Datenqualität bei Online-Fragebögen sicherstellen

Praktische Techniken zum Auffinden nicht ernsthaft ausgefüllter Fragebögen

Martin Schrepp

SAP SE

Zusammenfassung

Über Online-Fragebögen können Daten zur User Experience eines Produkts mit geringem Aufwand erhoben werden. Allerdings hat man hier in Bezug auf die Qualität der Daten wenig Kontrollmöglichkeiten. Insbesondere wenn die Teilnehmer mit einer Belohnung zur Teilnahme an einer Befragung motiviert werden, besteht die Gefahr, dass viele Fragebögen nicht sorgfältig ausgefüllt sind. Dies kann die Ergebnisse und die daraus abgeleiteten Interpretationen zur Produktqualität verfälschen. Wir beschreiben zwei Strategien, um diesem Problem zu begegnen. Die erste Strategie nutzt die Skalenstruktur des Fragebogens, um inkonsistente Antworten zu identifizieren. Treten in den Antworten einer Person zu viele Inkonsistenzen auf, ist dies ein Indikator für mehr oder weniger zufälliges Antwortverhalten. Die zweite Strategie vergleicht die Antwortzeiten einer Person für die einzelnen Items mit den aus einer kognitiven Modellierung geschätzten Zeiten. Hier sind viele zu schnelle Antworten ein Indikator für eine nicht sorgfältige Beantwortung.

1 Einleitung

Fragebögen sind eine einfache und kostengünstige Möglichkeit, die User Experience eines Produkts zu messen. Allerdings kann man nicht davon ausgehen, dass alle Teilnehmer die Fragen auch immer sorgfältig beantworten. Diese Gefahr besteht insbesondere dann, wenn ein Online-Fragebogen eingesetzt wird und die Teilnehmer durch eine Belohnung zur Teilnahme motiviert werden, z.B. wenn unter allen Teilnehmern ein Preis verlost wird. Hier steht manchmal ausschließlich die Teilnahme an der Verlosung im Fokus des Teilnehmers und die Beantwortung der Fragen wird einfach ohne großes Nachdenken schnell erledigt. Solche mehr oder weniger zufälligen Antworten enthalten natürlich keinen oder nur einen sehr geringen Informationsgehalt über die User Experience des Produkts und können die Ergebnisse verfälschen.

Veröffentlicht durch die Gesellschaft für Informatik e.V. und die German UPA e.V. 2016 in S. Hess & H. Fischer (Hrsg.): Mensch und Computer 2016 – Usability Professionals, 4. - 7. September 2016, Aachen.

http://dx.doi.org/10.18420/muc2016-up-0015

Wie kann man solche problematischen Antworten einfach finden und aus den Ergebnissen entfernen? Wir beschreiben am Beispiel des User Experience Questionnaire UEQ (Laugwitz et al. 2006 bzw. 2008) zwei Strategien, mit denen man dieses Problem lösen kann. Diese Strategien sind aber auf andere Fragebögen dieses Typs leicht übertragbar, d.h. sind nicht auf den UEQ beschränkt. Eine allgemeine Übersicht verschiedener Ansätze dieses Problem im Rahmen von psychologischen Online-Studien zu lösen, findet sich z.B. in Aust et al. 2013.

Das Problem nicht sorgfältig ausgefüllter Fragebögen ist natürlich besonders bei Online-Untersuchungen relevant, da man hier wenig Kontrolle über die Auswahl der Teilnehmer und deren tatsächliches Antwortverhalten hat. Aber auch wenn die Teilnehmer die Fragebögen in Papierform ausfüllen, kommen solche Verhaltensweisen vor. D.h. das Problem ist zwar bei Online-Fragebögen besonders relevant, aber auch bei Befragungen in Papierform kann eine nachträgliche Kontrolle der Daten durchaus Sinn machen.

