Bilingual Web Page and Site Readability Assessment

Tak Pang Lau
Dept. of Comp. Sci. and Eng.
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
tplau@cse.cuhk.edu.hk

Irwin King
Dept. of Comp. Sci. and Eng.
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
king@cse.cuhk.edu.hk

ABSTRACT

Readability assessment is a method to measure the difficulty of a piece of text material, and it is widely used in educational field to assist instructors to prepare appropriate materials for students. In this paper, we investigate the applications of readability assessment in Web development, such that users can retrieve information which is appropriate to their levels. We propose a bilingual (English and Chinese) assessment scheme for Web page and Web site readability based on textual features, and conduct a series of experiments with real Web data to evaluate our scheme. Experimental results show that, apart from just indicating the readability level, the estimated score acts as a good heuristic to figure out pages with low textual content. Furthermore, we can obtain the overall content distribution in a Web site by studying the variation of its readability.

Categories and Subject Descriptors

I.7.m [**Document and Text Processing**]: Miscellaneous - Text Readability

General Terms

Experimentation, Human Factors, Measurement

Keywords

Web Pages, Web Sites, English, Chinese, Readability

1. INTRODUCTION

The World Wide Web contains vast amount of valuable information, but there is not enough guidance for users to find information that is appropriate to their reading ability levels. When a user raises a query, existing search engines mainly return semantically-related materials, but whether the user has sufficient ability to understand the materials is often overlooked.

To enhance user experience, we investigate the use of readability assessment, which is a method to measure the difficulty of a piece of text material [5], in Web information retrieval. We propose a bilingual (English and Chinese) Web page and site readability assessment scheme based on textual features of the pages. As English and Chinese pages cover over 70% of the total Web pages [2], our scheme will have a high impact on the Internet community.

2. WEB READABILITY

Web Page Readability. For Web page readability, we adopt two existing assessments, Flesch [4] formula and Yang [6] formula, proposed in educational field originally as a baseline evaluation. The resulting page readability formula is as follow:

$$r_p = \begin{cases} -84.6X_{E_1} - 1.015X_{E_2} + 206.835 & \text{for English,} \\ 2 \times \{13.90963 + 1.54461X_{C_1} + \\ 39.01497X_{C_2} - 2.52206X_{C_3} - \\ 0.29809X_{C_4} + 0.36192X_{C_5} + \\ 0.99363X_{C_6} - 1.64671X_{C_7}\} & \text{for Chinese.} \end{cases} \quad (1)$$

where X_{E_i} and X_{C_i} are the factors:
- X_{E_1}: Average number of syllables per word;
- X_{E_2}: Average sentence length;
- X_{C_1}: Proportion of full sentence;
- X_{C_2}: Proportion of words in Chinese basic word list;
- X_{C_3}: Average number of stroke of characters;
- X_{C_4}: Num of chars with stroke = 5 (in 100 chars sample);
- X_{C_5}: Num of chars with stroke = 12 (in 100 chars sample);
- X_{C_6}: Num of chars with stroke = 22 (in 100 chars sample);
- X_{C_7}: Num of chars with stroke = 23 (in 100 chars sample);

The score of this formula is scaled between 0 and 100. The smaller the score, the more difficult the text passage is.

Web Site Readability. Web site readability is an indicator of overall difficulty level of a site, and it is defined over the proposed page readability. We first describe the following terms before continuing the discussion. *Web site* is a group of Web pages with the same domain name in the URLs. *Root page* of a Web site, is a user-specified page where the crawling of the site starts. *Page level* of a Web page within a site is the minimum number of traversal reaching it starting from the root page of the Web site through hyperlinks.

We propose three site readability assessments in terms of pages at different page levels: (1) *Exact-Level*, (2) *In-Level*, and (3) *Out-Level* (depicted in Figure 1), aiming at describing the site difficulty from different angles of pages composition.

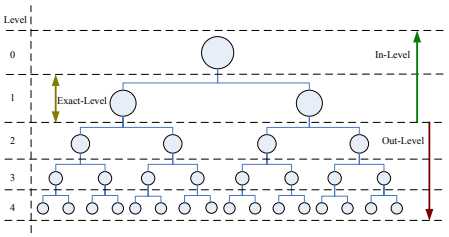

Figure 1: Illustration of Exact-Level, In-Level, and Out-Level Site Readability at level=1.

Exact-Level Readability indicates the average readability of pages at a particular level. By using this metric, Web authors can decide how the readability should change with levels. Take Online Teaching Site as an example. Teachers may want to teach some simpler things at the beginning, and then increase the difficulty level gradually. By analyzing the

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-332-9/06/0005.

