DEIM Forum 2019 I2-3

Measuring Beginner Friendliness of Chinese Web Pages explaining Academic Concepts using Deep Learning and Text/HTML Features

Bingcai HAN†, Hayato SHIOKAWA†, Shintaro OKADA††, Chiharu HIROHANA†, Takehito UTSURO†††, Yasuhide KAWADA††††, and Noriko KANDO†††††

† Grad. Sch. of Systems and Information Engineering, University of Tsukuba, Tsukuba, 305-8573, Japan
†† College of Engineering Systems, Sch. of Sci. and Eng., University of Tsukuba, Tsukuba, 305-8573, Japan
††† Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba, 305-8573, Japan
†††† Logworks Co., Ltd., Tokyo 151-0053, Japan
††††† National Institute of Informatics, Tokyo, 101-8430, Japan

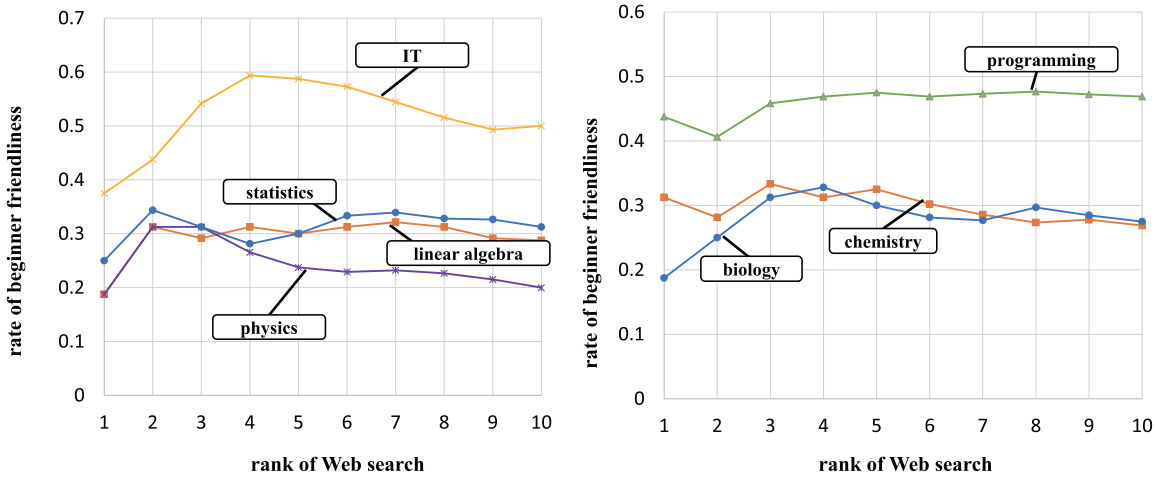
Abstract Search engine is an important tool of modern academic study, but the results are lack of measurement of beginner friendliness. For improving the efficiency of using search engine for academic study, it is necessary to find a method of measuring the beginner friendliness of Web pages explaining academic concepts and to build an automatic measurement system. In this thesis, we first formalize the measurement of beginner friendliness by several individual factors, including definition, formula and so on. We collect about 2,000 Web pages for manual measurement based on the individual factors and build a reference dataset. Then, we analyze the HTML data of the collected dataset, and extract specific features for measuring beginner friendliness. And we use a modified VGG16 model (a convolutional neural network model for image classification) to measure the layout of Web pages we have collected. The results are taken as features for further measurement. All the features are evaluated using SVM and the performance is shown in a recall-precision curve. Finally, we test about 300 Web pages and evaluate the performance of different features of HTML data and CNN measurement results of Chinese Web pages. The result of this thesis would be an important reference for further work of a practical assistance system on Web learning.

