Decision support for user interface design: usability diagnosis by time analysis of the user activity

Avi Harel
Ergolight Ltd, Haifa, Israel
ergolight@gmail.com

Ron S. Kenett*
KPA Ltd., Raanana, Israel
ron@kpa.co.il

Fabrizio Ruggeri
CNR IMATI, Milano, Italy
fabrizio@mi.imati.cnr.it

Abstract

This paper presents a methodology for setting up a Decision Support system for User Interface Design (DSUID). We first motivate the role and contributions of DSUID and then demonstrate its implementation in the case of usability diagnosis of web pages, based on time analysis of clickstream data. The resulting DSUID diagnostic reports enable website managers to learn about possible sources of usability barriers. The proposed DSUID analytic method is based on the integration of stochastic Bayesian and Markov models with models for estimating and analyzing the visitors' mental activities during their interaction with a website. Based on this approach, a seven-layer model for data analysis is suggested and an example of a log analyzer that implements this model is presented. We demonstrate the approach with an example of a Bayesian network applied to clickstream data and conclude with general observations on the generic role of DSUID and the implementation framework we propose.

1. Introduction

Operational failures, typically attributed to human errors, are a main factor of system failure. For example, 60-80% of car accidents are attributed to human errors.

Operational failures are not always identified correctly. For example, many calls for technical support about TV system malfunctioning turn out to be the result of user errors.

The term "Decision Support system for User Interface Design" (DSUID) refers to procedures employed during system development for preventing user interface design errors and for improving the system resistance to such errors. DSUID is about identifying patterns of operational failures and redesigning the user interface so that these failure patterns are eliminated. We can identify three main phases for the application of DSUID:

1. **Design phase** - At this phase DSUID provides system architects and designers with guidelines and operating procedures representing accumulated knowledge and experience on preventing operational failures. Such knowledge can be packaged in "patterns" or Standard Operating procedures (SOP).
2. **Testing phase** - operational reliability can be completely and effectively assessed only when testing with real users doing their real tasks, as opposed to testing by testing experts. Therefore, a key testing phase is the beta testing, when new versions are installed in real customer sites for evaluating the way they are actually being used.
3. **Tracking phase** - over time, systems are exposed to new user profiles. The original operational scenario might not be predictive of the evolving demands and, typically, ongoing DSUID tracking systems are required to adapt the system to the changing operational patterns. Statistical process control (SPC) may be employed to monitor the user experience, by comparing actual results to expected results and acting on the gaps. Tracking and on-going analysis provide opportunities for continuous improvements in the ways the system responds to unexpected events.

This paper presents the principles for setting up a DSUID using diagnostic reports based on time analysis of the user activity. The diagnostic reports enable product managers to learn about possible sources of dissatisfaction of site visitors.

Different types of operational design deficiencies are associated with different patterns of the user activity. The methods presented here are based on models for estimating and analyzing the user' mental

* corresponding author

0730-3157/08 $25.00 © 2008 IEEE
DOI 10.1109/COMPSAC.2008.61

activities during the system operation. Building on this approach, we propose a seven-layer DSUID model for data analysis, and present an example of a log analyzer that implements this model

2. A case study: web usability analytics

The goal of usability design diagnostics is to identify, for each site page, all the design deficiencies that hamper the positive navigation experience, in each of the evaluation stages. To understand the user experience we need to know the user activity, compared to the user expectation. Both are not available from the server log file, but can be estimated by appropriate processing. A Decision Support system for User Interface Design (DSUID) system flagging possible usability design deficiencies requires a model of statistical manipulations of server log data. The diagnosis of design deficiency involves certain measurements of the site navigation, statistical calculations and statistical decision.

How can we tell whether visitors encounter any difficulty in exploring a particular page, and if so, what kind of difficulty do they experience, and what are the sources for this experience? We assume that the site visitors are task driven, but we do not know if the visitors' goals are related to a specific website. Also, we have no way to tell if visitors know anything a priori about the site, if they believe that the site is relevant to their goals, or if they have visited it before; it may be that the visitors are simply exploring the site, or that they follow a procedure to accomplish a task. Yet, their behaviors reflect their perceptions of the site contents, and estimates of their effort in subsequent site investigation.

Server logs provide time stamps for all hits, including those of page html text files, but also of image files and scripts used for the page display. The time stamps of the additional files enable us to estimate three important time intervals :

- The time the visitors wait until the beginning of the file download, is used as a measure of page responsiveness
- The download time is used as a measure of page performance
- The time from download completion to the visitor's request for the next page, in which the visitor reads the page content, but also does other things, some of them unrelated to the page content.

