Automatisierte Erfassung von Nutzungsdaten mobiler Apps zur Verbesserung der App-Qualität – Ein Erfahrungsbericht

Frank Elberzhager[1], Britta Karn[2], Simon André Scherr[1], Thomas Immich[2]
[1]Fraunhofer IESE, Fraunhofer-Platz 1, 67663 Kaiserslautern
{frank.elberzhager, simon.scherr}@iese.fraunhofer.de
[2]Centigrade GmbH, Science Park 2, 66123 Saarbrücken
{britta.karn, thomas.immich}@centigrade.de

Zusammenfassung:
Nutzerfeedback gewinnt zunehmend an Bedeutung im Rahmen der App-Entwicklung. Entwickler können damit schnell erfassen, was Nutzer über die eigene App denken, wo Qualitätsprobleme liegen, und welche neuen Funktionen gewünscht werden. Im Rahmen des Opti4Apps Projekts wurde ein Prozess zur systematischen Nutzung unterschiedlichen Feedbacks in agilen Prozessen entwickelt und im Rahmen einer Studie evaluiert. In diesem Beitrag möchten wir Erkenntnisse aus der Studie zur automatisierten Erhebung von Nutzungsfeedback darstellen und aufzeigen, wie Verbesserungspotential aus dem erfassten Nutzerfeedback abgeleitet werden konnte.

Schlüsselworte: Qualitätssicherung, Mobile Apps, Nutzerfeedback, Experiment

1. Einleitung und Motivation

Entwickler mobiler Anwendungen kommen heute nicht mehr um die Berücksichtigung von Nutzerfeedback herum. Auf unterschiedlichen Kanälen bewerten Nutzer Apps und teilen mit, was ihnen gefällt und was sie stört. Insbesondere negatives Feedback, beispielsweise als geschriebenes Review oder auch als geringe Sternebewertung, kann schnell dazu führen, dass die betroffene App von anderen Nutzern nicht genutzt wird. Somit ist es zunehmend unerlässlich, solches Feedback einzusammeln, auszuwerten und Erkenntnisse in die eigenen Entwicklungsaktivitäten zurück fließen zu lassen. Neben dem schnellen Reagieren auf Probleme ist es aber auch notwendig, auf neue Trends und auf in den Markt eintretende Konkurrenz zu achten. Das führt zu der Anforderung, permanent die vom Nutzer gewünschte Qualität einzuhalten, da es ansonsten wahrscheinlich ist, dass die eigene App nicht mehr genutzt wird. Wir plädieren deswegen für einen Qualitätssicherungsansatz, der Nutzer enger einbindet und ihr Feedback stärker berücksichtigt. Dies ermöglicht nicht nur das Auftreten von Fehlern zu minimieren, sondern auch eine gute User Experience zu bieten und neue Nutzerwünsche umzusetzen.

2. Hintergrund

Ein Prozess, der Nutzerfeedback in der eigenen Entwicklung und der Qualitätssicherung berücksichtigt, muss sowohl effektiv als auch effizient sein, damit er in der Praxis eingesetzt wird.

In [1] haben wir unsere Vision der automatischen Einbindung von Nutzerfeedback, welcher im Rahmen von Opti4Apps entstanden ist, vorgestellt. Unser Fokus lag hierbei auf theoretischen Grundlagen sowie der Klassifikation von Feedback. Abbildung 1 stellt den Ablauf des Entwicklungsprozesses unter Verwendung des Ansatzes schematisch dar. Nutzerverhalten (automatisch erfasstes Feedback) und Nutzerfeedback (textuelles Feedback) werden im Opti4Apps Ansatz gesammelt, ausgewertet und Verbesserungspotential abgeleitet, welche einer Managerrolle zur Verfügung gestellt werden und welcher entscheidet, wie diese in die Weiterentwicklung fließen. Ergebnis ist dann eine weiterentwickelte und verbesserte App.

