LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
Department "Institut für Informatik"
Professur für Computational Social Science and Big Data
Prof. Jürgen Pfeffer

Masterarbeit

Not all those who wander are lost

Dynamiken bei der Interessensentwicklung in Online Communities

Oliver Baumann
<baumanno@cip.ifi.lmu.de>

Bearbeitungszeitraum:	30.04.2018 bis 29.10.2018
Betreuer:	Dr. Mirco Schönfeld
Verantw. Hochschullehrer:	Prof. Jürgen Pfeffer

Zusammenfassung

Die vorliegende Arbeit reiht sich in Forschungsliteratur zu interaktiven Tischen, interaktiven Arbeitsumgebungen, gekrümmten Multitouch-Displays und indirekten Multitouch-Mappings ein. Anhand einer Nutzerstudie wird die Wirkung zweier indirekter Eingabemodi auf den Nutzer untersucht. Dazu wurde für *Curve*, ein interaktiver Tisch mit gebogenem Display, eine prototypische Anwendung entwickelt, die entweder mit einer Maus oder über Multitouch-Gesten bedient werden kann. Im Gegensatz zu isolierten Tasks ermöglicht die Anwendung den von einer Desktopumgebung gewohnten Arbeitsablauf. Das System bietet für den Anwendungsfall "'Audio-Bearbeitung"' die Möglichkeit, in einem Audio-Sample zu navigieren und dieses zu modifizieren. Die beiden Interface-Varianten wurden auf ihre Wirkung auf das Nutzererlebnis und ihre Eignung zum Einsatz in interaktiven Arbeitsplätzen hin untersucht. Es wurde festgestellt, dass keine der beiden Varianten dabei übermäßig gut oder schlecht abschneidet. Beide Eingabetechniken sind dabei ähnlich gut für den speziellen Anwendungsfall geeignet. Ein Transfer zu anderen Einsatzmöglichkeiten schließt die Arbeit ab. Es sei darauf hingewiesen, dass die in dieser Studie präsentierten Ergebnisse anhand einer kleinen Stichprobe ermittelt wurden und möglicherweise nicht vollends generalisierbar sind.

Eidesstattliche Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbstständig angefertigt, alle Zitate als solche kenntlich gemacht sowie alle benutzten Quellen und Hilfsmittel angegeben habe.

. . . . . .

München, 21. November 2018

Inhaltsverzeichnis

1 Einleitung

2 Grundlagen und verwandte Forschung

2.1 Topic-Modelle

2.1.1 LDA

2.1.2 Verwandte Arbeiten

2.2 Soziale Netzwerkanalyse

2.2.1 Graphen und Netzwerke

2.2.2 Ego-Netzwerke

2.2.3 Verwandte Arbeiten

2.3 Reddit

2.3.1 Begriffsklärung

2.3.2 Verwandte Arbeiten

Tabelle 3.1: wichtige Schlüssel-Wert-Paare des Datensatzes

Schlüssel	Wert
author	Username des Kommentar-Erstellers
id	eindeutige ID des Kommentars
parent_id	eindeutige ID des Elements, auf das sich der Kommentar bezieht
subreddit	Name des Subreddits, in dem der Kommentar erstellt wurde

3 Datenanalyse

Dieses Kapitel liefert einen Überblick über die Methodik sowie die Ergebnisse der Datenanalyse. Zunächst wird der verwendete Datensatz präsentiert und Kritik daran erörtert. Weiterhin wird dargelegt, wie die betrachteten Topic-Modelle erzeugt werden und welche Methoden der sozialen Netzwerkanalyse Anwendung finden, sowie welche Software-Komponenten jeweils zum Einsatz kommen. Im zweiten Teil des Kapitels werden dann die Ergebnisse vorgestellt, ohne dabei jedoch einer Interpretation zu weit vorzugreifen.

