Are Ten Participants Enough for Evaluating Information Scent of Web Page Hyperlinks?

Christos Katsanos, Nikolaos Tselios, and Nikolaos Avouris

Human-Computer Interaction Group, Electrical and Computer Engineering Dept.,
University of Patras, GR-265 00 Rio Patras, Greece.
{ckatsanos, nitse}@ece.upatras.gr, avouris@upatras.gr

Abstract. Information scent of hyperlinks, that is the user's assessment of semantic relevance of navigation options in a webpage, has been identified as a critical factor in Web navigation. An important question in this context is to identify the minimum number of participants required to measure reliably information scent. A two phase study was conducted in an attempt to provide an answer to this question. In the first phase, involving 101 participants, ratings produced by different size subsets of participants were compared to those of the whole set. In the second phase, the ratings of these different size subsets of participants where compared with measures of behavior of 54 participants, who performed the same information navigation tasks using a typical web browser. Results indicate that representative estimates of information scent can be obtained from 10 participants in both cases. This finding has important implications for future scent-related studies.

Keywords: Information scent, Web usability study, cost-benefit analysis.

1 Introduction

Recent models of user behavior while foraging for information in the Web have contributed to the better understanding of human-information interaction. A key concept in these models (e.g. SNIF-ACT, CoLiDeS, MESA - see [1] for a review) is *information scent*, defined as the user's assessment of semantic relevance of the provided navigation options. Recent studies rendered information scent as the most important factor in Web navigation [1], [2].

Various semantic similarity algorithms, such as LSA and PMI-IR, have been proposed as a computational model of information scent [3] and have been used in order to facilitate the task of measuring information scent [4], [5]. However, in certain cases, such as modeling users with considerable background knowledge and/or expertise, or assessing similarity of 'rare' or informal words, computational techniques may yield misleading results.

Therefore, often human raters are called to evaluate information scent [6], [7]. In a typical study, participants are presented with hyperlink options as well as the information goal and are asked to evaluate semantic similarity among them. A key question in this context is how many people to involve in such a study in order to

obtain representative estimates of information scent. The current practice does not seem to follow a clear pattern on this issue. For instance, Miller and Remington [6] used the assessments of three judges and Brumby and Howes [7] reported involvement of 13 participants, without discussing the quality of the results with respect to the number of raters.

A similar question has been asked in the context of general usability evaluation. As concluded by a variety of related studies, the required number of participants to unveil a specific percentage of usability errors is ruled by a cumulative function of the geometric distribution [8]. The question is of significant importance towards understanding quantitative aspects of human-information interaction on the Web.

In the study presented in this paper, we attempt to identify the number of required participants to evaluate information scent in a reliable way, using as reference the ratings of a large set of users first and objective user behavior measures in the second phase of the study. Such a finding could help both practitioners and researchers to manage the available resources in a more efficient way.

2 First Phase of the Study

In the first phase of the study we compared the ratings of different size groups of raters with the ratings produced by a large pool of raters, considered as reference case. The study involved 101 University students, 39 female, with a mean age of 22.2, who were asked to rate on a 1-5 scale the semantic relevance of all the links of a menu to the associated given goal (1=poor relevance, 5=high relevance). All eight menus consisted of eight links each and were selected from actual websites[1]. A total of 6464 ratings (=8 goals x 8 links x 101 raters) were gathered during this phase of the study.

Then, subsets of various sizes were built and compared to the ratings of the whole group; an approach also used by Tullis and Wood [9 - page 223] who aimed at identifying the optimal number of users required for a card-sorting study. Ten subsets were randomly selected, of N raters each, for N=2, 5, 10, 15, 20, 25, 30, 40 and 50. Next, the average ratings of these subsets were compared to the ratings of the whole population of raters. The mean spearman correlation between the ratings of each sample size and the ratings of the 101 raters was calculated.

Fig. 1a presents the resulting total variance explained (R^2) as an increasing function of sample size. The error bars in the graph represent standard deviation of the values for the 10 random samples and were calculated as $(r_{MEAN} \pm r_{SD})^2$. As depicted in the graph, a sample size of 10 raters was found to explain 84-90% of the total variance of the ratings of all 101 participants. The lowest value observed for this sample size was 76% for the seventh goal, whereas the highest was 98% for the fifth goal[1]. After that point, there is a marginal gain in involving more participants. In specific, increasing to 15 or 20 participants does not have any impact and only when the raters are tripled the results get approximately 5% closer to the whole dataset. Thus, 10 raters appear to be a cost-effective solution to evaluate information scent without expense in the quality of results.

[1] Tasks and menus used can be found at http://hci.ece.upatras.gr/Katsanos_et_al_INT2009

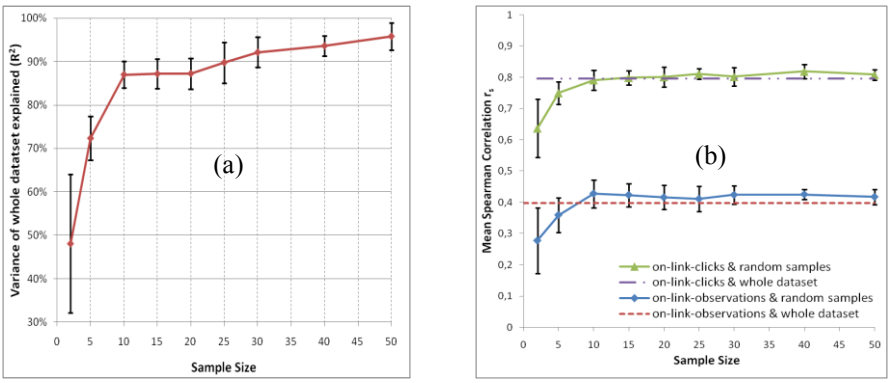

Fig. 1. (a) Total variance of 101 participants' scent-ratings explained as a function of sample size. (b) Mean spearman correlation between two measures of users' behavior (on-link-clicks, on-link-observations) and scent-ratings of random samples of raters. Note: Error bars represent standard deviation of the 10 random samples.

