Readability of web content

An analysis by topic

Hélder Antunes
MIEIC, Faculdade de Engenharia da Universidade do Porto
INESC TEC
Porto, Portugal
up201406163@fe.up.pt

Carla Teixeira Lopes
Faculdade de Engenharia da Universidade do Porto
INESC TEC
Porto, Portugal
ctl@fe.up.pt

Abstract — **Readability is determined by the characteristics of the text that influence their understanding. The web is composed of content on various topics and the results retrieved in the top positions by the main search engines are expected to be those with the highest number of views. In this study, we analyzed the readability of web pages according to the topic to which it belongs and their position in the search result. For that, we collected the top-20 results retrieved by Google to 23,779 queries from 20 topics and used several readability metrics. The results of the analysis showed that the content from organizations (like colleges and other institutions) and health-related content have lower readability values. Categories Games and Home are on the opposite side. For the categories identified as having less readability, tools can be developed that help the user understand their content. We also found that top-ranked pages have higher values of readability. One can conclude that, directly or indirectly, readability is a factor that seems to be being considered by the Google search engine or has an influence on page popularity.**

Keywords - Readability; World Wide Web; Web Search Engines; Ranking.

I. INTRODUCTION

The growth of the Web has been exponential over the last decade. Given the high usage of web content, it becomes important to analyze its readability. Most web access is done through web search engines. Within this market, Google stands out with more than 82 % market share [3]. A great diversity of users seeks information on the Web, so the search engines should adapt the retrieved results to the user's knowledge to guarantee that content is perceived.

Readability defines the ease in understanding a given text. It can be analyzed through sentence complexity parameters, like the number of words per phrase, and word complexity parameters, like the number of characters per word. Identifying topics that are initially expected to be less readable may be important in creating mechanisms to make content more readable to a wider range of people. Tools that evaluate the readability of a topic, taking advantage of its specific vocabulary and semantics, can help in writing text easier to read.

Our main goals are to identify the most readable topics on the web and to study the dynamic of the readability of web pages according to their rank in search result pages. We will begin by describing related works. We will then detail our methodology as well as present and discuss the results.

II. BACKGROUND AND RELATED WORK

Readability metrics are associated with the difficulty in perceiving certain textual content. Dale and Chall [22] defined readability as the "sum (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers has with it". The success of reading is characterized by "the extent to which they understand it, read it at an optimal speed, and find it interesting".

The first readability measures dated to last century. These traditional measures are easy to compute, using average sentence length to evaluate the syntactic difficulty and the average word length to evaluate the semantic difficulty. Some of them correlate the obtained score of a text with the degree of schooling necessary for the perception of it. One of the most used formulas is the Flesch-Kincaid [18], that calculates the required grade to read the text using the following formula:

$$0.39 \times \alpha + 11.8 \times \beta - 15.59. \quad (1)$$

In which α is the average number of words per sentence and β is the average number of syllables per word.

Classic readability measures have limitations, since used surface characteristics of text, ignoring other important aspects of text like cohesion or coherence. More recent approaches combine natural language processing and machine learning using more advanced features [23].

Concerning to readability across the web, in 2012, articles from Wikipedia [10] were analyzed using the Flesch reading ease metric and revealed that 75 percent of the articles presented resulted in lower readability levels than the standard difficulty (standard contents are easily understood by 13- to 15-year-old students).

Results from Google searches based on the now extinct Google Reading Level filter feature, showed that there was no correlation between the difficulty given by the tool (Basic, Intermediate, Advanced) and the score obtained by the Flesch Reading Ease and Flesch-Kincaid Grade Level metrics [5].

In addition to the use of readability formulas, others have used a machine learning approach (gradient boosted decision tree) [8] in small-sized summaries of a web page, presented in the results page of a search engine. This method proved to be

2019 14th Iberian Conference on Information Systems and Technologies (CISTI)
19 – 22 June 2019, Coimbra, Portugal
ISBN: 978-989-98434-9-3

more accurate than traditional readability formulas but was only used for small summaries of web pages, not knowing if the conclusions apply to more extensive content.

