Teaching HCI Methods: Replicating a Study of Collaborative Search

Max L. Wilson
Mixed Reality Lab
University of Nottingham, UK
max.wilson@nottingham.ac.uk

This paper was submitted to RepliCHI 2013, a CHI'13 workshop

Abstract

This paper describes the challenges experienced when replicating a user study that evaluated synergy in a collaborative search system. The original paper saw significant differences in collaborative performance, depending on the mode of collaboration. We were unable to replicate the findings, but experienced several challenges that created ambiguity and differences in the methods, which may have prevented us from doing so. These challenges and experiences, and their affect on our ability to replicate the findings, are described in detail.

Author Keywords

Collaborative search, Synergy, Replication

ACM Classification Keywords

H.5.3 [Group and Organization Interfaces]: Collaborative computing.; H.3.3 [Information Search and Retrieval]: Search Process.; H.3.7 [Digital Libraries]: User Issues.

Introduction

Hands on experience of replicating an experiment is often considered a good method of teaching [2]. For this reason, a cohort of 6 MSc students were asked to replicate a user study; to learn the methodological and analytical skills required to do so. Further, we hoped to confirm the findings for the benefit of the wider community. Based

Original Task Description

A leading newspaper has hired your team to create a comprehensive report on the causes, effects, and consequences of the recent gulf oil spill. As a part of your contract, you are required to collect all the relevant information from any available online sources that you can find.

To prepare this report, search and visit any website that you want and look for specific aspects as given in the guideline below. As you find useful information, highlight and save relevant snippets. Make sure you also rate a snippet to help you in ranking them based on their quality and usefulness. Later, you can use these snippets to compile your report, no longer than 200 lines, as instructed.

Your report on this topic should address the following issues: description of how the oil spill took place, reactions by BP as well as various government and other agencies, impact on economy and life (people and animals) in the gulf, attempts to fix the leaking well and to clean the waters, long-term implications and lessons learned.

upon the interests of the staff and students involved, we chose to replicate a user study of the synergetic effect experienced by users searching in collaboration, originally carried out by Shah and Gonzalez-Ibanez [5], herein referred to as the original researchers.

The original researchers studied their own collaborative search software (Coagmento[1]), which had been evaluated previously [6], to examine synergy between collaborators in different group orientations. These orientations, as the primary independent variable, were co-located (same computer), co-located (different computers), and remotely located (different computers); individual searchers, automatically paired post hoc, were used as a baseline. The paper further contributed to the issue of evaluating synergy in collaborative search, by presenting new applicable measures. This focus on measures provided additional learning benefit to the MSc students involved.

The MSc students were given an entire semester to coordinate and run the study, and had each had to write about the results and the experience for their primary assessment. Support from the original researchers had been previously arranged by the staff.

Challenges Faced and Decisions Made

Significant challenges were faced throughout the replication attempt, from setting up the study, running the study, and analysing the results. These are described in turn below.

Setup Challenges

There were three major challenges in the setup phase: software procurement, data capture, and task design.

- Software Procurement - Initially it was considered that the procurement of software would be very easy, as Coagmento can be easily downloaded from the website. After installing the software, however, we noticed several differences in the user interface to the system described in the original paper [5]. The original researchers told us their study was based on an earlier version of the software. At first, we decided to accept the difference in functionality and to report it as a limitation later if needed. The original researchers, however, agreed to try and roll-back their functionality and provide us with a version that matched the evaluated version. This was very generous of the original researchers, and not always an option for those wishing to replicate studies.

- Data Capture - After investigating which data must be captured for the study, we discovered that the original researchers captured the data at the server level. Again, we were faced with two options: video record the desktop and manually log the necessary data afterwards, or request access to the data from the original researchers. The original researchers were again generous and agreed to provide us with the logs.

- Task Design - One significant challenge we faced was task design. The study was based upon an open-ended exploratory recall task, based upon american political parties. Our third decision was whether we should keep the american political task focus, or choose a more temporally (since the political topic had become old) and culturally relevant task for the British university. Several alternatives were proposed before making the decision, and in the end a temporally and culturally relevant task was chosen that focused on the 2012 Olympics (see original and revised task descriptions in the margins). This decision was made because task relevance and

[1] http://www.coagmento.org/

Revised Task Description

A leading newspaper has hired your team to create a comprehensive report on he causes, effects and consequences of the Olympic Games. As a part of your contract, you are required to collect all the relevant information from any available online sources that you can find.

To prepare this report, search and visit any website that you want and look for specific aspects as given in the guideline below. As you find useful information, highlight and save relevant snippets. Make sure you also rate a snippet to help you in ranking them based on their quality and usefulness. Later, you can use these snippets to compile your report, no longer than 200 lines, as instructed.

Your report on this topic should address the following issues: Impact on economy of host countries (people and animals), long-term implications on the host country, conditions and voting policy to become hosting nation and the next host country and their preparations to host the games.

inherent motivation are considered key factors in creating good work tasks for user studies [7, 1].

Running the Study

There were three major challenges in the process of running the study: the experience of the research team, the financial support for incentives, and time limitations.

- Research Team - As this replication was being used to teach new MSc students about the process of running a study, the first and most obvious challenge is that the study is being run by inexperienced researchers. This challenge was further confounded by the necessity to teach many students at once. In this case, the original study was performed by one experienced phd student, but the replication was carried out by 6 novice MSc students. Each MSc student required experience at designing study materials (like questionnaires), handling participants, and analysing the results. This means that there was likely to be a high variance in each of the stages. To reduce variance, one final protocol was selected from each of protocols submitted by the students. However, there were not many constraints, apart from a default script, in terms of how, where, and when the researchers carried out the study with their participants.
- Financial Support for Incentives - As part of a taught module, rather than a funded research project, the students had to design alternative incentive methods. In the end, they choose a prize draw for a single prize (provided by the staff), but of a value much lower than a £10 voucher for each participant. There is some related work (e.g. [4]) into the style of different incentive structures, but the effect in this case was not clear.
- Time limitations - Also driven by the taught-module based constraints, the students had a limited amount of time to perform the study. Consequently, the students had to make a decision, also relating to the financial limitations, about how many participants to include in the study. The students managed 40 participants in the timeframe, rather than the 70 involved in the original research.

