
Karlsruhe Institute of Technology

Dezentrale Systeme und Netzdienste
Institut für Telematik

Lehrstuhl
Prof. Dr. Hannes Hartenstein

Fakultät für Informatik

Diplomarbeit
2014

Analyse internationaler Nachrichtenflüsse im Twitter-Netzwerk

Peter Michael Bolch

Mat.Nr.: 1345211

Referent:
Betreuer: Matthias Keller

Ich erkläre hiermit, dass ich die vorliegende Diplomarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Karlsruhe, 2014 Peter Michael Bolch

Inhaltsverzeichnis

Todo list

1 Einleitung

1.1 Motivation und Hintergründe

Über den Kurznachrichtendienst Twitter lassen sich in Echtzeit 140 Zeichen lange Textnachrichten veröffentlichen. Seit dem Start des Kurznachrichtendienstes im Jahr 2006 sind die Nutzerzahlen kontinuierlich angestiegen. 2010 konnte Twitter 75 Millionen aktive Nutzer verzeichnen [CCL10]. Im Jahr 2014 wird Twitter täglich von zirka 100 Millionen Menschen weltweit aktiv genutzt. Zur Gesamtanzahl der Nutzer-Konten gibt es von Twitter selbts keine verlässlichen Aussagen. Die Nutzer versenden täglich mehr als 500 Millionen Kurznachrichten über den Dienst [ti13]. Die meisten dieser Nachrichten sind öffentlich zugänglich und können von allen Twitter-Nutzern uneingeschränkt betrachtet werden. Twitter selbst bietet eine sogenannte Streaming-API an. Über diese lässt sich ein Echtzeit-Sample der Twitter-Kurznachrichten abrufen. [1] Die Streaming-API liefert maximal 1% aller versendeten Twitter-Kurznachrichten. Zudem lassen sich die Twitter-Kurznachrichten nach bestimmten Kriterien wie Nutzer-ID, geografischer Region oder Schlüsselwörtern Filtern. [MPLC13]

Im Text der Twitter-Kurznachrichten sind viele nutzergenerierte Informationen enthalten. Sakaki et al interpretieren die Twitter-Kurznachrichten als Sensor-Daten [SOM10]. Der Twitter-Nutzer fungiert dabei als Sensor, der ein beliebiges Ereignis erfährt oder erlebt. Möglicherweise berichtet der Twitter-Nutzer im Text der Twitter-Kurznachricht über dieses Ereignis. Damit kann der Text als Sensor-Datum interpretiert werden, wenn auch mit erheblichm Rauschen. Sakaki et al zeigen aber, dass mit diesem Vorgehen, Erdbebenzentren lokalisiert oder die Trajektorie eines Typhoons vorhergesagt werden können.

[1] https://dev.twitter.com/docs/streaming-apis

Aus der Masse an Informationen und Nachrichten die über Twitter-Kurznachrichten versendet werden, können noch weitere Erkentnisse gewonnen werden. Sozialwissenschaften und Meinungsforschung profitieren von dem enormen Informationsfundus der durch Twitter geboten wird. Die Kommunikation innerhalb des Twitter-Netzwerks kann neue Einsichten über die globale Kommunikation oder die Ausbreitung von Neuigkeiten liefern. Tumasjan et al. untersuchen in [TSSW11] wie sich die politische Landschaft im Twitter-Netzwerk wiederspiegelt. Die Wissenschaaftler haben zur Bundestagswahl 2009 100.000 Tweets analysiert und stellten fest, dass die Erwähnungen von Parteien und Politikern in Twitter, den Wahlausgang sehr genau wiederspiegelten. Garcia-Gavilanes et al. erforschen in [GGMQ14] die Kommunikation zwischen Ländern. Es wird gezeigt, dass die globale Kommunikation innerhalb des Twitter-Netzwerks nicht nur von der geografischen Distanz abhängig ist, sondern auch von sozialen, ökonomischen und kulturellen Attributen eines Landes. aber auch die Epidemieforschung kann von den Daten des Twitter-Netzwerks profitieren. So zeigten Szomsor et al. in [SKD11], dass die Vorhersage der Schweingrippe im Jahr 2009 durch die Analyse von Twitter Daten eine Woche früher möglich gewesen wäre als dies mit konventionellen Frühwarnsystemen der Fall war.

Diese Erkenntnisse und Informationen sind allerdings nur gewinnbringend einzusetzen, wenn der Standort des Twitter-Nutzers bekannt ist. Die Information, dass eine Krankheit ausgebrochen ist, ist mit einer Georeferenz wertvoller als ohne diese. Auch die Arbeit von Sakaki et al. sind auf eine Georeferenz angewiesen, wobei die Wissenschaftler ausführen, dass die ungefähre Position für ihre Anwendung ausreichend ist. Bei der Untersuchung internationaler Kommunikation wiederum, ist es wichtig zu Wissen aus welchem Land eine Twitter-Kurznachricht abgesetzt wurde. In diesem Fall kann die Georefrenz einen weiteren Raum umfassen und muss nicht GPS-Genauigkeit aufweisen. Wohingegen eine detaillierte Untersuchung des politischen Klimas innerhalb Deutschlands eine Auflösung auf Bundesländerebene erforderlich machen würde.

