UX Fragebögen und Wortwolken

Eine attraktive Möglichkeit zur Kommunikation von Ergebnissen

Bernard Rummel
SAP UX Enablement
SAP SE
Walldorf, Germany
bernard.rummel@sap.com

Martin Schrepp
SAP Cloud Platform Experience
SAP SE
Walldorf, Germany
martin.schrepp@sap.com

ABSTRACT

Wortwolken sind eine häufig verwendete Darstellungsform, um die Bedeutung von Begriffen in einem bestimmten Kontext zu visualisieren. Sie sind einfach zu interpretieren und bei geeignetem Layout ein ansprechendes und interessantes Gestaltungselement. Wir zeigen, wie man Wortwolken aus etablierten Fragebögen zur User Experience erzeugen und zur Kommunikation der Ergebnisse verwenden kann. Während die Skalenmittelwerte quantitative, statistisch auswertbare Ergebnisse repräsentieren, helfen die Wortwolken explorativ Ideen zur Interpretation dieser Ergebnisse zu entwickeln. Solche Interpretationen der subjektiven Wahrnehmung eines Produkts, die sich hinter den Skalenwerten verbergen, sind wichtig, um aus den Ergebnissen eines UX Fragebogens konkrete Schlussfolgerungen in Bezug auf die Verbesserung bzw. Weiterentwicklung eines Produktdesigns zu ziehen.

KEYWORDS

Wortwolken, Tag-Clouds, UX Fragebögen

1 Einleitung

Wortwolken (Schlagwortwolken, Word Clouds) sind eine beliebte Darstellungsform, um für eine Menge von Begriffen zu visualisieren, wie bedeutsam diese Begriffe im jeweiligen Kontext sind [1]. Die Größe, Farbgebung und/oder Position eines Begriffes wird dabei variiert, um dessen Bedeutsamkeit anzuzeigen. Typische Anwendungen sind Schlagwortwolken, in denen die Buchstaben-Größe eines Schlagwortes der Häufigkeit seiner Verwendung entspricht. Außer der Häufigkeit des Begriffs können auch andere, mit dem Begriff assoziierte Parameter herangezogen werden, um die Buchstabengröße zu variieren.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
MuC'19 Workshops, Hamburg, Deutschland
© Proceedings of the Mensch und Computer 2019 Workshop on Konstruktion und praktischer Einsatz von User Experience Fragebögen. Copyright held by the owner/author(s)
https://doi.org/10.18420/muc2019-ws-637

Bei der quantitativen Bewertung des Benutzungserlebnisses - welchen Eindruck haben Nutzer von einem Produkt, wie zufrieden sind sie damit, wie beurteilen sie bestimmte Produkteigenschaften - werden häufig UX Fragebögen herangezogen, deren Ergebnisse typischerweise in Form von Skalenwerten zu mehreren UX Dimensionen vorliegen. Bei der Kommunikation dieser Ergebnisse an Personen mit geringen Statistikkenntnissen ist es oft schwierig, eine korrekte und überzeugende Interpretation zu vermitteln. Für methodische Feinheiten wie Skalenvalidität, Faktorstruktur und Konfidenzintervalle besteht in der Regel wenig Interesse.

Wortwolken, als bekannte und einfach zu interpretierende Darstellung (auf der Wolke gibt es keine Zahlen und Formeln!), können helfen, die qualitativen Ergebnisse einer Befragung zu vermitteln. Durch Hervorhebung der relevanten Begriffe in einer Wortwolke treten diese buchstäblich hervor. An Stelle abstrakter Skalenwerte werden konkrete Items entsprechend ihrer jeweiligen Bedeutsamkeit für das Benutzungserlebnis herausgehoben, so dass die wichtigsten Aussagen der befragten Gruppe buchstäblich für sich selbst sprechen. Zusätzlich sind Wortwolken ein modernes, visuell ansprechendes und spannendes Darstellungsmittel [2], d.h. sie können auch genutzt werden, die Präsentation von Ergebnissen eines UX Fragebogens interessanter zu gestalten und damit auch die notwendige Aufmerksamkeit von Produktverantwortlichen auf die Ergebnisse zu lenken.

