NewsCube Replication: Experience Report

SidharthChhabra
School of Information
University of Michigan.
Ann Arbor, MI 48109 USA
sidc@umich.edu

Paul Resnick
School of Information
University of Michigan.
Ann Arbor, MI 48109 USA
prsenick@umich.edu

Abstract

We tested the robustness of a result demonstrated by Park et al, with the NewsCube system, that presenting suggested news articles related to a single story in clusters led to more exploration of articles and clusters [1].We adjusted the apparatus to control for one potential confound in the original experiment,modified the experimental design to a within-subjects comparison to increase statistical power and allow assessment of subjective preference between the treatment and control interfaces, and switched from Korean to U.S. subjects to test generality. The results were only partially in agreement with the previous study. We reflect on the difficulty of drawing definitive conclusions when the original study and the replication differ in multiple ways. We also reflect on the challenges and value of conducting the replication as a learning exercise for a first-year doctoral student.

Author Keywords

Recommender; Diversity; Clustering; Replication; Experiment; Experience

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction

We conducted a robustness test of a previous finding [1] for two reasons. First, we thought the previous finding had important design implications. Before it entered into designers' lore, we thought its reliability and generality should be checked. Second, one of the authors of this paper (Resnick) thought that a replication study would be a good learning exercise for the other author (Chhabra) as a first-year Ph.D. student. The idea was that it would be a chance to learn by example about experiment design, and about reporting findings in a scientific way, in the context of an interface that was known to be promising, rather than waiting to learn about good evaluation methods until after having devised an innovative interface of his own.

The Previous Study

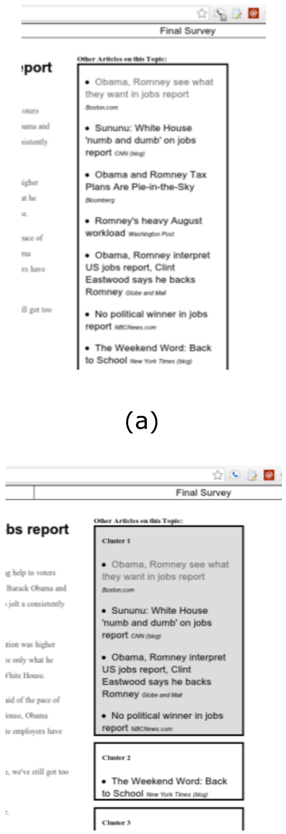
Park et al conducted a lab experiment to find if presenting people meaningful clusters, at the sub-topic level, can lead to opinion (/political) diversity in what people read[1,4]. This was an important finding because diversifying exposure is good for society [5,6] but difficult to achieve through interface design [7,8]. Park et al'sinterface recommended articles in a sidebar while the subject was reading an article. Recommended articles were grouped together into clusters based on text similarity. If the subject had read from a cluster, then that cluster was grayed to mark that it had already been explored. They compared three different presentation methods: clustered, randomly clustered (i.e., articles randomly assigned to clusters) and unclustered. They found that people read more articles and explored more clusters in the clustered presentation than the unclustered presentation. On average, people read 4 articles from 3 clusters with the meaningfully clustered interface and 2.5 articles from 1.1 (unidentified) clusters with the unclustered interface. Random clustering did not produce any statistically significant effect compared to theunclustered presentation. They concluded that clustering was effective at encouraging people to explore multiple clusters, but only when the clusters were meaningful.

Our Study

To test the robustness of the finding that presenting sub-clusters of articles leads users to read more articles and clusters, we conducted a follow-on study. We re-implemented the NewsCube clustering algorithm [1] and the experimental apparatus. We then conducted a test with U.S. rather than Korean subjects. We used English language articles and allowed them to pick among many topics from the news of the day they came to the lab. Rather than directly replicating the original study apparatus and experimental design, we attempted to improve them slightly: we tried to identify and eliminate any potential confounds; and we switched from a within-subjects to a between-subjects design, to increase statistical power and to allow for elicitation of subjective preference for clustered vs. unclustered presentation.

