The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

Measuring Beginner Friendliness of Chinese Web Pages Explaining Academic Concepts using HTML Structures

Bingcai HAN[*1] Takahide KASUGA[*1] Hayato SHIOKAWA[*1] Kota KAWAGUCHI[*1]
Takehito UTSURO[*2] Yasuhide KAWADA[*3]

[*1]Grad. Sc. of Systems and Information Engineering, University of Tsukuba
[*2]Faculty of Engineering, Information and Systems, University of Tsukuba
[*3]Logworks Co., Ltd.

Search engine is an important tool of modern academic study, but the results are lack of measurement of beginner friendliness. For improve the efficiency of using search engine for academic study, it is necessary to find a method of measuring the beginner friendliness of Web page explaining academic concepts and build an automatic measurement system. In this research, we formalize the measurement of beginner friendliness by several individual factors, including definition, formula and so on. We collect about 900 Web pages for manual measurement using the individual factors to build a reference dataset. Then we analyze the HTML structure of the collected dataset, and craft some rules for measuring four individual factors. The crafted rules are tested using program and the performance is evaluated by recall and precision. Finally, we test about 400 Web pages and find some special features of HTML structure of Chinese Web pages. The modified rules for the features perform better than general rules we first crafted. The results of this research would be the an important reference for further work for a whole automatic measurement system of beginner friendliness.

1. Introduction

Search engine is a quite important tool for getting concerned knowledge when it comes to the study of academic concepts. However, if we want to select the beginner friendly pages during using search engine, it is necessary to compare pages by manual work. The reason of ineffective manual comparison is that there is no systematization for beginner friendly Web pages in the results of search engine. Therefore, it comes up with us to find a method of measuring the beginner friendliness of Web pages explaining academic concepts automatically, and finally build a whole assisting system for promoting academic study using searching engine, which would improve the efficiency of web learning.

This research presents a method of formalizing the measurement of beginner friendliness by several individual factors. For collecting reference dataset, we analyze the Web pages explaining academic concepts based on HTML structures. And after collecting data, we select four factors to craft rules for measurement using HTML, and then build a Python program based on the crafted rules to verify the feasibility of measurement using HTML.

We use data from Baidu.com and Google.hk, which are two search engines mostly used for academic search in Chinese. The formalization of 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. And Section 4 describes the rules used in program of measuring beginner friendliness and the evaluation of the program. Section 5 introduces the related work of this research. Finally, Section 6 concludes this paper.

Contact: Takehito UTSURO, Graduate School of Systems and Information Engineering, University of Tsukuba, 305-8573 Japan, 029-853-5427

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 persons in our group.

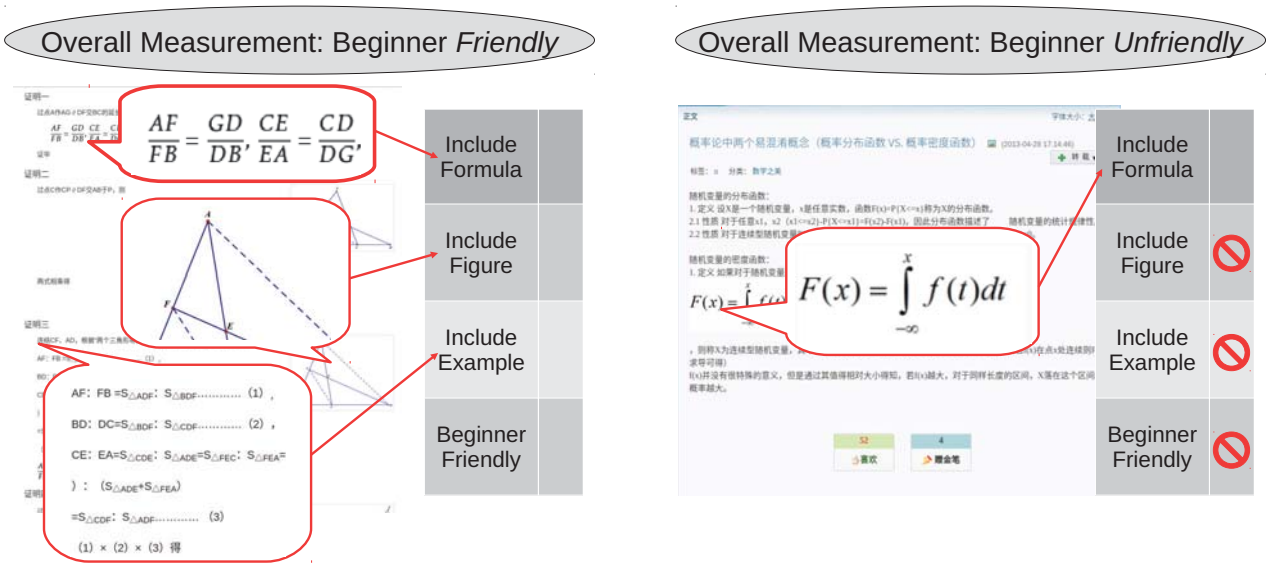
2.1 Individual Factors

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. For each factor, there are some basic rules for measurement by manual work.

