Was That Webpage Pleasant to Use? Predicting Usability Quantitatively from Interactions

Maximilian Speicher[1,2], Andreas Both[2], and Martin Gaedke[1,⋆]

[1]Chemnitz University of Technology, 09111 Chemnitz, Germany
{maximilian.speicher@s2013,martin.gaedke@informatik}.tu-chemnitz.de
[2]R&D, Unister GmbH, 04109 Leipzig, Germany
{maximilian.speicher,andreas.both}@unister.de

Abstract. Webpage usability is crucial for customer satisfaction and loyalty. Yet, evaluations of webpages are usually tedious or do not provide sufficient information. Thus, we aim at providing a novel layout-independent framework for automatically predicting a quantitative measure of usability from user interactions. A study has shown that it is necessary to take into account differences in user intention and structural features already for very similar webpages. We propose preprocessing steps in terms of structure-based clustering and determining user intention, which will make it possible to provide meaningful usability models that support satisfaction and loyalty.

Keywords: Classification, Interaction Tracking, Metrics, Usability.

1 Introduction

The usability of a webpage is a crucial factor for customer satisfaction and loyalty [9]. In today's IT industry, usability evaluations are commonly performed as lab studies, inspections by dedicated experts or split tests. While the first two options are costly and time-consuming, the latter is usually based on conversions (e.g., a completed checkout process) and cannot give insights into the actual behavior of the user [7]. Particularly, a higher conversion rate might even be *contradictory* to usability [7]. Yet, split tests are the most convenient and cost-effective way of evaluating an *online* webpage with real users. Thus, we aim at providing a novel quantitative measure for predicting usability based on user interactions. In this way, we can instantly measure the usability of a live webpage and split testing can be based on usability rather than conversions alone, among other things. This adds to customer satisfaction and loyalty.

Based on an instrument for measuring web interface usability [10], we have developed a webpage plug-in for tracking user interactions and asking for explicit usability ratings. Having obtained a set of training data from one or more webpages of the same type (e.g., news websites), it is possible to learn a statistical

⋆ This work-in-progress is part of a PhD thesis carried out in cooperation with *Unister GmbH* and supervised by Prof. Dr.-Ing. Martin Gaedke and Dr. Andreas Both.

Q.Z. Sheng and J. Kjeldskov (Eds.): ICWE 2013 Workshops, LNCS 8295, pp. 335–339, 2013.

model which predicts usability quantitatively from implicit user behavior alone. Given similar websites and normalized interaction features (e.g., based on the amount of text content or number of media elements), we expected the usability measure to be *layout-independent* to a certain degree. Thus, our initial hypothesis was that *an according model should be able to predict usability, not only for webpages that delivered training data, but also for other pages of a similar kind.* In particular, this would enable internet companies that run several websites of the same type to launch a new website and instantly measure its usability based on training data obtained from the established ones. Additionally, comparison to competitors' websites would be more easily possible.

For generating a first training set, we have conducted a user study featuring four specifically prepared online news articles from different sources. Results suggest that—despite normalization of the tracked interaction features—user behavior varies considerably already for very similar webpages of the same type. *This means that the desired model needs additional preprocessing steps, i.e., clustering pages by structure and providing different models for different user intentions to provide a reliable measure for usability.*

2 User Study

We recruited a total of 81 non-unique participants (66 male) at an average age of 28.43 (σ=2.37) via Twitter, Facebook and internal mailing lists. Each participant had to read one out of four online news articles (published by CERN, CNN, Yahoo! News, Scientific American) about the Higgs boson[1] and was asked to answer a specific question. Once the user found the desired answer or was absolutely sure the article did not contain it, they had to indicate that they finished the task. Subsequently, they were presented a questionnaire for rating the usability of the online news article based on yes/no questions for the usability items *informativeness*, *understandability*, *confusion*, *distraction*, *readability*, *information density* and *accessibility* [10]. This means we determined an overall *usability value* between 0 and 7 points. It was possible to take part in the study multiple times with a different article each time.

Only two of the articles contained the necessary piece of information to answer the question (CERN, CNN). Moreover, two of the articles featured a rather short text (CERN, Yahoo! News: $\sim$1 page) while the remaining two featured a longer text (CNN, Scientific American: $\geq$2 pages). Thus, the news articles constitute the four sets $ANSWER_{yes}$, $ANSWER_{no}$, $TEXT_{long}$ and $TEXT_{short}$.