3.1 Methodik

3.1.1 Datensatz

Die Grundlage der Analyse bildet ein frei zugänglicher Datensatz mit Reddit-Kommentaren. Jason Baumgartner, der unter dem Pseudonym *stuck_in_the_matrix*[1] selbst auf Reddit aktiv ist, stellt monatliche Zusammenfassungen aller erstellten Kommentare zum Download bereit [3]. Diese reichen zum gegenwärtigen Zeitpunkt von Oktober 2018 zurück bis Dezember 2005.

Struktur Die monatlichen Datensätze liegen in Form von Textdateien vor, in denen jede Zeile einen Kommentar sowie Metadaten enthält. Das maschinenlesbare JSON-Format, in dem die Daten abgelegt sind, ermöglicht dabei eine effiziente computergestützte Auswertung. Tabelle 3.1 führt die für diese Arbeit relevanten Schlüssel-Wert-Paare des Datensatzes auf. Der Schlüssel *parent_id* bezeichnet dabei das Element, auf welches sich der Kommentar bezieht. Dies können Beiträge oberster Ordnung sein, sog. „Links“, oder selbst Kommentare [5]. Zu beachten ist hier insbesondere, dass der eigentliche Textinhalt des Kommentars für diese Auswertung nicht genutzt wird.

Kohärenz des Datensatzes Im März 2018 haben Gaffney und Matias eine Analyse des Baumgartner-Korpus vorgelegt [4]. Der vollständige Korpus enthält neben Kommentaren auch Datensätze mit allen monatlich erstellten Beiträgen, im folgenden auch „Submissions“ genannt. Gaffney und Matias kommen zu dem Schluss, dass die Erfassung sowohl der Submissions als auch der Kommentare Lücken aufweist, also Elemente gänzlich nicht im Datensatz vorhanden sind. Für den Gegenstand der vorliegenden Arbeit ist dieser Umstand insofern von Bedeutung, als dass

[1] https://www.reddit.com/user/stuck_in_the_matrix

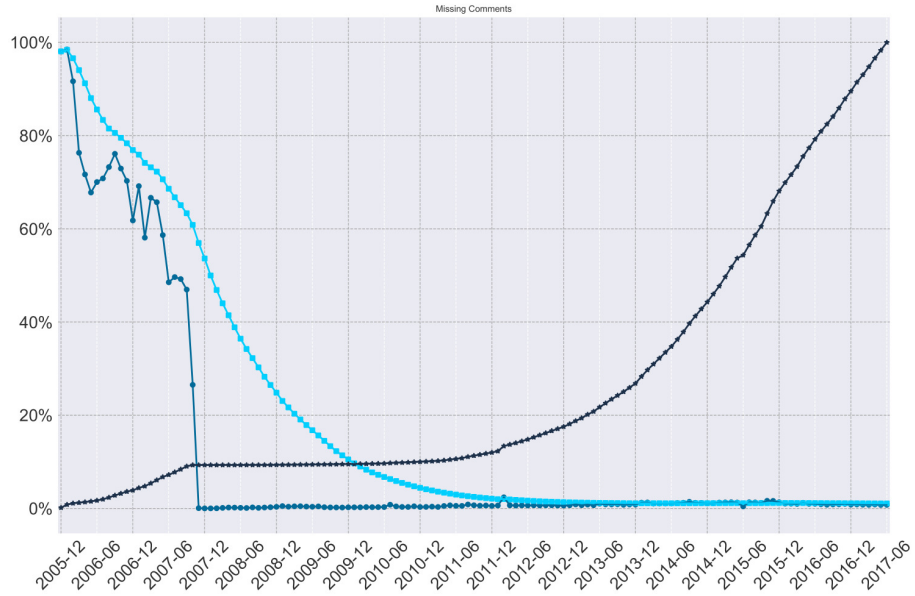

Abbildung 3.1: Anteil fehlender Kommentare. Die hellblauen Quadrate (obere Linie) stellen den gleitenden Mittelwert fehlender Kommentare in Prozent dar, die mittelblauen Punkte (mittlere Linie) den prozentualen Anteil fehlender Kommentare, und die dunkelblauen Kreuze (untere Linie) die kumlierte Gesamtzahl fehlender Kommentare [4]

