Tracking Online Reading of College Students

Andrew M. Olney
Institute for Intelligent Systems
University of Memphis
Memphis, TN 38152
aolney@memphis.edu

Eric Hosman
Department of Counseling, Educational Psychology and Research
University of Memphis
Memphis, TN 38152
ehosman@memphis.edu

Art Graesser
Institute for Intelligent Systems
University of Memphis
Memphis, TN 38152
graesser@memphis.edu

Sidney K. D'Mello
Departments of Psychology & Computer Science
University of Notre Dame
Notre Dame, IN 46556
sdmello@nd.edu

ABSTRACT

We conducted a pilot study that used kernel-level packet capture to record the web pages visited by college students and the reading difficulty of those pages. Our results indicate that i) no students were fully compliant in their participation, ii) the number of texts encountered by participants was highly skewed, iii) the reading difficulty of texts was about 7th grade, $M = 7.24$, $CI_{95}[7.04, 7.43]$, though difficulty varied by participant, and iv) the increasing use of encryption is likely a limiting factor for using kernel-level packet capture to measure online reading in the future.

Keywords

reading, Internet, measurement, text difficulty

1. INTRODUCTION

A recent survey revealed that approximately 90% of undergraduate respondents used laptops for their electronic course readings even though 68% did not prefer electronic textbooks to print [3]. The increase in online reading behavior has created new opportunities for researchers to track ecologically valid reading behavior. Online reading reflects true interests and goals (unlike artificial experimental paradigms) and further allows measures of the time spent reading and of the text itself over extended periods of time.

To better understand the online reading behavior of college freshmen, we conducted a pilot study using custom-designed online reading tracking software based on kernel-level packet capture. Tracking naturalistic online reading behavior appears to be novel to the literature, as most studies of online reading behavior either use lab-based methods like eye-tracking or self-report methods like surveys. Our main research objectives were to determine whether i) participants would comply with the tracking, ii) the reading behavior of participants was measured consistently, and iii) the text difficulty of measured texts was in a reasonable range.

2. METHOD

2.1 Participants

Participants (N = 7) were recruited through the psychology subject pool at an urban university in the southern United States. Self-reported ACT scores ($M = 21.29, SD = 3.64$) ranged from 18 to 29. Participants were required to own and bring a laptop to the study when they enrolled.

2.2 Materials

Kernel-level packet capture software for tracking online reading behavior was developed in C♯ using the WinPcap and PcapDotNet packet capture libraries. The resulting software, called Snarf, runs as a Microsoft Windows service in the background whenever the computer is turned on. Snarf monitored all http packet traffic on all network devices and sent anonymized timestamped records of web page URLs to an online Google Fusion Tables service for collection. Records were anonymized by using the media access control (MAC) address of the participant's network card as an identifier. To minimize data traffic, Snarf sent only URLs that did not match a blacklist of known non-reading-related URLs, such as Windows Update and image/audio/video filetypes. Also excluded from collection was any service using the encrypted https protocol. Encrypted traffic was excluded for two reasons. First, it is highly likely that encrypted traffic is of a personal nature that the participants would prefer not to share, e.g. email, banking, or health information. Secondly, breaking encryption could potentially introduce security vulnerabilities and put participants at significant risk.

2.3 Procedure

Approval for the research protocol was obtained from our institutional review board. Participants were enrolled in the study in the fall of 2015. After consent was obtained,

Table 1: Participant reading behavior

			Flesch-Kincaid Grade Level				Word Count			
					95% CI				95% CI	
Id	Texts	Days	M	(SD)	LL	*UL*	*M*	(*SD*)	*LL*	*UL*
1	1	0^{-}	-							
2	23	4^{-}	9.30	(8.05)	6.01	12.59	1137.10	(1985.10)	325.83	1948.30
3	170	100^{+}	6.98	(5.74)	6.12	7.85	509.72	(1578.30)	272.46	746.97
4	210	101^{+}	9.20	(6.67)	8.30	10.11	1152.50	(2086.00)	870.37	1434.60
5	829	94^{+}	7.15	(5.57)	6.77	7.53	963.39	(1778.20)	842.34	1084.40
6	4	50^{+}	7.28	(7.13)	0.29	14.26	14.00	(8.98)	5.20	22.80
7	3116	119^{+}	7.10	(6.76)	6.86	7.34	417.77	(1236.40)	374.36	461.18