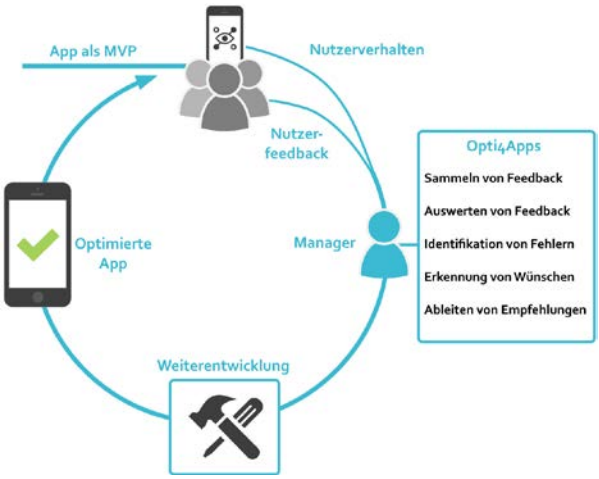

Abbildung 1: Opti4Apps im Entwicklungsprozess

Grundsätzlich gibt es zwei Hauptarten von Feedback. Zum einen gibt es *textuelles Feedback*, welches durch Datenkollektoren aus zahlreichen unterschiedlichen Datenquellen wie App Stores, sozialen Medien, Foren und

Supportbereichen gesammelt werden kann. Einen Entwicklungsprozess, der sowohl automatisch erhobenes als auch textuelles Feedback für die Qualitätssicherung in die Rollen einer Entwicklungsorganisation bringt, haben wir in [2] präsentiert. Bei der Ausgestaltung des Ansatzes lag unser Fokus bisher auf textuellem Nutzerfeedback.

Zum anderen gibt es automatisch erhobenes Feedback, bei dem ein Software-Agent in der App *Nutzungsdaten* sammelt, also das Nutzungsverhalten anhand verschiedener Sensoren aufzeichnet. Diesen Aspekt wollen wir in diesem Beitrag in Abschnitt 3 näher beleuchten.

Anspruch von Opti4Apps ist es, eine leichtgewichtige Integration von Feedback insgesamt in den Entwicklungsprozess zu realisieren. Die dadurch erfassten Daten werden analysiert, was einerseits Ideen für neue Features gibt, andererseits auch Fehler aufzeigt. Ziel ist es, Unterstützung für Entwicklungsorganisationen zu bieten und hierbei vor allem Produktmanagern in Form eines Web-basierten Dashboards die Möglichkeit zu geben, fundierte Entscheidungen zur Qualitätssteigerung der eigenen Produkte zu fällen.

3. Studie

Für den Bereich des automatisch erhobenen Feedbacks stellt sich die Frage, wie gut es möglich ist, Aussagen über die Qualität eines Produktes nur anhand von gesammelten Nutzungsdaten zu machen. Wir haben daher in der vorliegenden Arbeit evaluiert, wie stark die Aussagekraft des agentenbasierten Feedbacks ist. Dazu haben wir eine prototypische Anwendung mit einem Agenten ausgestattet, die App mit Nutzern evaluiert und dabei Videoaufzeichnungen der Personen und der Interaktionen gemacht (Vorstudie). Hieraus konnten wir wertvolle Handlungsempfehlung zur Produktverbesserung ableiten. Daneben haben wir Probanden die App ohne Videoaufzeichnung benutzen lassen (Hauptstudie). Dieser Studie (bestehend aus Vor- und Hauptstudie) lag die folgende Forschungsfrage zugrunde: Ist automatisch erhobenes Feedback alleine in der Lage, Empfehlungen zur Produktverbesserung in gleichem Maße zu ermöglichen wie eine traditionelle Nutzerbeobachtung?

Bei der prototypischen Anwendung handelte es sich um die sogenannte „Timely App“, eine Zeiterfassungsapplikation für Hochschulmitarbeiter. Damit ist es möglich, Arbeitszeiten auf dem Mobilgerät zu erfassen sowie Krankheiten und Urlaubszeiten mitzuteilen.

Im Vorfeld der Studie wurden elf User Stories definiert, die aus den Nutzerbedürfnissen der Hochschulmitarbeitern hinsichtlich der Zeiterfassung mit Hilfe des Continuous UX [3] Methodenbaukastens extrahiert wurden. Weiterhin wurden KPIs zur Messung der Güte der Umsetzung von User Stories festgelegt, wie beispielsweise „time on task“ für die Durchführung einer Aufgabe, mit deren Hilfe die Daten später analysiert werden sollten.