Twitter bietet seinen Nutzern die Möglichkeit ihren Standort im Nutzerprofil anzugeben. Hecht et al. stellten in [HHSC11] eine erste ausführliche Analyse der eingegebenen Standort-Daten bereit. Ab 2009 ermöglichte Twitter ein "per-tweet geo-tagging" [CCL10]. Dadurch können Anwendungen, auf Endgeräten mit GPS, Längen- und Breitengrad des aktuellen Standorts als Georeferenz an die Twitter-Kurznachricht anhängen. Nur ca. 1,7% der Twitter-Kurznachrichten enthalten allerdings eine konkrete Georeferenz in dieser Form.

1.2 Problembeschreibung

Um das volle Potenzial der Informationen in Twitter-Kurznachrichten auszuschöpfen ist es wichtig die Position des Twitter-Nutzers, beziehungsweise den Ort der Twitter-Kurznachricht bestimmen zu können. Die Anzahl der Twitter-Kurznachrichten die unmittelbar einem geografischen Ort zugeordnet werden können ist sehr gering.

Es ist also wichtig ein Verfahren zu finden um Twitter-Nutzer oder Twitter-Kurznachrichten mit Hilfe der in einem Tweet vorhandenen Daten einen möglichst geanuen geografischen Position. Dies soll auch möglich sein, wenn keine konkrete geografische Angabe in Form von Längen- und Breitengrad vorliegt.

1.3 Fragestellungen und Anforderungen

Folgende Fragestellungen sollen beantwortet werden:

Q1 Wie können Twitter-Nutzern, ohne Angabe von Längen- und Breitengrad, zuverlässig lokalisiert und direkt auf geographische Einheiten abgebildet werden?

Q2 Kann die Lokalisierung von Twitter-Nutzern durch die Anwendung von probabilistischen Modellen auf das Userlocation Feld verbessert werden.

1.4 Anforderungen

R1 Möglichst exakte Bestimmung des Ortes, von welchem die Twitter-Kurznachricht abgesetzt wurde. (R1)

R2 Unabhängig von kommerziellen Anbietern geografischer Informationen, oder sonstiger benötigter Daten. (R2)

R3 Das Ergebniss ist ein konkreter geografischer Ort, dabei soll zwischen folgenden Hierarcheibenen gewählt werden können (R3) :

a) Land

b) Verwaltungsebene erster Ordnung [2]

[2] in D bspsw. Länder, Baden-Württemberg, Bayern usw.

c) Verwaltungsebene zweiter Ordnung [3]

d) Stadt

R4 Die Lokalisierung soll unter Angabe einer maximalen Unsicherheit beziehungsweise einer minimalen Sicherheit erfolgen.

R5 Verfahren unabhängig von Sprache und Schriftzeichen weltweit einsetzbar.

1.5 Gliederung der Arbeit

Abschnitt 2: Grundlagen

In diesem Abschnitt sollen die Grundlagen für die entwickelte Methode vermittelt werden. Es wird auf den Kurznachrichtendienst Twitter eingegangen und es werden grundsätzliche Methoden und Verfahren vorgestellt welche zum Verständniss der entwickelten Methode benötigt werden. Ebenso werden geografische Grundbegriffe vermittelt, welche im in der arbeit häufig genutzt werden.

Abschnitt 3: Stand der Technik

Es werden aktuelle Ansätze betrachtet, eingeordnet und in Bezug auf die angegebenen Anforderungen untersucht. Es werden sowohl die Verfahren zur 'Änalyse'ünd Zuordnung als auch die Verfahren zum abbilden der geografischen Einheiten untersucht und eingeordnet.

Abschnitt 4: Lösungsansatz

In diesem Abschnitt wird die erarbeitete Methode erläutert und im Detail erklärt. Hier werde ich entweder einen Top-Down Ansatz oder einen Bottom Up Ansatz wählen.

Top-Down:

1. Genereller Aufbau der Wissensbasis [4]

[3] in D bspsw. Regierungsbezirke, Regierungsbezirk Stuttgart, Regierungsbezirk Karlsruhe usw.
[4] Datenbankschema oder Informationsschema

2. Lokalisierung von Social Media Daten (Lokalisierungsprozess)
3. Geografische Hierarchiebenen [5]
4. Sicherheit anhand der Verteilungswahrscheinlichkeiten
5. Einsatz der geografischen Hierarchiebenen zur Justierung der Sicherheit
6. NGramme zur Repräsentation der Indikatoren

Bottom-Up:

1. NGramme aus Indikatoren erzeugen
2. Geomapping
3. Datenstruktur
4. Treffer zählen (NGramm + Geoid gleich usw.)
5. Geografische Hierarchiebene
6. Unsicherheit bei Lokalisierung messen (neuer Daten)
7. Justierung der Lokalisierungsunsicherheit auf geografischen Hierarchiebenen

Abschnitt 5: Referenzimplementierung der entwickelten Methode

Es werden ausgewählte Auszüge, Probleme und Fallstricke der Referenzimplementierung erläutert und erklärt.

Abschnitt 6: Leistungsbewertung der entwickelten Methode

In diesem Kapitel werden die Ergebnisse der Refernzimplementierung bewertet und, soweit sinnvoll, gegenüber bestehenden Ansätze einer kritischen Betrachtung unterzogen.

