PROJEKT
Raumzeitliche Analyse von Twitter Daten

Jens Kersten und Friederike Klan

AG Bürgerwissenschaften
Institut für Datenwissenschaften, Jena
Deutsches Zentrum für Luft- und Raumfahrt (DLR)

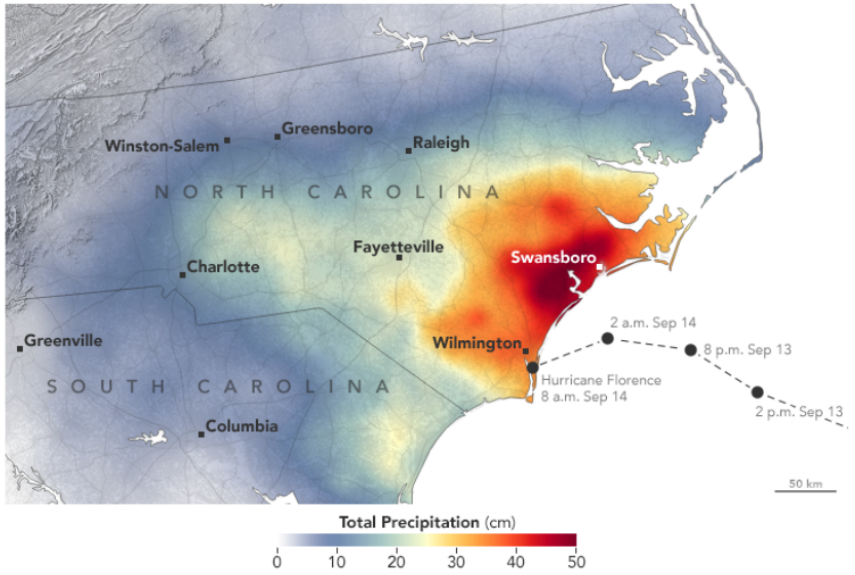
Jena, 10.04.2019

DLR Institut für Datenwissenschaften in Jena

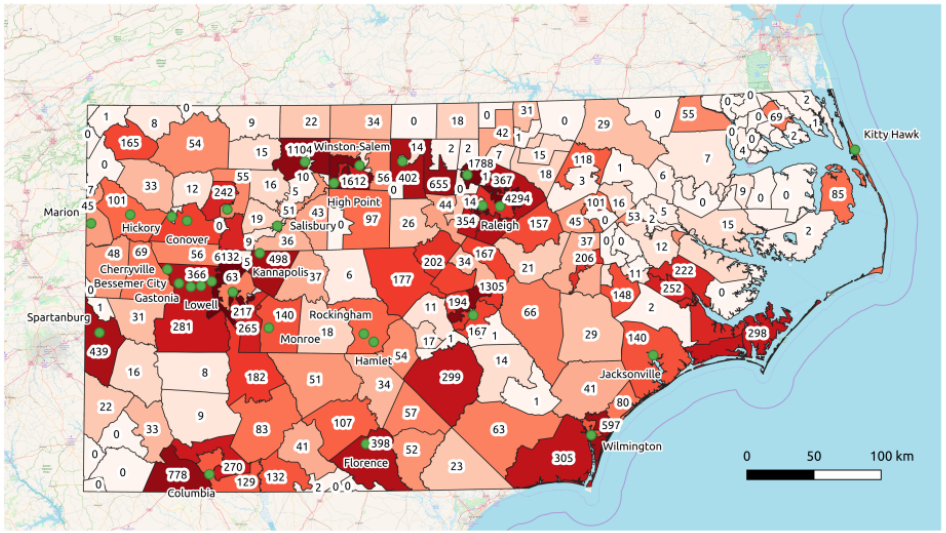
Ziel: Raumzeitliche Analyse von Titter Daten („Clustering“) zur Detektion und Analyse von Naturkatastrophen

Input: Geo-lokalisierte Tweets mit krisenbezug + Informationsklasse je Tweet (z.B. „affected individuals“)