Auch für erfahrene UX Researcher ist es oft nicht leicht, eine gute Erklärung für die Bedeutung einer abstrakten Fragebogenskala in einer konkreten Untersuchung zu geben. Die direkte Verwendung der Fragebogenitems umgeht dieses Problem.

Spätestens wenn es darum geht, welche Konsequenzen man aus einem Fragebogenergebnis für das evaluierte Produkt ziehen muss, helfen abstrakte Skalenwerte nur bedingt weiter. Hier braucht man eine Erklärung, welche subjektiven Wahrnehmungen ein Produkt bei seinen Anwendern auslöst und was man tun muss, diese in die gewünschte Richtung zu lenken. Auch hier können Wortwolken helfen, eine geeignete Interpretation zu finden und zu vermitteln, indem sie Fragebogenitems zum Einen entsprechend ihrer Bedeutsamkeit gewichten, zum Anderen in den Kontext der übrigen Items stellen.

Der UEQ (User Experience Questionnaire, siehe [3,4]) ist aufgrund seines Item-Formats besonders zur Erzeugung von Wortwolken geeignet. Befragungsteilnehmer geben in einem semantischen Differential von Produktattributen jeweils an, in welchem Maß ein Attribut bzw. dessen Gegenteil zutrifft. Für die 26 Gegensatzpaare des UEQ - insgesamt 52 Adjektive - liegen also Daten vor, die für die Erzeugung einer Wortwolke verwendet werden können: je stärker ein Adjektiv als zutreffend angegeben wird, desto mehr wird es in der Wolke hervorgehoben.

Wir zeigen am Beispiel des UEQ, wie man Wortwolken aus den Ergebnissen eines Fragebogens erzeugen kann, und besprechen kurz die Übertragung des Konzepts auf andere Fragebögen. Anhand von praktischen Beispielen wird demonstriert, welchen Mehrwert diese Vorgehensweise gegenüber der reinen Betrachtung der Skalenmittelwerte bietet.

2 Umsetzung der Ergebnisse in eine Wortwolke

Wortwolken sind besonders geeignet, einzelne Begriffe in einer Menge von Begriffen hervorzuheben [5]. Während die gesamte Begriffsmenge bzw. große Teile davon sichtbar sind, wird das Heraussuchen einzelner Begriffe zwar nur schlecht unterstützt. Dafür werden die als bedeutsam hervorgehobenen Begriffe von weniger relevanten Begriffen deutlich abgegrenzt und in ihrer Zusammenschau kontextualisiert. Damit sind Wortwolken für die Kommunikation von Befragungsergebnissen interessant, da sie sehr deutlich zeigen, welche Konstellation von Eigenschaften dem evaluierten Produkt in besonderem Maß zugeschrieben wird, so dass sie es für die Gruppe der Befragten charakterisiert.

In [5] wird eine Reihe von Merkmalen angegeben, die ein Wort in einer Wortwolke besonders effektiv hervorheben, unter anderem Wortgröße, zentrale Position und Farbe.

Es gibt eine Vielzahl von Tools, mit denen Wortwolken erzeugt werden können. Diese unterscheiden sich in den graphischen Möglichkeiten zur Gestaltung der Wortwolke, realisieren aber letztlich das gleiche Prinzip. Je bedeutsamer ein Wort ist, desto zentraler wird es platziert, und desto größer ist der Font.

Wir verwenden zur Erzeugung der Wortwolken dieses Beitrags SAP Lumira (Programmpaket zur Visualisierung von Business Daten) und den freien Wortwolken-Generator www.wortwolken.com. Die beschriebene Technik kann aber auch mit anderen verfügbaren Generatoren für Wortwolken umgesetzt werden. Letztlich ist hier nur entscheidend, dass die bedeutsamsten Begriffe ausreichend hervorgehoben werden. Wir verwenden aus diesem Grund für die Beispiele in diesem Beitrag bewusst Wortwolken aus zwei Generatoren und zusätzlich verschiedene Gestaltungsoptionen dieser Generatoren.