Key words Web pages explaining academic concepts, beginner friendliness, search engine, deep learning, support vector machine (SVM), convolutional neural network (CNN)

1 Introduction

A search engine is a quite important tool for obtaining concerned knowledge when it comes to the study of academic concepts. However, if we want to select the beginner friendly pages during using a search engine, it is necessary to compare and to manually select beginner friendly pages. The reason of the inefficient manual comparison is that the state-of-the-art search engines are not designed to selectively rank beginner friendly Web pages high in the searched results. For example, for Chinese Web pages explaining academic concepts, Figure 1 shows an evidence of non-existence of such systematic criterion on measuring beginner friendliness of Web pages ranked at 10th or higher by `Baidu.com` and `Google.hk`, in the case of the overall 56 queries of academic terms from the seven academic fields of linear algebra, physics, biology, programming, IT, statistics, and chemistry (listed in Table 1). In the figure, we plot the rates of the beginner friendly Web pages among those ranked at N-th or higher ($N = 1, \ldots, 10$). With this evidence, we claim that those search engines do not have any systematic criterion on measuring beginner friendliness of Web pages explaining academic concepts. Therefore, it comes up with us to invent a method of measuring the beginner friendliness of Web pages explaining academic concepts automatically, and finally to build a practical assistance system for promoting academic study using a search engine, which would improve the efficiency of Web learning.

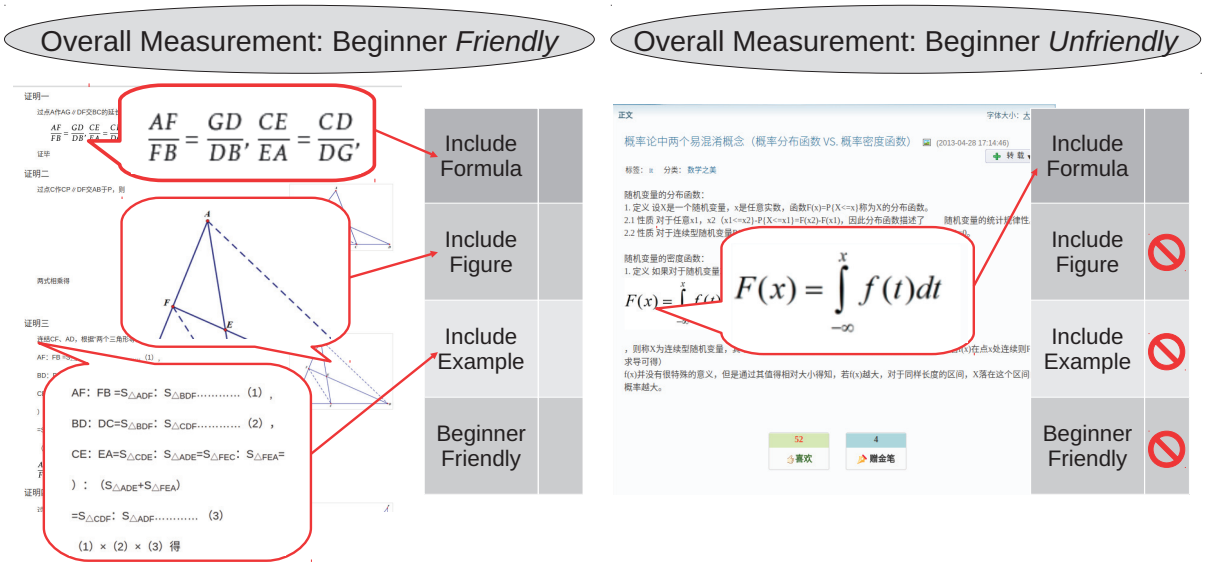
This thesis presents a method of formalizing the measurement of beginner friendliness considering several individual factors and automatic measurement of beginner friendliness using SVM based on HTML features and CNN measurement results. For collecting reference dataset, we analyze the Web pages explaining academic concepts based on HTML data, mainly the structures. And after collecting data, we extract specific features from HTML data for classifier training. As



(a) IT, statistics, linear algebra, and physics

(b) programming, biology, and chemistry

Figure 1 Rate of beginner Friendly Web Pages explaining Academic Concepts ranked at 10th or Higher in the Results of Search Engines



(a) Overall Measurement as Beginner Friendly

(b) Overall Measurement as Beginner Unfriendly

Figure 2 Examples of Web pages that explains an academic concept

a supplement, a modified CNN model based on VGG16 [10] is used for measurement of Web page layout, which is one of individual factors for the overall measurement. The features consist of HTML part and CNN part. At last, we evaluate the performance of feature combinations using recall-precision curves.