2 Interne Skalenstruktur

Bei den meisten Fragebögen zur User Experience sind die Items in Skalen gruppiert. D.h. mehrere Items sind jeweils einer Skala zugeordnet, die in gewisser Weise deren Gemeinsamkeiten beschreibt. Die Items einer Skala messen damit zumindest ähnliche Aspekte der User Experience eines Produkts. Wenn die Items einer Skala also von einer Person sehr unterschiedlich beantwortet werden, kann dies auf eine eher zufällige Beantwortung der Items hindeuten.

Die 26 Items des UEQ sind den 6 Skalen *Attraktivität* (6 Items), *Effizienz*, *Durchschaubarkeit*, *Steuerbarkeit*, *Stimulation* und *Originalität* (jeweils 4 Items) zugeordnet.

Pro Skala sind zwei Items positiv und zwei Items negativ gepolt (bei der Skala Attraktivität jeweils 3). Wenn ein Teilnehmer die Items nur oberflächlich oder völlig zufällig beantwortet, ist es recht wahrscheinlich, dass er oder sie für eine Skala widersprüchliche Antworten gibt. Betrachten wir als Beispiel die folgenden Antworten einer Person auf die Items der UEQ-Skala *Durchschaubarkeit*:

unverständlich	o o o o o x o	verständlich
leicht zu lernen	o o o o o o x	schwer zu lernen
kompliziert	o o o o o x o	einfach
übersichtlich	o o o o o x o	verwirrend

Offenbar sind diese Antworten nicht sehr konsistent. Das Produkt wird als sehr verständlich, schwer zu lernen, sehr einfach und sehr verwirrend eingeschätzt. D.h. hier bestehen in der Einschätzung des Produkts durch den Teilnehmer offensichtliche Inkonsistenzen.

Bringt man alle Items in der Reihenfolge negativ (1) nach positiv (7), so sieht man, dass die Bewertungen von 1 (schwer zu lernen/leicht zu lernen) bis 6 (unverständlich/verständlich) reichen, d.h. die Distanz zwischen der besten und der schlechtesten Bewertung beträgt 5. Wir verwenden diese Distanz zwischen der besten und der schlechtesten Bewertung eines Items

innerhalb einer Skala als Indikator, dass bei der Beantwortung der Items der Skala Inkonsistenzen vorliegen.

Wenn man davon ausgeht, dass alle Items einer Skala die gleichen Eigenschaften messen, sollte ein Teilnehmer für die Items auf einer Skala keine zu unterschiedlichen Bewertungen abgeben. D.h. die Distanz zwischen der besten und der schlechtesten Bewertung eines Items der Skala sollte nicht zu groß sein. Natürlich kann es immer vorkommen, dass ein Teilnehmer einen Begriff falsch interpretiert oder beim Ankreuzen die falsche Kategorie markiert, d.h. bei einer Abweichung dieser Art wird man nicht gleich die gesamte Antwort verwerfen wollen. Beobachtet man eine solche Abweichung aber für mehrere Skalen, ist dies ein deutlicher Hinweis darauf, dass der Teilnehmer die Items nicht sorgfältig gelesen hat.

Die Grundidee ist es, Antworten auszuschließen, die zu viele Inkonsistenzen zur angenommenen Skalenstruktur des UEQ aufweisen. Dabei muss man einerseits sicherstellen, dass man wirklich zufällige Antworten von Teilnehmern sicher als solche identifiziert und andererseits nicht unnötig viele Antworten verwirft.

Wir untersuchen im Folgenden 4 Heuristiken nicht sorgfältig ausgefüllte UEQ Fragebögen zu identifizieren. Mit *D* bezeichnen wir dabei die maximale Distanz zwischen der besten und der schlechtesten Bewertung für ein Item einer Skala.