Another experiment [11] attempted to modify traditional readability formulas to apply them to web pages by adding bias adjustments to these formulas. Also, they tested two known content extraction algorithms, Content Code Blurring [7] and Document Slope Curves [13], which were shown to lead to better estimates.

Predicting the readability of a web page depends on the syntactical features of its text [12] but also on its legibility as determined by its visual design. However, our work will only focus on purely textual characteristics. To our knowledge, there is no study of readability across the web by topic.

III. METHODOLOGY

In the impossibility to analyze the whole web, we decided to use a sample of web pages obtained from a random sample query collection of AOL Search in the Fall of 2004 [4], with 23,779 queries divided into 20 categories/topics. The query collection resulted from real searches from the AOL search engine and has been manually classified in different topics. For each query, we analyzed the web pages resulting from the first 20 search results given by Google's search engine. A study [2] shows that, in the set of searches with clicks in organic results, the probability of a click being on a result of the two first pages is 98%. We think that analyzing the top-20 results is representative as the probability of clicking a result after page 2 is only 1.6%. We collected the data in January 2018.

We used the Google Custom Search API [14] to obtain search results. By using that API, the search result is more neutral than the use of standard browsers, since it avoids custom user results due, for instance, to browser history [1]. From the obtained results we only considered results in HTML, PDF or doc format. In the HTML pages, it was necessary to remove the header, footer, navigation menus, leaving only the main content. It is also important to note that some HTML pages do not have any textual content, and only the pages with five or more sentences were analyzed. To extract the main content of an HTML page in the plain text form, we used the boilerpipe library for Java language [16]. The algorithms used by the library are based on shallow text features proposed by Kohlschütter [9].

To measure the readability of the returned text, we used the source code of an open source project called ReadabilityMetric [15]. In our study, we used the following metrics: Simple Measure of Gobbledygook (SMOG) [17], Flesch reading ease [18], Flesch-Kincaid Index [18], the Gunning Fog Index [21], Coleman-Liau Index [20], the Automated Readability Index (ARI) [19]. We chose these metrics because they are easy to compute and well known in the field of readability.

In this investigation, we hypothesize that not all topics have the same readability and that readability drops with the rank position.

IV. READABILITY OF WEB CONTENT

To compare the topics in more detail, we choose to analyze the SMOG metric across the topics ordered by more readable topic (details in Figure 1). We choose the SMOG metric because it was the metric that more correlates with the others across the collected web pages (details in Figure 2). If we use any other metric, the results would be similar. In simple terms, this metric associates readability with the number of complex words (words with 3 or more syllables) per sentence. In the interpretation of the SMOG metric value, note that the value means the required grade level to read a particular text being lower when the text is more readable. Of the 278,081 web pages analyzed totalizing 320830885 words and 14789284 sentences, it is estimated that, on average, at least 13 years of education will be required (the SMOG mean of all content is 13.06) to understand the content of those pages. This means that it's needed a level of schooling similar to a college student. By percentile analysis, it is required at least 11 years of schooling to understand 25% of content, 13 to understand 50% of content and 15 to understand 75% of the content.

Another correlation that caught our attention was the correlation between the web page position in Google's search result and the average readability grades by that position ($r = 0.96$, p-value = 1.791e-11). The results are presented in Figure 3 and show that documents get less readable as the page rank increases. If we do not consider the average readability by rank but all readability values of all web pages, the correlation is not significant ($r = 0.03$, p-value < 2.2e-16).

