Readability of the Web: A study on 1 billion web pages.

Marije de Heus
University of Twente
m.deheus@student.utwente.nl

Djoerd Hiemstra
University of Twente
hiemstra@cs.utwente.nl

ABSTRACT

We have performed a readability study on more than 1 billion web pages. The Automated Readability Index was used to determine the average grade level required to easily comprehend a website. Some of the results are that a 16-year-old can easily understand 50% of the web and an 18-year old can easily understand 77% of the web. This information can be used in a search engine to filter websites that are likely to be incomprehensible for younger users.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Selection process; H.1.2 [**User/Machine Systems**]: Human information processing

General Terms

Algorithms, Measurement, Human Factors

Keywords

Readability, ARI, Code Crawl, MapReduce

1. INTRODUCTION

The internet has users of all ages. Some texts are more easily readable by young users than others. In general, texts that have longer sentences with longer words which contain more syllables, are less likely to be easily understood by young users than texts with shorter sentences that consist of short words. This paper analyzes the readability of the web, as part of the Norvig Web Data Science Award[1].

There are several measures to compute the readability of a text, such as Flesch-Kincaid readability[10], Gunning Fog index[9], Dale-Chall readability[5], Coleman-Liau index[6], SMOG[11] and the Automated Readability Index[12]. Most of these use a formula that requires counting the number of syllables. Deciding where a syllable begins and ends is a difficult problem, depending on the language. Therefore we chose to use the Automated Readability Index, which was designed for real-time computation of readability on electronic typewriters and does not use the number of syllables. Instead it uses the average number of characters per word and the average number of words per sentence. The outcome represents the US grade level that is needed to easily comprehend the text.

DIR 2013, April 26, 2013, Delft, The Netherlands.

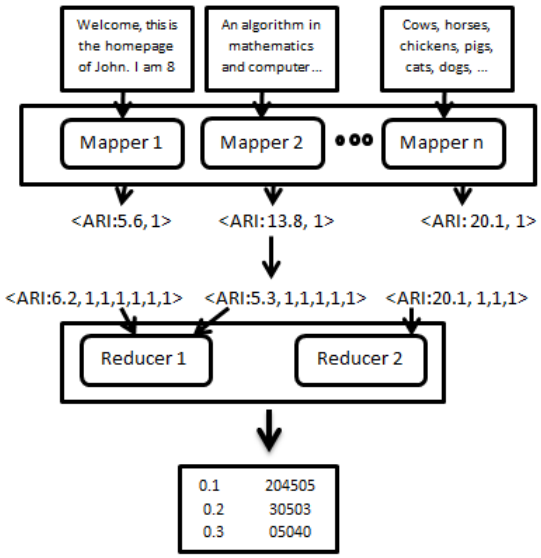

Figure 1: Visual overview of the MapReduce program

The ARI formula[12] is shown below.

$$ARI = 4.71 * \frac{characters}{words} + 0.5 * \frac{words}{sentences} - 21.43 \quad (1)$$

So far most of the research regarding readability of websites has focused on legal documents and health documents [2][8][3]. No previous experiments with readability large numbers of websites have been found. The goal of our research is to examine the readability of the web. For this purpose, we ran a MapReduce program on more than a billion webpages. The Common Crawl dataset consists among others of 61 million domains, 92 million PDF Docs and 7 million Word Docs. More than 60% of the data came from .com TLD's, with .org and .net on second and third place. Thereafter came .de, co.uk, .ru, .info, .pl, .nl et cetera[1]. We did not filter non-English websites.

2. IMPLEMENTATION

The program was implemented using MapReduce[7] on Hadoop[4]. Figure 2 provides a visual overview of our program. The mapper takes the text of a website without html tags. It computes the ARI of the text. It then emits this ARI and a count of 1. The reducer receives an ARI score