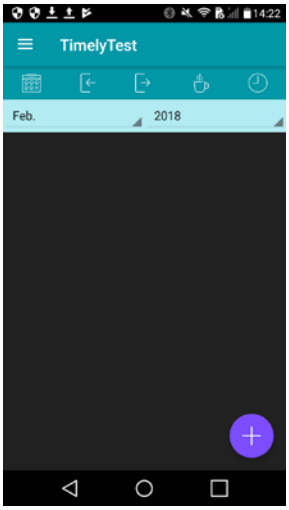

Abbildung 2: Startseite der Timely App

3.1 Vorstudie

Die erste Studie wurde mit vier wissenschaftlichen Mitarbeitern der Hochschule Heilbronn durchgeführt. Hierbei bekamen die Probanden ein Handy mit der Timely App zur Verfügung gestellt und wurden durch die Moderatorin anhand von Aufgaben durch die Applikation geleitet. Ziel war es, Fehlermuster in der Interaktion zu erkennen, sowie die Auswertungsmethode zur Vorbereitung der Hauptstudie zu evaluieren. Die Teilnehmer trugen eine GoPro Kamera auf dem Kopf, damit Bewegungen und Blickrichtungen aufgezeichnet werden konnten. Zudem wurde der Bildschirm des Handys mit einem Screencapturing-Tool aufgezeichnet, um später die Auswertung zu erleichtern. Die Teilnehmer arbeiteten circa 15 Minuten mit der App und sollten unter anderem ihre Arbeitszeiten für die zurückliegende Woche erfassen. Dabei wurden Kommentare und beobachtete Probleme von der Moderatorin mitnotiert.

Der Fokus der anschließenden Auswertung lag auf der Analyse der Güte der definierten User Stories, d.h. wie gut konnten Nutzer die Use Cases selbstständig durchführen und wo kam es zu Problemen.

Es fiel auf, dass einige der User Stories in dem vorliegenden Testkontext nicht erfasst und analysiert werden konnten. Bei vielen anderen Stories wurde der Parameter „time on task“ im Vergleich zu einem Referenzwert als KPI für die Güte verwendet. Die „time on task“ wurde anhand der aufgezeichneten Videos gestoppt.

Hier zeigte sich als Optimierungspunkt für die Hauptstudie, dass die Zeit der einzelnen User Stories automatisch mitgetrackt werden sollte.

Trotz allem konnte anhand der vorliegenden Daten bereits erkannt werden, dass fünf der elf User Stories nicht den geforderten Referenzwert der KPI erreichen konnten. Drei weitere User Stories waren erfolgreich (Referenzwert lag im Konfidenzintervall der „time on task“) und drei User Stories konnten im Rahmen der Studie nicht ausgewertet werden. Die Videos und teilweise auch die Daten boten außerdem erste Hinweise auf Fehlermuster in der Interaktion, wie beispielsweise der fehlende Aufforderungscharakter von Interaktionselementen. Auch hier war der Wunsch, dies zukünftig automatisch in den Daten zu erkennen.

Die Start- und Endpunkte der User Stories mussten in der Vorstudie noch manuell aus den Videos ausgelesen werden. In der Hauptstudie sollten dementsprechend Start- und Endpunkte der User Stories im Vorfeld definiert und vom Tracking Agent automatisiert erfasst werden, damit eine Auswertung der Daten ohne Hilfsmittel bzw. Vor-Ort-Tests möglich ist. Anhand der bereits erwähnten Optimierungsanforderungen an den Tracking Agent, wurde dieser überarbeitet und für die Hauptstudie vorbereitet.

