There's always one! Modelling outlying user performance.

Julie Schiller
System Concepts, Ltd
2 Savoy Court, Strand
London WC2R 0EZ
UK
julie.schiller@gmail.com

Paul Cairns
Dept of Computer Science
University of York
York YO10 5DD
UK
pcairns@cs.york.ac.uk

CHI 2008, April 5 – April 10, 2008, Florence, Italy
ACM 1-xxxxxxxxxxxxxxxxxx.

Abstract
Informal analysis of many usability tests suggests that there is regularly one participant that is substantially slower than all the others. Moreover, such outliers are more extreme and more frequent than would be predicted by a normal distribution. We propose using a rational model to explain the outliers and the work described here begins to parameterise the model based on empirical data to provide accurate analyses of user performance. This prediction appears to be correct and the model begins to reflect the outlying performance. Moreover, by using an executable model, we believe that it could be used in future as an analytical tool to help designers improve usability for those users who are struggling the most.

Keywords
Outliers, modelling, user performance, population distribution

ACM Classification Keywords
H.1.2 [Models and principles]: User/machine systems – Human Factors; I.6.4 [Simulation and modelling]: Model validation and analysis; G.3 [Probability and Statistics]: Distribution functions.

Outlying performances in usability tests

Many usability tests and studies tacitly assume that user performance, specifically, time to complete a task, is accurately represented by a normal distribution. This can be seen by the prevalence of t-tests and ANOVA to analyse the differences in task times between different interfaces or in different conditions [4]. However, informal analysis of usability tests suggests that there always seems to be one participant in a test that is substantially slower than all of the others. Moreover, this person is not only performing a bit worse but is more than 2.5 standard deviations from the average performance. The probability of this happening under the assumption of normality is $p<0.0025$ or less than one such user in 400. Thus, in modest usability tests involving tens of users, it would be rare to see such outlying performances at all, let alone regularly in most tests. This suggests that the distribution underlying user performance is not normal but has a strong positive skew resulting in a higher proportion of very slow times.

Many usability tests also focus on average performance where the mean task time across participants is compared for different designs. Obviously improvements in means should substantially affect any user but the total reduction in task times could be quite small and may not correlate to improvements in user experience, rather than user performance. Small changes in average task time may be irrelevant to users and so not worth the additional development effort. However, for the straggling, outlying users, improvements in design could result in a twofold improvement in performance. That is, there is room to substantially improve the happiness of a few individuals while having only a small impact on the performance of the masses. This surely must be something that human-computer interaction (HCI) should consider. Arguably, HCI is already considering such factors with the notions of accessibility and universal access but by considering them as special cases rather than features of a general population of users.

We therefore propose that the user performance is not fully modelled by the normal distribution but that finding a more appropriate distribution could lead to improvement in usability that, for some users, will be enormous. Undoubtedly, it is possible to find a distribution that fits the data that we observe. However, without an explanation of why this distribution is appropriate, such a distribution must always be considered as contingent. It will always be a possibility that some new or more accurate observations undermine the validity of the new distribution. A theoretical approach is necessary but in the face of such a complex phenomenon as users performing tasks, it would be hard to specify a theory in sufficient detail to infer an accurate statistical distribution. Therefore, in seeking a better statistical distribution, we have taken a modelling approach where a model is run many times to provide an approximation of a population distribution, akin to a Monte Carlo simulation. The theory behind the model provides a theoretical justification for the arising distributions.

There are some key features in choosing the model as discussed in the next section. Like many models there are parameters to the model that need to be chosen, in particular what constitutes a valid set of variations between individual runs of the model. We therefore undertook an empirical data gathering exercise,

primarily to determine suitable parameters, but also to provide data to confirm our informal observations and to help evaluate the model. This is described below followed by an analysis of the performance of the model which leads naturally to a discussion of future work and our hopes for this approach.

The TreeWalker model

In devising a suitable model to simulate users performing a usability test, we required a model that could: work over a variety of systems; perform a variety of tasks; and faithfully replicate user behaviour. The cognitive model of Cox and Young [6] is a rational model of users selecting items from a menu of items depending on the task given to the user. It has been shown to replicate many of the actual behaviours exhibited by real people [3] such as choosing an item without fully scanning the list, back tracking and skipping over items. This cognitive model has a clean interface in that the only information used in making decisions on actions to perform is based on the notion of an information scent [7] of the relevance of a menu item to the task in hand. For example, a perfectly designed menu would lead the user to consider only the required item to be relevant, and so be scented with a value of 5, whereas all other items are irrelevant and so scented with a 1. The cognitive model would reject all the 1-scented items and choose the 5-scented item. This contrasts with other cognitive modelling approaches such as ACT-R [2] where in addition to a generic cognitive architecture, the model also requires representation of knowledge specific to the tasks and the interfaces being studied.