Die Erzeugung einer Wortwolke aus UEQ-Ergebnissen ist leicht möglich. Die Adjektive der UEQ-Items liegen in Paaren vor, die in einem 7-stufigen semantischen Differential gemeinsam eingestuft werden. Ein Beispiel ist das folgende Item, bei dem der Teilnehmer das Produkt als eher unattraktiv bewertet:

unattraktiv o **x** o o o o o *attraktiv*

Für die Erzeugung der Wortwolke aus diesen Daten wird wie folgt vorgegangen. Für jedes Adjektiv wird der Wert 0 vergeben, wenn der Befragte eher das Komplement als zutreffend eingestuft hat. Hat der Befragte dagegen das Adjektiv als eher zutreffend eingestuft, ergibt sich aus dem UEQ-Skalenwert, als wie zutreffend er es eingestuft hat; entsprechend werden die Werte 1-3 vergeben. Im obigen Beispiel würde also das Adjektiv *attraktiv* den Wert 0 und das Adjektiv *unattraktiv* den Wert 2 erhalten.

Es wird also nur jeweils die Skalenhälfte verwendet, die näher an dem jeweiligen Adjektiv liegt; die Mitte und die Gegenseite werden mit 0 bewertet. Der Gesamtscore eines Worts wird als arithmetisches Mittel dieser Werte über alle Teilnehmer gebildet.

Wortgröße, Wortposition und Wortfarbe werden nun entsprechend des Gesamt-Scores des jeweiligen Adjektivs durch das verwendete Tool zur Erzeugung der Wolken gesteuert (im Gegensatz zu „klassischen" tag cloud, wo die Worthäufigkeit in der betrachteten Textmenge zugrunde gelegt wird). Hier hat der Anwender die durch das gewählte Tool vorgegebenen Freiheitsgrade bei der Gestaltung[1].

Die folgende Abbildung 1 veranschaulicht die Methode zur Berechnung eines Scores einer Person aus den UEQ-Werten:

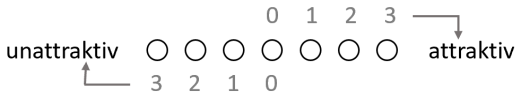

Abbildung 1: Umsetzung einer Bewertung im UEQ zum Scoring in der Wortwolke.

3 Beispiele

Abbildung 2 (links) zeigt eine Wortwolke aus Befragungsdaten zur spanischen Version von www.amazon.com (die spanischen Begriffe wurden für die Wortwolke durch die entsprechenden deutschen ersetzt).

Die Wortwolke zeigt sehr schön, dass www.amazon.com vor allem hinsichtlich der einfachen Zielerreichung (effizient, verständlich, schnell, sicher) sehr positiv bewertet wird. Hedonische Qualitäten (originell, innovativ, spannend, etc.) werden der Seite in geringerem Maße zugeschrieben. D.h. www.amazon.com wird offenbar sehr stark als praktisches und effizientes Tool gesehen, ein Einkaufserlebnis bzw. Spaß beim Bestellen und Stöbern, stehen hier nicht im Vordergrund.

[1] Da Wortwolken-Generatoren die Schriftgröße stufenweise anpassen, kann die Darstellung evtl. verbessert werden, indem der ursprüngliche Score mit einem Skalierungsfaktor multipliziert wird, so dass eine größere Anzahl Stufen verwendet wird. Beim Vergleich verschiedener Wolken muss man darauf achten, jeweils gleiche Skalierungsfaktoren zu verwenden.

Von einer nichtlinearen Skalierung (z.B. logarithmisch) raten wir ab. Die zentralen Begriffe können zwar besser hervorgehoben werden, Vergleiche werden jedoch erheblich erschwert.

Abbildung 2: Eine Wortwolke aus einer Befragung zu Amazon.com (Wortwolke mit www.wortwolken.com generiert).

Für qualitative Untersuchungen ist häufig die Frage interessant, inwieweit das Benutzungserlebnis einer Benutzergruppe homogen ist, und inwieweit es sich in Teilgruppen unterscheidet.

