Work in Progress: A Protocol for the Collection, Analysis, and Interpretation of Log Data from eHealth Technology

Floor Sieverink[1], Saskia Kelders[1], Saskia Akkersdijk[1], Mannes Poel[2],
Liseth Siemons[1], and Lisette van Gemert-Pijnen[1]

[1]Centre for eHealth and Wellbeing Research, Department of Psychology, Health and Technology, University of Twente, Enschede, The Netherlands
[2]Department of Human Media Interaction, University of Twente, Enschede, The Netherlands
{f.sieverink, s.m.kelders, s.m.akkersdijk, m.poel, l.siemons,
j.vangemert-pijnen}@utwente.nl

Abstract. Randomized controlled trials to evaluate the effectiveness of eHealth technologies provide only little understanding in why a particular outcome did occur. Log data analysis is a promising methodology to explain the found effects of eHealth technologies and to improve the effects. In this paper, we describe our experiences with the collection, analysis, and interpretation of log data from eHealth technology so far. It serves as a first step towards the development of a log data protocol to support eHealth research and will be extended and validated for different types of research questions and eHealth applications in the future.

Keywords: eHealth, evaluation, log data analysis, black box.

1 Introduction

Although persuasive eHealth technologies aim to support people changing their behaviour or attitudes, one of the main problems is that the adoption and subsequent use of such technologies remains low [1-3]. Moreover, most eHealth research is dominated by a classic conception of medical research where randomized controlled trials (RCTs) are the golden standard for measuring outcomes. This type of research focuses on the effectiveness of the technology, without divulging how and why the technology has contributed to this effectiveness. By conducting RCTs only, we do not really know why a particular outcome did occur and how the technology supports the users in healthier living, improving wellbeing, or conducting daily tasks [4, 5]. We call this the 'Black Box Phenomenon' [2, 5]. To open this Black Box, it is necessary to look for methodologies that go beyond the classic conception of effect evaluations where pretests and posttests are the standard.

The analysis of log data (anonymous records of real-time actions performed by each user) is a promising methodology to explain the found effects of a technology

[4]. This data has the potential to provide insight into the navigation process (what functionalities are used and in what order) and to look beyond just the amount of use. After all, longer exposure to an eHealth technology might be an indicator that the system fits the needs of its users, but it can also be an indication for unfocused and non-strategic use and inefficient systems [4, 6].

In our research, we therefore use log data to look among others at how users navigate through an application (what functionalities of the application are used and in what order?) at which points users drop out and when the technology is used [6-8]. This information provides input to improve the content, layout, as well as the underlying system of the application, and in turn, increase the persuasiveness of the application and its long-term usage.

In this paper, we describe our work in progress regarding the handling, analysis and interpretation of log data of eHealth technologies as a starting point for the development of a log data protocol for eHealth research.

2 The collection of log data

Depending on the research questions (information requests), there are different ways to collect log data. Google Analytics for example, can provide more information about the quantity of use by al users as a group. However, to gain more in-depth knowledge regarding the usage patterns of individual users, log data collected from the client side (actions from users are logged) or server side (requests to the server are logged) of the technology provide richer information.

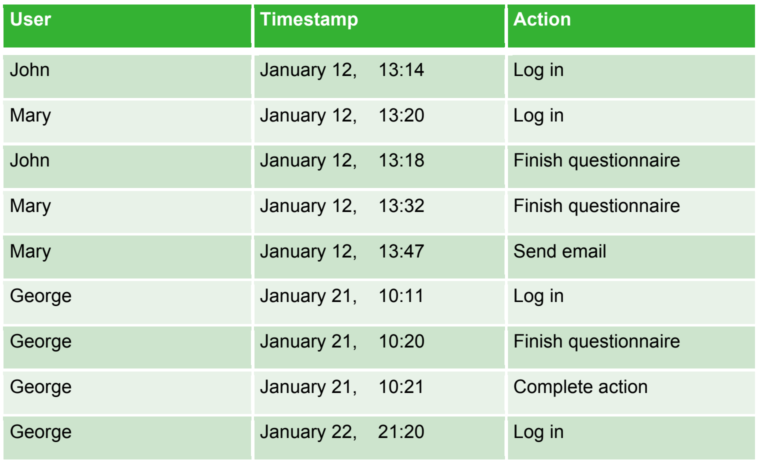
It is recommended to think about the research questions and the type of information that is needed before the application is build, since it is often less time (and money) consuming to build the logging functionality into a new application than into an existing application. Also, information that is not logged from the beginning can often not be recovered.