Note: CI = confidence interval; LL = lower limit; UL = upper limit; -/+ indicates under/over study length.

an experimenter installed the SNARF online reading behavior tracker onto the participant's laptop and confirmed that SNARF was logging data to the Google Fusion Table service. At the end of the study, each recorded URL was queried and, if it was accessible, downloaded. Text from downloaded files was extracted using the Apache Tika library, tokenized into sentences using the Stanford CoreNLP tools [2], and then measured for word count and text difficulty using the Flesch-Kincaid Grade Level metric [1].

3. RESULTS & DISCUSSION

Of the 327,179 timestamped URLs collected, only 87,029 were unique, and of those unique URLs, only 26,762 (31%) were downloadable at the end of the study. Inspection of the timestamped URLs revealed that, despite efforts to blacklist non-reading-related web traffic, many URLs were not reading-related, e.g. antivirus updates, ads, and video websites.

Texts from downloadable URLs had extreme Flesch-Kincaid Grade Level (FKGL) values ranging from -3.40 to 7431, and extreme word count values ranging from 0 to approximately 10 million. Inspection of the data revealed that the FKGL frequency distribution dropped precipitously at grade level 20 and that the word count frequency distribution likewise dropped at 10,000 words. These values would be possible if a participant read a document with an average sentence length of 22 and average syllables per word of 2.3 (FKGL) or a 20-page single spaced paper (word count); thus these values are plausible but may be overly generous. Descriptive statistics for the texts and downloadable URLs after applying these filtering criteria are shown in Table 1.

Table 1 presents evidence addressing our research objectives. First, participants did not comply with tracking: two participants uninstalled the software within a week (one within the same day) and the remaining five participants failed to uninstall the software or meet the experimenter to uninstall the software after being reminded by email. Secondly, participant's online reading behavior was not measured evenly: the number of texts (as measured by downloadable URLs) read by participants was highly skewed, ranging from 1 to over 3,000. This skewed distribution could be caused by some participants mostly using encrypted sites like Wikipedia or the New York Times which, by virtue of being encrypted, SNARF would not record. Finally, the reading difficulty of texts was in a reasonable range, generally 7th grade, $M = 7.24$, $CI_{95}[7.04, 7.43]$, and word count on average was comparable to a page of single spaced text, $M = 564$, $CI_{95}[521, 507]$, though both varied somewhat by participant as shown in Table 1. These results are slightly lower than might be expected when reading for academic purposes, but for general reading seem reasonable.

4. CONCLUSIONS

Our results indicate that kernel-level packet capture is a viable means for measuring online reading behavior save for the increasingly prevalent use of encryption on all web sites. While it would be possible to modify a browser to record the text displayed to the user, this alternative could inadvertently collect email, banking, or health information that should remain private. Thus it may be that the balance between privacy concerns and reading research is best struck by avoiding general purpose reading applications like web browsers and instead focusing on reading-specific applications that are not otherwise used to access personal information.

5. ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (NSF; 1235958 and 1352207) and Institute of Education Sciences (IES; R305C120001). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the author and do not represent the views of the NSF or IES.

6. REFERENCES

[1] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy enlisted personnel. Research branch report 8-75., Naval Air Station, Memphis, 1975.

[2] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[3] D. Mizrachi. Undergraduates' academic reading format preferences and behaviors. *The Journal of Academic Librarianship*, 41(3):301 – 311, 2015.