The DSUID challenge is to decide, based on statistics of these time intervals, whether the visitors feel comfortable during the site navigation; when they feel that they wait too much for the page download and how do they feel about what they see on screen.

How can we conclude that a time interval is acceptable by page visitors, is too short, or is too long? For example, consider an average page download time of five seconds. Site visitors may regard it as too lengthy, if they expect the page to load fast, for example, in response to a search request. However, five seconds may be quite acceptable if the user's goal is to learn or explore specific information, if they expect it to be related to their goal. The diagnostic-oriented time analysis is by observing the correlation between the page download time and the page exits. If the visitors are indifferent about the download time, then the page exit rate should be invariant with respect to the page download time. However, if the download time matters, then the page exit rate should depend on the page download time. When the page download time is acceptable, most visitors may stay in the site, looking for additional information. However, when the download time is too long, more visitors might abandon the site, and go to the competitors. The longer the download time, the higher is the exit rate.

To enable statistical analysis, we need to change our perspective on these variables, and consider how the download time depends on the exit behavior. We compare the download time of those visits that were successful, with that of those visits that ended up in the site exit. If download time matters, we should expect that the average download time of those visitors who abandoned the site will be longer than the average of those who continued with the site navigation. Otherwise, if the site visitors are indifferent about the download time, we should expect that the download time of the two groups would not be significantly different.

To decide about the significance of the usability barrier we compare the download time of two samples: one of page visits that ended up in site exit and the other of all the other page visits. The null hypothesis is that the two samples are of the same population. If the null hypothesis is rejected, we may conclude that the page visitors' behavior depends on the download time: if the download time of the first sample exceeds that of the second sample, and the difference is statistically significant, then we may conclude that the download time of the particular page is significantly too long. A simple two-tailed t-test is sufficient to decide whether the visitors' behavior is sensitive to the page download time.

The approach described above can be generalized in a comprehensive DSUID model: the Seven Layers Model described next.

3. A DSUID implementation framework: the seven layers model

In setting up a Decision Support system for User Interface Design we identify several layers of data collection and data analytics:

The lowest layer – user activity

This layer records correspond to significant user actions (involving screen changes or server-side processing).

The second layer – page hit attributes

This layer consists of download time, processing time and user response time.

The third layer – transition analysis

The third layer statistics are about transitions and repeated form submission (indicative of visitors' difficulties in form filling)

The fourth layer – User problem indicator identification

Indicators of possible navigational difficulty, including:

- Estimates for site exit by the time elapsed until the next user action (as no exit indication is recorded on the server log file)
- Backward navigation
- Transitions to main pages, interpreted as escaping current sub task.

The fifth layer – usage statistics

The fifth layer consists of usage statistics, such as:

- Average introduction time
- Average download time
- Average time between repeated form submission
- Average time on screen (indicating content related behavior)
- Average time on a previous screen (indicating ease of link finding).

The sixth layer – statistical decision

For each of the page attributes, DSUID compares the statistics over the exceptional page views to those over all the page views. The null hypothesis is that (for each attribute) the statistics of both samples are the same. A simple two-tailed t test can be used to reject it, and therefore to conclude that certain page attributes are potentially problematic. A typical error level is set to 5%.

The top layer – interpretation

For each of the page attributes DSUID provides a list of possible reasons for the difference between the statistics over the exceptional navigation patterns and that over all the page hits. However, it is the role of the usability analyst to decide which of the potential source of visitors' difficulties is applicable to the particular deficiency

4. Modeling site navigation

The discussion presented above about the relationships between time intervals and exit rates, demonstrates our methodology. Setting up a DSUID involves integration of two types of information:

- Design deficiencies, which are common barriers to seamless navigation, based on the first part of the model described above – visitor's page evaluation
- Detectors of these design deficiencies, common indicators of possible barriers to seamless navigation, based on the second part of the model – visitor's reaction.

The way to conclude that the download time of a particular page is too long is by a measure of potentially negative user experience, namely, the site exit rate. Exit rate is a preferred measure for deciding on the effect of download time, but is irrelevant to the analysis of the visitors' response time. The reason for it is that server logs do not record event of the visitor leaving the site, which means that we cannot measure the time intervals of terminal page views. Therefore, we need to use other indicators of potential negative navigation experience.

