

Estimating Web Site Readability Using Content Extraction

Thomas Gottron
Johannes Gutenberg University Mainz
Institute of Computer Science
55099 Mainz, Germany
gottron@uni-mainz.de

Ludger Martin
Johannes Gutenberg University Mainz
Institute of Computer Science
55099 Mainz, Germany
martin@informatik.uni-mainz.de

ABSTRACT

Nowadays, information is primarily searched on the WWW. From a user perspective, the readability is an important criterion for measuring the accessibility and thereby the quality of an information. We show that modern content extraction algorithms help to estimate the readability of a web document quite accurate.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; H.5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia—*User Issues*

General Terms

Experimentation, Measurement, Human Factors

Keywords

Readability, Usability, Content Extraction

1. INTRODUCTION

The WWW is a major resource for finding information. The problem is to find relevant information and to understand this information. Therefore a user has to first find an appropriate document and then identify, read and understand its content to satisfy his need for information.

Readability is commonly evaluated by using readability indices, such as SMOG [5] or FRE [1]. In a web environment, these indices are calculated on full web documents, e.g. in [4]. Hence, they consider navigation menus, layout elements, related links lists, copyright information, etc. in their evaluation. These elements typically do not consist of well formed language and distort the values of the readability indices. This might result in misleading values, especially, because users usually focus on the main article in a document and ignore all other parts. Accordingly, a readability analysis should be restricted to those parts the user actually reads, i.e. the main content of a document.

In this paper we will investigate the combination of content extraction and readability tests. Section 2 describes two readability indices, section 3 presents content extraction and in section 4 we combine the both concepts.

2. READABILITY INDICES

A readability index is a mean to express the complexity of written text. Based on simple features, such as sentence and word length, they indicate how easy it is to read and comprehend a text. Originally developed for text documents, they are nowadays applied to web documents, too [4, 3].

2.1 Flesch Reading Ease

One index to determine readability is the Flesch Reading Ease (FRE) [1]. It has a score between 0 and 100. A higher score indicates an easier text. E.g., a text with a score between 100 and 90 is understandable by 11-year old students. A score lower than 30 implies a college graduate. If y is the total number of syllables, w the number of words and s the number of sentences in the text, FRE is defined as $r_{FRE} = 206.835 - 84.6 \frac{y}{w} - 1.015 \frac{w}{s}$.

2.2 SMOG

McLaughlin [5] presented a different readability formula, the SMOG grading. This index indicates the educational level, i.e. the years of school education, required to understand a text. If p is the number of polysyllables (words with three or more syllables) and s again the number of sentences, SMOG is defined as $r_{SMOG} = 1.043 \sqrt{\frac{p}{s}} + 3.1291$.

3. CONTENT EXTRACTION

Content Extraction (CE) is the process of determining those parts of an HTML document which represent its main textual content.

3.1 Document Slope Curves

Pinto *et al.* [6] developed the Document Slope Curves (DSC) algorithm. It identifies those parts of a document which contain most of the text while excluding most of the tags. Employing a sliding window technique, DSC is capable to locate also distributed and interrupted parts of the document with these characteristics.

3.2 Adapted Content Code Blurring

The idea of content code blurring (CCB) [2] is to take advantage of typical visual features of the main and the additional contents. Additional contents are usually highly formatted and contain little and short texts, while the main text content is commonly long and homogeneously formatted. CCB identifies these homogeneously formatted regions in HTML documents by calculating local content code ratios on a character level. Adapted CCB (ACCB) additionally ignores anchor-tags in this process, which yields better results.

Table 1: Average SMOG values for different web sites, with and without CE.

Web-site	Number of Documents	r_{SMOG}			
		full document	main content	ACCB	DSC
BBC News	337	4.0569	4.8323	4.9360	4.8052
The Economist	53	4.2486	5.0578	5.1433	5.0835
Herald Tribune	300	4.0891	5.0477	5.0650	5.0412
MSNBC News	197	4.4675	4.8949	4.9050	4.8491
Yahoo News	227	4.2063	4.9416	4.7563	4.7670

4. ESTIMATING READABILITY

As mentioned initially, readability indices for web documents should focus in the main content. Hence, our idea is that we can find good estimates for the readability of a web page by first applying content extraction algorithms.