User Interaction Tracking. We used a specifically developed jQuery plug-in to track participants' interactions during the study. That is, we recorded low-level mouse events and determined a number of features from these on the client side. The features were chosen based on existing research (e.g., [2,5]) as well

[1] Our aim was to choose a topic an average user would most probably not be familiar with. The complete set-up of the study can be found at `http://vsr.informatik.tu-chemnitz.de/demo/inuit`.

as own experience with user interaction tracking (clicks, length of cursor trail and hovers, among others). Where appropriate, the features were determined separately for the whole page, the *area of interest* (AOI)[2], all media elements, all text elements and media/text elements in the AOI respectively.

The investigated news articles were slightly different concerning their structure (e.g., number of media elements or text length). Thus, the collected interaction features were normalized using certain features of the webpage to ensure comparability. For example, the page dwell time was normalized by the main article's word count (i.e., we assumed the dwell time to depend on the time needed for reading the article) and the total amount of scrolling was normalized by the height of the document.

3 Results

Let IF be the set of interaction features {"clicks", "hovers", ...}, UI be the set of usability items {"informativeness", "understandability", ...}, and $X(A)$ be the random variable X for the set of webpages A. Then,

$$\begin{aligned}\mathcal{NC}(A) \stackrel{\text{def}}{=} \big\{(if, ui) \in IF \times UI \mid corr\big(if(A), ui(A)\big) \geq 0.3 \\ \wedge\ \%RSD\big(if(A)\big) < 100 \ \wedge\ \%RSD\big(ui(A)\big) < 100\big\}\end{aligned} \tag{1}$$

is the set of noteworthy correlations for the set of webpages A.[3] We have computed $\mathcal{NC}(A)$ for five sets, i.e., the set containing all four articles ALL as well as $ANSWER_{yes}$, $ANSWER_{no}$, $TEXT_{long}$ and $TEXT_{short}$. Based on our initial hypothesis and the fact that all interaction features were normalized we expected large commonalities among all sets of webpages in this respect. However, out of 46 noteworthy correlations that were identified only five occured for more than one set. In fact, the largest set of common noteworthy correlations was $\mathcal{NC}(ALL) \cap \mathcal{NC}(TEXT_{short})$ with a size of only three. This result indicates that already for very similar webpages of the same kind, patterns of user interaction vary considerably.

Furthermore, patterns of user behavior vary, not only due to structural features of a webpage ($TEXT_{long}$ vs. $TEXT_{short}$), but also due to differences in users' intentions. These differences were "simulated" by providing only two articles containing the answer to the posed question. While users who can answer the question should act like a *fact finder* [3], users who cannot should behave more like an *information gatherer* [3]. This assumption is underpinned by the fact that $ANSWER_{yes}$ and $ANSWER_{no}$ have no common noteworthy correlation.

The complete list of noteworthy correlations per set of webpages can be found at `http://vsr.informatik.tu-chemnitz.de/demo/inuit`.

[2] The AOI, i.e., the main article text was annotated manually for each news article. Confer [2] for a different, more automatic approach.

[3] The thresholds of 0.3 for correlations and 100% for relative standard deviations (%RSD) were chosen after qualitative inspection of the data and are rather generous.

4 Implications for Future Work

The above results indicate that despite similarity in type and content, there are only very few common patterns of interaction across webpages. This rejects our initial hypothesis that it should be possible to predict a webpage's usability based on training data from different webpages of the same type (e.g., news websites). Instead, our results suggest that users' interactions on a webpage depend, not only on its *usability*, but also on *lower-level structure* and *user intention*. Thus, a framework for layout-independent prediction of usability must include two additional preprocessing steps. First, we need to cluster webpages according to their structure to minimize variations of user behavior in this respect. A different experiment (that we will not discuss in detail here) has shown that users indeed behave very similarly on similarly structured pages. Hence, based on the structure s of a page and the collected user interactions b, we can infer the user intention i [3] using an appropriate classifier I_s: $i = I_s(b)$.

Once we know both the structure of the webpage and the user's intention, it should be possible to predict the webpage's usability u with an according classifier $U_{s,i}$: $u = U_{s,i}(b)$.