- Heuristik 1 (H1): $D > 2$ für mindestens 2 Skalen des UEQ
- Heuristik 2 (H2): $D > 2$ für mindestens 3 Skalen des UEQ
- Heuristik 3 (H3): $D > 3$ für mindestens 2 Skalen des UEQ
- Heuristik 4 (H4): $D > 3$ für mindestens 3 Skalen des UEQ

Eine Antwort wird zum Beispiel bei Heuristik 2 als zufällig erkannt und eliminiert, wenn für mindestens drei Skalen eine Distanz von mehr als 2 zwischen der besten und schlechtesten Bewertung eines Items der Skala auftritt.

Eine gute Heuristik muss einerseits zufällig ausgefüllte Fragebögen mit hoher Sicherheit als solche erkennen und andererseits eine Antwort nicht schon bei einigen wenigen Antwortfehlern (falsche Kategorie angekreuzt, ein Item falsch interpretiert) oder Abweichungen zur angenommenen Skalenstruktur als zufällig klassifizieren.

Zur Beantwortung der ersten Frage kann man berechnen[1], wie groß die Wahrscheinlichkeit ist, dass die Heuristik einen zufällig ausgefüllten Fragebogen (alle Antworten der 26 Fragen mit Gleichverteilung auf den Antwortkategorien ausgefüllt) als solchen erkennt. Hier ergaben sich (für die Skalen mit 4 Items) folgende Werte: H1(99,98%), H2 (99,66%), H3 (99,06%) und H4 (93,3%). D.h. wirklich zufällige Antworten werden von allen Heuristiken mit hoher Sicherheit als solche identifiziert.

[1] Man kann diese Wahrscheinlichkeiten kombinatorisch berechnen. Da es in diesem Fall aber nicht um die exakten Werte, sondern um die Bestimmung der Größenordnung dieser Wahrscheinlichkeiten geht (und die Berechnung nicht so ganzeinfach ist), kann man die Werte auch einfach aus einer großen Menge zufällig generierter Antwortmuster schätzen.

Um die zweite Frage zu beantworten, wurden 190 Fragebögen untersucht, bei denen der UEQ im Anschluss an einen Usability Test ausgefüllt wurde. Die Teilnehmer bearbeiteten hier Aufgaben mit einem Produkt und wurden dabei von einem Moderator durch den Test geführt. Nach der Bearbeitung der letzten Aufgabe füllten die Teilnehmer im Beisein des Moderators eine Papierversion des UEQ aus. Natürlich können auch hier einige Teilnehmer beim Ausfüllen des Fragebogens nicht mit der notwendigen Sorgfalt vorgegangen sein, deren Zahl sollte aber wegen der Anwesenheit des Moderators (und da das Ausfüllen des UEQ als integraler Bestandteil des Usability Test angesehen wurde) eher gering sein.

Hier ergaben sich folgende Häufigkeiten als zufällig erkannter Fragebögen: H1(25,26%), H2 (8,42%), H3 (6,32%) und H4 (1,58%). Offenbar liefert H1 hier unrealistisch viele Treffer.

Nimmt man die hohen Erkennungsraten zufälliger Antworten und die Ergebnisse aus den Usability Tests zusammen, so sind offenbar Heuristik 2 und Heuristik 3 brauchbare Strategien, um nach zufälligen Antwortmustern zu suchen. Tendenziell scheint Heuristik 3 etwas besser geeignet zu sein (liefert bei den Tests etwas weniger Treffer, wo auch wenige zu erwarten sind), aber aufgrund der aktuellen Datenlage kann hier noch keine eindeutige Empfehlung abgeleitet werden. In jedem Fall sind die Unterschiede zwischen diesen beiden Heuristiken gering, so dass man im Prinzip mit beiden erfolgreich arbeiten kann. Für den UEQ wurde die Heuristik 3 im Datenanalyse Tool implementiert und steht damit bei Auswertungen direkt zur Verfügung.

