Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 100 (2016) 29 – 36

Procedia Computer Science

Conference on ENTERprise Information Systems / International Conference on Project MANagement / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2016, October 5-7, 2016

Predicting the comprehension of health web documents using characteristics of documents and users

Melinda Oroszlányová[a,*], Carla Teixeira Lopes[a,b], Sérgio Nunes[a,b], Cristina Ribeiro[a,b]

[a]*DEI, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal*
[b]*INESC TEC, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal*

Abstract

The Web is frequently used as a way to access health information. In the health domain, the terminology can be very specific, frequently assuming a medico-scientific character. This can be a barrier to users who may be unable to understand the retrieved documents. Therefore, it would be useful to automatically assess how well a certain document will be understood by a certain user. In the present work, we analyse whether it is possible to predict the comprehension of documents using document features together with user features, and how well this can be achieved. We use an existing dataset, composed by health documents on the Web and their assessment in terms of comprehension by users, to build two multivariate prediction models for comprehension. Our best model showed very good results, with 96.51% accuracy. Our findings suggest features that can be considered by search engines to estimate comprehension. We found that user characteristics related to web and health search habits, such as the success of the users with Web search and the frequency of the users' health search, are some of the most influential user variables. The promising results obtained with this dataset with manual comprehension assessment will lead us to explore the automatic assessment of document and user characteristics.

Peer-review under responsibility of the organizing committee of CENTERIS 2016

Keywords: Consumer Health Information; Information Retrieval; World Wide Web; Comprehension

* Corresponding author. Tel.: +351 22 508 14 00.
E-mail address: melinda@fe.up.pt

Peer-review under responsibility of the organizing committee of CENTERIS 2016
doi:10.1016/j.procs.2016.09.120

1. Introduction

The number of Internet users and the amount of information on the Web, including consumer-oriented health information, increased rapidly in the past decades. It has been observed that people prefer Internet as a source when searching for health-related information[18,14]. A recent survey in the USA reported that, in 2012, 72% of all adults looked online for health information[10].

Among health consumers, the Web is frequently used as a way to access health information. In the health domain, the terminology can be very specific, frequently assuming a medico-scientific character. This can be a barrier to users who may be unable to understand the retrieved documents. Therefore, it would be very useful to automatically assess how well a certain document will be understood by a certain user.

Previously, we showed that user, task and document characteristics are good descriptors and possible predictors of comprehension[3]. The goal of our current work is to investigate whether it is possible to predict comprehension using document and user characteristics. In the present study we use an existing dataset composed by a set of annotated web pages, users' characteristics and comprehension judgements[2]. Using this collection, we built a multivariate prediction model for the comprehension of the document. In the next two sections we review the related literature, and describe the dataset and statistical analysis used in the study. The following sections present the logistic regression models and the multivariate analyses. Finally, we discuss and summarize our main findings.

2. Literature Review

Since the introduction of a guide to the Internet by Pallen in the 90's[17], information on medical and health topics started to rise on the Web. As the availability of health information increased, users' health search on the Web started to have an impact on their health care routines[11,8,9,7].

The role and representation of comprehension have been studied in detail by Kintsch[15]. He defines "the mental representation of the text and actions based on this construction" as the product of the comprehension process. He names several elementary units which enter into this process (e.g.: perceptions, ideas, concepts, images, emotions) and studies the relations among them. According to Kintsch, a "comprehender", who is, in our context, a user, has specific goals, a given perceptual situation (e.g.: the words on a page of text), and a background of knowledge and experience. All these affect the information processing at the biological level. A more coherent combination of the above factors might imply that the reader can achieve a higher comprehension level in the context of online searching[22].

Search engines are mostly restricted to topical relevance. In health information retrieval, expert and average users will satisfy their information needs with very different texts[19] and comprehension can become a typical problem to average users[19], affecting decision making[4,16,12] and diminish the value of the document[21]. Several authors have shown that many users have difficulties understanding online resources for patients[6,20].

Studies about users' comprehension of health-related information available on the Web conclude that a high reading level is necessary to increase the comprehension of web-based health information[4,20]. Other authors concluded that users do not uniformly prefer simple texts, and that the text comprehensibility level should match the user's level of preparedness[5,19]. Collins-Thompson et al. showed that it works when addressed through user's reading difficulty[5]. Often, assessing the readability of medical documents is the first step to ensure that they are readable and are thus comprehensible when shared with patients and families[21]. Readability has also been shown to be useful in the evaluation of information retrieval systems for consumer health search and to contribute to system effectiveness compared to considering topicality alone[23].

3. Dataset

Our study is based on an existing dataset composed by an annotated sample of health web documents. These documents were initially collected for a user study[1] and were later automatically and manually annotated[2]. The dataset contains the Google top-30 ranked documents for 8 information situations using 4 different queries for each, with different language and medical terminology (lay or medico-scientific). The documents were assessed by a researcher, who also defined the metadata scheme. Part of the documents (10%) was also assessed by an external