fehlende Kommentare die Topic-Affinität von Nutzern verzerren können, etwa wenn ein Topic, in dem der Nutzer durchaus aktiv war, aufgrund fehlender Daten unterrepräsentiert ist. Auch Gaffney und Matias stellen fest, dass Studien, welche auf die vollständige Historie von Nutzern zugreifen, dem höchsten Risiko ausgesetzt sind, lückenhafte Daten zu betrachten [4].

Reddit weist jedem Kommentar eine eindeutige numerische ID zu. Das von Baumgartner eingesetzte System nimmt zusammenhängende Blöcke von jeweils 100 solcher IDs und versucht, die zugehörigen Kommentare über die Reddit-API[2] aufzulösen [2]. Da Reddit auch Anfragen nach gelöschten Elementen sinnvoll beantwortet, also nicht etwa mit einer Fehlermeldung, sollte ein Bereich von 100 sequentiellen IDs auch vollständig im Datensatz abgebildet sein, inklusive als gelöscht markierter Elemente. Gaffney und Matias stellen jedoch für den Zeitraum Dezember 2005 bis Februar 2016 fest, dass 943.755 Kommentar- und 1.539.583 Beitrags-IDs nicht in den Datensätzen enthalten sind. Als mögliche Gründe für das Fehlen nennen Gaffney und Matias dreierlei: sog. „dangling references“, also Verweise, bei denen das Element, auf das verwiesen wird, nicht auffindbar ist; öffentlich zugängliche Daten, die aus unbekanntem Grund nicht von Reddit an Baumgartners System übertragen wurden; oder Daten aus als privat eingestuften Communities, die nicht öffentlich sondern nur von Mitgliedern mit Zugangsberechtigung einsehbar sind [4].

In Abbildung 3.1 stellen die mittelblauen Punkte den Anteil fehlender Kommentare in Prozent dar. Ab etwa April 2006 beginnt dieser Anteil zu sinken, fällt ab etwa August 2007 stark ab und stabilisiert sich ab etwa November 2007 im niedrigen einstelligen Bereich. Um den Einfluss fehlender Kommentare so gering wie möglich zu halten, wurden daher für die vorliegende Arbeit die Datensätze beginnend mit November 2007 bis einschließlich Februar 2018 ausgewertet.

[2] https://api.reddit.com/

Tabelle 3.2: Zusammenfassung der Altersverteilung.