When log data is not collected from the beginning, it might be difficult to interpret the results, since valuable information is missing about how usage patterns developed over time. Depending on the research question, a solution might be to only follow the users that are logged from the first visit.

However, (eHealth) technologies are often updated after the first release. It is thus important to check after every system update whether the use of new and changed functionalities is logged, in order to get a complete picture of the (changes in) usage patterns.

In Figure 1, an example of a fictional log data file is given. In this data, every record (row) contains an (anonymous) user identity, a timestamp, and an identification of the action for every action of the user. To meet the information request, the data files should contain the needed information and the data should be available for analysis under the applicable privacy regulations. Furthermore, the data should be a good description of the future and be of sufficient quality, because as always, "garbage in, leads to garbage out".

Fig. 1. An example of fictional log data



User	Timestamp	Action
John	January 12, 13:14	Log in
Mary	January 12, 13:20	Log in
John	January 12, 13:18	Finish questionnaire
Mary	January 12, 13:32	Finish questionnaire
Mary	January 12, 13:47	Send email
George	January 21, 10:11	Log in
George	January 21, 10:20	Finish questionnaire
George	January 21, 10:21	Complete action
George	January 22, 21:20	Log in

3 Mathematical translation of log data

3.1 Data preparation

It is common that the log data that is needed for the analysis, consists of ten thousands of records. To handle large amounts of (unstructured) log data, we are currently developing a tool to generate data reports (e.g. number of logins, number of activities, order of the activities) that are ready to use for further analysis.

3.2 Data analysis

Once the log datasets are prepared, the files are ready for analysis. Up to now, our analyses mostly consists of counting the occurrence of different usage patterns [6-8]. However, more information can be obtained by applying machine learning algorithms, for example by transforming the data into a format that is readable for the Weka (Waikato Environment for Knowledge Analysis) tool [9] and use this tool for visualizations of the data and predictive modelling. The following methods for analysis can be used:

- Supervised learning: predicting adherence and effects from early use patterns, which enables early action for groups at risk [10].
- Unsupervised learning: what usage profiles appear from the log data and can we match those to a certain group of participants? [10]
- Markov modelling: What is the dominant path through the system? [11, 12]
- Market-basket analysis: What features are often used together? [13]

The results will provide scientific input as well as practical input for system improvements.

4 Scientific and practical translation of log data

4.1 Scientific translation

As stated in the introduction, most eHealth research is dominated by a classic conception of medical research where randomized controlled trials (RCTs) are the golden standard for measuring outcomes. However, this type of research focuses on the effectiveness of the technology, without divulging how and why the technology has contributed to this effectiveness (the 'Black Box Phenomenon') [2, 5]. The results of the log data analysis provide input for opening this black box.

However, a log data analysis does not give information about *why* certain results occur. For example, from previous research we know that almost all users of a Personal Health Record for patients with Type 2 Diabetes Mellitus drop out when they use the education service as a first step in their first session [8]. Additional research is needed to find out why: e.g. are users overwhelmed by the large amount of information, or did the information not meet the expectation of the users? Although a log data analysis in itself does not always provide a complete picture, it does provide specific targets for conducting additional research, for example via interviews, questionnaires, or usability tests.