Challenges

The first challenge was figuring out what the original study had done, to a sufficient level of detail to permit replication. Publications in the CHI conference are not expected to be replication ready: it is not among the review criteria and strict length limitations discourage inclusion of details that are important for replication but not for casual readers. For example, there were parameters in the clustering algorithm that were set



Final Survey

port

Other Articles on this Topic:

- Obama, Romney see what they want in jobs report *Boston.com*
- Sununu: White House 'numb and dumb' on jobs report *CNN (blog)*
- Obama and Romney Tax Plans Are Pie-in-the-Sky *Bloomberg*
- Romney's heavy August workload *Washington Post*
- Obama, Romney interpret US jobs report, Clint Eastwood says he backs Romney *Globe and Mail*
- No political winner in jobs report *NBCNews.com*
- The Weekend Word: Back to School *New York Times (blog)*

voters
ama and
sistently
igher
at he
e.
pace of
ma
rs have
ill got too

(a)

Final Survey

bs report

Other Articles on this Topic:

Cluster 1

- Obama, Romney see what they want in jobs report *Boston.com*
- Sununu: White House 'numb and dumb' on jobs report *CNN (blog)*
- Obama, Romney interpret US jobs report, Clint Eastwood says he backs Romney *Globe and Mail*
- No political winner in jobs report *NBCNews.com*

Cluster 2

- The Weekend Word: Back to School *New York Times (blog)*

Cluster 3

g help to voters
Barack Obama and
jolt a consistently
tion was higher
e only what he
White House.
aid of the pace of
louse, Obama
te employers have
, we've still got too
e.

(b)

Figure 1 The sidebars for reading an article with unclustered treatment (a) and clustered treatment(b). The sidebars shows the articles suggestions with the read article(s) faded. In the clustered condition, explored clusters are grayed.

heuristically and neither the process nor the actual parameter values were documented in the original paper. It was not clear how the topics were chosen. It was not clear how many items were included in each cluster in the clustered presentation. Finally, the unclustered presentation was described as using the "Google News" interface.It was unclear whether this simply meant that it was unclustered within a news topic, as Google News was, or whether it literally used the same layout and interface elements used in Google News.

Fortunately, Souneil Park, the lead author of the previous study, was helpful; heanswered many questions and discussed and critiqued our plans. We also had access to Park's thesis (the second author of this paper was an external committee member for Park's dissertation) which provided some details not found in the CHI publication. Without the additional documentation and help from Park, we would have been farther from a direct replication than we actually were, and in ways that neither we nor the scientific community would have been able to assess.

The second challenge was eliminating confounds. It turned out that Park's unclustered treatment did literally use the Google News layout and interface, which differed in several ways from the custom interface used for the clustered treatment. First, the custom design was plain and had the feel of being done in the lab. More importantly, the unclustered treatment did not make recommendations in the sidebar while the clustered treatment did. In either condition, subjects could go back to a top-level page to select an additional article, but only in the clustered treatment could they do so without returning to the top-level page. We eliminated this potential confound by adding a sidebar to the unclustered treatment. For the clustered condition, we picked thefour most recent from articles from each cluster to display, or all of the articles for clusters with fewer than four. For the unclustered condition, we picked the same articles but did not group them into clusters, instead simply sorting them in reverse chronological order. Fig 1 shows both sidebars for a story.

Unfortunately, in retrospect we realized that we did not eliminate all the important differences between the treatments. First, after a user read any article from a cluster, the entire cluster was grayed and the actual article was faded out. For the unclustered treatment, only the actual article was faded out. The intention was to draw attention away from the already-explored clusters, but the impact may very well have been to draw extra attention to them. Second, because not all of the articles in the sidebar fit on the screen, users had to scroll to seethem. Only a few clusters were "above the fold" in the clustered condition whereas in the unclustered condition articles from more clusters may have been visible without scrolling.

We corrected a second potential confound until just before our planned first lab session. In the top-level page for a story, for each cluster there was a short snippet for first article and for others only the title. In the unclustered treatment, a snippet was shown for only the very first article, rather than one for each cluster. We presented our design to the Michigan Interactive & Social Computing group (misc.si.umich.edu). Someone pointed out that a subject could read more snippets in the clustered treatment than the unclustered treatment, which might influence how many articles they clicked on. To overcome this confound, we eliminated the top-level page altogether: the user selected a topic from the list of possible topics by selecting an article and was sent directly to that article, thereafter using only the sidebar to select additional articles.

	Clustered	Unclustered	t-test
#Articles read	3.3	2.4	t(39)=3.0 p=0.003
#Clusters explored	1.7	1.5	t(39)=1.8p=0.24
Time spent	227 secs	183 secs	t(39)=2.1p=0.04.

Table 1.Reading results.

