Using Web Analytics to Investigate the Navigational Behavior of Users

Marc Ahrens\textsuperscript{1}, Ralph Kölle\textsuperscript{2}, Katrin Werner\textsuperscript{2}, Thomas Mandl\textsuperscript{2}

\textsuperscript{1}HIS Hochschul-Informations-System eG
\textsuperscript{2}Institut für Informationswissenschaft und Sprachtechnologie, Universität Hildesheim

Abstract

Today, web analytics and big data provide the basis for the evaluations of websites. However, the interpretation of this data in terms of Human Computer Interaction (HCI) can be challenging and there has been little progress in the systematic exploration of the potentials of web analytics within the usability evaluation domain. In this paper, we introduce a new approach to analyzing large amounts of usage data with the goal to detect usability problems in the use of a website. The underlying idea was that the analysis of patterns can support the HCI analysis of a website. To this end, we developed an extension for the open source web analytics package Piwik that allows investigating the navigational behavior of users through the abstraction of individual interaction paths. The developed system supports data mining on navigation logs and the discovery of interesting patterns.

1 Introduction

Today’s relevance of the internet for private and commercial use can hardly be overestimated. Today, more than 50 million people in Germany use the Internet as integral part of their daily lives (Bundesverband Digitale Wirtschaft 2013). This includes all age groups, whereas 14-29 year old Germans have the highest online penetration rate. According to Bundesverband Digitale Wirtschaft (2013) almost all internet users (97.1 percent) have already searched for information on a product or service on the internet. However, offering a web page alone is not enough to guarantee commercial success. Along with design and content the navigation of the web page plays an important role. The best way to investigate the usability of a website is to conduct a dedicated user test in order to understand how potential users cope with the navigation. Unfortunately, running such tests is associated with high costs which in particular small and medium-sized enterprises are not able or willing to take. Log files are a comparatively cheap alternative to access large amounts of usage data. In this paper we introduce an approach for the investigation of the navigational behavior of users that combines web analytics with knowledge about the structure of the website. The main goal was to take the navigational behavior of real users and to determine patterns that...
signify usability problems (e.g. disorientation on a web page). We wanted to use the structural information about the web page to supplement the log file analysis. This means linking the information of the log files to the structure of the page. In order to allow for an easy integration into existing web pages we decided to implement the software component as a plugin for the widely used open source web analytics software Piwik.

Web analytics is the process of automatically tracking and analyzing the behavior of users interacting with a web page. The use of web analytics enables a company to discover new insights about their customers, to attract more visitors or to increase the sales per customer. In fact, the analysis of log file data is an important complement to usability tests. Whereas laboratory tests can only appraise the behavior of a small part of consumers (who are often not even real users), the usage of log data allows taking a closer look at the entire user population over a considerable period of time. However, both approaches have their own deficiencies and therefore it is important to understand their potentials and limitations.

Related work can be broadly divided into research concerned with finding practical ways to identify usability issues using web analytics and research aimed at analyzing navigational trails in order to relate frequent patterns of user behavior to usability issues.

1.1 Identification of Usability Issues Using Web Analytics

Claypool et al. (2001) examined which indicators best predict users’ interest in the content of a website. In their study with 75 participants and 2267 analyzed web pages the authors used a browser specially developed for this investigation. Along with dwell time, mouse clicks and movements, information on scrolling results and the use of arrow keys, participants could communicate their interest in the web page content. Claypool et al. (2001) then examined correlations between the actions and the explicit statements of the participants. The results showed that both dwell time and the time that the participants used the scrolling function, were good indicators for the user’s interest in a web page. With the help of these implicit indicators the researchers could predict interest with 70 percent likelihood.

Burton and Walther (2001) used server logs in order to investigate a usability guideline of Jakob Nielsen¹ that claims that it is not good usability to use frames in websites because frames tend to confuse users. Consequently, Burton and Walther (2001) conducted an online experiment that used log file data to compare user responses to three different versions of the same website. In each version the content was identical, but the structural layout varied. The versions compared included a pop-up version, a typical frame layout as well as a single long web page. The results indicated two main findings. First, frames do not affect the amount of time users spend on a website which was interpreted by Burton and Walther (2001) as an indication that frames are not as confusing to users as Nielsen assumes. And second, frames make a difference in the amount of content users consume.

Jürgens et al. (2010) used the web analytics software Piwik to investigate a website that contains information about mobile phones. The purpose of this study was to determine the

¹ http://www.nngroup.com/articles/original-top-ten-mistakes-in-web-design/
importance of standard web analytics metrics for usability evaluations of websites. The study analyzed data over a period of at least two months, focusing on how to relate this data with the three usability indicators: effectiveness, efficiency, and satisfaction. The structure of the evaluation was threefold: user properties and behavior, referrer and keywords, content use, and navigational behavior. Overall, the study showed that in some instances, it is possible to identify usability issues. With respect to the navigational behavior of users, Jürgens et al. (2010) observed that landing pages that only result in few user interactions may reflect a low acceptance rate.

