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.

Aufgabenstellung

Lorem ipsum

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

3 Datenanalyse

In diesem Kapitel sollen das methodische Vorgehen bei der Datenanalyse sowie die Ergebnisse daraus vorgestellt werden. Zunächst wird der verwendete Datensatz präsentiert und Kritik daran erörtert. Weiterhin wird dargelegt, wie Topic-Modelle erzeugt wurden und welche Methoden der sozialen Netzwerkanalyse Anwendung finden, sowie welche Software-Komponenten dazu genutzt wruden. Der zweite Teil des Kapitels präsentiert die Ergebnisse, ohne dabei jedoch einer Interpretation zu weit vorzugreifen.

3.1 Methodik

3.1.1 Datensatz

Der Baumgartner-Korpus

Kohärenz der Daten

Im März 2018 haben Gaffney und Matias [3] 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 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

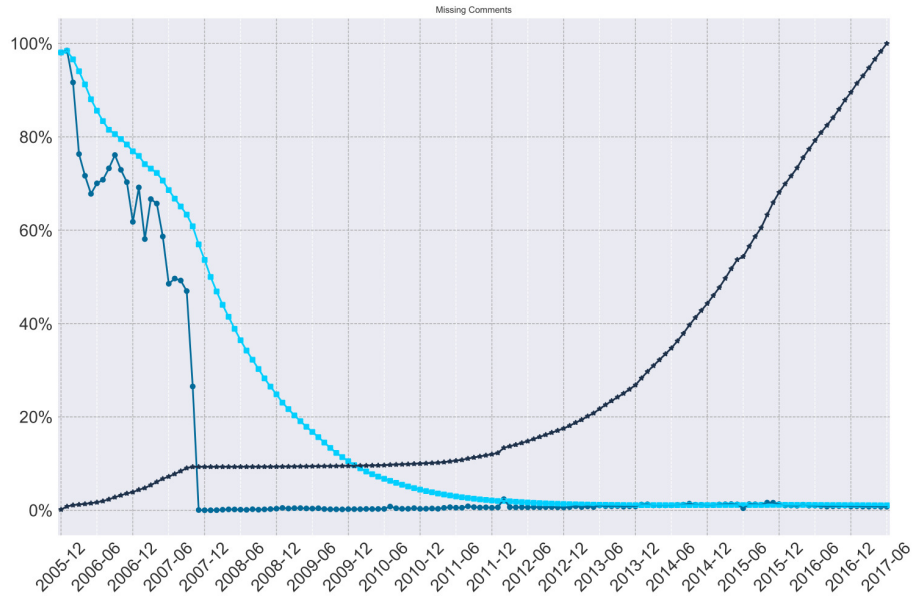

Abbildung 3.1: Verschiedene Maße zur Bestimmung fehlender Kommentare [3]

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] 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].