Analysing the Results

There were two major challenges in the analysis phase: data processing and data analysis.

- Data Processing - The main challenge experienced in the analysis section was around the pre-processing of log data for analysis. The original researchers, for example, removed search engine result pages from their analysis of diverse website coverage, but the exact set of URLs considered as search engine results pages was implicit rather than explicit. In fact, any form of log processing and filtering in such a study would be a possible source of variance in user studies, unless the exact rules are accessible to the replicating team. One challenging example is whether to include both a user's typo and then their correction in analysing log data. In our own experiment, we created filters to achieve the same goals as reported in the paper, but we could not guarantee the exact same data would be filtered as the original research, given the same log; these elements of research methods are extremely difficult to comprehensively report in research publications.
- Data Analysis - With many methods, there are many variations on how to apply methods. In the case of this study, it was ambiguous as to how the data from the NASA Task Load Index (TLX) [3] was analysed. Many studies remove physical effort from the scale, as using a computer does not lend itself to variation in the physical

effort questions. In this case, it was unclear as to exactly how the NASA TLX was applied, including as to whether pair-wise comparisons were made.

Study Outcome and Discussion

The outcome of our replication attempt was that we could not replicate any of the original findings, as we hope may be reported in detail in a future publication. In summary, we saw no difference between the different measures, where the original researchers found a number of differences. However, there are many possible reasons for the differences, where we'll begin with the limitations of our replication attempt.

Limitations of our Replication

Although we were somewhat privileged to have the support of the original authors, we also had several limitations in our attempt:

- Researchers - our study was performed by 6 novice researchers, who each took part in running the study, with different individual abilities
- Participants - we had fewer participants (40 instead of 70), but from a similar academic population
- Participant Motivation - as part of a teaching module, participants were volunteers found by the MSc students, and were not motivated in the same way as original study
- Software - although the original researchers provided rolled-back software for the study, the process of rolling back introduced bugs that sometimes made the software unresponsive

Possible Causes of Different Findings

There are many reasons, including those listed above, that may have affected the outcome of our results, and prevented us from getting the same findings. Reflectively, its hard to estimate which element would have likely had the biggest impact on our attempt to replicate the study. First, the performance of the software, after being rolled back, was not ideal and this alone may have obstructed the synergetic effect seen by the original researchers. Second, the study was performed by several novice researchers, who may simply not have performed the study effectively. Third, the differences in the number of participants and the lack of voucher-based motivation could have limited the performance of participants. Fourth, task design has been seen to have a large affect on task outcome, and so perhaps your culturally and temporary relevant task may have not have been suitable. Finally, the processing of data for the analysis could have been simply different. Having some different or more comprehensive filtering rules may have led to significant differences in the measures.

Implications for RepliCHI

We chose to report this HCI replication, despite being focused on a user study not published at an HCI venue, because of the sheer number of issues that it highlighted for a community that wants to better support replication. Our specific example leaves many open questions that we may wish to investigate:

- What should we do when presented with different software versions from the original study?
- Should we use original tasks? Or is it acceptable to replace them for increased temporal/cultural relevance?
- Where data processing is involved, how should we best support others who wish to replicate our studies?
- If we want to recommend replication as a form of

teaching, what are the consequences of using groups of novice researchers?

- If we can't overcome these challenges, is there any value in replicating the studies?

Overall, the students experienced many challenges in trying to replicate the study, but learned a lot about study design and paper writing by doing so. For these educational reasons, the replication attempt provided a lot of value to the students. In terms of confirming the original study, we were unable to confirm the results, but were of course unable to disprove them also. This is perhaps a final challenge and discussion point for replication in HCI: we need to decide what we take away from studies that cannot replicate findings, and what value we have from understanding them. From this experience report, we hope that researchers may learn about several decisions that they may likely have to make when performing replications, and perhaps make more informed choices when the time comes.

Acknowledgements

We'd like to thank the original authors, Chirag Shah and Roberto Gonzalez-Ibanez for their support: providing software and and advice for the replication.

References

[1] Borlund, P. The concept of relevance in ir. *Journal of the American Society for information Science and Technology 54*, 10 (2003), 913–925.

[2] Frank, M. C., and Saxe, R. Teaching replication. *Perspectives on Psychological Science 7*, 6 (2012), 600–604.

[3] Hart, S. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 50, SAGE Publications (2006), 904–908.

[4] Musthag, M., Raij, A., Ganesan, D., Kumar, S., and Shiffman, S. Exploring micro-incentive strategies for participant compensation in high-burden studies. In *Proceedings of the 13th international conference on Ubiquitous computing*, ACM (2011), 435–444.

[5] Shah, C., and González-Ibáñez, R. Evaluating the synergic effect of collaboration in information seeking. In *SIGIR11: Proceedings of the 34th annual international ACM SIGIR conference on Research and development in information retrieval, July 24*, vol. 28 (2011), 24–28.

[6] Shah, C., Marchionini, G., and Kelly, D. Learning design principles for a collaborative information seeking system. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, ACM (2009), 3419–3424.

[7] Wildemuth, B., and Freund, L. Search tasks and their role in studies of search behaviors. In *Third Annual Workshop on Human Computer Interaction and Information Retrieval, Washington DC* (2009).