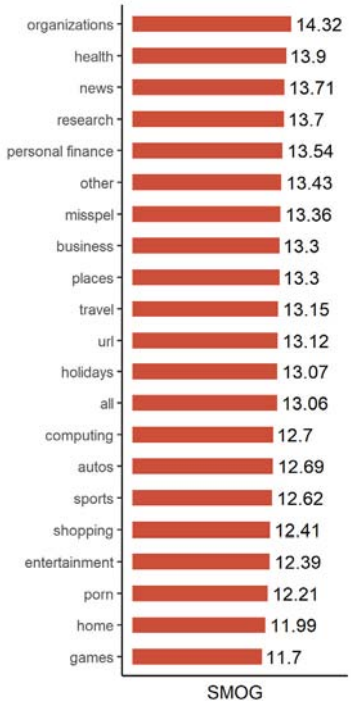

Figure 1. Mean SMOG by topic.

ISBN: 978-989-98434-9-3

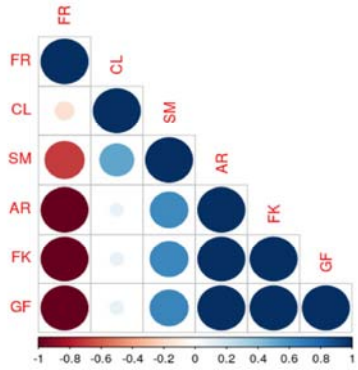

Figure 2. Correlation between readability metrics. Red color means a negative correlation and blue means a positive correlation. FR - Flesch reading ease, CL - Coleman-Liau Index, SM – SMOG, AR - Automated Readability Index, FK - Flesch-Kincaid Index, GF - Gunning Fog Index.

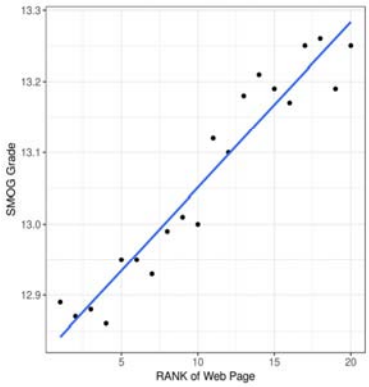

Figure 3. Correlation between rank and readability of web page in all topics.

TABLE I. TUKEY HSD FOR POST-HOC ANALYSIS FOR SMOG READABILITY LEVEL

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
autos (1)																			
business (2)	>>>																		
computing (3)	1	<<<																	
ent. (4)	<<<	<<<	<<<																
games (5)	<<<	<<<	<<<	<<<															
health (6)	>>>	>>>	>>>	>>>	>>>														
holidays (7)	>>>	0.20	>>>	>>>	>>>	<<<													
home (8)	<<<	<<<	<<<	<<<	>>	<<<	<<<												
misspell (9)	>>>	0.99	>>>	>>>	>>>	<<<	>	>>>											
news (10)	>>>	>>>	>>>	>>>	>>>	<	>>>	>>>	>>>										
org. (11)	>>>	>>>	>>>	>>>	>>>	>>>	>>>	>>>	>>>	>>>									
other (12)	>>>	0.15	>>>	>>>	>>>	<<<	>>>	>>>	0.96	<<<	<<<								
p. finance (13)	>>>	0.15	>>>	>>>	>>>	<	>>>	>>>	0.66	0.71	<<<	0.99							
places (14)	>>>	1	>>>	>>>	>>>	<<<	0.18	>>>	1	<<<	<<<	0.19	0.17						
porn (15)	<<<	<<<	<<<	>>	>>>	<<<	<<<	>	<<<	<<<	<<<	<<<	<<<	<<<					
research (16)	>>>	>>>	>>>	>>>	>>>	<<	>>>	>>>	>>>	1	<<<	>>>	0.75	>>>	>>>				
shopping (17)	<<<	<<<	<<<	1	>>>	<<<	<<<	>>>	<<<	<<<	<<<	<<<	<<<	<<<	>>	<<<			
sports (18)	0.99	<<<	0.99	>>	>>>	<<<	<<<	>>>	<<<	<<<	<<<	<<<	<<<	<<<	>>>	<<<	>		
travel (19)	>>>	0.54	>>>	>>>	>>>	<<<	1	>>>	<	<<<	<<<	<<<	<<<	0.50	>>>	<<<	>>>	>>>	
url (20)	>>>	<	>>>	>>>	>>>	<<<	1	>>>	<<<	<<<	<<<	<<<	<<<	<	>>>	<<<	>>>	>>>	1

>>> or <<< → p-value <0.001, >> or << → p-value <0.01, > or < → p-value <0.05. The < and > operators specify the direction of the difference. For instance, when a < is presented, the row category has a lower mean SMOG than the column category.

