Inferring Document Readability by Integrating Text and Eye Movement Features

Yongqiang Chen[1], Wenya Zhang[1], Dawei Song[1,2], Peng Zhang[1] ,Qingtao Ren[1], Yuexian Hou[1]
[1]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, China
[2]Department of Computing and Communications, The Open University, United Kingdom
dawei.song2010@gmail.com {cyq, wenyazhang, pzhang, qtr, yxhou}@tju.edu.cn

ABSTRACT

Latest research has shown that the readability of documents plays an important role in information seeking and acquisition, especially for non-domain-expert users. Classical document readability measures are based on surface text features, independent of users. In this paper, we propose to integrate text features with the users' eye movement features. The latter is expected to reflect a user's reading level, thus can be used to measure document readability in a personalized way. Based on the eye tracking data collected from our preliminary user evaluation, we investigated the impacts of different features on document readability prediction. The results show that the combination of text and eye movement features has a higher correlation with human judgments than using either of them individually.

Keywords: Information Retrieval, Readability Prediction, Eye movements

1. INTRODUCTION

There has been an increasing need for users, particularly non-domain-expert users, to find documents that are both relevant and readable [9]. However, conventional information retrieval (IR) systems do not explicitly consider the readability of the returned documents. In order to calculate document readability, various measures have been proposed based on surface text features. We agree that there exists an objectively inherent component of readability that is reflected by the text features. Nonetheless, we believe that, as a subjective concept, document readability should also be calculated in a personalized way, e.g., depending on individual users' background knowledge, interests, and reading habits.

Recent technological development has led to advanced eye tracking equipment that tracks users' eye movements in a natural setting. It becomes an effective way to investigate how users interact with documents when reading [3, 8]. As a way of implicit feedback, we can infer users reading levels from their eye movements. Thus, we propose to measure documents' readability from both objective and subjective perspectives, by combining text and eye movement features. Particularly, we investigate the effectiveness of various text and eye movement features, individually and in combination, for document readability prediction.

Related Work. Most classical readability measures predict documents readability based on text features, such as average sentence length, average word length, and ratio of complex words [2]. However, it has been shown that different users' reading levels can be different [6]. In fact, the same document can be of different readability levels for different users. Therefore, it is a reasonable to integrate user factors into readability measures. Eye tracking has been recognized as a new and effective way to capture users' implicit relevance feedback in IR. For instance, [5] used eye tracking to investigate user behavior in Web search. [4] demonstrated that users have a tendency to spend more time in examining highly ranked entries, especially the top two, in the search result list. In [1], eye movements have been used to analyze what a user has read or has skimmed, in order to improve query expansion. In addition, a probabilistic model has been proposed in [7] to learn implicit queries from eye movements data. Consistent with this trend, this paper presents a systematic investigation on combining text and eye tracking features to study document readability for IR.

2. METHODOLOGY AND EXPERIMENT

2.1 Methodology

We model readability prediction as a classification problem based on both text and eye movement features, and aim to find out which features are the most effective in predicting document readability. The *Random Forest (RF)* algorithm is employed for its ability to alleviate the over-fitting problem when dealing with small datasets (as in this paper) and to reveal the contributions from different features.

In our investigation, eight widely used text features are selected: (1) average word length; (2) average syllable length; (3) ratio of complex words; (4) ratio of long words (length > 5 characters); (5) ratio of stop words; (6) average sentence length; (7) average grammatical tree depth; (8) vocabulary size. Based on the assumption that gaze time reflects user's engagement in reading and pupil size refects the extent of user interest, seven eye movement features are investigated: (9) average gaze time per word; (10) gaze time per document; (11) pupil size of left eye; (12) pupil size of right eye; (13) average fixation duration; (14) average saccade duration; (15) skimming ratio.

The data we use are collected from a user study, detailed in the next subsection. In the user study, for each document, each user is required to assign a readability score, ranging from "1" to "6" (1 means the hardest to read, while 6 means the easiest to read). Three class labels are produced for each user. Specifically, labels "Class 1", "Class 2" and "Class 3" are assigned to documents scored "1" and "2", "3" and "4", "5" and "6", respectively. 10-fold cross validations are conducted, in which 40 documents are used for training and the rest 20 documents for testing.

2.2 User Study and Data Collection

We recruited 24 postgraduates as participants (12 males, 12 females) whose average age is 23.6. Each participant would be rewarded upon their successful completion of the experiment. We collected 60 documents from the Wikipedia and the OHUMED medical collection, which are grouped into three broad topic categories: Social Science (SS), Computer Science (CS) and Medical Treatment (MT), resulting 20 documents in each category. Generally speaking, the documents in SS tend to be the easiest to read, while documents in MT the hardest. For illustration, the average surface text readability scores calculated by "*Flesch Reading Ease*" for each category are shown in Table 1. Ranging from 1 to 100, the higher the score is, the more readable (surface text-wise) the document tends to be.