Die Wortwolken in Abbildung 3 entstammen einem Usability-Test. Eine genauere Analyse der Verhaltensdaten des Tests erlaubte eine Aufteilung in zwei Teilgruppen, für die separat Wortwolken erzeugt wurden. Die erste Benutzergruppe hatte mit den Aufgaben des Tests wenig Probleme, d.h. ein eher positives Benutzungserlebnis. Die zweite Benutzergruppe hatte mit erheblichen Usability-Problemen zu kämpfen; die im UEQ erfassten Benutzungserlebnisse sind entsprechend.

Die Wortwolken zeigen die unterschiedliche Wahrnehmung des Produkts in den beiden Nutzergruppen sehr deutlich.

Abbildung 3: UEQ-Wortwolken für zwei Gruppen aus einem Usability-Test (Wortwolken mit SAP Lumira generiert; automatische Skalierung und Farbsättigung der Begriffe).

Die Wortwolken in Abbildung 4 und 5 zeigen die Ergebnisse einer Bewertung von Skype mit dem UEQ durch deutsche und indonesische Studenten [8] (auch hier wurden die indonesischen Begriffe wieder durch die jeweils passenden deutschen Begriffe ersetzt, damit die Wortwolken vergleichbar werden).

Abbildung 5: Daten der deutschen Studenten als Wortwolke (erzeugt mit www.wortwolken.com).

Abbildung 6: Daten der indonesischen Studenten als Wortwolke (erzeugt mit www.wortwolken.com).

Hier wurden in beiden Wortwolken eine homogene Schriftfarbe und eine horizontale Ausrichtung der Wörter gewählt, was nicht ganz so attraktiv wirkt, aber den Vergleich der Wortwolken erleichtert.

Die Wortwolke der deutschen Studenten zeigt einen starken Focus auf der pragmatischen Qualität des Produkts (z.B. verständlich, einfach, schnell, leicht zu lernen, etc.). In der

Wortwolke der indonesischen Studenten finden sich diese Aspekte ebenfalls sehr dominant, aber hedonische Aspekte (z.B. interessant, wertvoll) sind hier stärker gewichtet.

Ansonsten sind die Ergebnisse aber sehr ähnlich, d.h. es bestehen offenbar keine gravierenden kulturspezifischen Unterschiede in der Wahrnehmung der Produktqualitäten.

Wie gut Wortwolken unterschiedliche Wahrnehmungen von Produkten darstellen und vermitteln können, zeigen die Wortwolken aus Abbildung 7. Hier wurden verschieden Anwendungen nach einem klassischen Usability Test von den Teilnehmern mit der Kurzversion des UEQ bewertet. Die Wolken geben sehr klar wieder, dass hier deutlich unterschiedliche Wahrnehmungen der UX Qualitäten vorliegen.

Alle Anwendungen werden in der Tendenz positiv beurteilt. Bei Anwendung 1 (erste Wortwolke von oben) sticht offenbar der Aspekt der einfachen Erlernbarkeit heraus (easy to learn, organized, understandable). Bei Anwendung 2 (zweite Wortwolke von oben) wird dies ebenfalls als positiver Punkt gesehen, hier ist aber die Wahrnehmung der Anwendung als effizient und praktisch noch wichtiger. Bei Anwendung 3 (unterste Wortwolke) wird ebenfalls die effiziente Bedienung und Nützlichkeit hervorgehoben. Hier kommt aber im Gegensatz zu den ersten beiden Anwendungen ein offenbar sehr positiver Eindruck in Bezug auf die Stimulation hinzu (leading edge, interesting, innovative, inventive). Die Anwendung wird als neuartig und innovativ empfunden (was es auch plausibel macht, dass die Aspekte der Erlernbarkeit hier nicht so positiv bewertet werden).

Abbildung 7: Wortwolken aus dem UEQ zu verschiedenen Anwendungen (erzeugt mit SAP Lumira).

Das Beispiel in Abbildung 7 zeigt sehr schön, dass die Wortwolken in der Lage sind zwischen verschiedenen Eindrücken, die Produkte beim Nutzer hinterlassen, zu differenzieren.

Fragebögen wie der UEQ [3,4] oder der AttrakDiff [6] sind natürlich wegen ihres bipolaren Item-Formats besonders für die Erzeugung von Wortwolken geeignet. Es ist allerdings durchaus möglich, diese Technik für andere Frageformate einzusetzen, wenn man die Items in geeigneter Weise durch einfache Schlagworte repräsentieren kann.