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

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

(c) Figure: measured positive when a Web page contains figures or pictures relevant to the academic



(a) Overall Measured Beginner Friendly

(b) Overall Measured Beginner Unfriendly

Figure 1: Examples of Web pages that explains an academic concept

concept explained in the Web page, except when the figure shows formula only.

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

(e) Beginner friendliness of text: measured positive when the text of Web page is considered beginner friendly by evaluator. The measurement should not take consider in the other factors.

(f) Web page layout: measured positive when the layout of Web page is considered easy to read by evaluator. The measurement should not take consider in the other factors.

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. The rules of overall measurement are modified by results of different members. Parts of overall measurement is shown in Figure 1. A Web page explaining academic concept would be measured as beginner friendly when the measurement of each individual factor is considered benefiting the beginner friendliness. The overall measurement is performed by not only the combination of individual factors but also which factor is measured positive. For positive situation, the Web page should contain at least one of Web page layout or beginner friendliness of text measured positive. If a Web page contains only one measured positive of layout or beginner friendliness of text, it would be measured positive for overall measurement when there are enough positive results for the other factors. And if the layout and beginner friendliness of text are both measured positive, the overall measurement would be positive when there is at least one positive measurement of the other factors.

3. Reference Dataset of Web Pages Explaining Academic Concepts

In this section, we collect the reference dataset of Web pages explaining academic concepts based on the rules determined in Section 1. The dataset would be used for building automatic measurement program and test.

3.1 Academic Fields and Concepts for Study

After prior investigation, we determine that the Web pages are collected on three academic fields, including statistics, physics and linear algebra. These filed have same features for measurement of individual factors. For each field, we choose 15 queries for academic concepts based on the teaching content of high school and collage in China, and then we collect the top 10 pages of each query. Table 1 shows the details about the collection.

3.2 Procedure

For building the reference dataset, we choose two search engines as Baidu and Google.hk for collecting Chinese Web pages. For each search engine, we collect the URL data of top 10 Web pages for each query determined before. Then we measure the individual factors of every Web page according to the rules determined in Section 2. The inaccessible and unrelated Web pages would be ruled out. The manual measurement results are sorted by academic fields and search engines for

Table 1: The Fields and Concepts of Reference Data

Fields	Queries
统计/統計/Statistics	惠斯通电桥/ホイートストンブリッジ/Wheatstone Bridge，离心力/遠心力/Centrifugal force，声波/音波/Sound wave，惯性定律/慣性の法則/Law of inertia，交流电/交流/AC，正电荷/正電荷/Positive charge，速度/速度/Speed，张力/張力/Tension，电力线/電気力線/Electric lines of force，电波/電波/Radio waves，电流/電流/Current，恢复系数/反発係数/Coefficient of restitution，变压器/变圧器/Transformer，万有引力/万有引力/Gravitation，互感/相互誘導/Mutual induction
物理/物理/Physics	回归分析/回帰分析/Round analysis，概率/確率/Probability，概率密度函数/確率密度関数/Probability density function，伽玛分布/ガンマ分布/Gamma distribution，主成分分析/主成分分析/Principal component analysis，自回归/自己回帰/Autoregression，置信区间/信頼区間/Confidence interval，相关系数/相関係数/Correlation coefficient，独立成分分析/独立成分分析/Independent component analysis，标准差/標準偏差/Standard deviation，方差/分散/Dispersion,正态分布/正規分布/Normal distribution，泊松分布/ポアソン分布/Poisson distribution，零假设/帰無仮説/Null hypothesis，协方差/共分散/Covariance
线性代数/線形代数/Linear algebra	阶数/階数/Order，行列式/行列式/Determinants，对角化/対角化/Diagonalization，梅涅劳斯定理/メネラウスの定理/Menelaus's theorem，正交矩阵/直交行列/Orthogonal matrix，共轭梯度/共役勾配/Conjugate gradient，范数/ノルム/Norm，克罗内克函数/クロネッカーのデルタ/Kronecker function，克莱姆法则/クラメルの公式/Clems law，三角矩阵/三角行列/Triangular matrix，标准正交基/正規直交基底/Standard orthonormal basis，特征多项式/特性多項式/Characteristic polynomial，雅可比矩阵/ヤコビ行列/Jacobian matrix，二次型/二次形式/Secondary type，线性变换/線形変換/Linear transformation

efficient analysis of HTML structures. For analysis of HTML structures, we collect HTML contents of the reference dataset from the recorded URL using program. The analysis results are used for crafting rules in next section.

3.3 Reference Dataset

The final reference dataset of measured Web pages explaining academic concepts contains about 900 Web pages. During the measurement, some Web pages showing only academic papers or books are considered as unmeasurable and ruled out from final dataset. The results of manual measurement are shown in Table 2. We use the dataset and the results of analysis for crafting rules of automatic measurement program and testing, which are described in next section.

4. Measuring Individual Factors by Manually Crafted Rules

According to the analysis results, we craft rules for four individual factors, including figure, formula, example and beginner friendliness of text. The rules are modified referring to the features of HTML structures appeared in Chinese Web pages.