We use data from `Baidu.com` and `Google.hk`, which are two search engines mostly used for academic search in Chinese. The formalization of manual measurement of beginner friendliness is described with details in Section 2. Section 3 shows the details about reference dataset of Web pages explaining academic concepts. Section 4 describes the details of building a modified CNN model and the measurement procedure. Section 5 describes the procedure of evaluation and results. Section 6 introduces the related work of this research. Finally, Section 7 concludes this thesis.

2 Factors of Beginner Friendliness of Web Pages Explaining Academic Concepts

The measurement of individual factors and overall measurement are binary decision, and the rules are modified according to the measurement results of same Web pages by 3 members in our group.

2.1 Individual Factors

We determine several individual factors to formalize the measurement of beginner friendliness of Web pages explaining academic concepts. After prior investigation, we abstract six individual factors including definition, formula, figure, example, beginner friendliness of text and Web page layout [1].

Table 1 The Details of Reference Dataset

Academic Fields		Number of Queries	Queries	Number of Training/Test Web Pages		
				Positive	Negative	**Total**
Training	线**性代数**/**線形代数**/ Linear algebra	8	阶**数**/**階数**/Order, **行列式**/**行列式**/Determinants, 对**角化**/**对角化**/Diagonalization, **内**积/**内積**/Inner Product **正交矩**阵/**直交行列**/Orthogonal matrix, **三角矩**阵/**三角行列**/Triangular matrix, **特征多**项**式**/**特性多項式**/Characteristic polynomial 标**准正交基**/**正規直交基底**/Standard orthonormal basis,	77	80	157
	物理/ 物理/ Physics	8	离心力/遠心力/Centrifugal force，声波/音波/Sound wave， 交流电/交流/AC，正电荷/正電荷/Positive charge， 速度/速度/Speed，电波/電波/Radio waves， 万有引力/万有引力/Gravitation，变压器/変圧器/Transformer	68	82	150
	生物/ **生物**/ Biological	8	DNA，果蝇/ショウジョウバエ/Drosophila， 叶绿体/葉緑体/chloroplast，原核生物/原核生物/Prokaryote， 减数分裂/減数分裂/Meiosis，光合作用/光合成/photosynthesis， 细胞/細胞/cell，RNA	61	84	145
	编**程**/ **プログラミング**/ Programming	8	C语言/C 言語/C language，指针/ポインタ/Pointer， 迭代处理/繰り返し処理/Iterative processing，Java， 结构/構造体/Structure，数组变量/配列変数/Array variable， 条件分支/条件分岐/Conditional branch，字符串/文字列/String	77	64	151
	IT	8	API, DBMS, HTML, SDK, SQL, Unicode, URL, IP地址/IP アドレス/IP Address	70	78	148
	Total	40		343	398	751
Test	统计/ 統計/ Statistics	8	回归分析/回帰分析/Round analysis, F分布/F分布/F-distribution, 伽玛分布/ガンマ分布/Gamma distribution, 概率/確率/Probability, 相关系数/相関係数/Correlation coefficient, 方差/分散/Dispersion, 正态分布/正規分布/Normal distribution, 协方差/共分散/Covariance	59	78	137
	化学/ **化学**/ Chemistry	8	化学反应式/化学反応式/Chemical reaction formula， 化学平衡/化学平衡/Chemical equilibrium， 过渡元素/遷移元素/Transition element，氧化还原/酸化還元/Redox， 天然聚合物/天然高分子/Natural polymer， 燃料电池/燃料電池/Fuel cell，酯/エステル/Ester， 合成聚合物/合成高分子/Synthetic polymer	66	76	142
	Total	16		125	154	279
Total		56		468	542	1010

Each factor is judged based on measurement by manual work according to several rules.

(a) Definition: measured as positive when a Web page contains a correct and precise definition of the explained academic concept.

(b) Formula: measured as positive when a Web page contains any formula whether in text or figures. The formulas should be relevant to the academic concept explained in the Web page.

(c) Figure: measured as positive when a Web page contains figures or pictures relevant to the academic concept explained in the Web page, except when the figure shows any formula only.