[5] In Grundlagen und Stand der Technik behandelt bei Geografie, hier nur erklären wie verwednet wird- Hier bin ich mir unsicher ob dies Sinn macht. Theoretisch könnte man hier die geografischen Standards und Grundbegriffe definieren sowie die genutzten Komponenten der Implemnetierung.

Abschnitt 7: Schlussfolgerungen

Unter besonderer Berücksichtigung der Ergebnisse des letzten Kapitels werden Schlussfolgerungen gezogen. Der Beitrag und nutzen der entwickelten Methode soll kritisch hinterfragt werden.

Abschnitt 8: Zusammenfassung und Ausblick

Zusammenfassung der Arbeit und kritischer Rückblick. Im Ausblick werden mögliche Verbesserungen und Ideen zur Weiterentwicklung gegeben.

2 Grundlagen

Um die Fragestellungen aus Kapitel 1.3 beantworten zu können und eine geeignete Methode zu entwickeln um die Anforderungen zu erfüllen, werden eine Reihe von grundlegenden Verfahren und Begriffen benötigt. Der Leser soll hier auf einen Stand gebracht werden, der es ihm ermöglicht die Ausführungen und Ideen nachvollziehen zu können. Es werden zunächst Begriffe der sozialen Medien im allgemeinen behandelt um dann den Terminus und die Mechanismen im Twitter-Netzwerk zu erläutern. Des weiteren werden besonders die geografischen Aspekte und Möglichkeiten geografischer Angaben in sozialen Medien und insbesondere Twitter erklärt. Zum Schluss wird die genutzte Datenbasis vorgestellt und erläutert.

2.1 Geografische Grundlagen und Begriffe

gesicherte Geoinformationen vs. ungesicherte Geoinformationen+ unmittelbar geografische Indikatoren mittelbar geografische Daten bspsw. Hashtags, Inhaltsanalysen ohne spezielle geografische Hinweise

geografische Position Unter geografischer Position wird hier ein konkreter Ort in

2.1.1 Geoinformationen in Daten sozialer Netzwerke

1.
2. konkrete Geolocations (bsp. Städte <–> Länder) [HHSC11]
3.

gesicherte und ungesicherte Geoinformationen definieren!

4. Lokalisierung von Social-Media Elementen (Videos, User, Nachrichten, Bilder) kleine Übersicht
5. Hinleitung zu Twitter

2.1.2 Twitter

Allgemeine Informationen zu Twitter.

1. Was ist Twitter -> Tweets/Mechanismen/"Wie wird Twitter genutzt"'
2. Einfluss von Twitter auf Weltbild/Meinung/ usw.
3. Twitter als Nachrichtenmedium (Can Twitter Replace Newswire (Petrovic et. al))
4. Anatomie eines Tweets
   a) Welche Informationen sind in einem Tweet enthalten?
   b) Konzentration auf Daten die Hinweise zur räumlichen Lage geben könnten aber auch allgemein auf die Daten eingehen.
5. Definition Twitter-Umfeld

Paper raussuchen -> Einfluss von Twitter auf Weltbild/Meinung/

2.2 Geographie Daten

2.3 Data Sample

3 Stand der Technik

Die Lokalisierung von Twitter-Kurznachrichten oder Twitter-Nutzern ist ein Feld an dem nach wie vor aktiv geforscht wird. Nicht zuletzt trägt auch die große Verfügbarkeit an Twitter-Daten zu dem Umstand bei, dass Twitter in den letzten Jahren Forschungsgegenstand zahlreicher Publikationen war.

In diesem Abschnitt sollen bestehende Ansätze zur Lokalisierung im Twitter-Umfeld untersucht werden. Es werden Kriterien zur Einordnung der bestehenden Ansätze erarbeitet und erläutert. Die Arbeiten werden mit Hilfe der Kriterien schematisch eingeordnet um einen Überblick zu erhalten. Zum Schluss wird untersucht ob die Arbeiten die bereits formulierten Anforderungen erfüllen, und wie sich die vorliegende Arbeit von den bestehenden Ansätzen abgrenzt.

verweis auf anforderungen

3.1 Kategorisierung bestehender Ansätze

In früheren Arbeiten wurde bereits versucht, eine Einordnung der bestehenden Verfahren vorzunehmen. Es ist interessant die Kategoriesierungsansätze und die verwandten Arbeiten einiger Autoren zu studieren. Es lässt sich dadurch die Entwicklung zum Thema Lokalisierung im Twitter-Umfeld beobachten. Einige Kategorisierungsansätze werden im folgenden aufgelistet und erläutert.

Sowohl in [HHSC11] als in [CCL10] beschränken sich die verwandten Arbeiten nicht auf die Lokalisierung im Twitter-Umfeld, es werden Arbeiten zur Lokalisierung von Web-Inhalten im Allgemeinen aufgelistet. Dies lässt darauf schliessen, dass sich vor den Jahren 2010/2011 nur wenige Arbeiten mit der Lokalisierung im Twitter-Umfeld beschäftigt haben.