Table 1: Average Text Readability for Categories

AvgLength (100-120 words)	SS	CS	MT
Readability	78.34	39.72	33.19

Before the user experiment, all participants were asked to fill in an entry-questionnaire aiming to collect their basic background information such as age, gender, and familiarity to the three document categories. A Tobii TX300 eye tracker was used. All the participants had to make calibration to ensure accurate eye movements recording. They were then asked to carefully read through all the 60 documents that randomly displayed with a relatively large font size of 24pt. Throughut the reading process, the brightness of the monitor remains steady. After reading each document, the participants were required to give a readability score, based on his/her judgement, ranging from 1 (the hardest to read) to 6 (the easiest to read).

In order to avoid errors caused by participants' tiredness after continuous reading for a long time, we separated the experiment into three sessions. Each session has 20 reading trials (one trial per document) and each trail takes about 1 minute. Between two sessions, the participants had 10 minutes to take a rest. Each participant's personal information including calibration was stored to avoid re-calibrating after each break. We recorded all participants' eye movements data during the experiment using the Tobii studio software and exported the data as excel format for further processing.

2.3 Classification Results and Analysis

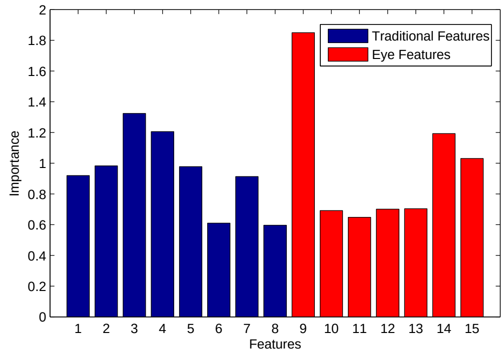
The classification performance (in term of average accuracy over participants) is summarized in Table 2. We can observe that integrating eye movement features with text features can significantly outperform the use of text features only, by 8.07% ((81.25-75.18)/75.18), with $p \leq 0.05$.

Table 2: Average Accuracy over Users

	Text Features	Eye Movement	Combination
Accuracy	75.18%	68.74%	**81.25%**

Figure 1 shows the importance of different features in classification. The average importance of text features (1-8) is 0.941, which is lower than that of eye movement features

Figure 1: Comparison among different features.



(0.974). Moreover, feature "9" (Average gaze time per word) contributes the most to the classification performance. The main reason may be that users tend to spend more time on the more readable documents according to their reading levels, which can not be captured by the text features.

3. CONCLUSIONS

We have proposed to integrate eye movement and text features for document readability prediction, which is investigated as a classification problem. Our experiment on a eye tracking data set collected from a user study shows that the combination of both kinds of features gained a better classification accuracy than using either of them individually. Our work demonstrates the potential of using eye movement features in personalized document readability computation.

4. ACKNOWLEDGEMENT

This work is funded in part by the Chinese 973 Program (grant no. 2013CB329304 and 2014CB744604), 863 Program (grant no. 2015AA015403), and the Natural Science Foundation of China (grant no. 61272265 and 61402324).

5. REFERENCES

[1] G. Buscher, A. Dengel, and L. van Elst. Query expansion using gaze-based feedback on the subdocument level. In *SIGIR*, pages 387–394. ACM, 2008.

[2] J. S. Chall and E. Dale. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books, 1995.

[3] M. J. Cole, J. Gwizdka, C. Liu, N. J. Belkin, and X. Zhang. Inferring user knowledge level from eye movement patterns. *IPM*, 49(5):1075–1091, 2013.

[4] E. Cutrell and Z. Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *SIGCHI*, pages 407–416. ACM, 2007.

[5] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in www search. In *ACM SIGIR*, pages 478–479. ACM, 2004.

[6] J. Y. Kim, K. Collins-Thompson, P. N. Bennett, and S. T. Dumais. Characterizing web content, user interests, and search behavior by reading level and topic. In *WSDM*, pages 213–222. ACM, 2012.

[7] K. Puolamäki, A. Ajanki, and S. Kaski. Learning to learn implicit queries from gaze patterns. In *ICML*, pages 760–767. ACM, 2008.

[8] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998.

[9] Y. Zhang, J. Zhang, M. Lease, and J. Gwizdka. Multidimensional relevance modeling via psychometrics and crowdsourcing. In *SIGIR*, pages 435–444. ACM, 2014.