**A Study of Web Page Understandability for Consumer Health Search**

Joao Palotti1,3; Guido Zuccon2; Allan Hanbury3

1Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

2Queensland University of Technology, Brisbane, Australia

3Vienna University of Technology, Vienna, Austria

# *Abstract*

**Background:** Understandability plays a key role in ensuring that people accessing health information are capable of gaining insights that can assist them with their health concerns and choices. The access to unclear or misleading information has been shown to negatively impact on the health decisions of the general public.

**Objective:** We investigated methods to estimate the understandability of health Web pages and used these to improve the retrieval of information for people seeking health advice on the Web.

**Methods:** Our investigation considered methods to automatically estimate the understandability of health information in Web pages, and it provided a thorough evaluation of these methods using human assessments as well as an analysis of pre-processing factors affecting understandability estimations, and associated pitfalls. Furthermore, lessons learnt for estimating Web page understandability were applied to the construction of retrieval methods with specific attention to retrieving information understandable by the general public.

**Results:** We found that machine learning techniques were more suitable to estimate health Web page understandability than traditional readability formulas, which are often used as guidelines and benchmarking by health information providers on the Web (p value). Learning to rank effectively exploited these estimates to provide the general public with more understandable search results (significant improvements over XXX, p-value).

**Conclusions:** The findings reported in this article are important for specialised search services tailored to support the general public in seeking health advice on the Web, as they document and empirically validate state-of-the-art techniques and settings for this domain application.

**KEYWORDS:** Consumer Health Search; Document Understandability; Document Readability; Learning to Rank

# *Introduction*

Search engines are concerned with retrieving relevant informa- tion to support a user’s information seeking task. Commonly, signals about the topicality or aboutness of a piece of infor- mation with respect to a query are used to estimate relevance, with other relevance dimensions like understandability, trust- worthiness, etc. [[1].](#_bookmark1) being relegated to a secondary position, or completely neglected. While this may be a minor problem for many information seeking tasks, there are some specific tasks in which dimensions other than topicality have an important role in the information seeking and decision-making process. The seeking of health information and advice on the Web by

the general public is one such task.

A key problem when searching the Web for health informa- tion is that this can be too technical, unreliable, generally mis- leading, and can lead to unfounded escalations and poor deci- sions [[2]](#_bookmark2) . Where correct information exists, it can be hard to find and digest amongst the noise, spam, technicalities, and ir- relevant information. In *high-stakes search tasks* such as this, access to poor information can lead to poor decisions which ul- timately can have a significant impact on our

health and well-being [[2,](#_bookmark2) [3]](#_bookmark3) . In this work we are specifi- cally interested in the understandability of health information retrieved by search engines, and in improving search results to favour information understandable by the general public. We

leave addressing reliability and trustworthiness of the retrieved information to future work; however, this can be achieved by extending the framework we investigate here.

The use of general purpose Web search engines like Google, Bing and Baidu for seeking health advice has been largely analysed, questioned and criticised [[4–9]](#_bookmark5) , despite the commendable efforts these services have put into providing increasingly better health information, e.g., the Google Health Cards [[10]](#_bookmark6) .

Ad-hoc solutions to support the general public in searching and accessing health information on the Web have been imple- mented, typically supported by government initiatives or medi- cal practitioner associations, e.g., *HealthOnNet.org* (HON) and *HealthDirect.gov.au*, among others. These solutions aim to provide *better* health information to the general public. For example, HON’s mission statement is "to guide Internet users to reliable, understandable, accessible and trustworthy sources of medical and health information". But, do the solutions these services currently employ actually provide this type of infor- mation to the health-seeking general public?

As an illustrative example, we analysed the top 10 search re- sults retrieved by HON on 01/10/2017 in answer to 300 health search queries generated by regular health consumers in health forums. These queries are part of the CLEF 2016 eHealth col- lection, which shall be extensively used in this article. The un- derstandability score of each one of the retrieved pages was es- timated with the most effective readability formula and prepro- cessing settings analysed in this article (low scores correspond to easy to understand Web pages). Figure [1](#_bookmark0) reports the cumu- lative distribution of understandability scores for these search results (note, we did not assess their topical relevance here). We report also the scores for the "optimal" search results (Or- acle), as found in CLEF 2016 (relevant results that have the highest understandability scores), along with the scores for the baseline method (BM25) and our best retrieval method (XGB). The results clearly indicate that, despite solutions like HON being explicitly aimed at supporting access to understandable health information, they often fail to do so.