Kategorisierung über die untersuchte Ressource

[HHSC11] nimmt deshalb eine Kategorisierung anhand der untersuchten Ressource vor. Es wird unterschieden zwischen Forschungen zur "Lokalisierung von Microblogging-Seiten und deren Inhalten" und der "Lokalisierung von Nutzern, welche Inhalte zu Web 2.0 Seiten beisteuern". Zusätzlich wird in dieser Arbeit das "Verhalten der Nutzer im Umgang mit der Veröffentlichung ihres aktuellen Standorts" und die "Vorhersage privater Informationen" betrachtet. Darauf soll hier allerdings nicht weiter eingegangen werden.

Kategorisierung über die verwendete Methode

[CCL10] klassifiziert die vorgestellten Arbeiten anhand der verwendeteten Methodik. Es wird auf Arbeiten zur Lokalisierung von Webseiten, Web-Logs, Suchanfragen und Web-Nutzern verwiesen. Diese werden in die folgenden drei Kategorien eingeteilt.

"Inhaltsanalyse mit Begriffen in einem geografischen Verzeichnis (Content analysis with terms in a gazetteer)" Es wird darunter eine einfache Datenbanksuche verstanden. Es werden einzelne Wörter in einer Datenbank nachgeschlagen um diese einem konkreten geografischen Ort zuweisen zu können. Dabei kann sowohl lokal auf eine Geo-Datenbank als auch auf Internet Ressourcen zurückgegriffen werden. In der Regel durchläuft der untersuchte Text eine manuelle oder automatische Vorverarbeitung um potenziell geografische Begriffe, sogenannte Toponyme, herauszufiltern.

"Inhaltsanalyse mit probabilistischen Sprachmodellen (Content analysis with probabilistic language models)" Dabei werden Texte oder Textteile einer Twitter-Kurznachricht zu vordefinierten geografischen Regionen wie Ländern oder Städten zugeordnet. Nach einer Vorverarbeitung des Textes erfolgt eine statistische Auswertung, um danach den Text oder einzelne Textteile, wie beispielsweise Wörter, einer geografischen Region zuzuordnen. Eine unbekannter Text kann dann mit Hilfe der zuvor gelernten Zuordnung einer geografischen Region zugeordnet werden.

"Schlussfolgerungen durch soziale Verbindungen (Inference via social relations)" es werden soziale Verbindungen, die in Netzwerken abgebildet sind, herangezogen um Rückschlüsse auf den geografischen Ort des untersuchten Inhaltes oder einer Person ziehen zu können.

Preidhorsky et al. schlagen in [PCV13] eine weitere Einteilung anhand der Methodik vor. Allerdings werden hier ausschließlich Arbeiten im Twitter-Umfeld betrachtet.

"Geocoding" Im wesentlichen entspricht dies der "Inhaltsanalyse mit Begriffen in einem geografischen Verzeichnis" aus [CCL10]. "Geocoding" wird als Begriff in vielen Fachrichtungen unterschiedlich definiert, was zu Missverständnissen führen kann. In [Gol08] wird genauer auf den Begriff des Geocoding und die Poblematik eingegangen und eine Definition des Begriffs vorgeschlagen. Im vorliegenden Kontext ist es präziser und weniger missverständlich die Methodik als "Inhaltsanalyse mit Begriffen in einem geografischen Verzeichnis" zu bezeichnen, anstatt den Begriff "Geocoding" einzusetzen.

"Geografische Themenmodelle (Geographic Topic Modeling)" wird definiert als die Verbindung von "Themenmodellierung" und "Standorterkennung (Location Awareness)". Durch klassisches "Themenmodellierung" lässt sich aus aus Texten eine Menge von Themen extrahieren. Durch eine Lernphase werden Wörterbücher zu den Themen erstellt. Mit Hilfe dieser Themen-Wörterbücher kann später das Thema eines Textes bestimmt werden. [BNJ12] Unter "Standorterkennung" wird hier verstanden, dass nicht nur das Thema sondern auch eine bestimmte Region extrahiert werden kann. Dies kann durch geografischen Koordinaten in Twitter-Kurznachrichten realisiert werden. Im Unterschied zur Kategorie "Inhaltsanalyse mit probabilsitischen Sprachmodellen" aus [CCL10] wird hier jedoch keine vorgegebene geografische Region gefordert. Vielmehr ergeben sich die geografischen Regionen aus den Themenmodellen und den zugehörigen geografischen Koordinaten. Es wird damit eine kontinuierliche Region beschrieben, welche nicht zwangsweise durch Stadt-, Staaten- oder Ländergrenzen beschränkt ist.

"Statistische Klassifizierung (Statistical classifiers)" Diese Kategorie entspricht der "Inhaltsanalyse mit probabilsitischen Sprachmodellen" wobei in [CCL10] nur eine Arbeit in dieser Kategorie betrachtet wird. [PCV13] listet mehrere Arbeiten auf, die sich in diese Kategorie einordnen lassen.

"Informationen aus sozialen Verbindungen (Social Network Information)" analog zu "Schlussfolgerungen durch soziale Verbindungen" aus [CCL10] werden soziale Verbindungen herangezogen um den Standort zu bestimmen.