3 Second Phase of the Study

In the second phase, a new set of users was asked to perform the navigation tasks of the first phase using a typical Web browser. Fifty-four University students, 11 female, with a mean age of 24, all proficient in English, took part in this phase. First, users were presented with a goal-description screen. Next, they were presented with the associated menu and were asked to select a link as they would normally do. The presentation of the menus and the order of links were randomized to avoid serial order effects. An unobtrusive 17'' Tobii T60 eye tracker with minimum fixation duration set to 100ms was used to record users' eye movements. Two measures of users' behavior were gathered: a) *clicks on each link* and b) *observations on each link*. Observations were used instead of simple on-link fixations to avoid bias of higher fixations counts due to lengthier text descriptions [10].

Next, the ratings of the different size sets of raters of the first phase were correlated with these two measures of users' behavior, using mean spearman correlation. A non-parametric measure of association was used since the assumption of normality was violated for all variables. The question in this case was to identify the number of human raters that were enough to reach an acceptable rate of correlation with the two measures of users' behavior that were used as a reference.

Fig. 2b presents graphs of the resulting mean correlations. The dotted lines represent the mean correlation between measures of observed behavior and all participants' scent-ratings for the eight goals. This correlation coefficient is high for the on-link-clicks measure (r_s=0.80, $p<0.01$, one-tailed) and medium for the on-link-observations measure (r_s=0.40, ns). As depicted in Fig. 2b, 10 participants are enough to reach these values with 0.7% deviation for the on-link-clicks measure and 7.4% deviation for the on-link-observations measure. However, given the medium overall correlation between scent-ratings and on-link-observations found in this study, scent-ratings should be used only as a rough indicator of users' distribution of attention on the available navigation options.

4 Conclusions

The goal of this paper was to investigate the minimum number of participants required to representatively evaluate information scent. Analysis of the data collected in the reported study, suggest that 10 human raters can be enough to obtain representative results of users' link-selection behavior and distribution of attention on the available links. Involving more users increases the resources spent with marginal gain in the quality of results.

This is an important finding for both researchers designing future Web interaction studies, and practitioners evaluating the semantic appropriateness of hyperlinks in a webpage. Furthermore, it contributes to the overall debate on suitable number of users for a Web usability study. In addition, it was found that scent-ratings should be used only as a rough indicator of users' distribution of attention regardless of the number of raters involved, due to their medium overall correlation with on-link-observations (r_s=0.40, ns). In such cases, an eye-tracking study would be more suitable. Furthermore, it should be noted that if strong statistical inferences about the user population are required, then additional participants should be recruited.

Future work includes investigating the influence of task complexity on the optimal number of participants required, as well as investigating the presented finding in the context of highly specialized domains and/or varied user group composition [8].

References

1. Fu, W., Pirolli, P.: SNIF-ACT: a cognitive model of user navigation on the world wide web. Human-Computer Interaction, 22(4), 355--412 (2007)
2. Spool, J., Perfetti, C., Brittan, D: Design for the Scent of Information. UIE Fundamentals, User Interface Engineering (2004)
3. Kaur, I., Hornof, A.J.: A comparison of LSA, wordNet and PMI-IR for predicting user click behavior. In: Proceedings of CHI 2005, pp. 51--60. USA, ACM (2005)
4. Blackmon, M., Kitajima, M., Polson, P.: Tool for accurately predicting website navigation problems, non-problems, problem severity, and effectiveness of repairs. In: Proceedings of CHI 2005, pp. 31--40. USA, ACM (2005)
5. Katsanos, C., Tselios, N., Avouris, N.: InfoScent evaluator: a semi-automated tool to evaluate semantic appropriateness of hyperlinks in a web site. In: Proceedings of OZCHI, pp. 373--376. Sydney, Australia, ACM (2006)
6. Miller, C.S., Remington, R.W.: Modeling information navigation: implications for information architecture. Human-Computer Interaction, 19(3), 225--271 (2004)
7. Brumby, D., Howes, A.: Strategies for Guiding Interactive Search: An Empirical Investigation into the Consequences of Label Relevance for Assessment and Selection. Human-Computer Interaction, 23(1), 1--46 (2008)
8. Caulton, D.A.: Relaxing the homogeneity assumption in usability testing. Behaviour & Information Technology, 20(1), 1--7 (2001)
9. Tullis, T., Albert, W.: Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics. Morgan Kaufmann (2008)
10. Poole, A., Ball, L.J: Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects. In: C. Ghaoui (ed), Encyclopedia of Human Computer Interaction. Idea Group Reference, pp. 211--219, (2006)