$$\mathrm{PDLN(d)} = avedl + slope \times (dl - avedl).$$

Symbols in the above equations are explained as follows.

- $\mathrm{TF(t|q)}$ and $\mathrm{TF(t|d)}$ represent term frequency of word t in query q and document d, respectively.
- aveTFq and aveTFd represent average term frequency in query q and document d, respectively.
- dl represents document length (the number of words).
- $avedl$ represents the average document length in the corpus.

- $slope$ is a constant value[2] that ranges from 0 to 1.

Generally, the scheme of PDLN is favorable if documents are long and a search query is short. A search query by the user of the syllabus search is expected to be short and vague. On the other hand, syllabi that the user is going to read are supposed to include detailed text content.

Depending on the purpose of search, the equation for similarity between d and q may be replaced by the cosine similarity, and the document-term matrix, D_{toy}, may be replaced by the document embedding, D_{emb}, that is learned from a large text corpus, as

$$D_{emb} = \left(\begin{array}{c|cc} & vector_1 & vector_2 \\ \hline doc_1 & x_{11} & x_{12} \\ doc_2 & x_{21} & x_{22} \\ doc_3 & x_{31} & x_{32} \\ doc_4 & x_{41} & x_{42} \end{array}\right).$$

As for the document-term matrix, the term-document matrix T_{toy} may be replaced by the word embedding T_{emb}, which is learned from a large text corpus as

$$T_{emb} = \left(\begin{array}{c|ccc} & vector_1 & vector_2 & vector_3 \\ \hline word_1 & y_{11} & y_{12} & y_{13} \\ word_2 & y_{21} & y_{22} & y_{23} \\ word_3 & y_{31} & y_{32} & y_{33} \end{array}\right).$$

The document or word embedding can be obtained from an arbitrary resource that is not necessarily a syllabus corpus. For example, a bibliography database or an online encyclopedia may have terms in common with a syllabus corpus (e.g., "aeronautical," "biological," and/or "computational"). Furthermore, there may be common document titles between a syllabus corpus and other text corpora. For example, lecture titles in syllabi, titles of Wikipedia articles, and book names in a bibliographic database may coincidentally include some common words, such as "Machine Learning," "Operating System," "Algorithm," and "Database."

In the next section, experiments based on the above preliminary knowledge are empirically conducted.

3 Experiments

First, we describe a collection of syllabi for the experiments. Next, we explain three experiments for the analysis of searchability.

3.1 Syllabus data

For the experiments in this study, we downloaded 6,493 online syllabi from The University of Tokyo Online Course Catalogue[3]. The downloading process was conducted from October 9 to 14 and October 23 to 28 in 2018. The University of Tokyo is one of the top ranked universities in the Japan University Rankings 2018[4] and 2019[5]. The quantity and

[2]To obtain the optimum value for *slope* (e.g., 0.2), a traditional way of parameter optimization (e.g., grid search) can be used.

[3]`https://catalog.he.u-tokyo.ac.jp/`

[4]`https://www.timeshighereducation.com/rankings/japan-university/2018`

[5]`https://www.timeshighereducation.com/rankings/japan-university/2019`

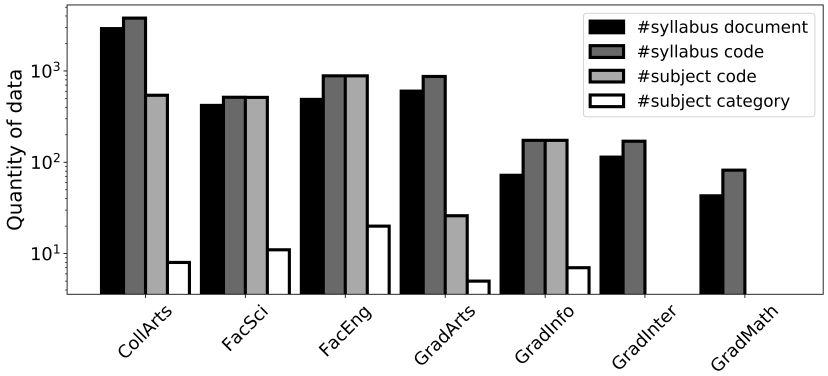

Figure 1: Syllabus data for the experiments.

Table 1: Average word numbers in syllabi.

Division	Title	Desc.	Topic	Concatenated	#syll.
CollArts	3.52	147.87	80.92	192.10	3795
FacSci	2.44	65.53	67.55	126.18	516
FacEng	2.92	59.52	88.58	91.21	886
GradArts	3.38	80.91	45.41	117.70	870
GradInfo	4.27	96.76	69.78	83.11	174
GradInter	4.05	99.11	81.65	158.67	170
GradMath	3.02	52.35	51.25	66.59	82
average	3.36	117.55	75.25	158.12	6493

quality of research achievements in Computer Science, Mathematics, and Engineering at this university were top-ranked in the benchmarking of Japanese universities in 2015[6]. We focused on these subject areas for a detailed analysis of certain experiments conducted in this study, which will be discussed subsequently in this section.

The syllabus collection consisted of the downloaded syllabus documents provided by seven divisions of the university, including: (a) College of Arts and Sciences, (b) Faculty of Science, (c) Faculty of Engineering, (d) Graduate School of Arts and Sciences, (e) Graduate School of Information Science and Technology, (f) Graduate School of Interdisciplinary Information Studies, and (g) Graduate School of Mathematical Science. Among the seven divisions, (a), (b), and (c) are undergraduate divisions and the others are graduate divisions. Hereafter, these divisions will be referred to as (a) CollArts, (b) FacSci, (c) FacEng, (d) GradArts, (e) GradInfo, (f) GradInter, and (g) GradMath.