Mit wie vielen nicht sorgfältig ausgefüllten Fragebögen kann man in einer Online-Untersuchung rechnen? Hierzu wurde ein großer Datensatz (Ilmberger et al., 2009) untersucht, bei dem 722 Teilnehmer den UEQ Online ausgefüllt hatten (als Motivation für die Teilnahme wurde die Verlosung eines Geldpreises unter allen Teilnehmern durchgeführt). Die Heuristik 3 identifiziert hier 106 der 722 Fragebögen (14,68%) als nicht sorgfältig ausgefüllt. Dieses Beispiel zeigt, dass hier in der Tat ein erhebliches Potential zur Verbesserung der Datenqualität besteht, d.h. es durchaus sinnvoll ist, die eingehenden Datensätze vor der Auswertung einer Prüfung zu unterziehen. Eine Untersuchung weiterer kleiner Online-Untersuchungen lieferte ebenfalls Werte in dieser Größenordnung.

3 Einsatz kognitiver Modellierung

Versucht ein Teilnehmer einen Fragebogen ohne großen Aufwand auszufüllen, z.B. um an einer Verlosung teilzunehmen, wird er oder sie erheblich schneller sein als ein sorgfältig antwortender Teilnehmer. Daher bietet die Messung der Antwortzeiten einen Ansatzpunkt Teilnehmer zu identifizieren, die die Fragen nicht sorgfältig bearbeitet haben, d.h. wesentlich schneller auf die Items geantwortet haben, als das zu erwarten ist.

Wie lange ein Teilnehmer mindestens brauchen sollte, um bei ernsthafter Bearbeitung eine Frage zu beantworten, kann man über eine kognitive Modellierung schätzen. Methoden der kognitiven Modellierung, z.B. GOMS (Card et al. 1983) oder CogTool (John et al. 2004), nutzen Modelle der menschlichen Informationsverarbeitung, um die bei der Abarbeitung einer Aufgabe erforderlichen kognitiven und motorischen Schritte zu beschreiben und daraus dann

eine Zeitdauer für die Bearbeitung einer Aufgabe (in unserem Fall die Beantwortung eines UEQ Items) abzuleiten.

Die Grundidee einer GOMS-Analyse (Card et al. 1983) ist es, die Benutzerinteraktion bei der Bearbeitung einer Aufgabe in elementare Operatoren zu zerlegen. Die Bearbeitungszeit der Aufgabe wird dann aus den bekannten Zeiten dieser elementaren Operatoren geschätzt. Operatoren sind grundlegende physische (z.B. Drücken einer Taste oder ein Mausklick) oder kognitive Prozesse (z.B. Abruf einer Information aus dem Gedächtnis oder mentale Vorbereitung für den nächsten Schritt in einer Handlungssequenz), die der Benutzer zur Erreichung des Ziels ausführen muss. Für eine elementare Einführung in die kognitive Modellierung siehe Schrepp & Held (2015).

Unterschiedliche Personen benötigen natürlich unterschiedliche Zeiten für die grundlegenden physischen oder kognitiven Operationen. Die GOMS Analyse abstrahiert von den Zeiten konkreter Personen durch die Verwendung typischer Durchschnittswerte (z.B. Tastendruck beim Tippen einer Zeichenkette 0,23 Sekunden, Positionieren des Mauszeigers 0,44 Sekunden, Mentale Vorbereitung 1,2 Sekunden, etc.). Diese Durchschnittswerte wurden in experimentellen Studien ermittelt (z.B. John & Kieras 1996; Olson & Olson 1990; Schrepp & Fischer 2007). In KLMGOMS wird zwischen verschiedenen physischen Operatoren unterschieden. Für alle kognitiven Operationen wird nur ein einziger Operator verwendet.

CogTool ist eine Software, die die kognitive Modellierung von computerbasierten Aufgaben ermöglicht (John et al. 2004, John & Salvucci 2005). Die Software wurde an der Carnegie Mellon Universität entwickelt und ist verfügbar unter cogtool.com. Letztlich realisiert CogTool die KLMGOMS Methode, wobei einige zusätzliche Operatoren und eine gewisse (wenn auch eingeschränkte) Berücksichtigung paralleler Informationsverarbeitung zur Verfügung stehen.