(d) Example: measured as positive when a Web page contains examples relevant to the explained academic concept, including examples of application, proof, explanation and so on. When the examples shown in figures, it would be measured as positive for both figure and example.

(e) Beginner friendliness of text: measured as positive when the text of Web page is considered as beginner friendly by the annotator. The measurement should not take other factors into consideration.

(f) Web page layout: measured as positive when the layout of the Web page is considered as easy to read by the annotator. The measurement should not take other factors into consideration.

2.2 Overall Measurement considering Individual Factors

The overall measurement of the beginner friendliness of Web pages explaining academic concepts is performed by each of our group members. A Web page could be measured as beginner friendly as in the case of the example shown in Figure 2(a), or beginner unfriendly as in the case of the example shown in Figure 2(b). The overall measurement is

Table 2 Details of Manual Measurement Rules

Factors	Absolute Rules	Optional Rules
Beginner Friendliness of Text	1. Description is in Order 2. Structure of text is clear 3. Use of concerned unknown concepts is no more than 3 times	1. Main contents contain a clear outline 2. Important parts are marked with obvious marks 3. Other language is used as little as possible
Web Page Layout	1. No Ads or unrelated blocks are contained in main contents 2. Main contents are appeared soon after first click 3. Design of color is comfortable for reading	1. Functional blocks are clear and obvious 2. Ads in the whole page are as few as possible
Overall Measurement of Beginner Friendliness	1. Factor definition is measured positive 2. Factor example is measured positive 3. Factor beginner friendliness of text is measured positive 4. Factor Web Page layout is measured positive	1. Enough useful information (include formula, figures and so on) 2. Information or explanation provided by official 3. Positive comments to main contents

performed not only based on the combination of individual factors but also based on which factor is measured as positive. A Web page explaining academic concept would be measured as beginner friendly when the result of the measurement of individual factors can be recognized as showing that the page contains necessary and enough information helpful for beginner learning.

3 Reference Dataset of Web Pages Explaining Academic Concepts

In this section, we collect the dataset of Web pages explaining academic concepts and develop a reference dataset by annotating them with beginner friendliness.

3.1 Academic Fields and Concepts for Study

After prior investigation, we determine that the Web pages are collected on 7 academic fields, including statistics, physics and so on. For each field, we choose 8 queries for academic concepts based on the materials of high school and collage in China, and then we collect the top 10 pages of each query from the two search engines `Baidu.com` and `Google.hk`. Table 1 shows the details about the collection.

3.2 Procedure

For building the reference dataset, we choose the two search engines `Baidu.com` and `Google.hk` for collecting Chinese Web pages. For each search engine, we collect the URL data of top 10 Web pages(注1) for each query determined before. Here, the inaccessible and unrelated Web pages would be ruled out.

Among the factors listed in Section 2, those other than Web page layout, beginner friendliness of text, and overall measurement, we measure the individual factors as described in Section 2. For Web Page layout, beginner friendliness of text, and overall measurement, we consider several absolute and optional rules, whose details are shown in Table 2. The measurement process using absolute and optional rules are shown below.

(a) Beginner friendliness of text: For positive measurement, the absolute rules must be all satisfied, and any one of three optional rules must be satisfied. The absolute rule 2 is subjective. If its measurement is not so clear, it should be determined by the optional rules 1 and 2, which means the structure of text is considered as clear if the optional rules 1 and 2 are satisfied.

(注1) Out of those top 10 Web pages, we exclude those from the Web sites `baike.baidu.com`, `www.baike.com`, `zh.wikipedia.org`, `wiki.mbalib.com`, and `zh.wikihow.com`, and we recollect top 11 to 20 searched Web pages. Those Web sites share many top ranked Web pages across those 7 academic fields. Each site mostly has Web pages of the same Web page layout. So, it is necessary to exclude Web pages of those five Web sites, since Web pages of those five Web sites tend to have mostly the same Web page layout, where those Web pages may appear both in the training data and the development/evaluation data.