Abbildung 8 zeigt eine aus Daten zum VISAWI [7] erzeugte Wortwolke. Der VISAWI enthält 18 Items, die die visuelle Ästhetik eines Produkts auf 4 Skalen messen. Die Items haben die Form von Aussagen, für die man seine Zustimmung oder Ablehnung auf einer 7-stufigen Likert-Skala ausdrücken kann.

Betrachten wir ein Beispiel:

Das Layout erscheint angenehm gegliedert
Stimme gar nicht zu o o o o o o o Stimme voll zu

Für die Darstellung in der Wolke kann man dieses Item z.B. durch das Schlagwort *angenehm gegliedert* repräsentieren. Da es hier keine Gegensatzattribut gibt, wurde der Score für die Gewichtung des Schlagworts in diesem Fall einfach durch Aufsummieren der Bewertung (1 = Stimme gar nicht zu, 7 = Stimme voll zu) gebildet.

Abbildung 8: VISAWI als Wortwolke (erzeugt mit www.wortwolken.com).

Auch hier sieht man aus der Wortwolke auf einen Blick, dass die Bewertung des Produkts sehr positiv ausfiel und vor allem eine professionelle und sorgfältige Gliederung und eine gelungene farbliche Gestaltung hervorgehoben werden.

4 Wie sehen User Researcher diese Darstellung?

Ziel dieses Beitrags ist es, Wortwolken als alternative Darstellungsform der Ergebnisse eines Fragebogens vorzuschlagen. Der Anwender dieser Technik ist also ein User

Experience Designer oder User Researcher, der Ergebnisse an Produktverantwortliche oder Entwickler kommuniziert.

Damit stellt sich natürlich die Frage, wie diese Zielgruppe diese Art der Darstellung einschätzt.

Als erste kleine Evaluation wurde eine Serie von Interviews mit 7 erfahrenen User Researchern durchgeführt. Hier wurde die Methode erklärt (der UEQ war den Teilnehmern schon bekannt), mit der aus den UEQ Daten Wortwolken erzeugt werden können und es wurde ein Beispiel gezeigt. Danach wurden die Teilnehmer um ihre Einschätzung der Methode gebeten.

Wir verzichten hier wegen der geringen Zahl an Teilnehmern auf statistische Auswertungen, sondern geben nur die qualitativen Ergebnisse wieder.

Als problematisch wurde angemerkt, dass diese Art der Darstellung wenig wissenschaftlich wirkt, was auch die Gefahr birgt unseriös zu wirken. Zusätzlich wurde angemerkt, dass die Darstellung vorhandene Unterschiede zwischen der Bewertung der Items eventuell überbetont: die Aufmerksamkeit wird stark von den großen Begriffen im Zentrum der Wolke gebunden. Dies lässt deren Bedeutung größer erscheinen, als es die reinen Scores wiedergeben

Als positiv wurde vermerkt, dass die Wortwolken gut verwendet werden können, um die Diskussion über die Ursachen der Bewertung anzukurbeln (als Teaser). Für bestimmte Zielgruppen (Personen mit wenig methodischen Kenntnissen, die man mit den klassischen Auswertungen schwer erreicht), können sie ein sehr passendes Kommunikationsmittel sein: Anwender haben die Möglichkeit, ihrer Kommunikation je nach Bedarf eine eher „wissenschaftliche" oder aber „designerische" Anmutung zu geben. Durch die demonstrative Wertschätzung qualitativer Erlebensaspekte könne vielfach erst ein grundsätzliches Interesse geweckt werden, sich mit seriösen quantitativen Auswertungen auseinanderzusetzen.

5 Zusammenfassung und Ausblick

Wortwolken sind eine in vielen Web-Seiten verwendete und dadurch sehr bekannte Darstellungsform. Sie sind einfach zu interpretieren und können damit gut als Einstieg in die Präsentation von Ergebnissen von UX Evaluationen verwendet werden. Zusätzlich liefert die Wortwolke, über die Interpretation von Item- und Skalenmittelwerten hinaus, eine ganzheitliche Sicht auf die Wahrnehmung des evaluierten Produkts durch die befragte Personengruppe.