Each syllabus includes a unique syllabus code (hereafter, syllabus code), a course title (hereafter, title), the name of the lecturer, the number of credits, the semester, the academic year, and the name of a lecture room. Some syllabi include a subject code and/or a subject category for grouping related syllabi. The lecturer is expected to provide optional information that includes a description of the lecture (hereafter, description), a course schedule associated with the lecture topics (hereafter, topics), methods of teaching and evaluation,

[6] `http://hdl.handle.net/11035/3116`

and information regarding the required textbook and reference books. Some syllabi include important points that have to be emphasized, such as do's and don'ts, precautions, and general information in the fields for "Notes on taking the course" or "Others.'

The number of documents can differ, depending on the syllabus database. Some syllabi include few or no Japanese words. The black and dimgrey bars in Fig. 1 show the number of syllabi for each division and the number of syllabi with 30 or more Japanese words, respectively. The dark grey and white bars in the figure indicate the number of subject codes and subject categories, respectively. Certain documents have neither a subject code nor a subject category. The word number (number of Japanese morphemes) in documents can differ, depending on the optional information provided by the instructors. Certain documents display moderate-length descriptions and schedules. Others have none of these. Table. 1 presents average number of words in the syllabi for the seven divisions and all of the syllabi for the experiments. The scatter plots and bars in Fig. 2 and Fig. 3, respectively, illustrate concrete examples of text lengths in syllabi. The word number in the syllabi can also be different, depending on the divisions.

3.2 Data applied to an experiment in machine learning

To verify whether the syllabus collection is sensible, we conducted a preliminary experiment involving automatic syllabus classification. The seven divisions were used as text categories. We used three widely recognized machine learning methods for this experiment: random forest [12], naive Bayes [8] [13], and support vector machine (SVM) [14] [15]. We used implementations in Python of random forest[7], naive Bayes[8], and SVM[9], respectively. The morphological analyzer MeCab[10] with mecab-ipadic-NEologd[11] was used for text processing.

Table. 2 and Table. 3 present the results of classification for titles and documents, respectively. In the tables, the values of precision, recall, and F1-score for each category and their averages are presented. As presented in the tables, classification for CollArts achieved the highest F1-scores while classification for GradInfo and GradMath were less effective among the seven divisions. This likely occurs because of the fewer number of syllabi for these divisions in comparison with the others. In summary, SVM was the most effective followed by random forest and naive Bayes methods. The syllabus collection contains reasonable words and categories based on the classification results of the documents.

It should be noted that syllabus classification in this experiment was designed to perform a sanity check on the syllabus collection for the following experiments. The results of automatic syllabus classification for documents showed remarkably lower effectiveness values for GradInfo and GradMath than others. Based on the results of this experiment, knowledge discovery from the syllabi is a non-trivial task because the syllabus collection is considered to be a sufficiently challenging dataset that contains natural characteristics of realistic data.

[7] https://scikit-learn.org/stable/modules/ensemble.html
[8] https://scikit-learn.org/stable/modules/naive_bayes.html
[9] https://scikit-learn.org/stable/modules/svm.html
[10] http://taku910.github.io/mecab/
[11] https://github.com/neologd/mecab-ipadic-neologd

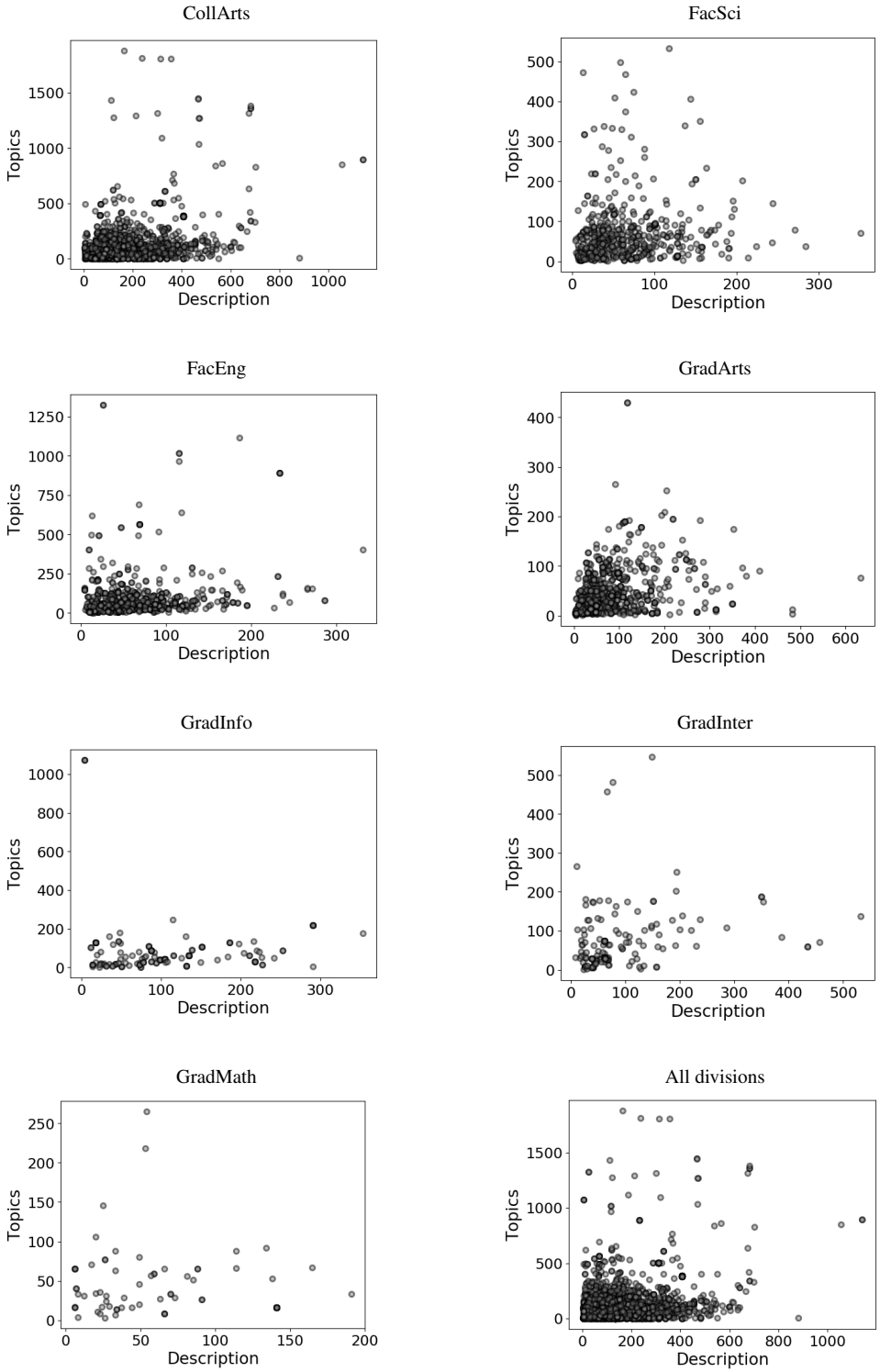

Figure 2: Distributions of word numbers in syllabus contents based on descriptions and topics.