V. READABILITY ANALYSIS BY TOPIC

The results show that Organizations is the category with less readability. This category is related to schools, colleges and other private or public institutions. After that, health-related contents are the ones with less readability. This category exposes specific terminology, what justifies this low readability value. The specific terminology also explains the low levels of readability of Research and Personal Finance. We didn't expect News content to have a low degree of readability. While trying to understand the causes for this, we found that the problem of News readability was already explored before [6]. The article blames not only the pressures of deadlines and format features but also the complexity of reporting the real world. Authors discovered that deceptive news, which did not express the real world, had higher readability values. On the other side of the readability score, the Games category was the one that obtained the highest readability value. The Games category has a young target audience, which probably justifies why its content tends to be more straightforward. Another group with high readability is the Home category. That can be explained because it contains a lot of advertising content (sales of articles for home and decoration), most of which have slogans which, by their nature, are short and simple phrases.

To analyze if there are significant differences between the several topics, we applied the ANOVA test. As we found significant differences between categories, we performed multiple pairwise comparisons using the Tukey HSD test between categories. The results are presented in Table 1. It is

ISBN: 978-989-98434-9-3

notorious that the Organizations, Games, and Health categories are significantly different from all others. For instance, the Games always has a higher value of readability. The Personal Finance, Research and, News categories do not show p-values below 0.05, so they are not significantly different from each other. It should also be noted that the Holidays, Places and Travel categories are not different from each other. That is understandable since those categories are closely related. The Google search results can vary over time, but the number of pages analyzed, and the statistical significance found lead us to believe that the readability results would be similar in another search period.

VI. Conclusions

The goal of this work was to study the readability of web contents, in general, by topic and by rank position in search result pages.

We estimate that, on average, at least 13 years of education will be required to understand the entire content of the pages. By percentile analysis, it is required 11 years of schooling to understand 25% of content, 13 years to understand 50% of content and 15 years to understand 75% of the content. We found that organizational contents are less readable, followed by health-related materials. On the other hand, games and home-related subjects are more readable. Using the SMOG metric in the tests performed after that, we show that there are in fact topics that differ significantly in the readability value. These results indicate that content simplification on some topics can have a significant impact on understanding by people with less knowledge in a given area.

Analyzing the variation of the SMOG metric as a function of the page rank in Google's search engine, it was shown that there is a correlation between the two variables and that, in general, the readability is higher in the first pages returned by Google. We do not know if there is a cause-effect relationship between these two variables, or if Google directly also uses readability as a criterion to rank results.

For future work, we plan to work on the automatic simplification of texts in the topics like health where readability is low.

Acknowledgment

Partially funded by the project "NORTE-01-0145-FEDER-000016" (NanoSTIMA), financed by the North Portugal Regional Operational Programme (NORTE2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF). We would also like to thank the Master in Informatics and Computing Engineering of the Faculty of Engineering of the University of Porto for supporting the registration and travel costs.