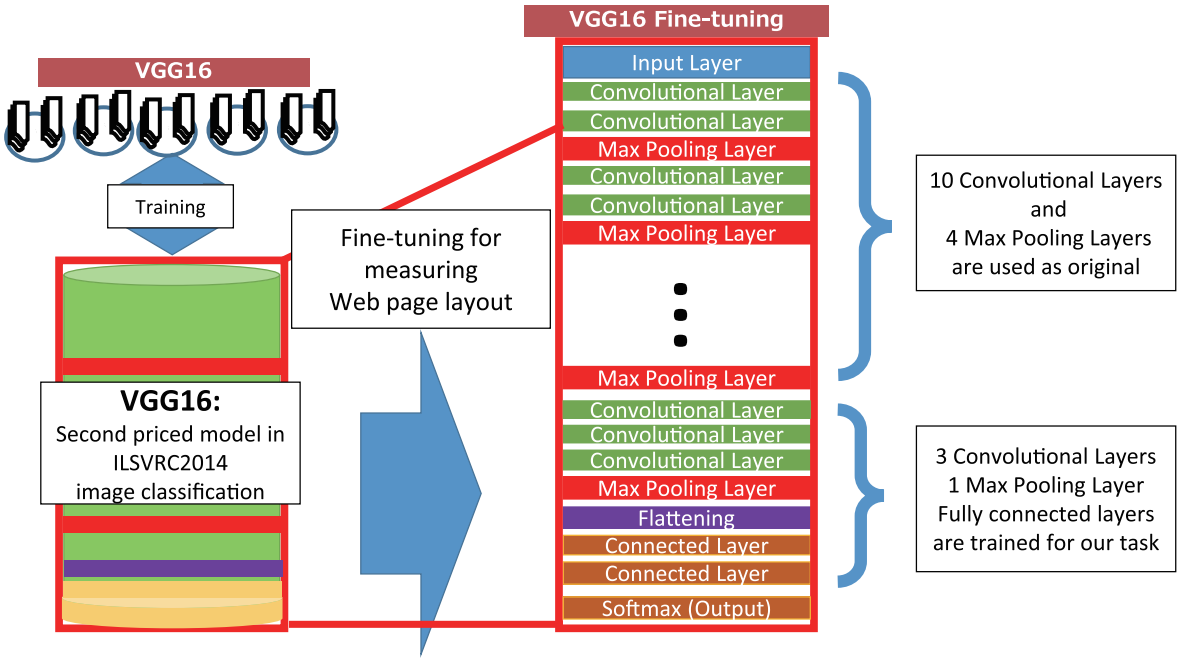


Figure 3 Procedure of Modifying VGG16 for Our Task

(b) Web Page layout: For positive measurement, the absolute rules must be all satisfied, and any one of the two optional rules must be satisfied.

(c) Overall measurement of beginner friendliness: For positive measurement, the absolute rules must be all satisfied, and any two of the three optional rules must be satisfied.

3.3 Reference Dataset

The reference dataset of Web pages explaining academic concepts contains over 1,000 Web pages. During the measurement, some Web pages showing only academic papers or books are considered as unmeasurable and ruled out from the final dataset. The results of manual measurement are shown in Table 1. We use the dataset and collected HTML files for extracting features, classifier training and evaluation, which are described in next section.

4 Measurement of Web Page Layout of Web Pages Explaining Academic Concepts by Deep Learning

4.1 VGG16 [10]

In recent years, deep learning shows quite high performance and reliability in the tasks of various fields. Especially in the field of image recognition, the appearance of convolutional neural network (CNN) and large scale datasets such as ImageNet [5,6] has demonstrated remarkable performance in various tasks. It is also known that the CNN parameters learned using large datasets such as ImageNet can be used in different domain tasks as high performance feature extractors. In this thesis, VGG16 model is used as a basic feature extractor dealing with the task of measurement of Web page layout of Web pages explaining academic concepts. VGG16 is a CNN model which has won the second prize for image classification and first prize for single-object location in ImageNet Large Scale Visual Recognition Challenge(ILSVRC) in 2014 [6]. This model consists of 13 layers of convolutional layers, 5 layers of max pooling layers, 3 fully connected layers (the first two layers have 4,096 units each and the last layer is a soft-max classification layer with 1,000 units representing the 1,000 ImageNet classes), and output layer. For the trained VGG16 model, the model is trained with the ImageNet 2014 dataset for the 1,000 ImageNet classification task, which has been publicly available. It is also known that this trained model can be widely used for other tasks. In this thesis, the experiment and evaluation are carried out using the model published in Keras(注2), which is a deep learning library of Python.

