

Lexically-based distinction of readability levels of health documents

Antoine BORST, Arnaud GAUDINAT, Natalia GRABAR, Célia BOYER
Health on the Net Foundation, SIM/HUG, 24 rue Micheli-du-Crest, Geneva, Switzerland

Abstract

A method for automatically estimating the readability of medical websites has been designed and tested.

Increasing numbers of Internet users look for online health information : 80% according to a study made in 2006 [1]. The issue is that easy to read health documents and very complex documents both co-exist on the web. In this work we aim to propose a method for the distinction of readability level of online health documents, in order to permit users to find documents they are able to understand.

Our work relies on the lexical level of documents combined with sentence length. Our hypothesis is that words' complexity grows with their rarity so we will try to evaluate words' frequency in the language. To evaluate these frequencies, we use lexicons with variable nature: the less exhaustive a general lexicon is, lesser the number of rare words it contains. On the contrary, words that appear only in very large lexicon are likely to be incomprehensible for most people, because they are not at all common.

We use freely available lexical and terminological sources: medical terminology (MeSH [2]) and common vocabularies, which represent both specialized and common language that are likely to appear in health documents. The smaller list contains about 30'000 words, including inflections and the larger, about 260'000. We also have about 9'000 MeSH terms appearing in no other lexicon, so they represent our most technical words.

Combining all these lexicons permits us to estimate the complexity of about 270'000 terms, which is quite exhaustive. Then the complexity of a document is thus derived from the complexity of the terms employed in addition to the length of the sentences used.

Evaluation of the system is performed with 1000 documents available in the Health on the Net Foundation database and selected from *Medline Plus*' encyclopedia (<http://www.nlm.nih.gov/medlineplus/encyclopedia.html>) and *Healthday's* news (<http://www.healthday.com/>) for easy documents, and from *OESO's* articles about esophagus diseases (<http://www.oeso.org/>) and *eMedecine* site (<http://www.emedicine.com/>) for experts documents.

Documents were automatically classified as easy or complex and the average accuracy of the proposed lexically-based method reached 92% .

References

- [1] Fox S. Online Health Search 2006. Tech. report, Pew Internet & American Life Project, Washington, 2006.
-

[2] Medical Subject Headings : <http://www.nlm.nih.gov/mesh/>