References

[1] Beitzel, S.M., Jensen, E.C., Frieder, O., Lewis, D.D., Chowdhury, A., Kolcz, A.: Improving automatic query classification via semi-supervised learning. In: Proceedings - IEEE International Conference on Data Mining, ICDM. pp. 42–49 (2005). https://doi.org/10.1109/ICDM.2005.80

[2] Bilal, D.: Comparing Google's readability of search results to the Flesch readability formulae: A preliminary analysis on children's search queries. In: Proceedings of the ASIST Annual Meeting. vol. 50 (2013). https://doi.org/10.1002/meet.14505001094

[3] Dalecki, L., Lasorsa, D.L., Lewis, S.C.: The news readability problem. Journalism Practice 3(1), 1–12 (2009). https://doi.org/10.1080/17512780802560708

[4] Gottron, T.: Content code blurring: A new approach to Content Extraction. In: Proceedings - International Workshop on Database and Expert Systems Applications, DEXA. pp. 29–33 (2008). https://doi.org/10.1109/DEXA.2008.43

[5] Kanungo, T., Orr, D.: Predicting the readability of short web summaries. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining - WSDM '09. p. 202 (2009). https://doi.org/10.1145/1498759.1498827

[6] Kohlschütter, C., Fankhauser, P., Nejdl, W.: Boilerplate detection using shallow text features. In: Proceedings of the third ACM international conference on Web search and data mining - WSDM '10. p. 441 (2010). https://doi.org/10.1145/1718487.1718542

[7] de Kunder, M.: Search engine market share. https://www.netmarketshare.com (February 2018), (Accessed on 02/10/2018)

[8] Lucassen, T., Dijkstra, R., Schraagen, J.M.: Readability of Wikipedia. First Monday 17(9) (2012). https://doi.org/10.5210/fm.v0i0.3916

[9] Martin, L., Gottron, T.: Readability and the web. Future Internet 4(1), 238–252 (2012). https://doi.org/10.3390/fi4010238

[10] McEvoy, M.: 7 reasons google search results vary dramatically. https://www.webpresencesolutions.net/7-reasons-google-search-results-varydramatically/, (Accessed on 05/06/2018)

[11] Nielsen, J.: Legibility, Readability, and Comprehension: Making Users Read Your Words (2015), https://www.nngroup.com/articles/legibility-readabilitycomprehension/

[12] Petrescu, P.: Google organic click-through rates in 2014 - moz. https://moz.com/blog/google-organic-click-through-rates-in-2014 (February 2018), (Accessed on 02/10/2018)

[13] Pinto, D., Branstein, M., Coleman, R., Croft, W.B., King, M., Li, W., Wei, X.: QuASM: a system for question answering using semi-structured data. Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries pp. 46–55 (2002). http://doi.acm.org/10.1145/544220.544228

[14] Google: Custom search — google developers. https://developers.google.com/custom-search/, (Accessed on 02/16/2019)

[15] Ipeirotis, P. A web service that computes a set of readability metrics for text. https://github.com/ipeirotis/ReadabilityMetrics, (Accessed on 02/16/2019)

[16] Kohlschtter, C.: Google code archive - long-term storage for google code project hosting. https://code.google.com/archive/p/boilerpipe/, (Accessed on 02/16/2019)

[17] Harry G. McLaughlin. 1969. SMOG grading - a new readability formula. Journal of Reading 12, 8 (May 1969), 639–646.

[18] J.P. Kincaid. 1975. Derivation of New Readability Formulas: (automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Chief of Naval Technical Training, Naval Air Station Memphis. https://books.google.pt/books?id=4tjroQEACAAJ

[19] Eric A. Smith and R. Senter. 1967. Automated readability index. AMRLTR. Aerospace Medical Research Laboratories (1967), 1–14.

[20] Meri Coleman and T L. Liau. 1975. A Computer Readability Formula Designed for Machine Scoring. 60 (04 1975), 283–284.

[21] Robert Gunning. 1952. The technique of clear writing. McGraw-Hill New York. 289 p. pages

[22] E. Dale and J. S. Chall. The Concept of Readability. Elementary English, 1949.

[23] Kevyn Collins-Thompson. Computational assessment of text readability: A survey of current and future research. ITL - International Journal of Applied Linguistics,165(2):97–135, Jan 2015.

ISBN: 978-989-98434-9-3