A model of user's mental activities in website navigation lists the most likely visitors' reaction to exceptional situations. Based on this model, we can list the following indicators about the visitor's tolerance to web page design deficiencies:

- The visitor returning to the previous page may indicate that the current page was perceived as less relevant to the goal, compared to the previous page .
- The visitor linking to a next website page may indicate that the link was perceived as a potential bridge to a goal page
- The visitor activating a main menu may indicate that the visitor is still looking for the information, after not finding it in the current page
- The visitor exiting the site may indicate that either the goal has been reached, or the overall site navigation experience became negative.

Accordingly, besides site exit, other indicators of potential usability problem are the rates of navigation back to a previous page and the visitor escaping from the page to a main menu. Once we have an estimate of the task-related mental activities we can adapt the method for deciding about problematic performance, to decide about problematic task-related mental activities. Visitors who feel that the information is irrelevant to their needs are more likely to respond quickly, and are more likely to go backward, or to select a new page from the main menu. Therefore, the average time on page over those visitors who navigated backwards or retried the main menu should be shorter than that of the average of all visitors. On the other hand, visitors who believe that the information is relevant to their needs, but do not understand the page text very easily, are likely to spend more time than the average reading the page content, and the average time on page should be longer.

The time that visitors spend reading a page depends on various perceptual attributes, including the page relevance to their goals, the ease of reading and comprehending the information on page, the ease of identifying desired hyperlinks, etc. Assume that for a particular page, the average time on screen before backward navigation is significantly longer than the average time over all page visits. Such case may indicate a readability problem, due to design flaws, but it can also be due to good page content, which encouraged users who spent long time reading the page to go back and reexamine the previous page.

5. DSUID modeling techniques

In implementing the seven layers DSUID system we consider four types of modeling techniques described in detail by Harel et al. [1].

- Model 1 - **Markov Processes:** Markov processes models of the time dimension in server requests, in which the states represent the web pages and the transitions between states represent the hypertext activity.
- Model 2 - **Mental Activity:** To draw cause-and-effect conclusions we need a model of how certain situations evolve to certain effects. To identify barriers to seamless navigation, we need to have a model of normal mental activities involved in page handling, and possible deviations from the normal activities, due to design deficiencies.
- Model 3 - **Bayesian Networks:** Bayesian networks map cause and defect relationships between key variables.
- Model 4 - **Statistical Analysis:** Specific measurements of the site navigation, statistical calculations and rules for statistical decision.

We illustrate our approach with an implementation example of Model 3. A Bayesian Network is a graphical model representing cause and effect relationships between variables (Kenett, [2], Ben Gal [3]). As an example, we show in Figures 1 and 2 a Bayesian Network derived from analysis of web log analyzers. The network indicates impact of variables on others derived from an analysis of conditional probabilities. The variables have been discretized into categories with uniform frequencies. The network can be used to represent posterior probabilities, reflecting the network conditioning structure, of the 1-5 categorized variables. We can use the network for predicting posterior probabilities after conditioning on variables affecting others or, in a diagnostic capacity, by conditioning on end result variables. In our example, we condition the network to low and high seek time, Figures 1 and 2 respectively.

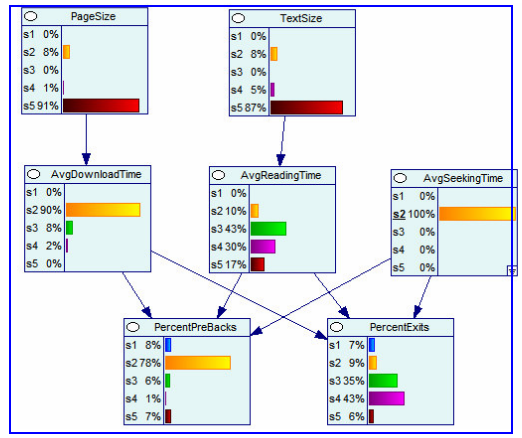

Figure 1. a Bayesian network of web log data, conditioned on low seek time.

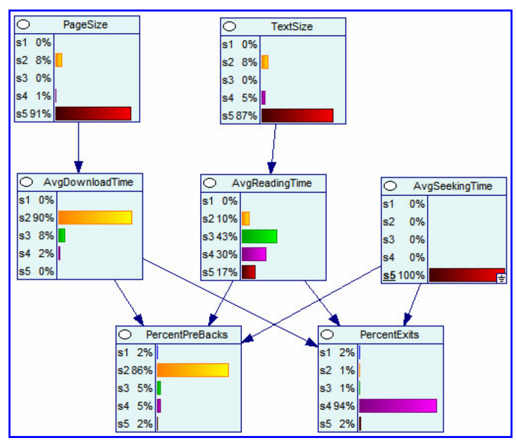

Figure 2. a Bayesian network of web log data, conditioned on high seek time.