Priedhorsky et al. wählen eine ähnliche Einteilung wie vormals Cheng et al. in 2010, die verwandten Arbeiten stammen allerdings aus dem Twitter-Umfeld. Dabei ist zu bemerken, dass sich die verwendeten Methoden zur Lokalisierung im Twitter-Umfeld nicht wesentlich von denen in anderen Bereichen unterscheiden. Um die Arbeiten im Twitter-Umfeld sinnvoll voneinander abgrenzen zu können muss die Kategorisierung mehr Dimensionen umfassen. Es müssen mehr Kriterien zur Kategorisierung herangezogen werden als die reine Methodik.

Mahmud et al. betrachten in [MND12] hauptsächlich Arbeiten im Twitter-Umfeld. Diese werden in die folgenden Kategorien unterteilt.

1. "Inhaltsbasierte Standortschätzung von Tweets (Content-based Location Estimation from Tweets)"
2. "Inhaltsbasierte Standortextrahierung von Tweets (Conetnt-based Location Extraction from Tweets"
3. "Standortschätzung ohne den Tweet Inhalt zu nutzen (Location Estimation without using Tweets Content)"

"Inhaltsbasierte Standort-Schätzung von Tweets (Content-based Location Estimation from Tweets)" hier wird die geografische Position durch eine Inhaltsanalyse der Twitter-Kurznachricht geschätzt. Die Schätzung erfolgt dabei durch probabilistische Modelle. Diese Kategorie vereint damit "Geografische Themenmodelle", "Statistische Klassifizierung" aus [PCV13] mit "Inhaltsanalyse mit probabilsitischen Sprachmodellen" aus [CCL10] und ist damit als genereller anzusehen, als die vorgenannten Kategorien.

"Inhaltsbasierte Standort-Extrahierung von Tweets (Conetnt-based Location Extraction from Tweets" die verwandten Arbeiten in dieser Kategorie versuchen direkte Hinweise auf einen geografischen Ort aus einer Twitter-Kurznachricht zu extrahieren. Diese Kategorie ähnelt dem "Geocoding" beziehungsweise der "Inhaltsanalyse mit Begriffen in einem geografischen Verzeichnis".

"Standortschätzung ohne den Tweet Inhalt zu nutzen (Location Estimation without using Tweets Content)" hierunter versteht der Autor alle Informationen die nicht unmittelbar in der Twitter-Kurznachricht enthalten sind. Dazu zählen Informationen aus dem Nutzerprofil oder Informationen über die sozialen Verbindungen des Nutzers.

[MND12] nutzt ebenfalls die Methodik um die Arbeiten zu kategorisieren. Allerdings wird hier eine generellere Einteilung vorgenommen. So wird unterteilt, ob der Standort geschätzt oder extrahiert wurde. Mahmud et al. bringen aber auch eine weitere Dimension ein. Es wird hier zusätzlich unterschieden ob das angewendete Verfahren den Tweet-Inhalt nutzt oder andere Informationen.

Dies ist sinnvoll, denn die genannten Methoden lassen sich sowohl auf den Tweet-Inhalt als auch auf andere Informationen, beispielsweise aus dem Nutzerprofil, anwenden.

Frühere Arbeiten verweisen auf ein weiteres Spektrum an Arbeiten aus anderen Bereichen, wie Lokalisierung von Flickr Bildern oder Web-Log Einträgen. Arbeiten zur Lokalisierung im Twitter-Umfeld werden hier seltener erwähnt. In späteren Arbeiten, wie in [PCV13], wird hingegen fast ausschließlich auf Arbeiten aus dem Twitter-Umfeld verwiesen. Dies spiegelt die steigende Anzahl der Arbeiten zur Lokalisierung im Twitter-Umfeld wieder. Betrachtet man die Ausarbeitungen zur Lokalisierung im Twitter-Umfeld genauer, wird allerdings schnell klar, dass die Kategorisierung der Arbeiten anhand der verwendeten Methodik, dem Umfang nicht mehr gerecht wird.

Bei genauerer Betrachtung der Arbeiten stellt man allerdings fest, dass diese Klassifizierungen dem Umfang der Arbeiten nicht gerecht wird. [HHSC11] verweist auf ähnliche Ansätze mit einem anderen Untersuchungsgegenstand. [CCL10] kategorisiert die Arbeiten anhand der Methodik, und verweist ebenso auf andere Untersuchungsgegenstände. [PCV13] verweist ausschliesslich auf Arbeiten im Twitter-Umfeld und kategorisiert diese anhand der verwendeten Methodik. Die Methodeneinteilung ist aufgrund der Begriffswahl missverständlich und kann somit zu Problemen führen.

3.1.1 ttt<sss

In [SHP+13] werden die folgenden Dimensionen zur Abgrenzung herangezogen.