In conclusion, we see that even though the study of Claypool et al. (2001) shows a connection between dwell time and user interest, according to Burton and Walther (2001), no direct inference about the information consumption can be drawn. Jürgens et al. (2010) on the other hand found that also in the case of content sites, web analytics data can be related to usability metrics if the navigational behavior is taken into account, as well.

1.2 Identification of Usability Issues Using Navigational Trails

Botafogo et al. (1992) analyzed websites in terms of the graph structure given by the links between subpages. In order to classify these hierarchical structures, the authors introduced different measures such as compactness and stratum. Compactness specifies the connectedness of different subpages. For example, long click paths between different nodes decrease the compactness of the website. Stratum, on the other hand, measures the linearity of a webpage: the more options are available for the traversal of a web document, the lower its stratum value. Botafogo et al. (1992) do not recommend a desirable level of compactness or stratum, but understand these measures as a reference point to support the author by the design of a webpage.

There are also studies dealing with log file information in order to predict user behavior and interests. Pitkow and Pirolli (1999) studied the possibility to predict the next interaction of a user with a webserver conditioned on his previous interactions. In particular, they investigated how to reduce the model complexity of such a predictor without sacrificing the accuracy of the model. The key assumption in order to extract predictive information from the log files was to focus exclusively on the most frequent interaction patterns. In fact, their experiments show that compared to various Markov models, the most frequent longest subsequences have the potential to achieve the same predictive power with a reduced computational complexity.

A study conducted by White and Drucker in 2007 tracked the interaction of 3291 participants with their browser for a period of five months and analyzed their search and browsing behavior. For the purpose of this study, they represented the search trails of their participants as character strings and compared these with the help of the Levenshtein distance. The researchers were able to identify two different user types, which can be distinguished by their search and navigation behavior. While for example, members of the first group tended to enter a search query, visit one result page and navigate through this page extensively before running another search, the behavior of members of the other group was less linear and
showed a larger variety of interaction patterns which includes going back and forth between the results and in-between search queries.

The third study was part of the Logclef track of the Cross Language Evaluation Forum (CLEF) in 2009. Lamm et al. (2010) investigated the occurrence of interaction patterns within interaction logs from the European Library (TEL) website collected over a period of 18 month. The focus of this study lay on the evaluation of user sessions with respect to search success. Lamm et al. (2010) propose the use of a hyperbolic tree view to visualize interaction trails. This visualization provides a simple means of determining the behavior of users within an IR system such as TEL. The start node of the tree is the point of departure of all interaction trails covered in the tree view and each hierarchy level within the tree represents one interaction within TEL. Furthermore the size of the edges visualizes which percentage of the sessions contains these interactions. The working hypothesis here was that the analysis of log files may offer clues about success or failure of a user session. In this context, Lamm et al. (2010) assumed that certain user interactions as e.g. printing, saving or forwarding a search result are preferably made in successful sessions and the tree visualization allowed them to interactively explore those supposedly successful sessions.

Most recently, Han et al. (2013) used a process mining approach to analyze navigational patterns of users of a university website in order to increase its usability and efficiency. Using the open source process mining tool ProM, originally developed at Eindhoven University of Technology, the authors investigated the click behavior of 551 users. The proposed approach includes heuristic as well as fuzzy mining techniques and investigates the navigation patterns on a content level rather than regarding the hierarchical structure of the website. From their analysis Han et al. (2013) observed five major navigation patterns on the basis of which they were able to suggest possible improvements to the university website.

In summary, we have described four approaches to utilize navigational information in the context of usability. This allowed the authors either to identify typical user behavior (White & Drucker 2007; Lamm et al. 2010), typical information needs (Han et al. 2013) or to predict subsequent user behavior (Pitkow & Pirolli 1999). In constrast, Botafogo et al. (1992) investigated structural properties of websites without an explicit connection to usage data. In this work we want to combine these two approaches and relate the observed navigational patterns with structural properties of the website.