4.2 Procedure of Automatic Measurement of Web Page Layout

In this thesis, we measured the Web page layout of Web pages explaining academic concepts using the trained VGG16 model as a base model. For building the dataset for VGG16 model training, we take the screen shots of Web pages we have collected . The details of the collected dataset are shown in Table 1. After preparing the dataset, we moved to prepare the CNN model for the measurement task. As shown in Figure 3, we use the trained VGG16 model as a base model. Since the model is trained for 1,000 ImageNet classification task, it is necessary to modify the trained model.

(注2) https://keras.io/ja/

First, we remove the 2 fully connected layers and output layer which are designed and applicable for 1,000 classification task, and replace them with new 2 fully connected layers and output layer for 2 classification task which is to measure a Web page in good layout or not in this parer. And then we modify the whole model into a new model, so called model with fine-tuned VGG16. For this model, we keep the original 10 convolutional layers and first 4 max pooling layers with their parameters as trained for ImageNet, and then we train the remain 3 convolutional layers, 1 max pooling layer and fully connected layers, then set new parameters with our prepared dataset. The training and test data details are shown in Table 1.

5 Classifier Learning using Features from HTML Data and CNN Model and Evaluation

In this section, we use the collected HTML data and measurement results by modified VGG16 of Web pages explaining academic concepts for training the classifier and evaluate the performance of various features.

5.1 Whole Procedure of Evaluation

Since we would take the CNN measurement results as a feature, which would be taken in SVM training and evaluation, we first train the modified VGG16 model with the collected dataset(注3). After the preparation of dataset with all the features built in, the evaluation would be performed by single feature and feature combination separately for comprehensiveness and reliability.

5.2 Features

The features used for classifier learning are described as below, including two parts as features from HTML data and measurement results form the CNN model. And for features from HTML data, there are two source of features as HTML structure and text contents.

5.2.1 Extracted from HTML Structures

The features extracted from HTML structures include three parts as height of Web page (in pixel), HTML tag and HTML tag attribute. The details of the features are shown as below.

(a) Height of Web page: the height of body contents of Web page (in pixel). This feature is multi-valued.

(b) HTML tag: the total number of times of use of the HTML tag in the page. The measured tags include *media, iframe, div, sizes, meta, nav, required, td, h1, h2, h3 and small*. This feature is multi-valued.

(c) HTML tag attribute: the total number of times of use of the HTML tag attributes in the page. The measured tag attributes include *index, allowtransparency, data-layout, frameborder, method, data-screen-name and onload*. This feature is multi-valued.

As supplement, the features of URL are also used. Here, we use a feature which represents whether the URL includes the string "news". This is a binary feature.

5.2.2 Features Extracted from Text Contents

We also extract features from text contents which are contained in HTML structure. The details are shown below.

(a) Text Character: the difference in the number of characters, which equals the result of the number of Chinese characters minus the number of alphabets in each page's main text. This feature is multi-valued.

(b) Include "応用 (application)": the frequency of strings "応用 (application)" in the HTML of each page. This feature is multi-valued.

(c) Include "如何 (how)": the frequency of characters "如何 (how)" in the HTML of each page. This feature is multi-valued.

(d) Include "· · · 法 (principle)": the frequency of string patterns "· · · 法 (principle)" in the HTML of each page. This feature is multi-valued.

(e) Include "首頁 (home)": whether there is a string "首頁 (home)" in the HTML of each page. This feature is a binary feature.