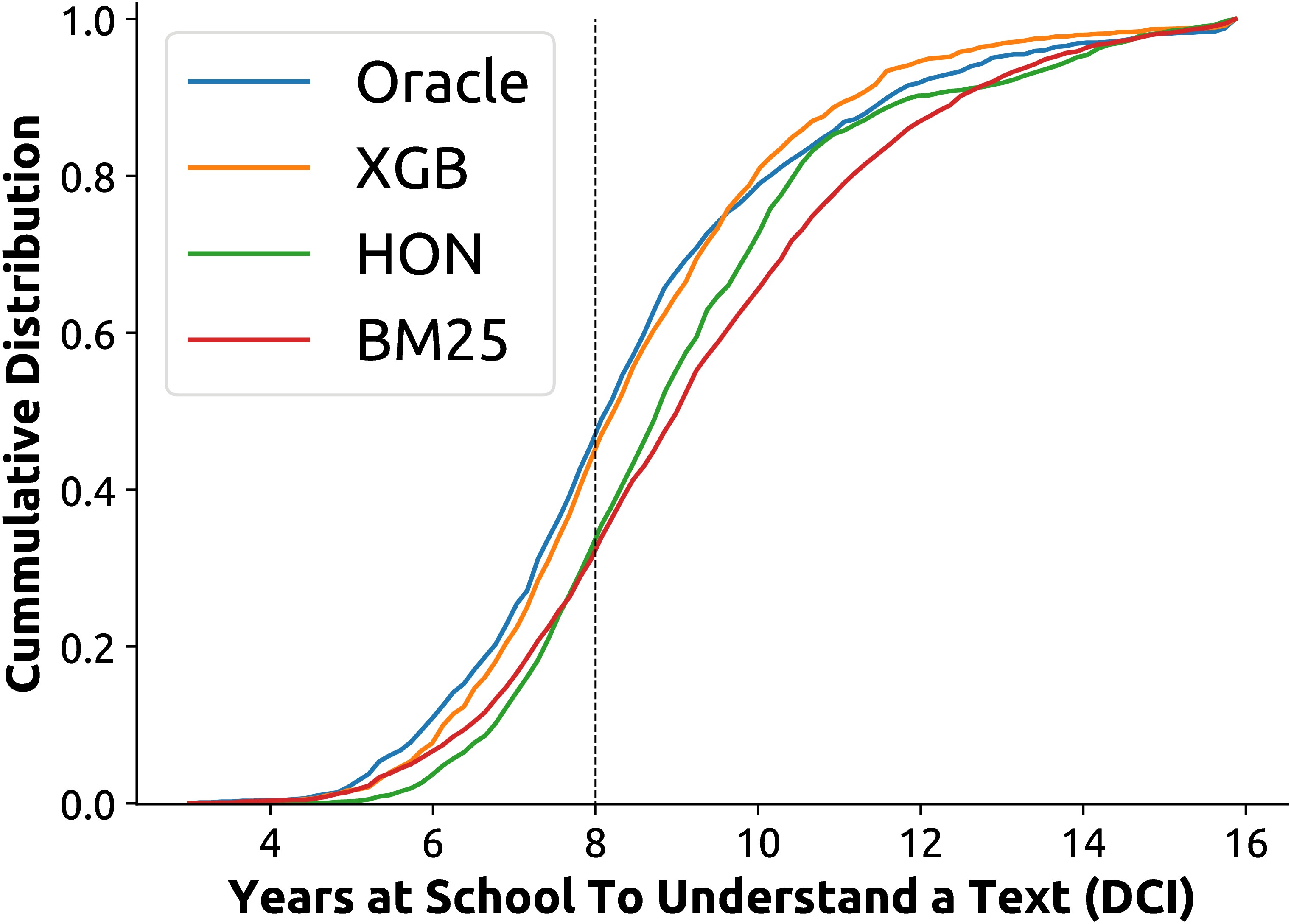
In this article, we aim to establish methods and best practice for developing search engines that retrieve *relevant and under- standable* health advice from the Web. The overall contribu- tions of this article can be summarised as:

1. We propose and investigate methods for the estimation of the understandability of health information in Web pages: a large number of medically-focused features are grouped in meaningful categories and their contribution to the un- derstandability estimation task is carefully measured;
2. We further study the influence of HTML processing meth- ods on these estimations and their pitfalls, extending our previous work that has shown how this often ignored as- pect greatly impacts effectiveness [[16]](#_bookmark10) ;
3. We further investigate how understandability estimations can be integrated into retrieval methods to enhance the quality of the retrieved health information with particu- lar attention to its understandability by the general public. New models are explored in this article, also extending our previous work [[17];](#_bookmark11)

This paper makes concrete contributions to practice, as it informs health search engines specifically tailored to the general public (for example the HON or HealthDirect services referred to above) about the best methods they should adopt, but they currently don’t. These are novel and significant contributions, as no previous work has systematically analysed the influence of the components at play in this study and we show that these greatly influence retrieval effectiveness and thus delivery of relevant and understandable health advice.

**Related Work**

Understandability refers to the ease of comprehension of the information presented to a user. Put in other words, health in- formation is understandable “when consumers of diverse back- grounds and varying levels of health literacy can process and explain key messages”~\cite{shoemaker2014development}. Often the terms understandability and readability are used in- terchangeably: we use readability to refer to formulas that es- timate how easy is to understand a text, usually based on its words and sentences. We use understandability to refer to the broader concept of ease of understanding: this is affected by text readability (as increasing readability tends to improve un- derstanding), but may also be influenced by how legible a text is and its layout, including e.g., the use of



**Figure 1.** Distribution of Dale-Chall Index (DCI) of search results. DCI measures the years of schooling required to understand a document. The average US resident reads at or below an 8th grade level (dashed line) [[11–14]](#_bookmark8) , which is the level suggested by NIH for health information on the Web [[15]](#_bookmark9) . The distribution for HON is similar to that of the baseline used in this article (BM25). Our best method (XGB) re-ranks documents to provide more understandable results; its distribution is similar to that of an “Oracle” system.

# *References*

IT’S TIME TO UPGRADE

Dear user, you are able to see document only till the word limit of 1000 words. To view entire document upgrade your plan.

[Click to Upgrade](https://typ.st/2noUNUS)

[www.typeset.io](https://typ.st/2xUBN7Q)

* 1. Zhang Y, Zhang J, Lease M, Gwizdka J. Multidimen- sional relevance modeling via psychometrics and crowd- sourcing. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in informa- tion retrieval; 2014. p. 435–444. Available from: 10.1145/ 2600428.2609577.
  2. White RW, Horvitz E. Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search. ACM Transactions on Information Systems. 2009 November;27(4):23:1–23:37. Available from: 10.1145/1629096.1629101.
  3. White R. Beliefs and biases in web search. In: Proceed- ings of the 36th international ACM SIGIR conference on Research and development in information retrieval. New

York, NY, USA: ACM; 2013. p. 3–12. Available from: 10.1145/2484028.2484053.