Wir verwenden CogTool nun zur Schätzung der zu erwartenden Antwortzeit auf ein Items des UEQ. Die Items des UEQ sind Gegensatzpaare, mit einer 7-stufigen Antwortskala, z.B.:

Gut o o o o o o o Schlecht

Für die Beantwortung dieses Items muss der Nutzer:

- seine visuelle Aufmerksamkeit auf den linken Begriff fokussieren (Look at),
- die Bedeutung des Begriffs erfassen (Think),
- seine visuelle Aufmerksamkeit auf den rechten Begriff fokussieren (Look at),
- die Bedeutung dieses Begriffs erfassen (Think),
- eine Entscheidung treffen welcher Begriff besser passt (Think),
- die Maus zum entsprechenden Radio-Button bewegen (wobei man annehmen kann, dass die Hand bereits auf der Maus ist)
- und einen Mausklick durchführen.

Die Begriffe in Klammern bezeichnen dabei die entsprechenden Operatoren in Cogtool (John et al., 2004). Eine Schätzung der notwendigen Bearbeitungszeit mit CogTool ergibt eine Gesamtzeit von 5,6 Sekunden pro Item (unter der Annahme, dass zumindest die beiden Begriffe gelesen und eine Entscheidung getroffen wird). Legt man das für den gesamten

Fragebogen zugrunde, ergibt sich eine Gesamtbearbeitungszeit von 145,6 Sekunden für die 26 Items. In gewisser Weise ist dies die mindestens zu erwartende Zeit. In der Realität wird die Zeit etwas höher liegen, da Teilnehmer bei einigen der Begriffe noch über deren genaue Interpretation im Kontext der Untersuchung nachdenken und daher in diesen Fällen noch zusätzliche Zeit verbrauchen. Im praktischen Einsatz beobachtet man Zeiten im Bereich von 3 Minuten.

Klickt der Teilnehmer dagegen nur schnell durch den Fragebogen, so entfallen zumindest die 3 oben erwähnten Think-Operatoren und die Zeit pro Item sollte eher 2 Sekunden liegen (d.h. hier wird im Prinzip nur pro Item eine Antwortkategorie willkürlich gewählt (Think) und geklickt (Mouse Move und Click). D.h. bei einem rein zufälligen Durchklicken ist eher mit einer Gesamtzeit von 52 Sekunden zu rechnen.

Die Idee ist also im Online-Fragebogen zu messen, wie lange der Teilnehmer für die Bearbeitung der Fragen braucht, d.h. die Zeiten zwischen zwei Klicks zu messen. Sind diese für mehrere Fragen deutlich kürzer, als der durch die kognitive Modellierung ermittelte Wert, ist dies ein klarer Hinweis darauf, dass Items nicht sorgfältig bearbeitet wurden.

4 Zusammenfassung

Wir haben zwei Methoden vorgeschlagen, um nicht sorgfältig ausgefüllte Fragebögen zu identifizieren.

Als erste Methode haben wir mehrere Heuristiken untersucht, die die Skalenstruktur des UEQ nutzen, um problematische Antwortmuster zu entdecken. Zwei der untersuchten Heuristiken erlauben es, wirklich zufällig erzeugte Antwortmuster sicher als solche zu erkennen und liefern in Situationen, in denen man von einer sorgfältigen Bearbeitung der Items ausgehen kann, auch nur wenige Treffer, d.h. die Wahrscheinlichkeit von Fehlalarmen ist offenbar gering. Welche dieser beiden Heuristiken besser geeignet ist, muss in Folgeuntersuchungen noch geklärt werden.

Diese Methode ist von der Art der Datenerfassung unabhängig, d.h. kann auch verwendet werden, wenn die Daten über ein Papierformular erfasst werden.

Die zweite Methode misst die Antwortzeiten der Items und vergleicht diese mit theoretisch geschätzten Antwortzeiten aus einer kognitiven Modellierung der Aufgabenbearbeitung. Diese Methode ist theoretisch gut abgesichert, muss für den UEQ aber noch praktisch erprobt werden. Diese Methode ist natürlich nur anwendbar, wenn die Daten über eine Online-Studie oder eine andere Art der Datenerfassung erhoben werden, die die exakte Messung von Antwortzeiten erlaubt.