Beispiel: Hurrikan Florence, September 2018

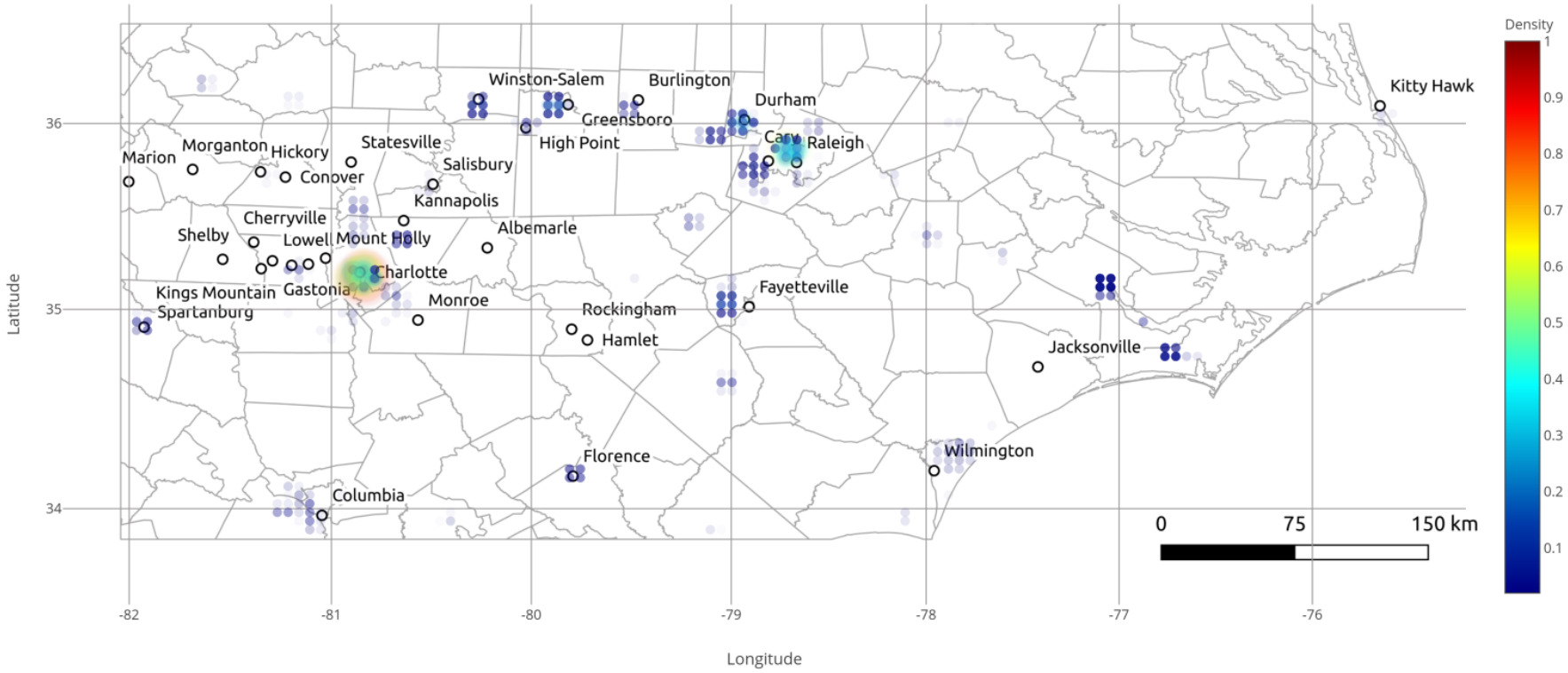


Vorhersage: Zu erwartender Niederschlag und Trajektorie des Hurrikan-Zentrums