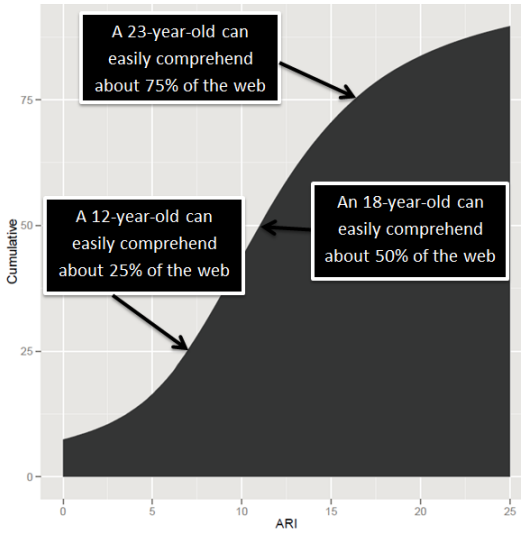

Figure 2: Cumulative results

and a number of counts. It sums the counts and writes the ARI and the sum to one line of the output file.

3. RESULTS

Figure 3 shows the cumulative results. This graph answers questions such as 'how much of the web can a 12-year-old (grade 6) easily comprehend?' (answer: about 20%).

4. DISCUSSION, CONCLUSION AND FUTURE WORK

4.1 Discussion

Very low results.

7% of the websites received a score below 0. 1.7% of the websites was empty. These results cannot be interpreted in terms of a US grade level. However, we can infer that these websites are probably easily readable for all ages, because such websites must have very short sentences and very short words.

Very high results.

13.3% of the websites received a score higher than 22. This means that a person would need more than 22 years of education to easily comprehend the website. Some of these even got scores above 100. A lot of these websites consist of enumerations of items, dates, addresses et cetera, which are not stripped. It is not clear what effect such items have on the readability. Maybe they should be ignored when computing the readability, or maybe they do influence readability. Some of these enumerations may be detected by certain html list tags, while others may not be removed as easily.

Non-English Languages.

In our analysis, we did not filter non-English websites. Automated Readability Index was not designed for English specifically, but Smith and Senter [12] only experimented with the English languages. We did not find studies on how accurate ARI is for other languages.

4.2 Conclusion

This paper presented an anlysis of the readability of the web using ARI and MapReduce. The results (presented in figure 3) depend on the reliability of ARI for web pages of different languages and can be used in a search engine to adjust search results to a user's education level.

4.3 Future Work

ARI for non-English texts.

We did not find literature on the accuracy of ARI for non-Enlgih languages. This needs to determined before ARI can be used in (multilingual) practice.

Readability of web pages.

Some of the high ARI scores may be due to the structure of some websites, e.g. long enumerations and lists of items. A readability measure like ARI may not be reliable on such websites. More research can be done on how the readability of a web page can be accurately determined.

5. REFERENCES

[1] Norvig Web Data Science Award. `http://norvigaward.github.io/index.html`, 2013.

[2] G.K. Berland, M.N. Elliott, L.S. Morales, J.I. Algazy, R.L. Kravitz, M.S. Broder, D.E. Kanouse, J.A. Muñoz, J.A. Puyol, M. Lara, et al. Health information on the internet. *JAMA: the journal of the American Medical Association*, 285(20):2612–2621, 2001.

[3] E.V. Bernstam, D.M. Shelton, M. Walji, and F. Meric-Bernstam. Instruments to assess the quality of health information on the world wide web: what can our patients actually use? *International journal of medical informatics*, 74(1):13–20, 2005.

[4] D. Borthakur. The hadoop distributed file system: Architecture and design. *Hadoop Project Website*, 11:21, 2007.

[5] J.S. Chall. *Readability revisited: The new Dale-Chall readability formula*, volume 118. Brookline Books Cambridge, MA, 1995.

[6] M. Coleman and T.L. Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283, 1975.

[7] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[8] M.A. Graber, C.M. Roller, B. Kaeble, et al. Readability levels of patient education material on the world wide web. *The Journal of family practice*, 48(1):58, 1999.

[9] R. Gunning. The fog index after twenty years. *Journal of Business Communication*, 6(2):3–13, 1969.

[10] J.P. Kincaid, R.P. Fishburne Jr, R.L. Rogers, and B.S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document, 1975.

[11] G.H. Mc Laughlin. Smog grading-a new readability formula. *Journal of reading*, pages 639–646, 1969.

[12] R.J. Senter and E.A. Smith. Automated readability index. Technical report, DTIC Document, 1967.