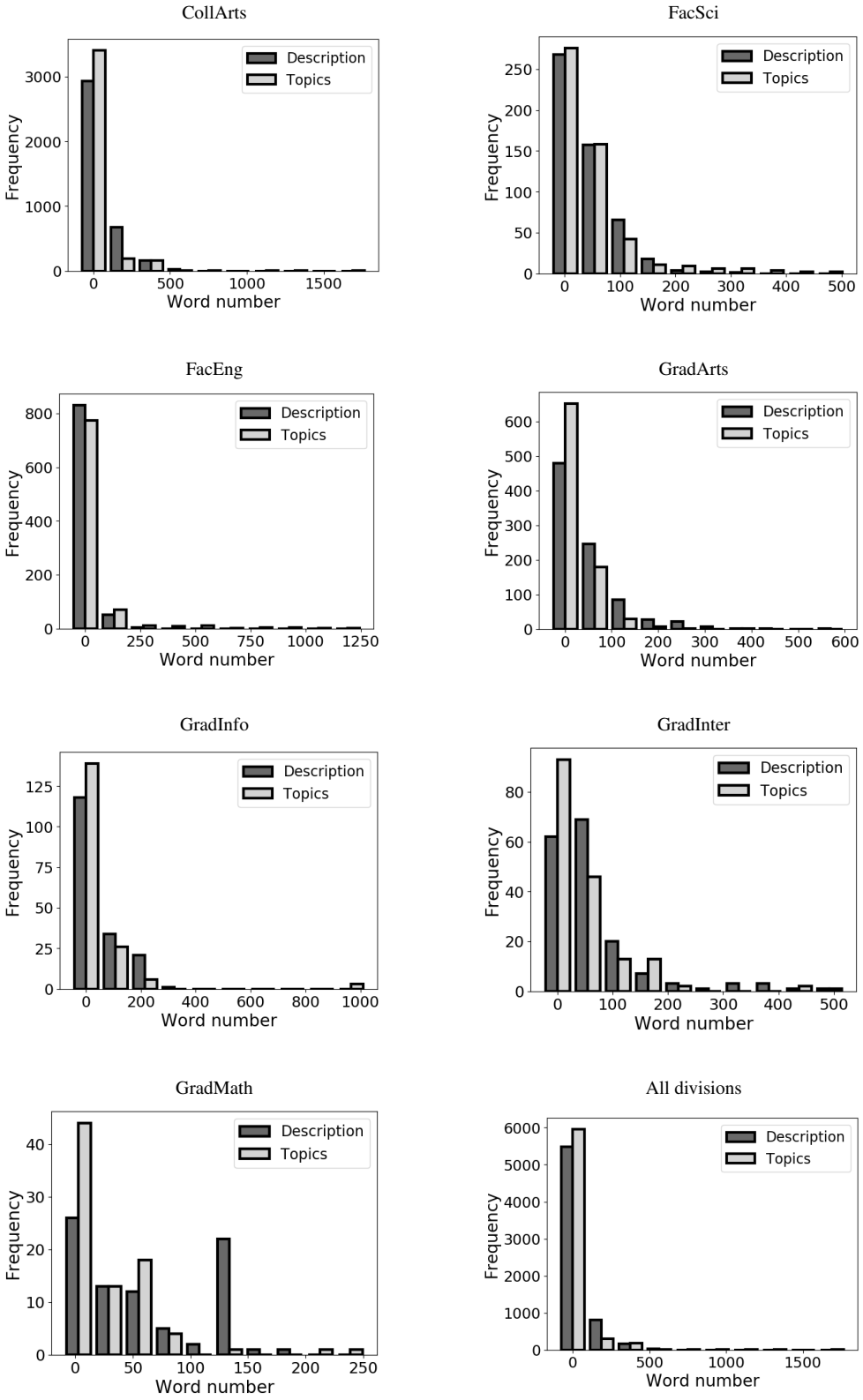

Figure 3: Frequency of word numbers in syllabus contents based on descriptions and topics.

Table 2: Automatic syllabus classification for titles.

Division	Random Forest			Naive Bayes			SVM		
names	prec.	recall	f1	prec.	recall	f1	prec.	recall	f1
CollArts	0.91	0.98	0.95	0.91	0.95	0.93	0.94	0.97	0.95
FacSci	0.70	0.63	0.66	0.66	0.58	0.62	0.69	0.67	0.68
FacEng	0.83	0.68	0.75	0.72	0.69	0.71	0.84	0.73	0.78
GradArts	0.89	0.82	0.86	0.80	0.87	0.83	0.86	0.89	0.88
GradInfo	0.44	0.31	0.36	0.67	0.15	0.25	0.57	0.31	0.40
GradInter	0.84	0.76	0.80	0.70	0.33	0.45	0.80	0.57	0.67
GradMath	1.00	0.43	0.60	1.00	0.14	0.25	0.75	0.43	0.55
average	0.88	0.88	0.88	0.85	0.85	0.84	0.89	0.89	0.89

Table 3: Automatic syllabus classification for documents.