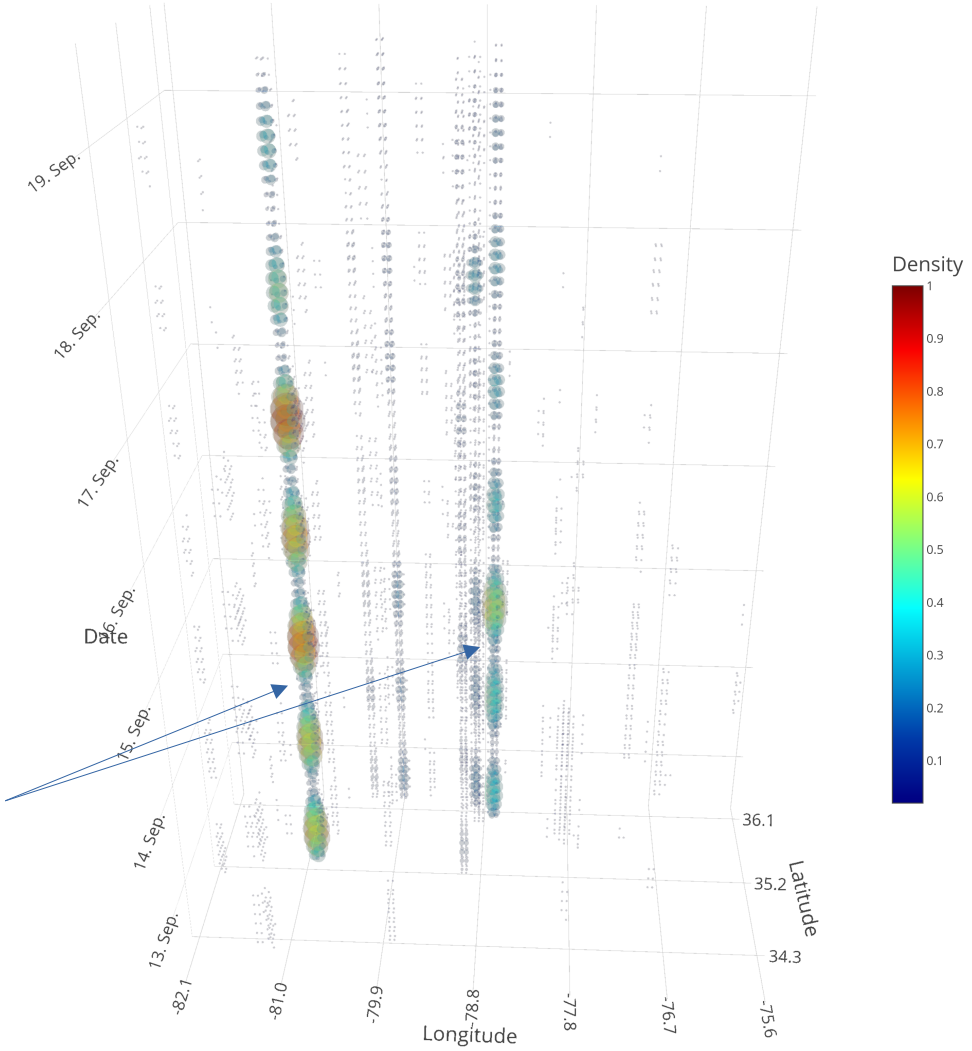


Räumliche Verteilung der ~30,000 relevanten Tweets (12.-18. September 2018)

Mögliche Analysen 1/3: 2D Dichteschätzung (Ort) der krisenrelevanten Twitter-Aktivitäten