* 1. Graber MA, Roller CM, Kaeble B. Readability levels of patient education material on the World Wide Web. Jour- nal of Family Practice. 1999;48(1):58–59. Available from: [https://www.ncbi.nlm.nih.gov/pubmed/9934385.](https://www.ncbi.nlm.nih.gov/pubmed/9934385)
  2. Fitzsimmons PR, Michael BD, Hulley JL, Scott GO. A readability assessment of online Parkinson’s disease in- formation. The journal of the Royal College of Physi- cians of Edinburgh. 2010;40(4):292–296. Available from: [https://www.ncbi.nlm.nih.gov/pubmed/21132132.](https://www.ncbi.nlm.nih.gov/pubmed/21132132)
  3. Wiener RC, Wiener-Pla R. Literacy, pregnancy and poten- tial oral health changes: The internet and readability levels. Maternal and child health journal. 2014;18(3):657–662. Available from: [https://www.ncbi.nlm.nih.gov/pubmed/](https://www.ncbi.nlm.nih.gov/pubmed/23784613) [23784613.](https://www.ncbi.nlm.nih.gov/pubmed/23784613)
  4. Patel CR, Cherla DV, Sanghvi S, Baredes S, Eloy JA. Readability assessment of online thyroid surgery patient education materials. Head & neck. 2013;35(10):1421– 1425. Available from: [https://www.ncbi.nlm.nih.gov/ pubmed/22972634.](https://www.ncbi.nlm.nih.gov/pubmed/22972634)
  5. Meillier A, Patel S. Readability of Healthcare Litera- ture for Gastroparesis and Evaluation of Medical Termi- nology in Reading Difficulty. Gastroenterology Research. 2017;10(1):1–5. Available from: [https://www.ncbi.nlm.](https://www.ncbi.nlm.nih.gov/pubmed/28270870) [nih.gov/pubmed/28270870.](https://www.ncbi.nlm.nih.gov/pubmed/28270870)
  6. Ellimoottil C, Polcari A, Kadlec A, Gupta G. Read- ability of websites containing information about prostate cancer treatment options. The Journal of urology. 2012;188(6):2171–2176. Available from: [https://www.](https://www.ncbi.nlm.nih.gov/pubmed/23083852) [ncbi.nlm.nih.gov/pubmed/23083852.](https://www.ncbi.nlm.nih.gov/pubmed/23083852)
  7. Evgeniy G. Cura Te Ipsum: answering symptom queries with question intent. In: Second WebQA workshop, SI- GIR 2016 (invited talk); 2016. Available from: [http://plg2. cs.uwaterloo.ca/avtyurin/WebQA2016/.](http://plg2.cs.uwaterloo.ca/%20avtyurin/WebQA2016/)
  8. Cowan CF. Teaching patients with low literacy skills. Jones & Bartlett Learning; 2004.
  9. Wallace LS, Lennon ES. American Academy of Family Physicians patient education materials: can patients read them? Family medicine. 2004;36(8):571–574. Available from: [https://www.ncbi.nlm.nih.gov/pubmed/15343418.](https://www.ncbi.nlm.nih.gov/pubmed/15343418)
  10. Davis TC, Wolf MS. Health literacy: implications for family medicine. Family Medicine. 2004;36(8):595–598. Available from: [https://www.ncbi.nlm.nih.gov/pubmed/](https://www.ncbi.nlm.nih.gov/pubmed/15343422) [15343422.](https://www.ncbi.nlm.nih.gov/pubmed/15343422)
  11. Stossel LM, Segar N, Gliatto P, Fallar R, Karani R. Readability of patient education materials available at the point of care. Journal of general internal medicine. 2012;27(9):1165–1170. Available from: [https://www. ncbi.nlm.nih.gov/pubmed/22528620.](https://www.ncbi.nlm.nih.gov/pubmed/22528620)
  12. National CI. Clear & Simple: Developing Effective Print Materials for Low-literate Readers. National Institutes of Health; Accessed: 2017-09. Available from: [https://www.nih.gov/institutes-nih/nih-office- director/office-communications-public-liaison/clear- communication/clear-simple.](https://www.nih.gov/institutes-nih/nih-office-director/office-communications-public-liaison/clear-communication/clear-simple)
  13. Palotti J, Zuccon G, Hanbury A. The Influence of Pre- processing on the Estimation of Readability of Web Docu- ments. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. New York, NY, USA: ACM; 2015. p. 1763–1766. Avail- able from: 10.1145/2806416.2806613.
  14. Palotti J, Goeuriot L, Zuccon G, Hanbury A. Ranking health web pages with relevance and understandability. In: Proceedings of the 39th international ACM SIGIR conference on Research and development in information retrieval; 2016. p. 965–968. Available from: 10.1145/ 2911451.2914741.