2 Piwik Plugin to Analyze Navigational Trails

With a focus towards privacy, Piwik is an open source alternative to Google Analytics. As other web analytics products Piwik provides the means to collect and analyze website usage data. Piwik is developed in PHP and stores data in a MySQL database. This facilitates the use on any web server configured accordingly. Piwik provides many statistical key figures, such as the referrer URL, visits over time, conversion rates and for us most importantly the subpages requested by a visitor.
The main goal of this research was to illustrate and provide some direction for dealing with the complexities of log file analysis in the context of usability evaluation. In the course of this, a new plugin for Piwik has been developed which visualizes the navigational behavior of website users in order to obtain additional usability data that can be used to identify root causes of usability problems. Drawing together the findings of section 1.1, we conclude that it is possible to identify implicit indicators of usability. Furthermore, the research described in section 1.2 implies that it is useful to look at the navigational trails of users in order to discover meaningful patterns and rules that regulate the behavior of users traversing a website. In comparison with tools like the Transitions Report in Piwik or the Visitors Flow in Google Analytics, which provide path analysis on the level of each single webpage, we want to abstract from this notion and focus on the navigational patterns with respect to the structural levels of a website. Figure 1 illustrates the hierarchical structure of a web page. This means that we ordered the subpages of a website into a tree according to their click distance from the start page. Here for example we see a web page with three clearly separated hierarchy levels.

The developed software consists of two parts: The Hierarchy Crawler which constructs the hierarchy tree of a website and the new Piwik plugin which links this information with the available log file data. The two sections below describe how these two components work. The log for testing the software has been collected at the usability-toolkit.de website in the period from September 2010 until August 2012. All figures below are based on our test data.

2.1 The Hierarchy Crawler

The Hierarchy Crawler was developed to determine the hierarchical structure of a website on the basis of the address path of individual web pages. Static websites consist of actual HTML files stored in folders on a web server. In this case, the URL of a website contains its relative location on the server. However, if a website relies on a dynamical content management system (CMS), the web content is not stored in individual HTML files, but in a database. Hence, there is no folder path and the URL is created dynamically.

So far the Hierarchy Crawler requires static HTML files that are hierarchically stored in folders in order to determine the hierarchical structure of a website. The Hierarchy Crawler is
written in Java. The only requirement for running the Hierarchy Crawler is the Java run time environment (JRE) in version 7. Figure 2 shows the Hierarchy Crawler after the crawling process is completed. The result set now contains a hierarchy tree for each sub-domain.

![Figure 2: Result page of Hierarchy Crawler with hierarchy tree](image)

### 2.2 The Hierarchy Plugin for Piwik

This new Piwik plugin links the usage data with the hierarchical topology provided by the Hierarchy Crawler. The Hierarchy Plugin extends the list of Piwik statistics by a new analysis option which is available in the Piwik dashboard as soon as the plugin has been activated and the import of the hierarchy data is completed. In the following, we briefly present how this new feature can be used.

![Figure 3: Table of evaluated datasets](image)
Figure 3 shows the Hierarchy Plugin in action. This view provides access to all available log file records. In particular we can explore the click paths taken by different users supplemented with the hierarchical information. Aside from this table, the plugin offers further ways to visualize this data. In the left panel of figure 4 we see for example frequencies of different click path length.

The table in figure 4 allows us to find the most frequently used sub-click path for a given minimum and maximum click path length. In this result table every click path is linked with the corresponding data set and allows for an easy identification of typical click paths. The integrated search function also allows searching for a specific click path. In addition, the click path data can be displayed as table, tag cloud, pie or bar chart.

3 Results from an Information Website

In this section we share some first experiences with our Piwik plugin. The log file from the usability-toolkit.de website mentioned previously provided the basis for this evaluation. First, we report some descriptive figures of our sample. We then focus on the analysis of the hierarchical information offered by the Piwik plugin. As already mentioned, the log file covers the period from September 2010 until August 2012. In this period 24793 visitors entered the website with an average length of stay of 2.67 minutes. The average click path length is 3; the average number of hierarchy level changes within a click path amounts to 1.

In order to investigate the navigational behavior of visitors to this website, we performed a sub-group analysis with respect to different click path lengths (CPL). In order to be able to make generalizable observations, we only took click paths (CP) into account which occurred with a relative frequency larger than 0.5 %. To this end, we only considered click paths with a CPL below 20. Table 1 shows the eight most frequent click path subsequences (CPS) of total length 2, 3 and 5. Note that in this case every occurrence of a pattern is counted regardless of its position in the overall click path. This means in particular that a single click path might contribute several times if it contains several instances of the same subsequence. We see that the most frequent behavior consists of a horizontal navigation within the third
level of the hierarchy. It is also striking that hierarchy level zero (which refers to the start page of this website) is not contained within the eight most frequent subsequences.

\[
\begin{array}{l|c|c|c|c}
  \text{CPL} & 2 & 3 & 5 \\
  \text{N} & 35603 & 26410 & 15763 \\
\end{array}
\]

Table 1: Most frequent click path subsequences of length 2, 3 and 5 (starting anywhere in the CP)

In addition, also the pages of the first hierarchy level are rarely accessed by the visitors. If we increase the CPL from 2 to 3, it becomes even more evident that visitors tend to stay within one hierarchy level rather than switching between hierarchy levels. If we interpret frequent changes of the hierarchy level as a search for the right information, one possible interpretation of this behavior would be that the web content of this website is logically structured so that frequent switches are not necessary to find what you seek. Qualitatively, this behavior remains stable if we further increase the CPL of the subsequences to 5.