3.2 Hauptstudie

Die Hauptstudie wurde mit zehn wissenschaftlichen Mitarbeitern der Universität des Saarlandes durchgeführt. Die Teilnehmer bekamen erneut ein Handy mit derselben Version der Timely App zur Verfügung gestellt wie in der Vorstudie und wurden ebenfalls durch die Moderatorin anhand von Aufgaben durch die Applikation geleitet. Im Gegensatz zur Vorstudie bekamen die Teilnehmer nun keine Kamera mehr aufgesetzt, da der Fokus auf der Auswertung der automatisiert erfassten Daten lag. Die Durchführung erfolgte ansonsten wie in der Vorstudie.

Zur Auswertung der Studie wurden ausschließlich die Daten der automatisierten Erhebung betrachtet und analysiert, es gab also kein Videomaterial mehr. In den Daten waren nun die vorher definierten Start- und Endpunkte der User Stories ersichtlich sowie ein Zeitstempel für jedes Ereignis in den Daten gegeben. Es wurden nur noch acht User Stories evaluiert, da die drei User Stories, die schon in der Vorstudie nicht auswertbar waren, hier weggelassen wurden.

Es konnte anhand der Daten erkannt werden, dass sieben der acht User Stories nicht erfolgreich waren. Als Beispiel sei hier das Anlegen des Working Profiles genannt (siehe Abbildung 3 und 4). Das Working Profile dient in der App dazu, die eigenen Arbeitszeit, die pro Tag gearbeitet werden muss, zu hinterlegen. Bevor der Nutzer dies nicht getan hat, kann er auch keine Arbeitszeiten erfassen. Hier hatten die Nutzer große Probleme, überhaupt den Menüpunkt zu erreichen. In den Diagrammen ist die „time on task“ für das Anlegen des Working Profiles aufgezeigt, wenn der Nutzer den Menüpunkt bereits gefunden hat. Hierbei zeigte sich, dass die Teilnehmer im Mittel, aber auch jeder Einzelne (siehe Abbildung 3) deutlich über dem Sollwert von 20 Sekunden lag, der für die erfolgreiche Ausführung dieser Aufgabe erwartet werden konnte. Die zweite Grafik zeigt auch, dass der Sollwert nicht im Konfidenzintervall der time on task der Teilnehmer lag und somit die User Story nicht erfolgreich war.

Die Zeiten pro User Story mussten in einer Excel Tabelle mit den erfassten Daten in diesem Fall noch händisch berechnet werden. Hierbei ist es aus Analysesicht wünschenswert, wenn in Zukunft der Timer beim Start einer neuen User Story immer auf null zurückgesetzt

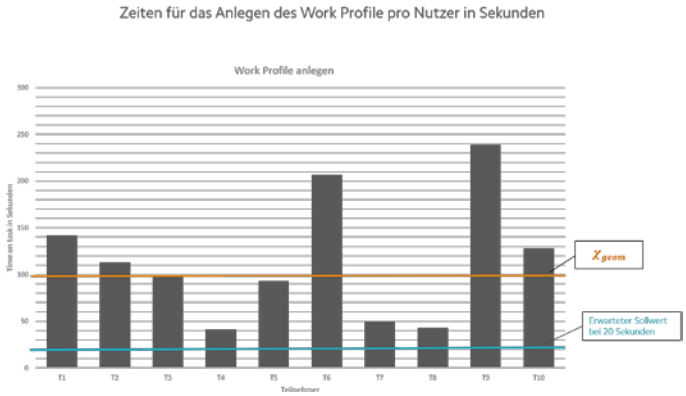

Abbildung 3: Zeiten pro Teilnehmer für das Anlegen des Working Profiles in Sekunden

Abbildung 4: Konfidenzintervall um das geometrische Mittel der „time on task“ für das Anlegen des Working Profiles, gemittelt über alle Probanden

werden würde und bis zum Ende der User Story hochzählt, sodass man sofort die „time on task“ erkennen kann. Weiterhin wurde erkannt, dass User Stories, die eine Abhängigkeit voneinander haben, noch nicht ausreichend ausgewertet werden können (z.B. Endpunkt von User Story 1 ist Startpunkt von User Story 2). Eine solche Transition zwischen User Stories konnte mit dem aktuellen Tracking Agent technisch noch nicht erfasst werden und sollte für folgende Studien definiert und eingebaut werden.