With low seek time we experience 49% of high and very high exit rates. With high seek time these numbers jump to 96%. Clearly seek time has affected the behavior of the users. We can even quantify this phenomenon and link, quantitatively, the impact of a second of seek time on exit rate percentages.

The proposed approach is flexible since it can handle problems either by different models (as discussed above) and combine different statistical approaches. For example, a t-test has been described earlier to compare download times of successful and unsuccessful visits. The problem leads naturally to a Bayesian approach (see e.g. Robert, [4]) which combines data and expertise. Along with analysis of the log, we think it is important to elicit the opinion of experts on possible user's reactions when navigating through web sites. Independent usability engineers can provide opinions on the usability of the web pages (from used language to graphical presentation), and describe the most likely reactions to uneven events like high download times. These different, non homogeneous sources of information can be combined by statisticians into prior assessment about probability of failure/success when some patterns are followed by the users and under different system answers, e.g. download times. The process might lead to some uncertainty in the assessment of a prior distribution; we strongly suggest performing a sensitivity analysis as described by Rios Insua and Ruggeri [5].

As an example of Bayesian approach, consider the Markov processes mentioned earlier. Transactions, as described in [1], can be thought as Markov chains with states given by the web pages in the web site. Two absorbing states are present (fulfillment of the task and unsuccessful exit) and our interest is in their probability. Transitions between one state and all the possible others are modeled as a multinomial distribution with probability p_{ij} of going from state i to state j. A Dirichlet distribution is a natural, conjugate choice for the vector $\mathbf{p}_i = (p_{i1} \ldots p_{ik})$, denoting the probabilities of all possible transitions from state i. The choice of the hyperparameters is a tough, although typical, problem in Bayesian analysis, where the statistician has to transform experts' opinions into numbers. Web log provides the number of transitions from state i to state j and data can be combined with the Dirichlet prior via Bayes' theorem, so that a posterior distribution, again a Dirichlet one, is obtained. Based on the posterior distribution on the transition probabilities, operational reliability can therefore be analyzed looking at the predictive probability of ending in the unsuccessful state, given a pre-specified pattern, i.e. a sequence of transitions between web pages. With such techniques, one can combine in a DSUID, expert opinions with data to model transitions between states and describe cause and effect relationships.

6. Summary and Conclusions

In this work we describe a Seven Layer Model for setting up a Decision Support system of User Interface Design (DSUID). We demonstrate the concept by focusing on an eCommerce application. The model has an impact on current effort to improve system usability through international standards ([6]). In more general terms, proactive interventions, such as design changes are, de facto, empirical experiments. The basis of statistical inference used in Model 4 and the statistical methodology for the design of experiments using appropriate combinations of experimental runs is described in Kenett and Zacks [7]. For an introduction to Markov models used in Model 1 see Karlin and Taylor [8].

This paper was designed to be introductory and conceptual in scope, leaving out the technical details and more elaborate examples of DSUID. Our objective here is to lay out the foundations for a generic *Decision Support system for User Interface Design*, in order to tackle usability issues of modern systems and interactive devices.

7. References

[1] Harel A., Kenett R. and Ruggeri, F., "Modeling Web Usability Diagnostics on the basis of Usage Statistics" in *Statistical Methods in eCommerce Research*, W. Jank and G. Shmueli editors, Wiley, 2008.

[2] Kenett, R.S., "Cause and Effect Diagrams" in *Encyclopaedia of Statistics in Quality and Reliability*, Ruggeri, F., Kenett, R. S. and Faltin, F. (editors in chief), Wiley, 2007.

[3] Ben Gal, I., "Bayesian Networks" in *Encyclopedia of Statistics in Quality and Reliability*, Ruggeri, F., Kenett, R. S. and Faltin, F. (editors in chief), Wiley, 2007.

[4] Robert, C. The *Bayesian Choice: from Decision-Theoretic Motivations to Computational Implementation*, Springer-Verlag, New York, 2001.

[5] Rios Insua, D. and Ruggeri, F. *Robust Bayesian Analysis*, Springer-Verlag, New York, 2000.

[6] International Standards Organization, ISO 9241-11, "Guidance on Usability", 1998.

[7] Kenett, R., Zacks, S., *Modern Industrial Statistics: Design and Control of Quality and Reliability*, Duxbury Press, San Francisco, 1998. Spanish edition, 2000, 2nd edition 2003, Chinese edition 2004.

[8] Karlin, S. and Taylor, H. M. *A First Course in Stochastic Processes*. (2nd Ed.) Academic Press, New York, 1975.