\[
\begin{array}{l|c|c|c|c|c|c|c}
  \text{CPS} & \text{Freq} & \text{CPS} & \text{Freq} & \text{CPS} & \text{Freq} \\
  1 & 8044 & 3; 3; 3 & 3969 & 3; 3; 3; 3 & 1482 \\
  2 & 2559 & 4; 4; 4 & 1188 & 4; 4; 4; 4; 4 & 458 \\
  3 & 2082 & 3; *; 3 & 1126 & 3; 2; 3; 3 & 266 \\
  4 & 2072 & 3; 2; 3 & 1103 & 3; *; 3; *; 3 & 251 \\
  5 & 1907 & 2; 3; 2 & 829 & 3; 4; 3; 4 & 247 \\
  6 & 1742 & 3; 4; 2 & 761 & 2; 3; 2; 2 & 217 \\
  7 & 1647 & 2; 2; 2 & 661 & 2; 2; 2; 2 & 210 \\
  8 & 1496 & 2; 3; 3 & 623 & 2; 3; 3; 3 & 208 \\
\end{array}
\]

Table 2: Most frequent click path subsequences of length 2, 3 and 5 (starting at the root of the CP)

\footnote{The * within a click path refers to a page view of an external website.}
The picture does also not change if we restrict the selection to subsequences at the beginning of the click path. For starting sequences of length 2, we still find that the 3;3 click path is by far the most frequent hierarchy pattern (see table 2). Aside from the already mentioned horizontal navigational behavior observed within the overall click paths this tells us something about how people arrive at the pages of this website. Click paths, which have hierarchy, level 0 as their root element show up only with the fourth highest frequency, even though one might have it expected to be the most frequent pattern. Since with increasing hierarchy level the page URL becomes more complicated it is unlikely that users would type in these web addresses by hand. Hence, it seems plausible that external pages link directly to the desired hierarchy level and not to the start page of the website. Looking at the distribution of the referrer types agrees with this interpretation, but our Piwik plugin gives us the additional information that visitors are led directly to the third, the content containing, hierarchy level. From a usability perspective such user behavior implies for example that news and recent developments which are mentioned on the starting page won’t usually be registered by the visitors and should therefore be directly accessible also from the third hierarchy level.

4 Discussion and Conclusions

This section contains an overview of what should be done in the next step. First of all, until now the Hierarchy Crawler cannot cope with arbitrary dynamical and static websites, hence there is room for improvement. One solution could be to define custom variables in Piwik that are assigned to the different hierarchy levels or page types. By integrating these variables into each subpage of a website, it should be possible to obtain the equivalent hierarchical statistics.

As already mentioned in the previous section, so far we did not have the resources to test the hierarchy approach as extensively as we would have liked. Thus, future studies will have to focus more closely on the evaluation of the developed Piwik plugin. One option would be to propose a framework to support and enhance log file analysis in the context of usability evaluation. Such a framework should specify metrics for comparing navigational trails, precise methodologies for interpreting these metrics as well as scenarios for integrating both log file analysis and user testing.

From a more empirical perspective it seems interesting to adopt the approach proposed by White and Drucker (2007) also for the analysis of navigational trails within the hierarchical structure of a website. On the one hand, it would be interesting to see if the two user types identified by White and Drucker (2007) are applicable to website usage. On the other hand, such an analysis could help to decide whether an unusual navigational behavior really points toward a usability problem or is only part of the typical behavior patterns of navigators or explorers. A similar starting point could be to further analyze the hierarchical structure of a website with respect to the interconnectedness. This would allow to access metrics such as compactness and stratum of Botafogo (1992) and open the possibility to connect these measures to the navigational behavior of the users.
Whereas log file data is sometimes hard to interpret, because the analysts do not know what users are thinking, laboratory user tests are often conducted with small sample sizes, which make it difficult to generalize the findings to the larger population of users. As a result of this research a new usability evaluation approach has been developed, which combines web analytics data with usability questions. In addition, two software applications were developed. A system to determine the hierarchical levels of a web page (The Hierarchy Crawler) as well as a plugin for Piwik (The Hierarchy Plugin), which allows to combine this hierarchical data with the data collected by Piwik and presenting the results in different display formats such as table, tag cloud, pie or bar chart. As a first application of these tools we analyzed log files of the website usability-toolkit.de. The most frequent click path subsequences showed a tendency for horizontal navigation behavior within the third hierarchy level. In addition, we could determine that the start page is seldomly visited, because external websites and search engines link directly to deeper pages.

References

Kontaktinformationen
{koelle, katrin.werner, mandl}@uni-hildesheim.de