The switch to a within-subjects design also created a new potential confound: order and contamination effects. For example a plausible order effect would be that users would explore more articles and aspects on the first topic that they read about than the second, simply because the first topic was more interesting. To control for that confound, we counter-balanced the order: subjects picked whatever topics they wanted to read, but some subjects got the clustered treatment for their first topic and some for the second. A contamination effect would occur if subjects' usage of say, the unclustered treatment, was different depending on whether they had already experienced the clustered treatment. This is always a risk in a within-subjects design, and we did not rule out this potential confound.

A third challenge was power calculation, estimating how many subjects were needed in order to have, say, a 90% chance of finding a statistically significant difference between the two conditions (at the .05 significance level), given an assumption about the size of the true difference in outcomes, the effect size. In principle, for a replication, the previously detected effect size and the observed variances in outcomes should have provided the needed inputs for a power calculation. In practice, however, since out design was within-subjects and the original was between-subjects, we had to resort to the same kind of guesswork that is usually done in power estimations. We ended up recruiting 40 subjects, 20 in each experimental condition.

Results

Table 1 presents the key findings. Subjects read more articles and spent more time using clustered than the unclustered presentation. This confirmed the generality of one part of the previous finding, with testing across many more sets of stories, with articles in English rather than Korean. Subjects also preferred the clustered presentation, though not overwhelmingly so (26-14).

However, subjects did not explore significantly more aspects of a story in the clustered than the unclustered treatment. By aspects, we mean clusters produced by the NewsCube algorithm. Thus, our results were not consistent the most important finding of the original paper.

Publishing a Replication

We wrote a full paper about the replication, with more details about the apparatus, results, and limitations described above and submitted it to CHI 2013. It received several high-quality, thoughtful reviews, none of which recommended it for publication. Reviewers picked up on the potential remaining confounds that we reported (graying and scrolling). More generally, given that our results were not fully consonant with the original findings, and that we had changed many things, from language and subject pool to specifics of the interface, we could not make a firm conclusion about the correctness or generalizability of the main finding. Reviewersargued that in order to make a real contribution, further work is needed (they had different suggestions about what further work), and that we should really consider this a work in progress rather than a completed study. We found this argument persuasive, submitted a work-in-progress paper/poster for CHI 2013, and have plans for follow-up studies to yield a more conclusive result.

One reviewalso argued, essentially, that neither the original study nor ours has yet demonstrated that the original finding was replication-worthy. We framed it as replication-worthy because of the prospect thatusing clustered presentation couldnudge people towards exposure to diverse viewpoints, which is a valuable social goal. In fact, however, the clusters of articles represented different textual aspects of a news topic (i.e., clustering was based on text similarity), which might not necessarily represent different viewpoints. In future work, either an argument needs to be made that it's valuable to nudge people towards exploring multiple textual aspects of a story, orwe will need to demonstrate that clustering on text similarity naturally leads to clustering on viewpoint similarity.

Discussion

Replication is animportant ideal guiding the advance of science in any field. However, CHI papers are not yet ready for it. There is no expectation and no space in CHI publications for reporting sufficient detail to permit replication. Calculating heuristic parameters, and providing test databases and experiment observations are a few of such details. Perhaps, a norm and mechanism for published supplements providing fuller details would be good.

We also suspect that if our results had confirmed the original findings, a report of the study also would not have been accepted, because it would not have been novel enough. Thus, for replication work to make a publication-worthy contribution in this field, it either needs to replicate and extend it, or it needs to show non-replicability and identify exactly why the original result did not replicate. This provides a limited incentive for any researcher to replicateand check a previous work. It's certainly not an easy path to a first publication for a student.

A replication study is still a valuable educational exercise for a first-year PhD student, and, if followed up, can yield a real contribution to accumulating generalizable knowledge. Throughthisreplication, the first author gained a few lessons which he would not have learned otherwise. Hegained an understanding of how to write research that can be replicated, providing every detail such that anyone can walk on the same path and conduct a similar experiment. It also taught him the importance of making available test cases and experiment data so that if in the future somebody wants to extend his work, he/she would be able to re-implement with confidence. Finally, we teach our PhD students about biases and threats toresearch validity. These concerns were driven home, however, to both student and advisor, when even after scrutinizing a previous study design for the better part of a year there were still potential confounds we identified after the fact in our own design.

References

[1] Park, S., et al., *NewsCube: delivering multiple aspects of news to mitigate media bias*, in *Proceedings of the 27th international conference on Human factors in computing systems*2009, ACM: Boston, MA, USA. p. 443-452.