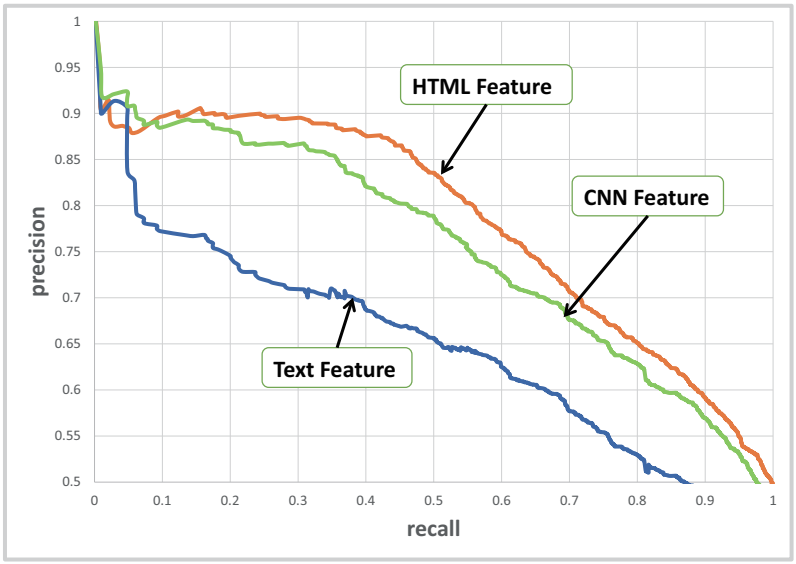
5.2.3 Features Obtained from CNN Model Measurement Results

From the measurement results of section 4.2, we could obtain the confidence value of each page on factor of Web page layout. The confidence value is within the range of [0,1]. Taking 0.5 as a midpoint naturally, the page has higher possibility in good Web page layout when the confidence value is close to 1, and the page has higher possibility in not good Web page layout when the confidence value is close to 0. The confidence value would be taken as a feature for the overall measurement.

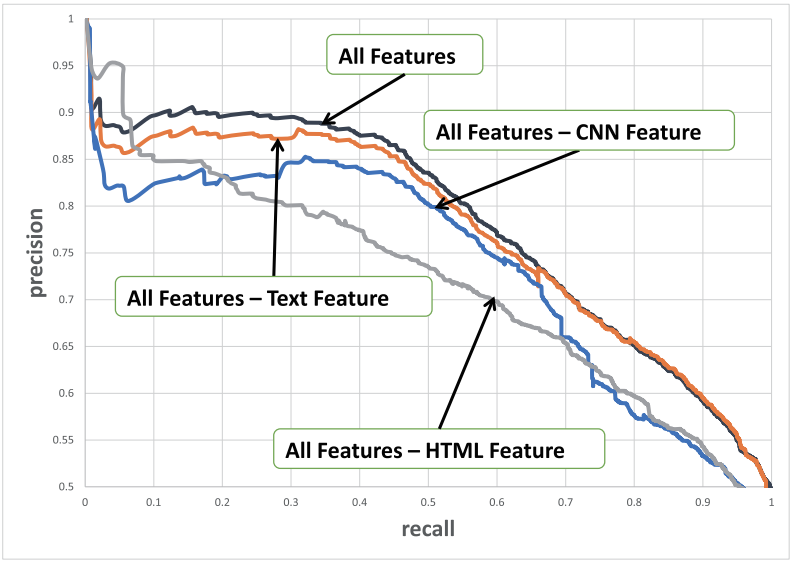
5.3 Detailed Procedure of SVM Training

For classifier, we used SVM (sklearn.SVM.SVC tool) [4] in the scikit-learn package. For features that take multi-valued values of three or more values, we have evaluated various types of method for setting value as binary feature with consideration of the distribution of value, such as binary features taken from multiple ranges without overlap, binary features taken from multiple ranges with overlap and etc. As a result, a feature setting method is used, which sets left end fixed and right end with 20 to 40 different ranges, since this setting method achieves the best performance. We use the

(注3) Out of the five academic fields for training, we use four for training the modified VGG16 model and the remaining one for obtaining the SVM feature by applying the modified VGG16 model to the remaining one academic field. We repeat this procedure five times and obtain the SVM feature for all the five academic fields.



(a) Performance of Single Features



(b) Performance of Various Feature Combinationsw

Figure 4 Evaluation Results

RBF kernel as the kernel function of the SVM and optimize the cost parameter C (1 or 10) and the parameter γ (0.01, 0.001 and 0.0001) of the RBF kernel by grid search, with the area of the recall-precision curves taken as the objective function of optimization. Then we trained the SVM with the labeled data of overall measurement results.

