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, 19. 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
foo	bar

3 Datenanalyse

Diess 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 Topic-Modelle erzeugt werden und welche Methoden der sozialen Netzwerkanalyse Anwendung finden, sowie welche Software-Komponenten 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.

3.1.2 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. Einen Auszug der für diese Arbeit relevanten Schlüssel-Wert-Paare der Datensätze zeigt Tabelle 3.1.

Kohärenz der Daten

Im März 2018 haben Gaffney und Matias [4] eine umfassende Analyse des Baumgartner-Korpus vorgelegt. Sie kommen zu dem Schluss, dass die Erfassung sowohl der Beiträge (*submissions*) als auch der Kommentare (*comments*) Lücken aufweist, also Elemente gänzlich nicht vorhanden sind.

Da jedes Datum auf Reddit, Beiträge wie Kommentare, eine eindeutige numerische ID besitzt, nimmt Baumgartners System Blöcke von jeweils 100 solcher Identifikatoren und stellt zu jedem davon eine Anfrage an die Reddit-API [2]. Da Reddit auch auf Anfragen nach gelöschten Elementen mit einem sinnvollen Objekt antwortet, insbesondere aber nicht mit einer Fehlermeldung, sollte

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

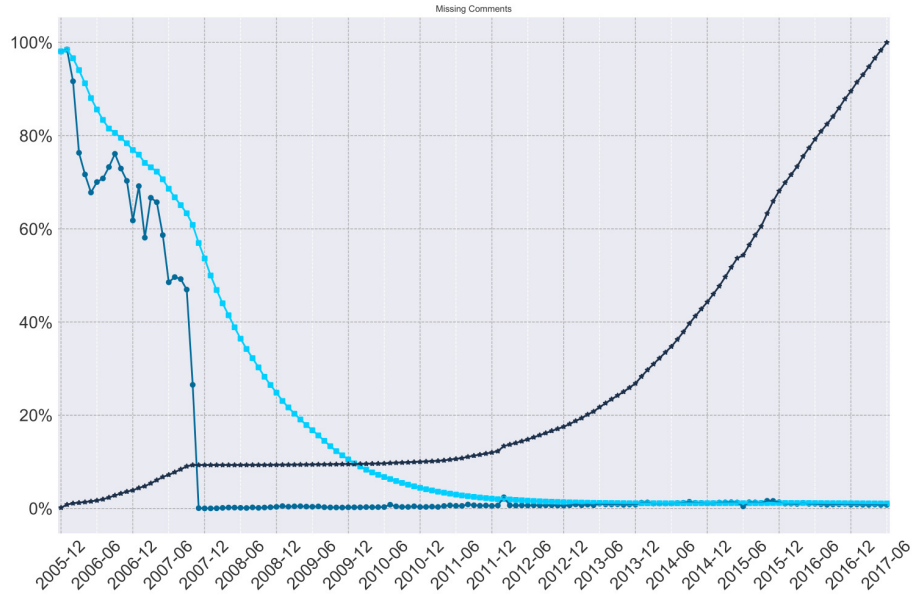

Abbildung 3.1: Verschiedene Maße zur Bestimmung fehlender Kommentare [4]

dieser Bereich von 100 sequentiellen IDs vollständig im Datensatz abgebildet sein, inklusive als gelöscht markierte Elemente. Gaffney und Matias machen jedoch für den Zeitraum Dezember 2005 bis Februar 2016 943.755 fehlende Kommentar- und 1.539.583 fehlende Beitrags-IDs aus.

Die mittelblauen Punkte (bis Juni 2007 die „mittlere“ der drei Linien) in Abbildung 3.1 zeigen den prozentualen Anteil fehlender Einträge an der Gesamtzahl aller Kommentare für einen Monat. Ab August 2007 fällt diese Linie stark ab und stabilisiert sich ab November 2007 im niedrigen einstelligen Bereich, was darauf hindeutet, dass ab diesem Zeitpunkt die Erhebung der Kommentare nahezu vollständig verläuft und kaum noch Lücken aufweist. Obgleich die fehlenden Kommentardaten in der Folge der Veröffentlichung von Gaffney und Matias nachgepflegt werden [1] wird sich die vorliegende Analyse auf den Zeitraum beginnend mit November 2007 bis April 2018 beschränken.

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.

3.2.1.1 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 3.2: 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].