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

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 analyse the Web pages explaining academic concepts based on HTML structures. And after collecting data, we select four factors to craft rules for measurement based on HTML, and then build a Python program based on the crafted rules to verify the feasibility of measurement based on 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.

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.
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
Table 1: The Fields and Concepts of Reference Data

<table>
<thead>
<tr>
<th>Fields</th>
<th>Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>统计/数学</td>
<td>Wheatstone Bridge, 离心力/向心力/Centrifugal force,</td>
</tr>
<tr>
<td>(Statistics)</td>
<td>声波/音波/Sound wave, 惯性定律/惯性/力学的法则/Law of inertia, 交流电/交流/AC, 正电荷/正电荷/Positive charge,</td>
</tr>
<tr>
<td></td>
<td>速度/速度/Speed, 张力/张力/Tension, 电力线/电力线/Electric lines of force, 电波/电波/Radio waves, 无线电/电波/Current, 恢复系数/反感系数/Coefficient of restitution,</td>
</tr>
<tr>
<td></td>
<td>变压器/变压器/Transformer, 万有引力/万有引力/Gravitation, 互感/互感/Coupling, 互感/互感/Mutual inductance</td>
</tr>
<tr>
<td>理论/物理</td>
<td>回归分析/回归分析/Regression analysis, 概率/概率/Probability, 概率密度函数/概率密度函数/Probability density function, 伽玛分布/伽玛分布/Gamma distribution, 主成分分析/主成分分析/Principal component analysis, 自回归/自回归/Autoregression, 置信区间/置信区间/Confidence interval, 相关系数/相关系数/Correlation coefficient, 独立成分分析/独立成分分析/Independent component analysis, 标准差/标准差/Standard deviation, 方差/方差/Variance, 正态分布/正态分布/Normal distribution, 梯形分布/梯形分布/Trapezoidal distribution, 零假设/零假设/Null hypothesis, 协方差/共分散/Covariance</td>
</tr>
</tbody>
</table>

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 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 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 measuring the four factors based on HTML structure is feasible and the method could be applied and tested for the other two factors.

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 based on 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.

Reference


Table 3: Example of Measurement of Individual Factors and Overall Measurement
Figure 3: Examples of Individual Factors