For this purpose, we analysed a total of 1114 documents from five different web-sites for which we manually outlined the main content. We then compared the readability indices computed on the actual main article with those computed on the full document and those based on a document version in which the main content was extracted via the ACCB or the DSC algorithm. The results for the SMOG index are listed in table 1. The figures show, that both CE approaches estimate the readability indices for the main article far more accurate than an index calculated on the full document.

We will now take a more detailed look at the first 100 pages of the Yahoo news data. Figure 1 and 2 show a plot of the readability scores using the FRE and SMOG index. In both cases the readability estimates involving CE correlate very well with the index of the hand extracted main content.

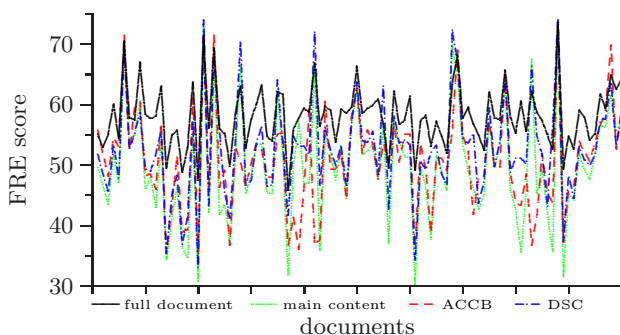


Figure 1: Yahoo analysed with FRE

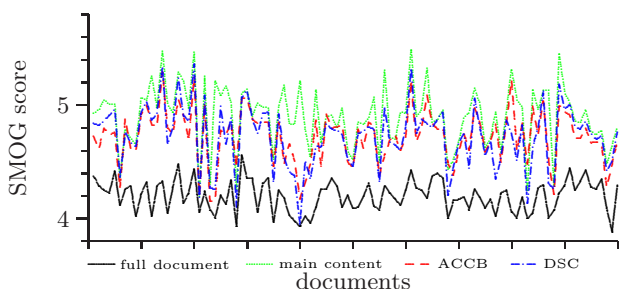


Figure 2: Yahoo analysed with SMOG

Finally, figure 3 shows the difference between the SMOG score of the actual main article and the three estimates. In the case of Yahoo the average difference between the score

of ACCB and the actual main content is 0.185, for DSC it is 0.179 while the difference to the full document is 0.704.

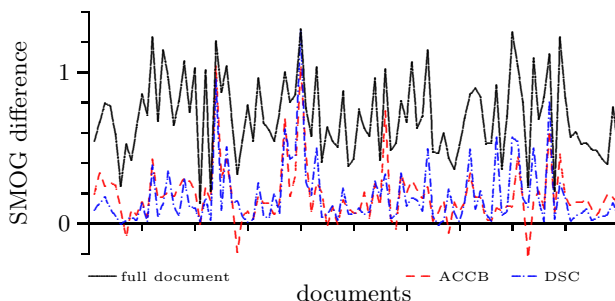


Figure 3: Differences to SMOG of main article

5. CONCLUSIONS

The paper demonstrates the improvements in estimating readability of web documents when applying content extraction algorithms prior to index calculation. We observed the SMOG and the FRE index to be far more accurate in combination with CE in comparison to calculating them on the full document. Further investigations need to be done to find the most suitable CE algorithm for a downstream readability analysis.

6. REFERENCES

- [1] R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948.
- [2] T. Gottron. Content code blurring: A new approach to content extraction. *Database and Expert Systems Applications, International Workshop on*, pages 29–33, 2008.
- [3] H. Kienle and C. Vasiliu. Evolution of legal statements on the web. In *10th IEEE International Symposium on Web Site Evolution*, pages 73–82, October 2008.
- [4] T. P. Lau and I. King. Bilingual web page and site readability assessment. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 993–994, New York, NY, USA, 2006. ACM.
- [5] G. H. McLaughlin. Smog grading: A new readability formula. *Journal of Reading*, 12 (8):639–646, 1969.
- [6] D. Pinto, M. Branstein, R. Coleman, W. B. Croft, M. King, W. Li, and X. Wei. Quasm: a system for question answering using semi-structured data. In *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 46–55, New York, NY, USA, 2002. ACM.