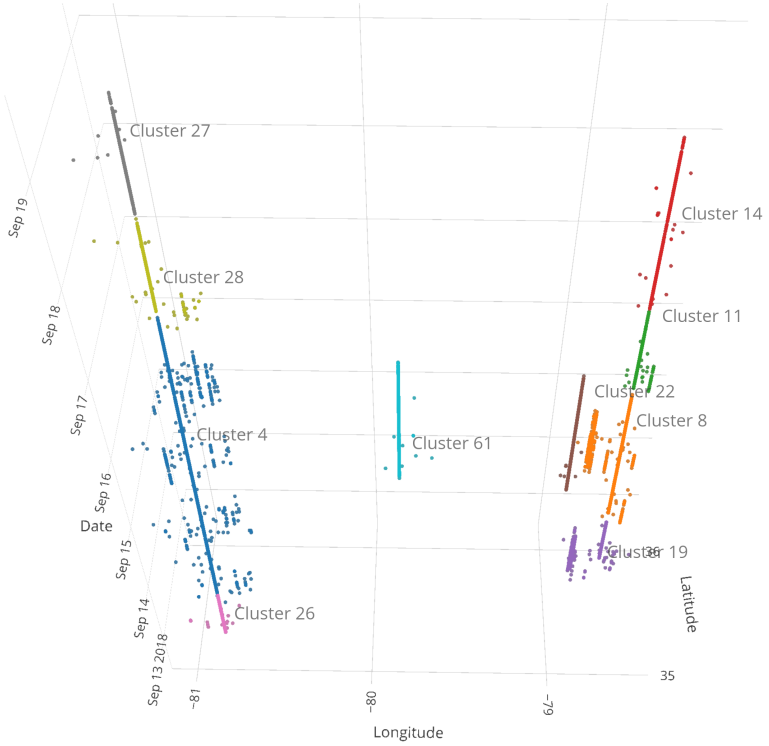


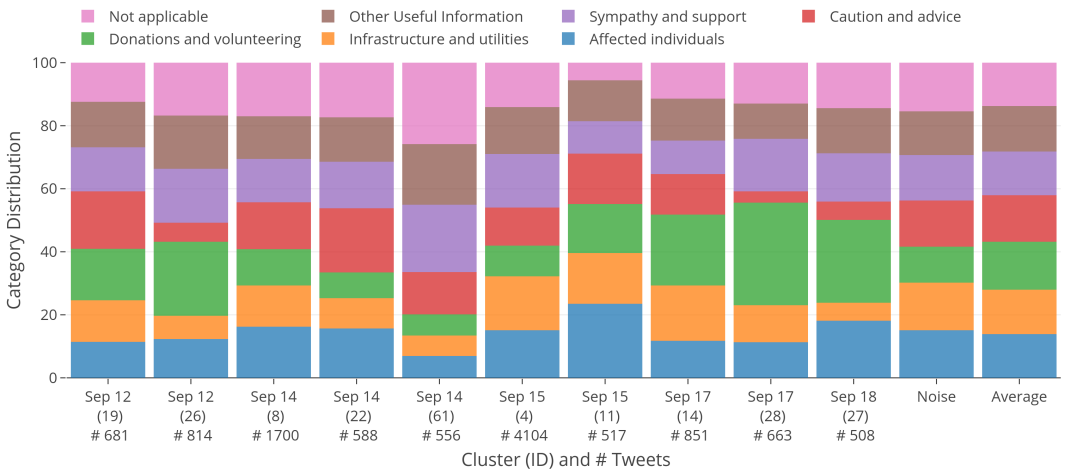
Mögliche Analysen 2/3: 3D Dichteschätzung (Ort und Zeit)



Landfall

Mögliche Analysen 3/3: Clustering (Ort und Zeit) + Analyse der Tweet-Klassen und top Keywords je Cluster





Herausforderungen: Wahl der Clustering...

- ...Methode
- ...Parameter
- ...Dimensionen (Raum, Zeit, thematische Klassen, …)

Daher hier: Implementierung eines flexiblen Clustering-Ansatzes, z.B.

Applied Intelligence (2019) 49:1228–1244
https://doi.org/10.1007/s10489-018-1324-x

FGCH: a fast and grid based clustering algorithm for hybrid data stream

Jinyin Chen[1] · Xiang Lin[1] · Qi Xuan[1] · Yun Xiang[1]

Published online: 30 October 2018

Abstract
Streaming large volumes of data has a wide range of real-world applications, e.g., video flows, internet calls, and online games etc. Thus, fast and real-time data stream processing is important. Traditionally, data clustering algorithms are efficient and effective to mine information from large data. However, they are mostly not suitable for online data stream clustering. Therefore, in this work, we propose a novel fast and grid based clustering algorithm for hybrid data stream (FGCH). Specifically, we have made the following main contributions: 1), we develop a non-uniform attenuation model to enhance the resistance to noise; 2), we propose a similarity calculation method for hybrid data, which can calculate the similarity more efficiently and accurately; and 3), we present a novel clustering center fast determination algorithm (CCFD), which can automatically determine the number, center, and radius of clusters. Our technique is compared with several state-of-art clustering algorithms. The experimental results show that our technique can achieve more than better clustering accuracy on average. Meanwhile, the running time is shorter compared with the closest algorithm.

http://www.readcube.com/articles/10.1007%2Fs10489-018-1324-x

Research Article

Fast Density Clustering Algorithm for Numerical Data and Categorical Data

Chen Jinyin,[1] He Huihao,[1] Chen Jungan,[2] Yu Shanqing,[1] and Shi Zhaoxia[1]

[1]*Zhejiang University of Technology, Zhejiang 310023, China*
[2]*Electrical Engineering Department, Ningbo Wanli University, Ningbo 310023, China*

Correspondence should be addressed to Chen Jinyin; chenjinyin@zjut.edu.cn

Received 20 August 2016; Revised 2 January 2017; Accepted 15 January 2017; Published 26 March 2017