Young and Cox's model was not intended for anything more than item selection but it is easily extended to make repeated choices over a network of nodes and hence to simulate a user navigating a more complex menu hierarchy. Moreover, this overall set up gives a clear separation of system and user: the menu hierarchy is represented by a graph of connected nodes [8], the task is represented by an information scent overlying the graph and the user model makes decisions based solely on the information scent. The overall executable model was therefore a combination of these three elements. We implemented it in Java 2.0 as the TreeWalker system.

A large class of interactive applications can be modelled using TreeWalker, for example, finding items on a website or activating functions in a mobile phone menu. The model, however, is rational so when faced with the same scent structures on a given menu hierarchy it always behaves the same way. In order to simulate distributions of user performances, therefore, the scent structures are varied reflecting the variation that people perceive in the relevance of menu items to a particular task. That is, while some people may consider an item highly relevant for a task, others may consider it to be less relevant. In terms of the model, the person rating the item relevance as highly relevant would associate a scent of 5 to the item where the other person would perhaps give it a scent of 4 or 3. These variations are sufficient to produce radically different selection behaviours while using the same underlying rational process.

Thus, through perturbing the scent structures when TreeWalker is navigating a menu hierarchy, it is possible to produce a wide range of behaviours and in particular ones that take circuitous routes to achieving the task, if they achieve it at all. The question now is,

what is an appropriate range of variations to the scent structure so that is reflects the variations seen between people? To this end, we conducted a study to gather data on how people perceive the scent of menu items.

Gathering user data

The goal of the data gathering study was to get users to rate menu items for relevance to achieve a task and also as a comparison with the model performing the tasks. A secondary goal was to replicate the initial, informal observation that "there is always one!" To fill out the picture of a statistical distribution, a large number of participants was required. The empirical data gathering was therefore done using an online questionnaire (n=190) that took scent ratings and also performance measures of the subjects.

Methodology

The questionnaire asked information seekers to assess common, public-access webpages to accomplish two randomly tasks chosen from a set of five different tasks on four different websites. This number of tasks and websites were chosen to avoid issues specific to one task or tasks on a particular type of website. Tasks were intended to be realistic and appropriate to the website such as finding a present for a girl on a department store website.

After completing the demographic section, the participant used the page to view a set of 2 to 4 stored pictures, depending on the first task they were given. These pictures appeared automatically and represented menus that an information seeker would see following a direct route through the menu hierarchy to achieve the first task. Participants were asked to assess the relevance of each menu item in each picture for relevance to the task on a scale of 5 (very relevant) to 1 (not at all relevant). For the second part of the questionnaire, the participants were directed to complete the second task on the actual, live website. They were given a free text entry field to respond with the desired information and timed. The timing began when the participant opened the live site via a hyperlink and ended when an answer was entered and the survey submitted.

All participants were regular Internet users and received the survey via the Internet. Information seekers were solicited anonymously via distribution on the web, news groups, and social networking sites. There were 68 female respondents and all respondents were proficient with using the Internet.

Occurrence of outliers

First, it is useful to examine whether outlying performance was occurring more frequently than we would expect with a normal distribution. Outliers are usually identified in a known normal distribution in terms of being a certain proportion of standard deviations away from the mean. For our purposes we took an outlier to be a point more than 2.33 standard deviations from the mean as this corresponds to a probability of occurring in a normal distribution of $p<0.01$. In collected samples though, the means and standard deviations are estimated from the sample and any outliers contribute to these estimates. Thus they can distort the estimates leading to themselves looking less extreme. We therefore estimated the number of outliers for each task in two ways. The first way used a standard estimate of mean and standard deviation, and we name such outliers to be standard outliers, the second method used a robust method using medians

and deviations from the median [AMC report] that is less influenced by outliers and these was have called robust outliers. Because the tasks were chosen at random, some did not have a large sample so those below a sample size of 5 or less were excluded. This left a total of 16 tasks across all four websites. Of the 158 included participants, there were 8 standard outliers (5.1%) and 29 robust outliers (18.4%). Of the 16 tasks, 6 tasks produced at least one standard outlier and 12 robust ones. Counting only tasks where sample sizes were at least 10, there were 7 such tasks done by 90 participants resulting in 7 standard outliers (7.8%) and 19 robust outliers (21.1%). Thus, regardless of how measured, outliers are far more prevalent than the 1% level that the normal distribution would suggest.