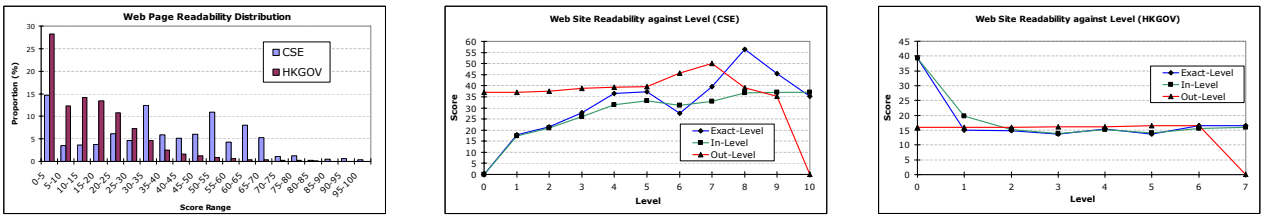

Figure 2: Experimental Results: (a) Page Readability Distribution (Left), (b) Site Readability of CSE (Middle), and (c) Site Readability of HKGOV (Right)

changes of Exact-Level readability along levels, teachers can then prepare and arrange materials in the proper order.

In-Level Readability of a site gives the average readability of Web pages starting from root page up to pages at specified level. This is an overall indicator of a site difficulty. By using this metric, users can get a general idea of whether the site is suitable to them before start browsing it.

Out-Level Readability of a site gives the average readability of Web pages starting one level upper than the specified one, up to pages with *maximum available level*, which is determined by the depth of crawling. In another words, it is a difficulty indicator of remaining pages after browsing a site for some times. Users may make use of this metric as one of factors in deciding whether he or she should continue browsing the site.

3. EXPERIMENTS AND DISCUSSION

We carry out two experiments to evaluate our proposed Web readability assessment on the following sites: (1) Department of CSE, CUHK (CSE) [1], and (2) The Hong Kong Government (HKGOV) [3]. All experiments are run on a Pentium 4 3.4GHz, 4GB memory machine with Red Hat Fedora Core 4 as operating system.

Experiment 1: Web Page Readability. Figure 2(a) shows the Web page readability distribution of the two Web sites being tested. The first observation is that both CSE and HKGOV have large portion of pages with scores ranged from 0 to 5. To investigate this phenomenon, we manually examine those pages and find that they mainly belongs *index pages*, which are introductory pages having hyperlinks to internal pages. Index pages generally receive low readability score than passage pages. It is because only index terms will remain after extracting raw texts from those pages, and a long sentence will form because there are no separators such as full stop to delimit the index terms. As a result, a long sentence will reduce the readability scores of index pages, although they should not be that difficult apparently.

The second observation is that CSE has larger portion of high-scored (less difficult) pages than HKGOV. It is because CSE contains personal pages of staffs and students, in which the contents are easier to be comprehended than formal articles published in HKGOV.

Conclusion: In addition to measuring difficulty of pages, page readability can also be a good indicator for figuring out low textual content pages such as index pages.

Experiment 2: Web Site Readability. Figure 2(b) and (c) show the site readability against levels of the two Web sites. The two graphs exhibit two types of readability variation: CSE has a dramatic change on readability, while HKGOV has a gradual change on it.

For CSE, we find that Exact-Level score gradually increases from level 2 to 5, and has a drop in level 6. It is because the pages located in level 4 and 5 are personal pages. Following our explanation in last section that they are easier to be comprehended, the site readability increases. For level 6, we find that as this level is right after personal pages, authors would like to put more non-textual information such as images, videos etc., causing the drop in score. In-Level score generally follows the trend of Exact-Level, but with a smoother variation due to accumulation effect of scores starting from level 0. For Out-Level score, although level 6 has relatively low Exact-Level and In-Level scores, it has high Out-Level score, which indicates the contents at later levels will be relatively easier. This may motivate users to continue browsing the site.

For HKGOV, it has a gradual change in Exact-Level score, indicating that pages at different levels have similar difficulty. It is because as a government organization, its Web site is well-organized to give pages with similar difficulties. Due to the gradual change of Exact-Level score, In-Level and Out-Level scores behave nearly the same.

Conclusion: By analyzing the Web site readability variation graph, we can get the content structure and distribution of a site. Web authors can then make use of this information to make the site organization better.

4. CONCLUSION AND FUTURE WORK

In this paper, we propose a bilingual readability assessment scheme for Web page and site in English and Chinese languages. Experimental results show that, for page readability, apart from just indicating difficulty, it can also serve as a good indicator of low textual content pages such as index and multimedia pages. For site readability, we can get an overall picture of content structure by studying the score variation. The future work includes establishing better Web readability formulas and performing experiments in larger scale. We also plan to apply the proposed scheme in other Web-related applications.

5. ACKNOWLEDGMENTS

The work described in this paper is supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK4235/04E) and is affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing & Interface Technologies.

6. REFERENCES

[1] Department of C.S.E., CUHK. *http://www.cse.cuhk.edu.hk*.
[2] Global Internet statistics: sources and references. *http://global-reach.biz/globstats/refs.php3*.
[3] The Hong Kong Government. *http://www.gov.hk*.
[4] R. F. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233, 1948.
[5] T. S. Hansell. Readability, syntactic transformations, and generative semantics. *Journal of Reading*, 19(7):557–562, 1976.
[6] S. J. Yang. A readability for Chinese language. *Ph.D. Thesis for Mass Communication, University of Wisconsin*, 1971.