N	arithm. Mittel	SD	Min	Q1	Median	Q3	Max
28.029.716	6,67	12,41	1	1	2	6	124

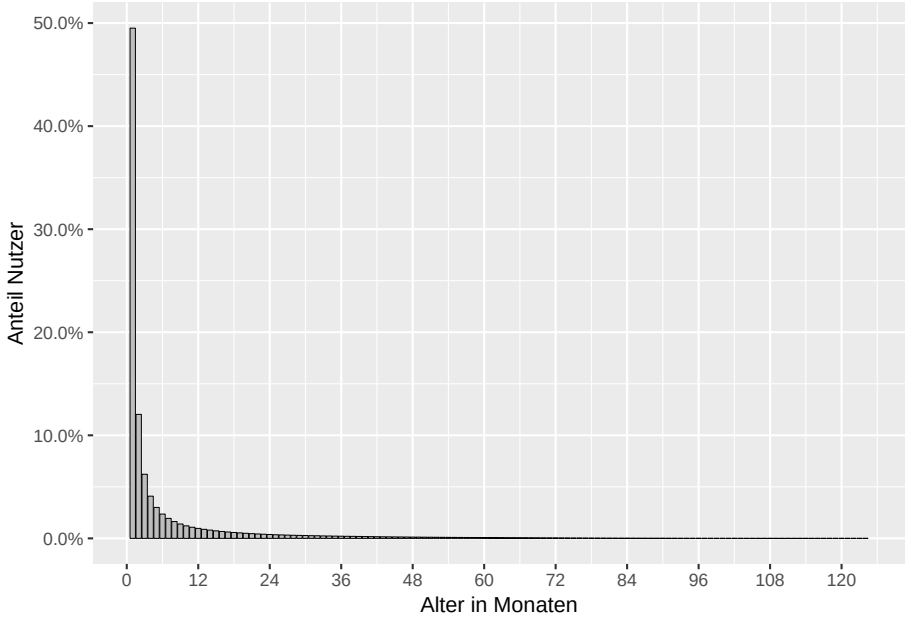

Abbildung 3.2: Verteilung der Alterswerte.

Jason Baumgartner hat als Folge der Veröffentlichung von Gaffney und Matias angekündigt, fehlende Kommentare und Beiträge nachträglich zu erfassen [1].

3.1.2 Stichprobe von Nutzern

Nachdem der Datensatz einer zeitlichen Einschränkung unterworfen wurde, müssen Kriterien für die Auswahl von Nutzern herangezogen werden. Dieser Abschnitt gibt Aufschluss über die Altersverteilung von Nutzern im Datensatz und beschreibt die Ziehung einer Stichprobe für die Datenanalyse.

Altersverteilung Nach der zeitlichen Eingrenzung der Daten liegen für den betrachteten Zeitraum von 124 Monaten etwa 3,6 Milliarden Kommentare vor, verfasst von ca. 28 Millionen Nutzern. Für jeden dieser Nutzer wurde bestimmt, in wie vielen Monaten er insgesamt im Datensatz enthalten ist; nachfolgend wird dies als „virtuelles Alter“ oder schlicht „Alter“ des Users bezeichnet.

Tabelle 3.2 bietet eine Übersicht über die Verteilung der Alterswerte. 50% der Nutzer sind zwischen einem und sechs Monaten auf Reddit aktiv (unteres resp. oberes Quartil), und 50% sind in zwei Monaten oder weniger enthalten (Median). Der arithmetische Altersdurchschnitt liegt bei etwa sieben Monaten. Da ein User mindestens einen Kommentar verfasst haben muss, um gezählt zu werden, liegt das minimale Alter bei einem Monat.

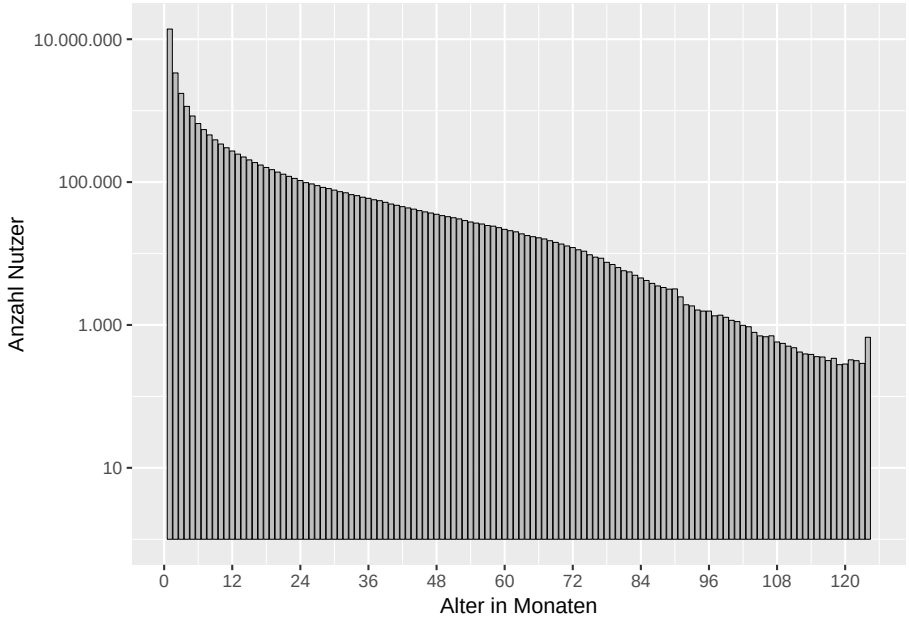

Abbildung 3.3: Verteilung der Alterswerte, logarithmische Darstellung.