Allerdings lassen sich noch andere Dimensionen zur Klassifizierung der Arbeiten heranziehen. Wird besipielsweise der Text einer Twitter-Kurznachricht durch eine einfache

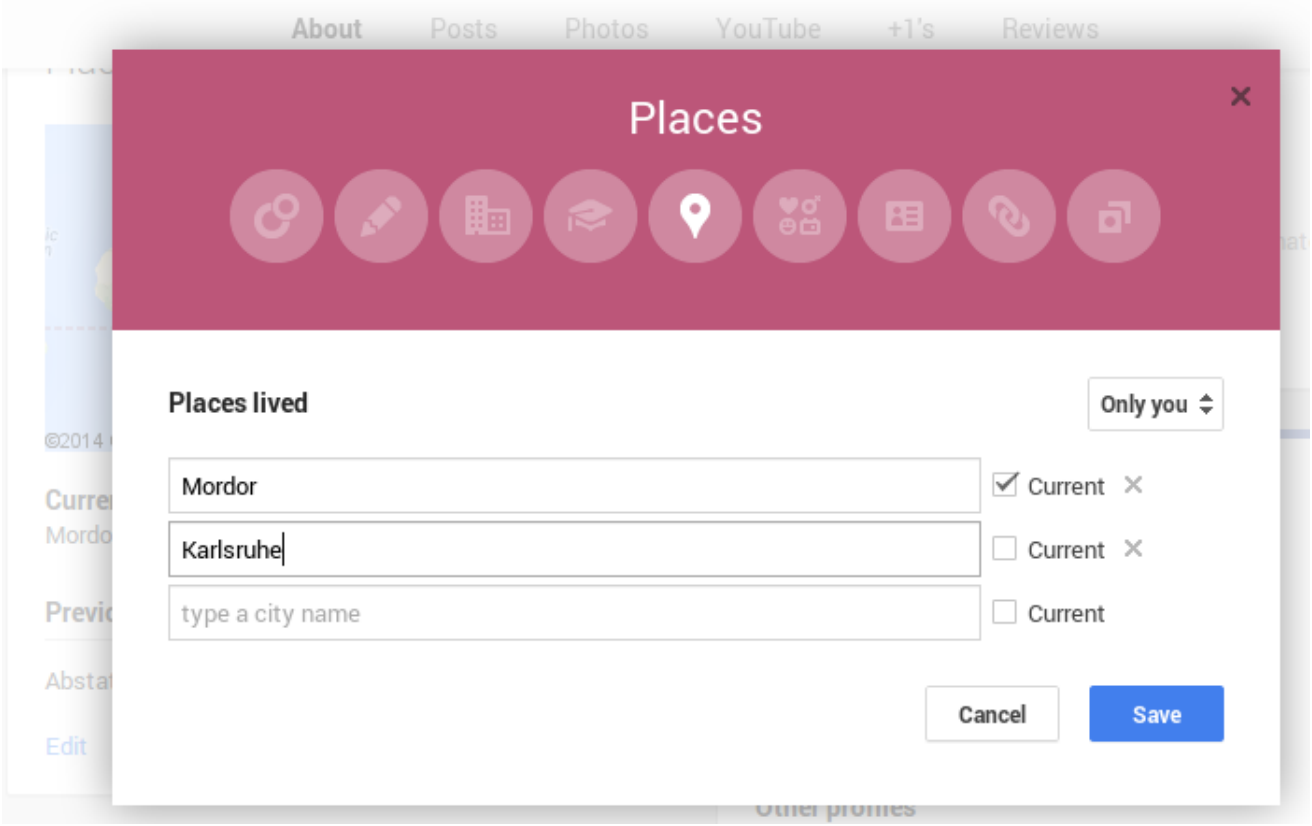
Geokodierung untersucht wird dies andere Ergebnisse liefern als eine Untersuchung auf Basis eines geografischen Themenmodells.

[HHSC11] nutzen diese Methode um eine Ground-Truth zu bestimmen indem das Userlocation-Feld in Wikipedia nachschlagen wird. Wikipedia bietet zu vielen Artikeln eine grografische Position in Form von Längen- und Breitengrad an, diese werden dann der untersuchten Twitter-Kurznachricht zugeordnet. [HGG12] nutzen die Yahoo und die Google Geocoding Api um das Userlocation-Feld eingehender zu untersuchen.

Eine weitere zu betrachtende Dimension stellt daher der konkrete Untersuchungsgegenstand in Form des Indikators dar.

Betrachtet man die Gesamtheit an arbeiten im Bereich der Lokalisierung im Twitter Netzwerk drängen sich noch mehr Dimensionen zur Klassifizeirung der arbeiten auf.

1. Räumliche Indikatoren
2. Techniken
3. Fokus der Lokalisierung



1. Naiver Ansatz -> Geocoding mit Google Maps API V3, nur Indikatoren die geografische Namen enthalten. Prinzipiell einfache Datenbankabfrage mit ein wenig semantik. Keine Jargon Namen wie Big Apple etc.
   a) Funktion der GMaps Api V3
   b) Einschränkungen der GMaps Api V3
   c) zurückgelieferte Daten der GMaps Api V3
   d) Kurze Beschreibung wie ich die API genutzt habe
2. aktuelle Ansätze
   a) allgemeiner Ansatz : Geotagged Tweets analysieren (Inhalt/andere Indikatoren usw.), zuordnen zu geografischen Bereichenünd daraus lernen.
   b) Verfahren mit Inhaltsanalysen

Ist das grundsätzliche Verfahren, analysieren Inhalt/Indikatoren -> Zuordnen auf geografische Angaben und danach Clustern tatsächlich immer gleich bei allen Arbeiten? kontinuierlich vs diskrete geografische Daten

c) Verfahren mit Indikatoren einzelne oder mehrere

d) Welche Verfahren kommen beim mapping auf geografische Entitäten zum Einsatz

geografische Entität definieren

3.1.2 Probleme früherer Ansätze

1. Genutzte API's und Indikatoren nur in bestimmten Sprachen verfügbar
2. keine Schätzung für Genauigkeit auf verschiedenen geografischen Hierarchieebenen verfügbar