Division	Random Forest			Naive Bayes			SVM		
names	prec.	recall	f1	prec.	recall	f1	prec.	recall	f1
CollArts	0.84	0.96	0.90	0.91	0.88	0.89	0.95	0.96	0.95
FacSci	0.61	0.44	0.51	0.53	0.74	0.62	0.67	0.67	0.67
FacEng	0.65	0.56	0.60	0.72	0.78	0.75	0.76	0.82	0.79
GradArts	0.77	0.61	0.68	0.64	0.73	0.68	0.85	0.80	0.82
GradInfo	0.50	0.06	0.11	0.00	0.00	0.00	0.83	0.31	0.45
GradInter	0.62	0.26	0.37	0.40	0.11	0.17	0.61	0.58	0.59
GradMath	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.14	0.15
average	0.77	0.79	0.77	0.79	0.80	0.79	0.88	0.88	0.87

3.3 Experiment in information retrieval

To investigate the searchability of the documents, we conducted experiments on information retrieval using words and documents as illustrated in Figure. 4. First, all documents in the syllabus collection were parsed and included in an inverted index for a full text search. The morphological analyzer MeCab with mecab-ipadic-NEologd was used for text processing. Then, a document was picked from the syllabus collection, a query sample was obtained from the document, a search was conducted using the query, and a list of top-k ranked documents was obtained. In Figure. 4, the value for k was set to 5 for simplicity while k in our experiment was set to 10 for practicality. We used a widely accepted document ranking model[11], as described in Section 2.

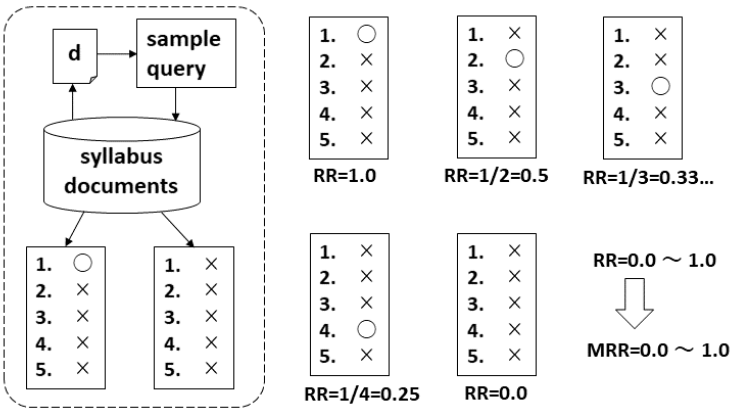

Figure 4: Syllabus search and analysis of the searchability.

The above search process was repeated on all the documents in the syllabus collection. The sample query consisted of (i) title in each syllabus, (ii) title and the most particular word in each syllabus, or (iii) title and the 10 most particular words in each syllabus. The most particular words were obtained by using PDLN that has been described in Section 2. The sampled query for document search encompassed all of the text segments in the syllabus document. After obtaining the search result, the effectiveness of the search processes was measured using the mean reciprocal rank (MRR).

If the query was a perfect document descriptor, the target document was ranked as 1, and the reciprocal rank was set to 1.0 for the search. If the target document was ranked as 2, 3, 4, and 5, the reciprocal rank was set to 0.50, 0.33, 0.25, and 0.20, respectively. If the query was not a good descriptor of the document, the target document was excluded from the results, and the reciprocal rank was set to 0.0 for the search. The MRR is the mean value of the reciprocal rank among a given set of queries.

The MRR was calculated for each division, and the results are illustrated in Figure. 5. The black, dark grey, and white bars in the figure represent the MRR values for the search process by title, QE1, QE10, respectively. QE1 indicates query expansion using the most particular word in each syllabus. QE10 indicates query expansion using the 10 most particular words in each syllabus. As a result, query expansion using the most particular word(s) was effective when search by title was not effective. Unfortunately, query expansion caused an adverse effect when search by title was adequately effective. Moreover, it should be noted that query expansion with the top-10 words was not 10 times more effective in comparison with query expansion with the top-1 word. A take-home message that we can learn from this experimental result is explained as "the quality of the additional word is more important than the quantity of the words for increasing the text length of a syllabus."

To investigate the details of particular words in syllabi, the search effectiveness measured by Rank@k, which means "the target document was ranked at k," for each of the divisions was examined. The results are shown in Figure. 6. Each matrix in the figure visualizes the correlation between the search by title and QE1, which indicates query expansion using the most particular word in each syllabus. In the matrix, each column and row indicate the successfulness of search by title and QE1, respectively. Each of the cells represents the correlation that is the correspondence of the associated column and row. The heat map in the matrix provides a graphical representation of the occurrence of searches ranging in

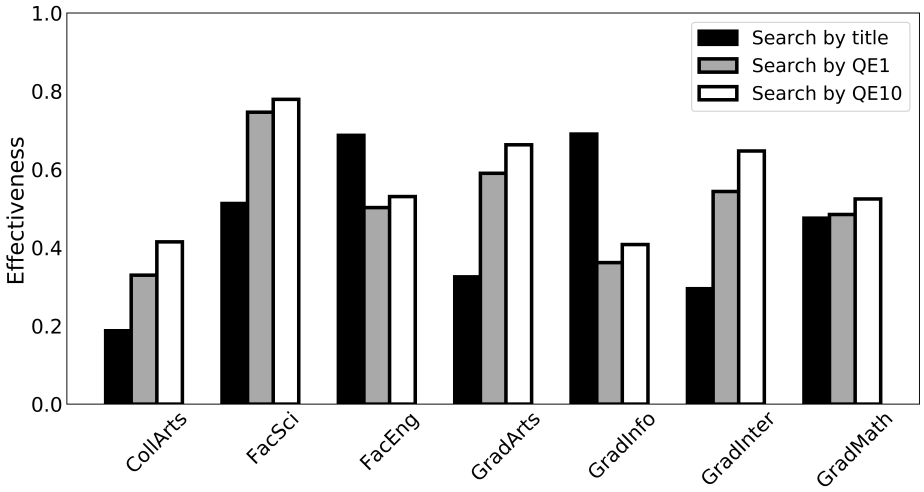

Figure 5: Search effectiveness measured by Mean Reciprocal Rank (MRR).