Interessant war bei der Datenauswertung darüber hinaus, dass in den Daten erkennbar war, wenn User Stories begonnen, aber nicht abgeschlossen wurden. An den entsprechenden Stellen konnten im Kontext der weiteren Interkationen Vermutungen darüber aufgestellt werden, warum der Nutzer die User Story startet, aber nicht beendet hat (sucht z.B. eine Funktion, die er nicht findet und klickt sich dabei durch alle Bereiche).

Über die Auswertung der User Stories anhand der „time on task“ hinaus konnten in den Daten weitere Muster erkannt werden, die auf Fehler in der Interaktion hindeuteten. Beispielsweise konnte ein Muster erkannt werden, bei dem die Teilnehmer nach dem Speichern der eingetragenen Zeiten erneut oder mehrfach den Speicherbutton betätigten. Hier ist die Vermutung, dass das Feedback auf das Speichern unzureichend bzw. die Erwartung der Nutzer eine andere war, was nach dem Speichern passiert („Landen auf Startseite“). Eine weitere Erkenntnis war, dass die Teilnehmer insgesamt 117-mal auf nicht klickbare Elemente des Interfaces drückten. Hierbei scheint der Aufforderungscharakter der Elemente die Nutzer irrezuführen, sodass sie hinter Bildern Funktionen vermuteten. Ähnliche „Tap“- Muster könnten in Zukunft direkt vom Agenten erkannt und als Fehler klassifiziert werden.

Die oben beschriebene Auswertung der automatisierten Daten deutet darauf hin, dass die Forschungsfrage hinsichtlich der Ableitung von Verbesserungsempfehlungen aus automatisierten Daten bejaht werden kann, vergleich man die Ergebnisse der Vorstudie mit der Hauptstudie. Es konnten aus den Daten bereits Handlungsempfehlungen abgeleitet werden, ohne dass der Produktentwickler bei der Interaktion anwesend sein muss. Die Erhebung von größeren Mengen von Nutzungsdaten gibt dem Entwicklungsteam Hinweise drauf, wo optimiert werden sollte. Einschränkungen ergeben sich, wenn es um Informationen über die Persona oder den Nutzungskontext geht, die man im klassischen Test Kontext bekommen würde. Auch die Interpretation der Muster der Daten lässt an manchen Stellen natürlich Spielraum zu.

4. Fazit und Ausblick

Bei den Forschungen zu Opti4Apps haben wir die Notwendigkeit erkannt, leichtgewichtige Feedbackanalysemethoden zu entwickeln. Dazu haben wir einen Qualitätssicherungsprozess definiert.

Im weiteren Projektverlauf haben wir sowohl unterschiedliches textuelles Feedback hinsichtlich der Aussagekraft für Verbesserungen analysiert, als nun auch automatisch aufgezeichnetes Feedback eines Software-Agenten. Beide Arten von Feedback können frühzeitige Rückmeldungen für App-Entwickler ermöglichen und somit die Qualität der App verbessern.

Einer der nächsten Schritte ist die Auswertung von manuell textuellem und automatisch aufgezeichnetem Feedback zu verbinden. Offen ist hierbei die Frage in welchen Bereichen sich beide besonders gut ergänzen und wie eine ganzheitliche Visualisierung erreicht werden kann, beide Feedback-Arten zusammenführt.

Danksagung

Dieser Beitrag wurde erstellt im Kontext des Projekts Opti4Apps, gefördert durch das Bundesministerium für Bildung und Forschung (Förderkennzeichen: 02K14A182).

Referenzen

[1] F. Elberzhager, K. Holl und S. A. Scherr, „Mobile App Testing? Nutzerfeedback automatisiert miteinbeziehen,“ *Softwaretechnik-Trends*, Bd. 37, Nr. 1, 2017.

[2] S. A. Scherr, F. Elbertshager und K. Holl, „An automated feedback-based approach to support mobile app development,“ in *Proceedings - 43rd Euromicro Conference on Software Engineering and Advanced Applications, SEAA 2017*, Vienna, 2017.

[3] T. Immich, „Continuous UX - Lean und Large unter einem Dach,“ in *Tagungsband der Mensch & Computer 2018*, Dresden, 2018.