Eine interessante Frage, die in Zukunft noch untersucht werden soll, ist das Zusammenspiel der beiden beschriebenen Methoden. Wenn für einen Datensatz gemessene Antwortzeiten vorliegen, kann man untersuchen, ob beide Methoden die gleichen Antworten als problematisch einstufen. Durch einen Einsatz beider Methoden kann vermutlich die Aufdeckungsquote problematischer Antworten noch etwas erhöht werden.

Kontaktinformationen

Martin Schrepp: martin.schrepp@sap.com

Literaturverzeichnis

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. Behavior research methods, 45(2), 527-535.

Card, S., Moran T.P. & Newel A. (1983). The Psychology of Human Computer Interaction. Mahwah:Lawrence Erlbaum Associates.

Ilmberger, W.; Theo Held, T.; Schrepp, M. (2009). *Was macht studiVZ attraktiv?* In: H. Wandke; S. Kain & D. Struve (Eds.): Mensch & Computer 2009. Oldenbourg Verlag, S. 323-332.

John, B.E. & Kieras, D.E. (1996): The GOMS family of user interface analysis techniques: Comparison and Contrast. ACM Transactions on Computer-Human Interaction 3(4), S. 320-351.

John, B., Prevas, K., Salvucci, D. & Koedinger, K. (2004) Predictive Human Performance Modeling Made Easy. In Dykstra-Erickson, E. & Tscheligi, M. (Hrsg.), Proceedings of CHI Conference on Human Factors in Computing Systems. New York: ACM Press. S. 455 – 462.

John, B. E. & Salvucci, D. D. (2005). Multi-Purpose Prototypes for Assessing User Interfaces in Pervasive Computing Systems. IEEE Pervasive Computing 4(4), S. 27-34.

Laugwitz, B.; Schrepp, M. & Held, T. (2006). *Konstruktion eines Fragebogens zur Messung der User Experience von Softwareprodukten.* In: A.M. Heinecke & H. Paul (Eds.): Mensch & Computer 2006 – Mensch und Computer im Strukturwandel. Oldenbourg Verlag, S. 125 – 134.

Laugwitz, B., Schrepp, M. & Held, T. (2008). *Construction and evaluation of a user experience questionnaire.* In: Holzinger, A. (Ed.): USAB 2008, LNCS 5298, S. 63-76.

Olson, J.R. & Olson, G.M. (1990): The growth of cognitive modelling in human-computer interactions since GOMS. Human-Computer Interaction, 5, S. 221-265.

Schrepp, M. & Fischer, P. (2007). GOMS models to evaluate the efficiency of keyboard navigation in web units. Eminds – International Journal of Human Computer Interaction 1(2), S. 33-46.

Schrepp, M. & Held, T. (2015). Wie effizient ist mein User Interface? - Bearbeitungszeiten mit GOMS und CogTool schätzen. In: Endmann, A.; Fischer, H. & Krökel, M. (Eds.), Mensch und Computer 2015 – Usability Professionals, S. 393-400, DE GRUYTER 2015.

Autoren

Schrepp, Martin

Dr. Martin Schrepp studierte Mathematik und Psychologie an der Universität Heidelberg. 1990 Abschluss als Diplom-Mathematiker. 1990 – 1993 Promotion in Psychologie. Seit 1994 bei der SAP AG tätig. Bisherige Tätigkeitsfelder waren hier die Konzeption technischer Dokumentation, Software-Entwicklung, User Interface Design und Barrierefreiheit. Hauptinteressen sind die Anwendung kognitionswissenschaftlicher Erkenntnisse auf das Design interaktiver Anwendungen, Barrierefreiheit und die Entwicklung von Methoden zur Evaluation und Datenanalyse.