Variation in information scents

When assessing the relevance of a given item to a task, the mean assessment was used to indicate the average relevance to the task. The variation from this average relevance was then considered. As you might expect, there were far more items rated as irrelevant to a given task than relevant. Moreover, where an item was on average very relevant or not at all relevant, there was very few disagreeing with this whereas where the average assessment was 3, the actual ratings given were in roughly similar proportions across the full range. The overall scent distributions are summarized in Table 1.

Modelling results

For each task, the TreeWalker model of the menu hierarchy was scented with the average relevance rating. Note, only the path to the task was modelled as incorrect menu choices were not rated and therefore it was not possible to model them. The model was then run 1,000 times so that variations from the average relevance produced overall variations in the proportions presented in Table 1.

Median rating	1s	2s	3s	4s	5s
1	1508	123	49	31	6
2	1082	260	230	153	113
3	235	145	256	259	168
4	41	42	105	202	295
5	2	1	10	51	210

Table 1: The numbers of each relevance rating given to all items broken down by the median rating for the items.

Over all of the sixteen tasks analysed for user performance, 2.7% of the runs were standard outliers and 4.4% of the runs were robust outliers. Over such a large sample of 16,000 runs in total, if the model were producing a normal distribution, the number of outliers ought to be very close to 1%.

Further support for the model comes from comparing it to the overall performance distributions. By comparing the TreeWalker performance to the data collected, the Kolmogorov-Smirnoff test indicates that the model using the empirically based scent variations provides a good fit ($D=0.211$, $p=0.752$). Thus, TreeWalker is representing real users' behaviour.

Conclusions and future work

In summary, studies have shown that there is always one, indeed, usually several individuals outlying the rest of the population. Their existence is unlikely to be found as part of a normal distribution, reflecting a skewed performance distribution. Furthermore, the data gathered supports the conclusions of a model

simulating rational human search behaviour in hierarchical menus. The model behaves this way due to its foundations in bounded rationality supporting the work of Cox and Young [5]. It predicts the existence of outliers as an outcome of menu structure and relevancy scent maps. By reflecting individual differences in relevancy it becomes yet more likely to accurately reflect the overall population. Interestingly, while predicting more outliers than a normal distribution, TreeWalker's proportion of outliers is still lower than found in our study. This requires further work to better relate the measure of "cognitive cost" used in the model to the actual time taken to evaluate menu items. Also fully modelling a menu hierarchy would allow for the model to pursue completely incorrect paths so increasing the possibility of outliers.

This work proposes a model for understanding and modelling patterns, not individual users. Other research into the effects of expertise and movement [4] show that there may be more parameters that are not currently explicitly addressed in TreeWalker. However, by proposing scent variation as a unitary view of relevancy, many of these parameters may already encapsulated by the model. This needs to be studied in greater detail.

The implications of this work do not only affect websites (as tested) but all hierarchical menus such as information kiosks or software interfaces. Also, the current wisdom in our field tries to guide users to a correct "golden" path. This model suggests that, as users may have different perceptions of relevancy, there ought to be several correct paths. Efforts to improve total user experience should reflect these varying, but equally valid, perceptions.

The applications of a rational model are wide ranging. One could foresee automated testing of information architectures very early in the design process by allowing TreeWalker to crawl through different structures for the data. Using this model would give a quantitative method for comparing menu structures before costly mistakes are made. Furthermore, the model could suggest improvements to menu architectures to include those with extremely poor performance. By improving their experiences we may provide a more valid improvement in user experience than making small changes in the mean.

References

[1] Analytical Methods Committee, Robust statistics. *Royal Soc. Of Chemistry, technical brief 6 (2001)*

[2] Anderson, J. and Lebiere, C. The Newell test for a theory of mind. *Behavioral and Brain Science26 (2003), 587-601*

[3] Brumby, D. and Howes, A. Good enough but I'll just check. *6th Int. Conf. on Cognitive Modelling* (2004)

[4] Cairns, P. HCI… not as it should be: inferential statistics in HCI research. In *Proc. of HCI 2007, vol 1* BCS (2007), 195-201

[5] Cockburn, A., Gutwin, C. and Greenberg, S. A predictive model of menu performance. *Proc. CHI 2007*

[6] Cox, A.L. and Young, R.M. A rational model of the effect of information scent on the exploration of menus. *6th Int. Conf. on Cognitive Modelling* (2004).

[7] Pirolli, P. and Card, S.K. Information foraging. *Psychological Review,* 106 (1999), 643-675

[8] Thimbleby, H. User interface design with matrix algebra. *ACM Trans. on CHI*, 11(2) (2004), 181-236