4.2 Practical translation

Besides the scientific value, log data analysis can be of added value for designers and healthcare providers as well. For example, information about the dominant path through the system can be used as input for adapting the system design to the users and increasing the match between these two, in order to make the technology more persuasive. Also, the results of a market-basket analysis can be used to give suggestions to the users regarding their follow-up actions on the system (e.g. "You have added a goal, other users have added their current weight as well. Click here to add your weight"). This can also be incorporated to the system to give real-time feedback to the users.

Because log data analysis via (un)supervised learning can provide information about users that will probably drop out from the intervention, healthcare providers have the opportunity for intervening and stimulating users to continue using the system. Also, healthcare providers can use log data analysis to see what the effects of response time on messages are on the adherence of users to the system, making log data analysis of added value in composing protocols for (blended) care via eHealth technologies.

5 Conclusion

The analysis of log data can be of great value for scientists, designers, as well as caregivers and policy makers for opening the black box of eHealth technology. However, from the collection of log data to translating the results into valuable information, some steps need to be taken, each with their own considerations. This paper serves as a first step towards the development of a log data protocol for data collection, analysis, and interpretation to support eHealth research and will be extended and validated for different types of research questions and eHealth applications in the future.

References

1. Nijland, N., van Gemert-Pijnen, J.E., Kelders, S.M., Brandenburg, B.J., Seydel, E.R: Factors influencing the use of a Web-based application for supporting the self-care of patients with type 2 diabetes: a longitudinal study. Journal of medical Internet research 13(3) (2011)
2. Van Gemert-Pijnen, J.E.W.C., Peters, O. Ossebaard, H.C.: Improving eHealth. Eleven International Publishing, The Hague (2013)
3. Black, A.D., Car, J, Pagliari, C., Anandan, C., Cresswell, K., Bokun, T., McKinstry, B., Procter, R., Majeed, A., Sheikh, A.: The Impact of eHealth on the Quality and Safety of Health Care: A Systematic Overview. PLoS Med, 2011. 8(1): p. e1000387.
4. Han, J.Y., Transaction logfile analysis in health communication research: Challenges and opportunities. Patient Education and Counseling 82(3), 307-312 (2011)
5. Resnicow, K., Strecher, V., Couper, M., Chua, H., Little, R., Nair, V., Polk, T.A. Atenza, A.A.: Methodologic and design issues in patient-centered e-health research. American journal of preventive medicine 38(1), 98-102 (2010)
6. Kelders, S.M., Bohlmeijer, E.T., Van Gemert-Pijnen, J.E.W.C.: Participants, usage, and use patterns of a web-based intervention for the prevention of depression within a randomized controlled trial. Journal of Medical Internet Research 15(8), e172 (2013)
7. Van Gemert-Pijnen, J.E.W.C., Kelders, S.M., Bohlmeijer, E.T.: Understanding the Usage of Content in a Mental Health Intervention for Depression: An Analysis of Log Data. J Med Internet Res 16(1), e27 (2014)
8. Sieverink, F., Kelders, S.M., Braakman-Jansen, L.M., van Gemert-Pijnen, J.E.: The Added Value of Log File Analyses of the Use of a Personal Health Record for Patients With Type 2 Diabetes Mellitus: Preliminary Results. Journal of Diabetes Science and Technology, 8(2), 247-255 (2014)
9. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explor. Newsl. 11(1), 10-18 (2009)
10. Han, J., Kamber, M., Pei, J.: Data mining: concepts and techniques. Elsevier (2011)
11. Seneta, E.: Markov and the Birth of Chain Dependence Theory. International Statistical Review/Revue Internationale de Statistique 64(3), 255-263 (1996)
12. Borges, J., Levene, M.: Evaluating Variable-Length Markov Chain Models for Analysis of User Web Navigation Sessions. IEEE Transactions on Knowledge and Data Engineering,. 19(4), 441-452 (2007)
13. Anand, S.S., Patrick, A.R., Hughes, J.G., Bell, D.A.: A Data Mining methodology for cross-sales. Knowledge-Based Systems, 10(7), 449-461 (1998)