To summarize, the hypothesis we derive from the above is as follows: *Within a cluster of webpages, we can provide a common model to predict a quantitative measure of usability for a given user intention (e.g.,* fact finder *or* information gatherer*).* Investigating this hypothesis and providing an according layout-independent framework is currently our main direction of future work.

5 Related Work

Our research is related to a variety of existing work in the fields of automatic usability evaluation, page clustering and prediction of user tasks. In [6], Nebeling describes usability metrics for large screens that are of a rather static nature and do not depend on user interactions. He also proposes an automatic approach for detecting potentially usability-critical components of webpages on touch devices. However, no quantitative measure for usability is provided. In [5], the authors aim at measuring user experience based on mouse tracking. Yet, their work is focused on the effect of advertisements/images on user attention. Again, no quantitative measure is provided.

In terms of page clustering, [4] describe an approach that is already based on user interactions. However, for clustering based on structure, it could also be possible to use existing approaches for page segmentation (e.g., [8]). We intend to build on these starting points to realize the structure-based clustering of webpages.

Regarding the prediction of user tasks, [3] presents an approach distinguishing between three kinds of user intentions based on the analysis of client logs. This approach can reach "accuracy values of up to 95% of correctly identified user tasks" [3]. As a more specific use case, [1] engage mouse movements to determine searcher intention on web search results pages. We intend to build on these starting points for determining user intention.

6 Conclusion

This work-in-progress paves the path to automatically predicting the usability of a webpage based on user interactions rather than questionnaires or tedious evaluations. Our intended approach bears numerous advantages concerning the evaluation and comparison of webpages and helps to ensure visitor satisfaction and loyalty. An initial study has shown that type-similarity of webpages and normalization of interaction features are not sufficient for providing a common and layout-independent usability model. Thus, we aim at providing a framework that involves additional necessary preprocessing steps, i.e., a) structure-based clustering of webpages and b) determining user intention. Moreover, we want to reinvestigate normalization of interaction features since our current approach might not be optimal yet. Grouping certain features of interaction (e.g., using an exploratory factor analysis) for finding stronger correlations with usability items might be an additional way of optimizing our desired model for usability prediction.

Acknowledgments. We would like to thank our industry partner *Unister GmbH*. This work has been supported by the ESF and the Free State of Saxony.

References

1. Guo, Q., Agichtein, E.: Exploring Mouse Movements for Inferring Query Intent. In: Proc. SIGIR (Posters), pp. 707–708. ACM, Singapore (2008)
2. Guo, Q., Agichtein, E.: Beyond Dwell Time: Estimating Document Relevance from Cursor Movements and other Post-click Searcher Behavior. In: Proc. WWW, pp. 569–578. ACM, Lyon (2012)
3. Gutschmidt, A.: The Prediction of Web User Tasks by Analyzing Client Logs. IADIS Int. J. on WWW/Internet 7(1), 79–93 (2008)
4. Leiva, L.A., Vidal, E.: Assessing Users' Interactions for Clustering Web Documents: A Pragmatic Approach. In: Proc. HT (Posters), pp. 277–278. ACM, Toronto (2010)
5. Navalpakkam, V., Churchill, E.F.: Mouse Tracking: Measuring and Predicting Users' Experience of Web-based Content. In: Proc. CHI, pp. 2963–2972. ACM, Austin (2012)
6. Nebeling, M.: Lightweight Informed Adaptation: Methods and Tools for Responsive Design and Development of Very Flexible, Highly Adaptive Web Interfaces. PhD thesis, ETH Zurich (2012)
7. Nielsen, J.: Putting A/B Testing in Its Place, `http://www.nngroup.com/articles/putting-ab-testing-in-its-place/`
8. Sano, H., Swezey, R.M.E., Shiramatsu, S., Ozono, T., Shintani, T.: A Web Page Segmentation Method by using Headlines to Web Contents as Separators and its Evaluations. IJCSNS 13(1), 1–6 (2013)
9. Sauro, J.: Does Better Usability Increase Customer Loyalty? `http://www.measuringusability.com/usability-loyalty.php`
10. Speicher, M., Both, A., Gaedke, M.: Towards Metric-based Usability Evaluation of Online Web Interfaces. In: Mensch & Computer Workshopband, pp. 277–281. Oldenbourg, Munich (2013)