scale from 0 to 1. To be more specific, the normalized value nv_{xy} in each cell is defined as

$$nv_{xy} = \frac{\log(1+v_{xy})}{\log(1+v_{max})}.$$

In the above equation, v_{max} indicates the greatest value of v_{xy}, which indicates the number of searches corresponding to x on the horizontal axis and y on the vertical axis. Let us consider an example case, wherein 1,000 searches in the matrix consists of 900, 90, 9, 1, 0 searches in the cells. By the above equation, these values are normalized as 1.000, 0.663, 0.338, 0.102, 0, respectively. The diagonal cells in the matrix indicate that the corresponding searches by title and QE1 have the same successfulness in search. For example, the cell in the upper left corner indicates search by title and search by QE1 were both successful and the target syllabus was ranked at 1. Colored cells above and below the diagonal cells indicate ameliorated and deteriorated searches by QE1, respectively. When k is large, users become impatient and stop looking at the search result. For this reason, the search wherein the target document is not highly ranked is considered as "Failure." In the matrix visualization, the search wherein the target document is ranked at k ($k \leq 10$) is considered as "Success," otherwise it is considered as "Failure."

While certain unsuccessful results of the title search tended to shifted to Success, the same document rankings were retained in many cases in the search for CollArts, as illustrated by the darker color of the diagonal elements in the matrix. From this result, it can be noted that the most particular word in the syllabus documents for CollArts do not contain effective document descriptors. Lecture courses of CollArts are for a freshman or sophomore and the same or similar subjects, such as English and basic mathematics, are provided in different lecture courses. As a result, query expansion for CollArts was not as effective as the query expansions for the others.

Conversely, the GradArts division presents darker colors on the top and right side of the matrix. While the title search for GradArts was unsuccessful because the query was ambiguous and did not improve the ranking of the target document, the words obtained from the syllabus document were effective document descriptors and improved the search results (the dark shaded cell in the upper right corner).

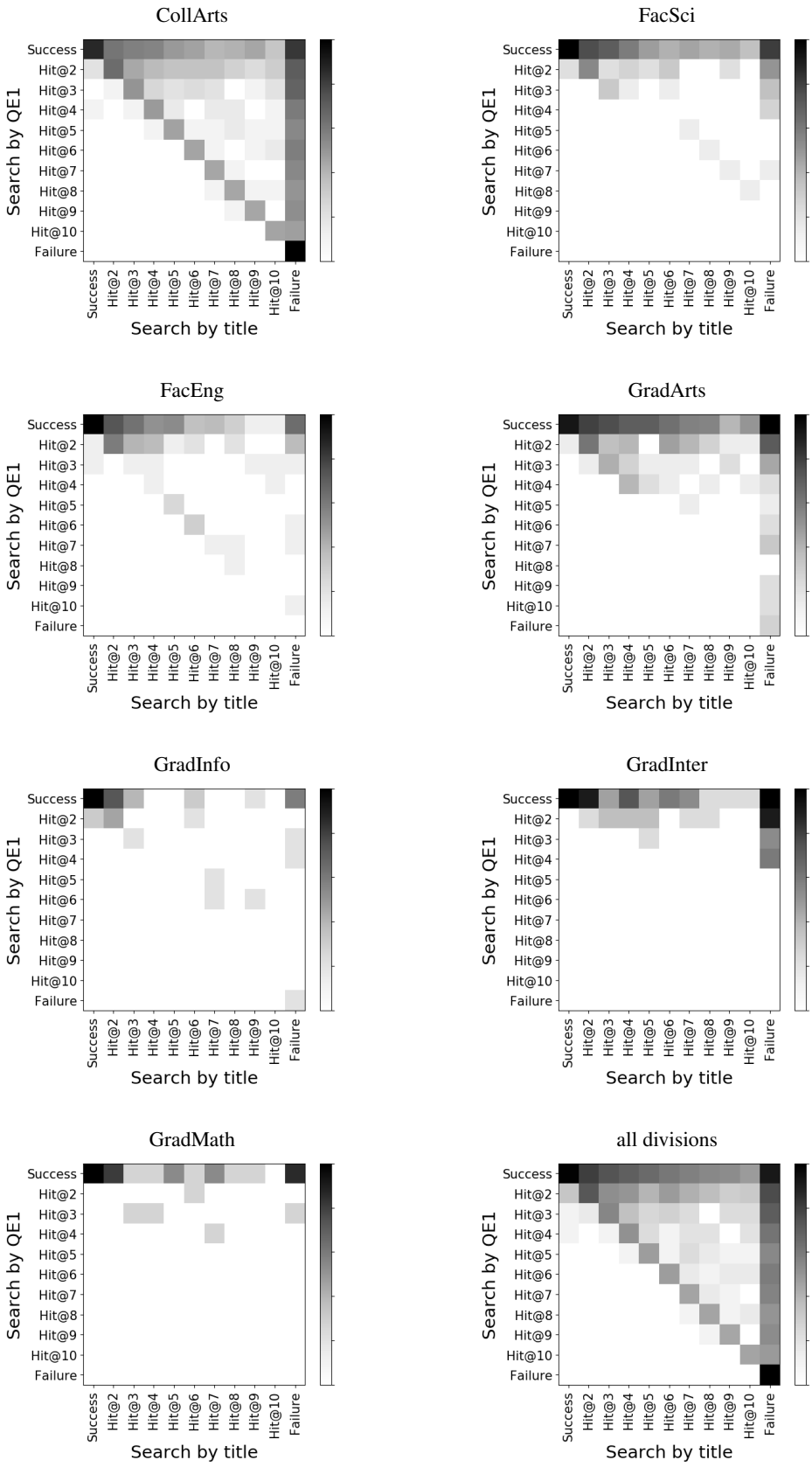

Figure 6: Search effectiveness measured by Rank@k.

Table 4: Examples of words in syllabi (XX_s) and suggested words ($XX_{b,d,w}$).