4 Lösungsansatz

4.1 Indikatoren zur Ortsbestimmung

1. Training und Validierungsdaten erklären
2.

u.U. “unmittelbar und mittelbar geografische Indikatoren“ “Entwurf“ -> “Grundlagen und Stand der Technik verschieben“

4.1.1 unmittelbar geografische Indikatoren

1. Mögliche Alternativen
2. Begründung warum Userlocation und Timezone
3. Beispiele und Auswertungen (manuell getaggter Datensatz)
4. [HHSC11]

4.1.2 mittelbar geografische Indikatoren

1. bspsw. Hashtags, Inhaltsanalysen ohne spezielle geografische Hinweise,

Wie detailliert hier auf Framework eingehen? Präprozessor-Konzept zur universellen Vorverarbeitung, oder eher in Implementierung

4.1.3 Vorverarbeitung der Indikatoren (Präprozessor-Konzept)

1. geonames matching (geonames tree) für geografische Namen bestehend aus mehreren Wörtern
2. Eliminierung von Sonderzeichen
3. Tokenizing
4. Ngram Erzeugung allgemein
5. Zeitzone als "schärfenden Indikator für doppeldeutige Namen"

Was bringt die Zeitzone als zusätzlicher Indikator? Verbesserun messen

4.1.4 Encoding

Problematik unterschiedlicher Sprachen, url-encoding sinnvoll als Vorbereitung auf Webservice.

4.2 Geolocation Mapping

4.2.1 nearest neighbour mapping

1. Wie genau kann gemappt werden? Fehler Durchschnitt.
2. Mapping auf cities 1000/1000/15000 mit Daten zu durchschnitll. Abstand
3. Hier ist noch Verbesserungspotenzial -> wenn Mapping Distanz zu weit entfernt -> verwerfen!

Welches Fehlermaß kann für das mapping angewandt werden? auf Städteebene gut möglich mit geografischen Distanzen, admin2,amdin1, Land schlecht möglich mit Distanzen

4.3 Verknüpfung von Indikatoren und geografischen Lokationen zur wiedergewinnung des erlernten Wissens

4.3.1 Generierung eines Wissendatensatzes

4.3.2 Verknüpfung mit Geodaten

4.3.3 Auflösen auf Administartionsebenen, Länder

4.4 Lokalisieren von Tweets ohne konkrete geografische Daten

4.4.1 Ablauf der Lokalisierung

4.4.2 Lokalisierungssicherheit durch Ausnutzung der geografischen Hierarchiebeziehungen

einbauen!!!

4.4.3 Geografische Grundbegriffe und Geografiedaten

Geografische Grundbegriffe

Geonames.org

[1] Allgemeines zu geonames.org, was ist geonames.org.

1. Woher stammen die Daten?
2. Umfang und Informationen
3. Aktualität
4. Hierarchiebeziehungen im geonames.org Datensatz

[1] eventuell erst in Implemetierung darauf eingehen

4.4.4 ???

N-Gramme

1. NGramme allgemein, Verwendung, Beispiele.
2. Zusammenhang zwischen Länge/Grad eines N-Grammes und Wahrscheinlichkeiten.
   -> mathematische Herleitung?!

NGramme -> Nochmal genau prüfen, Zusammenhang zu Markov Modell und NGram Statistik herausstellen

5 Implementierung

6 Leistungsbewertung

7 Schlussfolgerungen, Ausblick und Fragen

8 Zusammenfassung

9 Ideen und Notizen

9.1 Stakeholder analyse

Welche potenziellen Stakeholder profitieren von der Arbeit? Was benötigt jeder dieser Stakeholder? Bedürfnisse analysieren und Begründen.

1. Marketing Professionals
2. Statistiker allgemein
3. Sozialwissenschaftler -> Analyse von Informationsströmen

9.2 Fragen an Matthias

9.2.1 Strukturell

1. Soll ich noch auf die Messung eines Informationsflusses eingehen? Wenn ich keine Informationsflüsse untersuche hängt dieses Thema ein wenig in der Luft.
2. ???

9.2.2 Inhalt

9.3 Ideen

1. Voraussetzungen zur Anwendung des Verfahrens — In Einleitung
    a) Lerndaten mit konkreten geografischen Angaben

b) Indikatoren in Lerndaten, welche auch in Datensätzen ohne konkrete geografische Angaben vorkommen (hier eventuelle Diskrepanzen zwischen geogetaggten und nicht geogetaggten tweets + Mentalität in bestimmten Ländern)

c) Indikatoren mit geografischem Bezug, oder hinreichendem geografischen Bezug, Mittelbar oder unmittelbar

2. Auf Jargon Namen für Städte eingehen, wie bspsw. the big apple -> New York City
3. Landesgrenzen-Problematik wird durch meine Lösung obsolet -> auf stakeholder eingehen
4. Wahrscheinlichkeiten für korrekte Lokalisierung kann angegeben und justiert werden

Korrelation zwischen Lokalisierungungssicherheit und tatsächlichem Match berechnen

5. Wenn Wahrscheinlichkeiten auf best. Ebene nicht hoch genug dann verschieben auf Admin2 -> Admin1 -> Länderebene
6. mit vorherigem werden Unsicherheiten bei Lokalisierung abgebildet (Wichtig für Informationsflüsse)
7.