4.1 Manually Crafted Rules

The rules are all crafted based on the features of HTML structures. During the analysis of HTML structures, we find that some features of HTML of Chinese Web pages are totally different with other Web pages, mainly from Baidu.com and zhihu.com, which make up almost half of the collected Web pages. Referring to the HTML features of Chinese Web pages, the crafted rules of the selected four factors are shown as blow.

(a) Formula: measured positive when a Web page includes tag “math” in HTML, mathematical symbols and numbers in any attribute value in HTML. It is also measured positive when an equal symbol(=) is included in the text of HTML.

(b) Figure: measured positive when a Web page contains typical and specific tags of picture in HTML, like “pic”, “figure” and so on. And the picture should also be larger than 100×100 (in pixel) without link embedded in it.

(c) Example: measured positive when a Web page contains specific keywords or word patterns of example in HTML, like “例”, “Example” and so on.

(d) Beginner friendliness of text: the measurement of this factor is performed by a combination of several rules of HTML, including pattern of attribute values, size of “body” component and text composition in HTML. In general, it is measured positive when a Web page is in a suitable size with simple text in HTML.

4.2 Evaluation Procedure

The automatic measurement program is programmed with Python. For test data, we select 300 Web pages, 100 Web pages for each field with 5 queries averagely. The manual measurement is the reference for the evaluation of program. We record the measurement results of the measurement program and compare it with the manual results, and then calculate the recall and precision of the results to evaluate the performance of our crafted rules based on HTML structures.

4.3 Evaluation Results

The evaluation results are shown in Figure 2. From the figure, it is seen that the measurement program achieves an acceptable performance, especially for beginner friendliness of text. The results show that mea-

Table 2: Details of Manual Measurement Results

Fields	Number of Queries	Total Pages	Measured Pages	Include Definition	Include Figure	Include Formula	Include Example	Web Page Layout	Beginnner Friendly Text	Positive Overall Measurement
统计/ 統計/ Statistics	15	300	287	64%	66%	78%	50%	76%	80%	61%
物理/ 物理/ Physics	15	300	263	56%	50%	74%	42%	80%	88%	63%
线性代数/ 線形代数/ Linear algebra	15	300	279	58%	56%	50%	30%	75%	68%	74%
Summation	45	900	829	59%	57%	67%	41%	77%	79%	66%

suring the four factors using HTML structure is feasible and the method could be applied and tested for the other two factors.

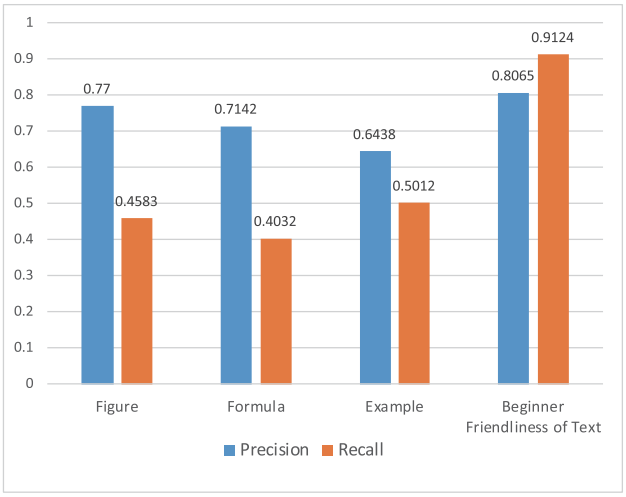


Figure 2: Evaluation Results of Automatic Measurement of Beginner Friendliness

5. Related Work

As a concerned field with this research, the community Question-Answering selection is focused by some researchers [石川 11, Sakai11]. These researches present methods of evaluating and selecting good answers in Question-Answering system, which has much same with the measurement of beginner friendliness discussed in our research. For Japanese Web pages explaining academic concepts, similar methods for the measurement using HTML are discussed in the studies by other members in our group, including using SVM to perform the measurement based on HTML structures [春日 18] and image recognition using deep learning for measuring the Web page layout [塩川 18].

6. Conclusion

This research presents a method of formalizing the measurement of beginner friendliness of Web pages by individual factors, and evaluates the method of measuring four individual factors using HTML structures. The evaluation results of the measurement program show that the crafted rules perform quite well. And it is proved that HTML structure is applicable for measuring the individual factors of beginner friendliness of Web pages explaining academic concepts. The method presented in this research would be an inspiration of finding a method for measuring the beginner friendliness of Web pages automatically, and it would also be an important reference for further work of building a complete assisting system for academic study using search engine.

References

[石川 11] 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).

[春日 18] 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).

[Sakai11] Sakai, T., Ishikawa, D., Kando, N., Seki, Y., Kuriyama, K. and Lin, C.-Y.: Using Graded-Relevance Metrics for Evaluating Community QA Answer Selection, *Proc. 4th WSDM*, pp. 187–196 (2011).

[塩川 18] 塩川隼人, 春日孝秀, 韓炳材, 宇津呂武仁, 河田容英：深層学習を用いた学術用語解説ウェブページの見易さの自動評定, 第 10 回 DEIM フォーラム論文集 (2018).