Das Histogramm in Abbildung 3.2 veranschaulicht das hohe Vorkommen eher kleiner Werte. Die ALtersverteilung weist einen sog. „Long Tail“ auf: sehr viele Nutzer sind eher kurz aktiv, während Nutzer mit eher langer Aktivität nur einen gerinen Anteil ausmachen. Die lineare Skala des Histograms macht es schwierig, die Werte des Long Tail sinnvoll darzustellen. Abbildung 3.3 nutzt daher eine logarithmische Darstellung.

Markant ist bei dieser Visualisierung der Ausschlag am äußersten rechten Rand. Offenbar entfallen auf die Altersgruppe „124 Monate“ noch einmal deutlich mehr Nutzer als noch auf die vorhergehenden Gruppen. In der Tat beträgt der Unterschied zwischen den beiden letzten Altersgruppen 382 Nutzer Obwohl an dieser Stelle keine Erklärung für diese Beobachtung geliefert werden kann, ist es denkbar, dass es einen „harten Kern“ von Usern gibt, die Reddit seit langer Zeit, möglicherweise sogar von Anfang an nutzen, und regelmäßig aktiv sind.

Kriterien für die Stichprobe Um Nutzer für die weitere Datenanalyse auszuwählen, wurden zwei Kriterien festgelegt. Zum einen sollte ein Nutzer über einen ausreichend langen Zeitraum hinweg aktiv sein. Damit wird sichergestellt, dass zeitliche Verläufe keine Lücken aufweisen und genügend Daten vorliegen, um Trends zu identifizieren. Zum anderen sollte der User ein Mindestmaß an Interaktionen pro Monat hervorbringen. Da diese Arbeit Interaktionsgraphen analysiert, sollten di

Für die Analyse treffen wir eine Zufallsauswahl aus den ältesten 1000 Usern. Dadurch stellen wir sicher, dass wir User über einen genügend langen Zeitraum beobachten können. Das folgende Histogramm zeigt die Verteilung dieses Ausschnitts aus dem Long Tail.

3.1.3 Topic-Modelling

Für die Berechnung des Topic-Modells wurden für eine Auswahl an Subreddits die Titel von 50 Beiträgen aus der *Top*-Kategorie

Für jeden Kommentar im Datensatz wurde festgehalten, in welchem Subreddit er erstellt wurde. Kommentare, deren Schlüssel `author` den Wert `[deleted]` enthielten, wurden dabei übersprungen; Reddit nutzt diesen Wert, um gelöschte Inhalte zu kennzeichnen.

3.1.4 Ego-Netzwerke und soziale Netzwerkanalyse

3.2 Ergebnisse

3.2.1 Das Alter von Usern

In den vorangehenden Kapiteln wurde erläutert, welche Daten gesammelt wurden und wie eine erste Vorauswahl getroffen wurde. Die detaillierte Analyse soll Gegenstand dieses Kapitels sein.

Virtuelles Alter Nach der Vorauswahl der Daten ergibt sich ein Analysezeitraum von 124 Monaten. Da nicht alle Nutzer über diesen gesamten Zeitraum hinweg aktiv gewesen sein werden, ist eine weitere Auswahl nöitg. Zudem ist es denkbar, dass ein Teil der Nutzer nach nur wenigen Monaten die Plattform wieder verlassen hat.