9.4 Formulierungen

9.4.1 unmittelbare ungesicherte geografische Indikatoren

Das “userlocation“ Feld in einem Tweet kann durchaus eine konkrete Lokation beinhalten, jedoch wird auch oft irgendetwas eingetragen. [HHSC11] Es kann sich dabei um beliebige Wörter oder Sätze handeln, die einzige Limitierung ist die Anzahl zur Verfügung stehender Zeichen. Nichtsdestotrotz ist es das Ziel dieses Feldes seinen eigene Standort anzugeben. Dabei kann allerdings nicht davon ausgegangen werde, das der eingetragene Wert nicht doch in einem Zusammenhang mit einer geografischen Lokation steht. Bezeichnungen von Städten in Umgangssprache wie besipielsweise "The Big Apple"für New York City oder Motown für Detroit, sind für einige Personen nicht unmittelbar zuzuordenen, geben allerdiings eine konkrete Lokation an. Da die Masse an Bei bzw.

Spitznamen für Städte nicht überschaubar ist und auch sprachliche Probleme bestehen ist es sinnvoll alle userlocation Einträge gleich zu behandeln und diese in erster Linie als Lokationsangaben zu behandeln. Durch die Einschränkung auf eine Geolocation werden einzelne gleich lautende Einträge, welche aber nicht auf einen konkreten Ort hinweisen in einzelnen Datensätzen abgelegt.

9.5 Datenbasis

1. Welche Datenbasis wurde genutzt
    a) Streaming API
    b) Is the Sample good enough (Morstatter et al 13)
    c) When is it biased? (Morstatter et al)
    d) How does the Data sampling Startegy Impact the Discovery of Information Diffusion in Social Media (De Choudhurry, 1)
2. Lerndatensatz
3. Kontrolldatensatz
4. Manuell getaggter Datensatz
5. Google Maps getaggter Datensatz

9.6 Vorteile neuer Ansatz bei Mapping auf Geografische Daten

Notwendigkeit/Vorteile von Hierarchiebeziehungen im Mapping auf Geograohie Daten

Literaturverzeichnis

[BNJ12] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2012.

[CCL10] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, page 759, 2010.

[EOSX10] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, 2010.

[FVMF13a] Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. Traveling trends: social butterflies or frequent fliers? *... conference on Online social ...*, 2013.

[FVMF13b] Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. Traveling trends: Social butterflies or frequent fliers? *CoRR*, abs/1310.2671, 2013.

[GGMQ14] R Garcia-Gavilanes, Y Mejova, and D Quercia. Twitter ain't without frontiers: Economic, social, and cultural boundaries in international communication. *... cooperative work & social ...*, pages 1511–1522, 2014.

[Gol08] Daniel W Goldberg. *A Geocoding Best Practices Guide*. North American Association of Central Cancer Registries (NAACCR), 2008.

[HGG12] S Hale, D Gaffney, and M Graham. Where in the world are you? geolocation and language identification in twitter. *Proceedings of ICWSM'12*, (2013), 2012.

[HHSC11] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. Tweets from justin bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 237–246, New York, NY, USA, 2011. ACM.

[KA08] Balachander Krishnamurthy and Martin Arlitt. A few chirps about twitter. In *Proceedings of the first workshop on Online social networks (WOSP '08)*, pages 19–24, 2008.

[KCLC13] Krishna Y. Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. Spatio-temporal dynamics of online memes: A study of geo-tagged tweets. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 667–678, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[MND12] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Where is this tweet from? inferring home locations of twitter users. *ICWSM*, pages 511–514, 2012.

[MPLC13] Fred Morstatter, J Pfeffer, H Liu, and KM Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. *Proceedings of ICWSM*, pages 400–408, 2013.

[PCV13] Reid Priedhorsky, Aron Culotta, and Sara Y. Del Valle. Inferring the origin locations of tweets with quantitative confidence. *CoRR*, abs/1305.3932, 2013.

[POM+13] S. Petrovic, M. Osborne, R. Mccreadie, C. Macdonald, and I. Ounis. Can twitter replace newswire for breaking news? In *ICWSM - 13*, 2013.

[SHP+13] Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Mühlhäuser. A multi-indicator approach for geolocalization of tweets. *Seventh International AAAI Conference on Weblogs and Social Media*, pages 573–582, 2013.

[SKD11] Martin Szomszor, Patty Kostkova, and Ed De Quincey. #swineflu: Twitter predicts swine flu outbreak in 2009. In *Lecture Notes of the Institute*

for Computer Sciences, Social-Informatics and Telecommunications Engineering, volume 69 LNICST, pages 18–26, 2011.

[SOM10] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. *Proceedings of the 19th international conference on World wide web*, pages 851–860, 2010.

[ti13] twitter inc. Final initial public offering(ipo) prospectus, 11 2013.

[TSSW11] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Election forecasts with twitter: How 140 characters reflect the political landscape, 2011.