Ein gewisser Nachteil einer Wortwolke ist es, das geringe Unterschiede in der Bewertung einzelner Items durch die Darstellungsform größer erscheinen, als sie in Wirklichkeit sind. Tatsächlich ist der für die Erzeugung der Wolke verwendete Score eines Items nur so reliabel wie dieses einzelne Item, so dass mit Zufallsschwankungen zu rechnen ist, die sich visuell erheblich auswirken können. Zusätzlich erlauben die verfügbaren Wortwolken-Generatoren eine Vielzahl von visuellen Gestaltungsmöglichkeiten, mit denen man als Anwender absichtlich oder unabsichtlich gewisse Wörter visuell stärker hervorheben kann, als es ihrer eigentlichen Bedeutsamkeit entspricht. Eine Wortwolke sollte daher nicht als einzige Art der Ergebnisdarstellung, sondern immer im Kontext der anderen Untersuchungsergebnisse präsentiert werden. Bei der visuellen Gestaltung der Wolke sollte man darauf achten, keine Fehlinterpretationen zu produzieren, die nicht durch die Daten gedeckt sind.

Informelle Befragungen mit 7 Usability Professionals ergaben, dass Wortwolken als wirkungsvolles Kommunikationsmittel angesehen werden, das jedoch explizit „nicht wissenschaftlich" wirke. Dies könne je nach Kommunikationskontext durchaus ein strategisches Stilmittel darstellen.

Die wirklich belastbaren Ergebnisse eines Fragebogens sind natürlich die Skalenwerte. Wortwolken sollten nur genutzt werden, um die für Veränderungen oder Weiterentwicklungen eines Produkts notwendige Aufmerksamkeit für die Ergebnisse eines Fragebogens herzustellen. Zusätzlich können sie in Diskussionen helfen, geeignete Interpretationen der Ergebnisse zu entwickeln.

Wortwolken erlauben es, die Ergebnisse eines Fragebogens originell zu präsentieren und damit die Aufmerksamkeit der Zuhörer zu gewinnen. Es muss aber noch weiter untersucht werden, ob die Art der Darstellung evtl. typischen Fehlinterpretationen Vorschub leistet und wie gut diese Art der Darstellung angenommen wird. Hier sind verschiedene Untersuchungen geplant, um diesen Fragen nachzugehen.

REFERENCES

[1] Viegas, F.B. & Wattenberg, M. (2008). Tag Clouds and the Case of Vernacular Visualization. Interactions, Vol. 15(4), S. 49-52.

[2] Hearst, M.A. & Rosner, D. (2008) Tag Clouds: Data Analysis Tool or Social Signaller? In: Proceedings of the 41st Annual Hawaii International Conference on System Sciences, S. 160-170.

[3] 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.

[4] Laugwitz, B., Held, T. & Schrepp, M. (2008). Construction and Evaluation of a User Experience Questionnaire. In Holzinger, A. (Hrsg.): HCI and Usability for Education and Work, LNCS 5298, Berlin, Heidelberg: Springer, S. 63–76.

[5] Steffen Lohmann, Jürgen Ziegler & Lena Tetzlaff (2009). Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration https://www.uni-due.de/~s400268/Lohmann09-interact.pdf.

[6] Hassenzahl, M.; Burmester, M. & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität In: Ziegler, J. & Szwillus, G. (Hrsg.), Mensch & Computer 2003. Interaktion in Bewegung, S. 187-196, Stuttgart, Leipzig: B.G. Teubner.

[7] Moshagen, M. & Thielsch, M. T. (2010). Facets of visual aesthetics. International Journal of Human-Computer Studies, 68 (10), S. 689-709.

[8] Santoso, H., Schrepp, M., Hinderks, A. & Thomaschewski, J., (2017). Cultural Differences in the Perception of User Experience. In: Burghardt, M., Wimmer, R., Wolff, C. & Womser-Hacker, C. (Hrsg.), Mensch und Computer 2017 - Tagungsband. Regensburg: Gesellschaft für Informatik e.V. (S. 267-272).