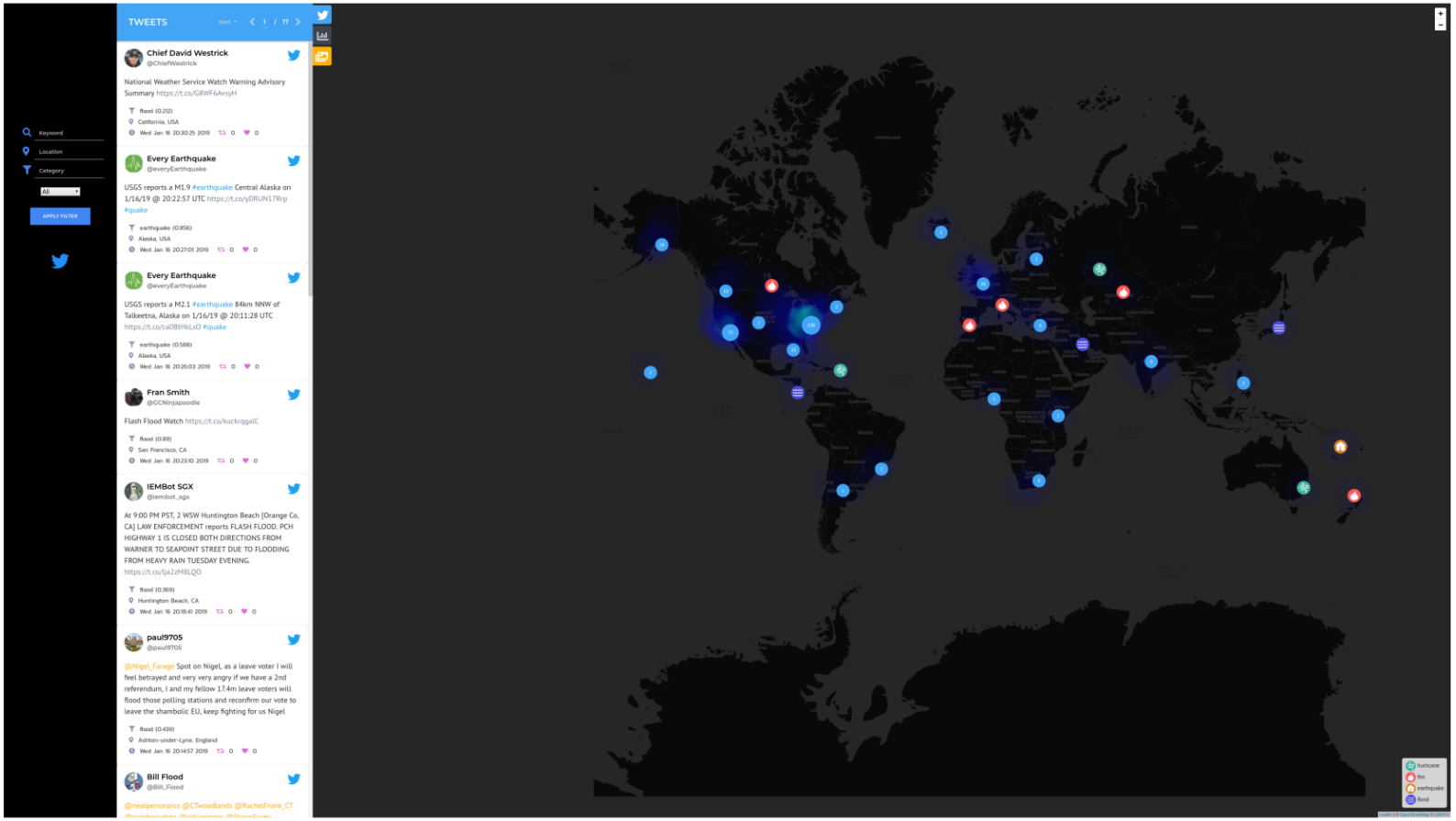
Academic Editor: Erik Cuevas

Data objects with mixed numerical and categorical attributes are often dealt with in the real world. Most existing algorithms have limitations such as low clustering quality, cluster center determination difficulty, and initial parameter sensibility. A fast density clustering algorithm (FDCA) is put forward based on one-time scan with cluster centers automatically determined by center set algorithm (CSA). A novel data similarity metric is designed for clustering data including numerical attributes and categorical

https://www.hindawi.com/journals/mpe/2017/6393652/

Startpunkt: Prototypischer Workflow zur Detektion von Krisenereignissen

Aufgabe: Ersetzen des derzeit genutzten, einfachen Clustering-Verfahrens durch neues, flexibleres Verfahren



Projektziele

- Identifikation eines geeigneten Ansatzes zur raumzeitlichen Analyse (Clustering) von Twitter Daten mit krisenkontext
- Konzeption und Implementierung der Methode (Python)
- Test und Validierung einzelner Funktionen
- Exemplarische Anwendung auf Florence-Daten
- Einpflegen der Implementierung in prototypischen Workflow (= Ersetzen des derzeitigen Verfahrens)

Projektrahmen

- Kunde: DLR
- Betreuer: Jens Kersten, Friederike Klan
- Softwareentwicklung, Programmierung und Tests
- 2-3 Personen
- Fortführung des Projektes im Rahmen einer Projekt-, Bachelor- oder Masterarbeit möglich

KONTAKT

Friederike Klan
friederike.klan@dlr.de

Jens Kersten
jens.kersten@dlr.de