Set ID	Suggestions by word association
ML_s	data, algorithm, model, data mining, artificial intelligence, analysis, ...
ML_b	bayes, data mining, pattern recognition, object-orientation algorithm, ...
ML_d	software agent, support vector machine, model checking, neural ...
ML_w	neural network, natural language processing, data mining, pattern ...
OS_s	process, scheduling, memory management, input and output, system, ...
OS_b	compiler, lisp, algorithm, protocol, cpu, electronic computer, image ...
OS_d	memory protection, windows, system request, advanced scsi ...
OS_w	device driver, virtual memory, virtual machine, unix, micro kernel, ...

Table 5: Similarity between words in syllabi and suggested words.

Search word	CiNiiBooks			JawikiDoc			JawikiWord		
from title	Jacc.	Dice	Simp.	Jacc.	Dice	Simp.	Jacc.	Dice	Simp.
Machine Learning	0.12	0.21	0.28	0.13	0.24	<u>0.34</u>	0.15	0.26	0.30
Operating System	0.11	0.20	<u>0.27</u>	0.10	0.19	0.19	0.11	0.19	0.25

3.4 Experiment in deep learning

While some particular words in syllabus documents are surmised to be effective, certain syllabus documents do not include sufficient words. In such cases, obtaining searchable information from external resources may be useful. Now, let us assume the following two extreme cases: (a) the most ideal syllabus includes all relevant words for the lecture content, and (b) the least ideal syllabus includes no relevant words for the lecture content. While we did not know all relevant words for each lecture course, syllabi in the experimental dataset were actual syllabus documents downloaded from a highly evaluated university in Japan. For this reason, we considered each syllabus in the dataset as a pseudo ideal syllabus. Then, searchable information from external resources could be judged by using the experimental syllabus collection.

Based on this assumption, we conducted a search experiment using words from a given text database. In particular, we used word embedding [16] [17] [18] and document embedding [19] [20] with large text databases. We used implementations of the methods in Python[12]. Parameter settings[13] were determined by reference to related works [18] [20]. A version of the Japanese Wikipedia's XML database (hereafter, Jawiki) and the CiNii Books database were used to train the models. Jawiki contained 2,351,545 articles. In this experiment, the morphological analyzer MeCab with mecab-ipadic-NEologd was used for text processing.

In the following experiments, we focused on the syllabus documents of the Department of Information Science, which is a department under FacSci, and the syllabus documents of the Department of Electrical and Electronic Engineering, which is a department under FacEng. These departments are referred to as FSC-IS and FEN-EE, respectively. We fo-

[12] `https://radimrehurek.com/gensim/models/doc2vec.html`

[13] Specifically, our parameter settings were as dm = 1, vector_size = 300, window = 10, alpha = 0.05, min_count = 2, and epochs = 20.

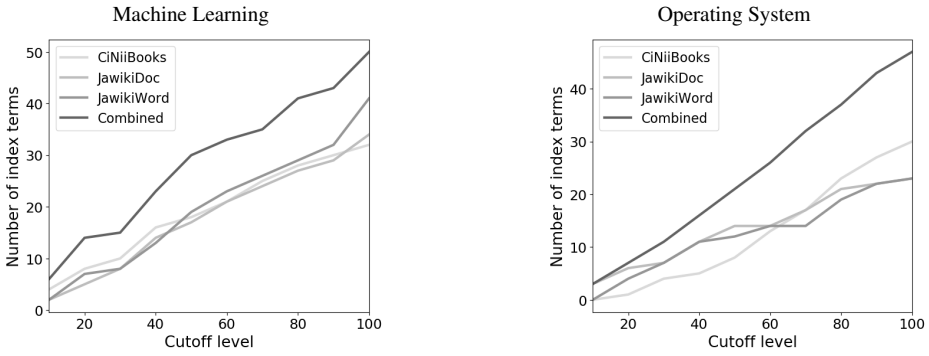

Figure 7: Correlation between cutoff level and number of index terms.

cused on two common words in the titles among FSC-IS and FEN-EE: "Machine Learning" and "Operating System." We used these words as search queries for the word search. Table. 4 presents excerpted examples of the search results using different methods. The methods include a gold standard method using the syllabus corpus, indicated as ML_s and OS_s; a baseline method using document embedding with CiNii Books, indicated as ML_b and OS_b; a method using document embedding with Jawiki, indicated as ML_d and OS_d; and a method using word embedding and Jawiki, indicated as ML_w and OS_w. In the table, ML indicates "Machine Learning" and OS indicates "Operating System." The obtained Japanese words were translated to English by Google Translate[14].

As presented in Table. 4, some suggested words among the methods were common. For example, "data mining" was included in ML_s, ML_b and ML_w. To investigate the commonality of these different groups, the similarity between syllabus words (internal resources) and the suggested word sets using external resources were measured using Jaccard's similarity coefficient, Dice's coefficient, and the overlapping coefficient (the Szymkiewicz–Simpson coefficient). The results of the experiment are listed in Table. 5. The word cutoff level was set at 1,000. Consequently, the suggested words obtained using the doc2vec learned model for Jawiki (JawikiDoc in Table. 5) for the search word "Machine Learning" produced the most similar search words based on the overlapping coefficient. On the other hand, the suggested words obtained using the word2vec learned model for CiNiiBooks (CiNiiBooks in Table. 5) were observed to be the most similar to the index terms for the search word "Operating System" based on the overlapping coefficient.

When fewer words are taken from the suggested words, the number of words included in the index terms of the syllabus search decreases. For this reason, it is considered to be beneficial to use external text resources with a comprehensive vocabulary. Fig. 7 presents the correlation between the cutoff level of the word search by word and the number of index terms in the syllabus search. The darkest grey line represents combined word sets (hereafter, Combined) from CiNiiBooks, JawikiDoc, and JawikiWord, indicating the largest number of coincided index terms at each cutoff level.