5.4 Evaluation Results

For now, we have obtained features sorted into 3 categories as shown below.

(a) HTML features: features obtained from the HTML structure, mainly the tags.

(b) Text features: features obtained from text contents in the HTML structure.

(c) CNN feature: taken from the CNN measurement results as the feature directly.

And we evaluate the performance of using each feature only to build a baseline for further evaluation. The results are shown in Figure 4(a) below. And it is seen that the feature from HTML structure has achieved the best performance among the single features.

To evaluate the best performance of our method, we evaluate the following feature combinations.

(a) All features

(b) All features − CNN feature

(c) All features − Text feature

(d) All features − HTML feature

Evaluation results are shown in Figure 4(b). It is seen that when all the features are used, the method achieves the best performance. Among the three of CNN, Text, and HTML features, the HTML feature has the most influence on the

performance since the combinations (a) and (d) have the maximum difference.

6 Related Work

As a concerned field with this research, the community Question-Answering selection is focused [2, 7], where methods of evaluating and selecting good answers in Question-Answering system are studied, which is related to the measurement of beginner friendliness discussed in this thesis to a certain extent. For Japanese Web pages explaining academic concepts, related methods are studied by other members in our group, including using SVM to perform the measurement based on HTML structures [3] and image recognition using deep learning for measuring the visual intelligibility of the Web page layout [8, 9].

7 Conclusion

This thesis presents a method of automatic measurement of beginner friendliness of Web pages explaining academic concepts using SVM, and evaluates the performance of various features from HTML data and CNN measurement results. The evaluation results of SVM show that it has the best performance when all the features are used for the classifier learning. And it is proved that HTML data is applicable for measuring the beginner friendliness of Web pages explaining academic concepts. The use of HTML structure is quite important for the automatic measurement according to the comparison figure. And the image recognition by deep learning is a reliable method to solve the measurement of Web page layout. It does benefit the overall measurement of beginner friendliness of Web pages explaining academic concepts. The method presented in this thesis would be a base idea of measuring the beginner friendliness of Web pages automatically, and it would also be an important reference for further work of building a complete assistance system for academic study using a search engine.

References

[1] B. Han, H. Shiokawa, K. Kawaguchi, T. Utsuro, and Y. Kawada. Measuring beginner friendliness of Chinese Web pages explaining academic concepts using HTML structures. 第 32 回人工知能学会全国大会論文集, 2018.

[2] 石川大介, 酒井哲也, 関洋平, 栗山和子, 神門典子. コミュニティ QA における良質回答の自動予測. 情報知識学会誌, Vol. 21, No. 3, pp. 362–382, 2011.

[3] 春日孝秀, 塩川隼人, 韓炳材, 宇津呂武仁, 河田容英. HTML 構造上の特徴を利用した学術用語解説ウェブページの分かり易さの自動評定. 第 10 回 DEIM フォーラム論文集, 2018.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830, 2011.

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, Vol. 115, No. 3, pp. 211–252, 2015.

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F.-F. Li. ImageNet large scale visual recognition challenge. *CoRR*, Vol. abs/1409.0575, , 2014.

[7] T. Sakai, D. Ishikawa, N. Kando, Y. Seki, K. Kuriyama, and C.-Y. Lin. Using graded-relevance metrics for evaluating community QA answer selection. In *Proc. 4th WSDM*, pp. 187–196, 2011.

[8] 塩川隼人, 春日孝秀, 川口輝太, 韓炳材, 宇津呂武仁, 河田容英. 異種の素性を併用した学術用語解説ウェブページの分かり易さの自動評定. 第 32 回人工知能学会全国大会論文集, 2018.

[9] H. Shiokawa, K. Kawaguchi, B. Han, T. Utsuro, Y. Kawada, M. Yoshioka, and N. Kando. Measuring beginner friendliness of Japanese Web pages explaining academic concepts by integrating neural image feature and text features. In *Proc. 5th NLPTEA*, pp. 143–151, 2018.

[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. 3rd ICLR*, 2015.