Zur Bestimmung des virtuellen „Alters“ der Nutzer, also wie lange sie sich aktiv an Reddit-Communities beteiligen, wurde für jeden Monat erfasst, in welchen Subreddits sie aktiv sind und wieviele Kommentare sie dort verfassen. Diese monatlichen Daten wurden dann über den gesamten Zeitraum aggregiert, sodass am Ende für jeden Nutzer ersichtlich war, in wie vielen monatlichen Zusammenfassungen er auftaucht, sprich: wie viele Monate er sich aktiv beteiligt hat.

PLACEHOLDER USER AGESPLACEHOLDER USER AGES

Wie in Abbildung~**??** ersichtlich liegen 75% der Daten unterhalb des Mittelwerts, viele User sind also weniger als 6 Monate auf Reddit aktiv. Um sinnvolle Aussagen über die zeitliche Entwicklung treffen zu können, werden nur Nutzer betrachtet, die mindestens 6 Monate aktiv waren. Damit verbleiben $7_{047}535$ Nutzer im Datensatz, was etwa 25% der Gesamtzahl entspricht. Abbildung~**??** zeigt die Verteilung für diesen neuen Datensatz.

Für diese Nutzer werden die Subreddits erfasst, in denen sie aktiv sind. Über die Reddit-API werden zu jedem dieser Subreddits 50 Beiträge aus der *Top*-Kategorie abgerufen, welche die höchstbewerteten Posts dieser Community auflistet, unabhängig vom Erstellungsdatum des Beitrags. Diese Top-Beiträge dienen im weiteren Verlauf dazu, ein Subreddit inhaltlich zu charakterisieren; ihre Titel werden einer Topic-Analyse unterzogen.

PLACEHOLDER USER AGES

	N	Mittel	Median	Min	Max
voller Datensatz	28 029 500	6,665±12,407	2	1	124
mindestens 6 Monate aktiv	7 047 535	21,593±17,634	15	6	124

Tabelle 3.3: Altersverteilungen

Beim Abrufen der insgesamt 419~616 Subreddits antwortete die Reddit-API in XY Fällen mit „404 Not Found“, und in XY Fällen mit „403 Forbidden“. Die genaue Ursache dieser Fehler kann nicht bestimmt werden, aber 404 lässt vermuten, dass das Subreddit gelöscht wurde, und 403 deutet auf eine zugangsbeschränkte Community hin, die nicht frei einsehbar ist.

Wegen eines Programmierfehlers wurden zu 499 Usern nicht die Subreddits abgerufen, in die sie gepostet hatten. Eine nachträgliche Analyse ergab, dass diese Nutzer im Mittel einen Monat aktiv waren und 1,17 Kommentare verfassten. Aufgrund dieser geringen Aktivität wurde keine Nacherfassung der Subreddits durchgeführt.

3.2.2 Verteilung von Topics

3.2.3 Fallstudie

4 Diskussion

5 Zusammenfassung und Ausblick

Literatur

[1] Jason Baumgartner. *"... anticipate that it will take between 4-6 weeks to fill in the largest gaps for missing comments. I will then rescan all missing ids in the sequential areas (ids over 27 billion for comments) and ingest the missing data there. Probably 1-2 months before complete." 6. April 2018, 20:13 Uhr*. URL: https://twitter.com/jasonbaumgartne/status/982456309726547968. Tweet.

[2] Jason Baumgartner. *Ingesting Data — Using high performance Python code to collect Data*. 7. Mai 2018. URL: https://pushshift.io/ingesting-data%E2%80%8A-%E2%80%8Ausing-high-performance-python-code-to-collect-data/ (besucht am 07. 05. 2018). Blog-Post.

[3] Jason Baumgartner. *pushshift.io*. URL: https://files.pushshift.io/reddit/comments/ (besucht am 23. 04. 2018).

[4] Devin Gaffney und J. Nathan Matias. „Caveat Emptor, Computational Social Science: Large-Scale Missing Data in a Widely-Published Reddit Corpus“. In: (13. März 2018). arXiv: 1803.05046v1 [cs.SI].

[5] Reddit. *reddit.com: api documentation*. 20. Nov. 2018. URL: https://www.reddit.com/dev/api.