Finally, we performed an experiment to measure searchable information in syllabus documents from the viewpoint of word suggestions. In this experiment, we used 236 lecture titles that were contained in the vocabulary of CiNiiBooks, JawikiDoc or JawikiWord. It should be noted that only 236 out of 6,493 lecture titles were included in the vocabulary

[14] `https://translate.google.com/`

Table 6: Efficacy of word suggestions measured by MRR, MAP, Recall and relevant words.

Data	MRR	MAP	Recall	#rel.
CiNiiBooks	0.3416	0.0614	0.708	4127
JawikiDoc	0.0854	0.0118	0.159	928
JawikiWord	0.2600	0.0367	0.461	2689
Combined	0.2991	0.0654	1.000	5832

of ether of the text resources because of the diversification of written expressions in the syllabus collection. Each of the lecture titles was used as a search word for word suggestions. Up to 1,000 suggestions were obtained from each of the trained models. The obtained suggestions were judged as relevant if the suggestions were included in the actual syllabus document that was associated to the lecture title. It should be noted that the search results in this experiment were ranked lists of words, rather than ranked lists of documents. Although the output in this experiment did not conform to the conventional notion of the search effectiveness measured by MRR, each word can be considered as a special case of a document that contains exactly one word. Hence, we compared MRRs of the obtained word lists to investigate efficacy of word suggestions. To obtain a combined list of search results, we calculated the sum of normalized weights of words that were obtained from CiNiiBooks, JawikiDoc and JawikiWord. For the normalization, a standard min-max normalization was adopted. To be more specific, the weight was scaled between 0 to 1 by using the maximum and minimum weight values.

Fig. 8 presents efficacy of the word suggestions that were obtained from each of the models. The black, grey, and white bars in the figure represent the MRR, MAP and recall values for word suggestions, respectively. As a result, suggestions obtained from CiNiiBooks achieved the highest MRR than the other suggestions. This result indicates that the bibliographic information of university libraries contains more helpful information to suggest searchable information when creating an ideal syllabus. Our finding in this experiment is considered to be surprising and non-trivial because trained models from Japanese Wikipedia have been suggested to be exploited in previous works [18] [20] and generally used. On the other hand, a trained model from CiNiiBooks has been rarely used.

To make a thorough study, we also calculated the MAP and recall values for the word suggestions. Table. 6 presents the obtained values. As a result, the combined word suggestions of CiNiiBooks, JawikiDoc and JawikiWord achieved the highest recall value than the other suggestions. If instructors would like to refer to a larger number of suggested words, the combined suggestions would be helpful because a wide range of vocabulary is used to obtain suggestions.

4 Discussion

With respect to the concept of searchability, Onaifo[21] and Ivanovic[22] discuss the findability of library contents from the viewpoint of search engine optimization. Larsson[23] discusses the retrievability of Ph.D. dissertations.

Regarding syllabus analysis, examining general eudcational syllabi for a better understanding of their attributes and characteristics was reported by Everly et al.[24] Their study used a smaller number of syllabi (n=145) in comparison with our study (n=6493). In our

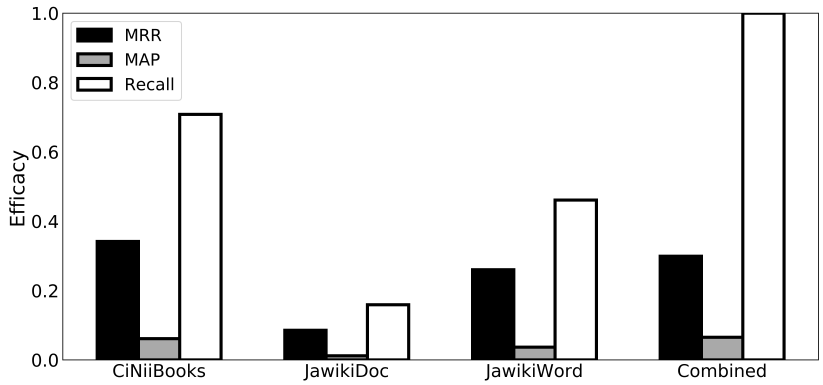

Figure 8: Efficacy of word suggestions measured by MRR, MAP and Recall.

study, we performed objective analysis of syllabi by using document analysis approaches, rather than a human-in-the-loop analysis. On the other hand, some previous studies on sysllabus analysis adopted human assessors. For example, syllabus assessment from the viewpoint of the wording of syllabi was studied by Ishiyama et al.[25] Comparison between a content-focused syllabus and a learning-focused syllabus was carried out by Palmer et al.[26] Subjective analysis on syllabi from the viewpoint of the need of students in the lecture course was conducted by Keller et al.[27]

As our technological challenge in this study, we have achieved a proof of concept that demonstrates how to measure the searchable information in syllabi and how to suggest effective words that should be included in syllabi. We have conducted a meticulous analysis on an actual collection of syllabus documents and discovered novel pieces of knowledge regarding the searchable information in syllabus documents. Specifically, we have obtained the following answers (A1, A2, A3-1, A3-2) to demystify the previously described research questions (RQ1, RQ2, RQ3-1, RQ3-2) in Section 1.

A1: Knowledge discovery from the syllabus collection is a non-trivial task. Simply applying established methods to a text-mining task in syllabi does not solve the problem automatically.

A2: Even one particular word can become a piece of searchable information to increase the search effectiveness. Increasing the quantity without considering the quality of words in a syllabus would not be a practical solution.

A3-1: Word suggestions should be obtained from a bibliographic database of university libraries, rather than a Wikipedia, due to the characteristics of the vocabulary.

A3-2: When a larger number of word suggestions are preferred, combined word suggestions are more effective than either of the word suggestions.

The main contribution in this study is that we have demonstrated how to select and combine reliable methods for achieving the research goal. Outperforming the experimented methods in our study is an open problem in the research areas of interest in institutional research.

Based on the experimental results, we propose the following guiding principles for assisting instructors in preparing a better syllabus.

- Searchable information in a syllabus can improve information access; however, inputting detailed information is time-consuming. If web user interface for syllabus input is provided, it should be combined with auto input suggestions. Such mechanisms can help to reduce human errors, such as Kanji conversion errors.

- While some neologisms may be included in the dictionary of morphological analyzer, word segments can be mistakenly inserted, depending on the expressions or usage of words. During the editing process of syllabus documents, it could be helpful for the instructor to verify the effectiveness of automatic classification and information retrieval. Then, he/she could reconsider the choice of words before entering the syllabus information into a database.

5 Conclusion

In this study, we investigated syllabus documents of a national university in Japan. While lecture titles can provide meaningful words to represent the lecture contents, they are not always adequate to provide searchable information in syllabus search. Our experimental results revealed that simply adding one particular word to the lecture title can be already effective to improve the searchability in syllabus search. Our finding is non-trivial knowledge for faculty development at universities. If it is a matter of "the more searchable, the better", instructors do not have to be troubled with the myth of "the longer, the better" and they are recommended putting their effort in including concise and yet effective information when preparing syllabus documents. This innovative concept should be beneficial for reducing the workload of instructors and increasing the educational quality at universities. We also studied a method to obtain a group of suggested words from external resources, such as the Wikipedia XML database and the CiNii Books database. We have found that a rich and comprehensive vocabulary in an external text resource is advantageous for effective word suggestions. Our discovery has a potential for enlightening policy-makers and data providers who seek practical measures that improve higher education. To ameliorate the efficacy of our word suggestion approach, we will study syllabus mining on a larger scale in our future work.

Acknowledgments

The experiments for this study used The University of Tokyo Online Course Catalogue. This work was supported by the ISM Cooperative Research Program (2020-ISMCRP-0012), JSPS KAKENHI Grant Number JP18K11986 and the Smart SE program of the enPiT-Pro project. We would like to thank the anonymous reviewers for their comments that have been stimulating and thought-provoking.

References

[1] M. Yasukawa, H. Yokouchi, and K. Yamazaki, "Syllabus mining for faculty development in science and engineering courses," in *8th International Congress on Advanced Applied Informatics, IIAI-AAI 2019, Toyama, Japan, July 7-11, 2019*. IEEE, 2019, pp. 334–341.

[2] B. G. Davis, *Tools for teaching*. John Wiley & Sons, 2009.

[3] H. M. Walker, "What should be in a syllabus?" *SIGCSE Bull.*, vol. 37, no. 4, pp. 19–21, Dec. 2005. [Online]. Available: http://doi.acm.org/10.1145/1113847.1113859

[4] H. Mima, "Mima search: a structuring knowledge system towards innovation for engineering education," in *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, 2006, pp. 21–24.

[5] Y. Yaginuma, "Visualization method of web pages based on syllabus," in *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE, 2017, pp. 1009–1010.

[6] Y. Matsuda, T. Sekiya, and K. Yamaguchi, "Curriculum analysis of computer science departments by simplified, supervised LDA," *Journal of Information Processing*, vol. 26, pp. 497–508, 2018.

[7] T. Upstill, N. Craswell, and D. Hawking, "Buying bestsellers online: A case study in search & searchability," in *ADCS 2002, Proceedings of the Seventh Australasian Document Computing Symposium, Sydney, Australia, December 16, 2002*, 2002. [Online]. Available: http://www.cie.ict.csiro.au/adcs2002/papers/upstill-craswell-hawking.pdf

[8] C. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," *Natural Language Engineering*, vol. 16, no. 1, pp. 100–103, 2010.

[9] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975.

[10] A. Takano, Y. Niwa, S. Nishioka, M. Iwayama, T. Hisamitsu, O. Imaichi, and H. Sakurai, "Information access based on associative calculation," in *International Conference on Current Trends in Theory and Practice of Computer Science*. Springer, 2000, pp. 187–201.

[11] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization," in *SIGIR '96*, 1996, pp. 21–29.

[12] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[13] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752, no. 1, 1998, pp. 41–48.

[14] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992, pp. 144–152.

[15] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space proceedings of workshop at iclr," 2013.

[17] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality in: Nips," 2013.

[18] M. Suzuki, K. Matsuda, S. Sekine, N. Okazaki, and K. Inui, "A joint neural model for fine-grained named entity classification of wikipedia articles," *IEICE Transactions on Information and Systems*, vol. 101, no. 1, pp. 73–81, 2018.

[19] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, 2014, pp. 1188–1196.

[20] J. H. Lau and T. Baldwin, "An empirical evaluation of doc2vec with practical insights into document embedding generation," *arXiv preprint arXiv:1607.05368*, 2016.

[21] D. Onaifo and D. Rasmussen, "Increasing libraries' content findability on the web with search engine optimization," *Library Hi Tech*, vol. 31, no. 1, pp. 87–108, 2013.

[22] L. Ivanovic, B. Dimic Surla, D. Surla, D. Ivanovic, Z. Konjovic, and G. Rudic, "Improving the discoverability of Ph.D. student work through a cris system," *The Electronic Library*, vol. 36, no. 3, pp. 471–486, 2018.

[23] J. Larsson, "The retrievability of a discipline: a domain analytic view of classification," *INFORMATION RESEARCH-AN INTERNATIONAL ELECTRONIC JOURNAL*, vol. 12, no. 4, 2007.

[24] M. B. Eberly, S. E. Newton, and R. A. Wiggins, "The syllabus as a tool for student-centered learning," *The Journal of General Education*, pp. 56–74, 2001.

[25] J. T. Ishiyama and S. Hartlaub, "Does the wording of syllabi affect student course assessment in introductory political science classes?" *PS: Political Science & Politics*, vol. 35, no. 3, pp. 567–570, 2002.

[26] M. S. Palmer, L. B. Wheeler, and I. Aneece, "Does the document matter? the evolving role of syllabi in higher education," *Change: The Magazine of Higher Learning*, vol. 48, no. 4, pp. 36–47, 2016.

[27] C. E. Keller Jr, J. G. Marcis, and A. B. Deck, "A national survey on the perceived importance of syllabi components: Differences and agreements between students and instructors in the principles of accounting course." *Academy of Educational Leadership Journal*, vol. 18, no. 3, 2014.