

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

1	Einleitung	1
1.1	Motivation und Hintergründe	1
1.1.1	alte Motivation	1
1.2	Problembeschreibung	2
1.3	Fragestellungen und Zielsetzungen	2
1.4	Gliederung der Arbeit	2
2	Grundlagen	5
2.1	Soziale Medien	5
2.1.1	Geoinformationen in Daten sozialer Medien	5
2.1.2	Twitter	5
3	Stand der Technik	6
3.1	Stand der Technik	6
3.1.1	Probleme früherer Ansätze	6
4	Lösungsansatz	7
4.1	Indikatoren zur Ortsbestimmung	7
4.1.1	unmittelbar geografische Indikatoren	7
4.1.2	mittelbar geografische Indikatoren	7
4.1.3	Vorverarbeitung der Indikatoren (Präprozessor-Konzept)	7
4.1.4	Encoding	7
4.2	Geolocation Mapping	8
4.2.1	nearest neighbour mapping	8
4.3	Verknüpfung von Indikatoren und geografischen Lokationen zur wiederge- winnung des erlernten Wissens	8
4.3.1	Generierung eines Wissensdatensatzes	8
4.3.2	Verknüpfung mit Geodaten	8
4.3.3	Auflösen auf Administartionsebenen, Länder	8
4.4	Lokalisieren von Tweets ohne konkrete geografische Daten	8
4.4.1	Ablauf der Lokalisierung	8
4.4.2	Lokalisierungssicherheit durch Ausnutzung der geografischen Hier- archiebeziehungen	8
4.4.3	Geografische Grundbegriffe und Geografiedaten	8
4.4.4	???	9
5	Implementierung	10

6	Leistungsbewertung	11
7	Schlussfolgerungen, Ausblick und Fragen	12
8	Zusammenfassung	13
9	Ideen und Notizen	14
9.1	Stakeholder analyse	14
9.2	Fragen an Matthias	14
9.2.1	Strukturell	14
9.2.2	Inhalt	14
9.3	Ideen	14
9.4	Formulierungen	15
9.4.1	unmittelbare ungesicherte geografische Indikatoren	15
9.5	Datenbasis	15
9.6	Vorteile neuer Ansatz bei Mapping auf Geografische Daten	16
	Literaturverzeichnis	17

■ gesichert und ungesicherte Geonformationen definieren!	5
■ Paper raussuchen -> Einfluss von Twitter auf Weltbild/Meinung/	5
■ 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	6
■ geografische Entität definieren	6
■ u.U. “unmittelbar und mittelbar geografische Indikatoren“ “Entwurf“ -> “Grund- lagen und Stand der Technik verschieben“	7
■ Wie detailliert hier auf Framework eingehen? Präprozessor-Konzept zur univer- sellen Vorverarbeitung, oder eher in Implementierung	7
■ Was bringt die Zeitzone als zusätzlicher Indikator? Verbesserung messen	7
■ Welches Fehlermaß kann für das mapping angewandt werden? auf Städteeebe- ne gut möglich mit geografischen Distanzen, admin2,amdin1, Land schlecht möglich mit Distanzen	8
■ NGramme -> Nochmal genau prüfen, Zusammenhang zu Markov Modell und NGram Statistik herausstellen	9
■ In Einleitung	14
■ Korrelation zwischen Lokalisierungssicherheit und tatsächlichem Match be- rechnen	15

1 Einleitung

1.1 Motivation und Hintergründe

Die Auslandsnachrichten in Fernsehen und Zeitungen bestimmen das Weltbild der Menschen. Für viele Menschen ist es die einzige Möglichkeit, sich ein Bild von der Welt zu machen. In den Kommunikationswissenschaften wird im Teilgebiet der Nachrichtengeographie untersucht, welche Nachrichtenflüsse zwischen Ländern bestehen. Es wird betrachtet, über welche Länder in den klassischen Nachrichtenmedien, wie Fernsehen oder Zeitungen, berichtet wird.

Die Verbreitung von Nachrichten und Informationen findet immer stärker auch in sozialen Netzwerken wie Twitter statt. Längst ist Twitter zu einem Massenphänomen geworden und kann die Rolle eines Nachrichtenmediums übernehmen [POM⁺13]. Twitter bietet seinen Nutzern, im Gegensatz zu klassischen Nachrichtenmedien, die Möglichkeit, direkt Einfluss auf die Verbreitung von Informationen zu nehmen. Diese direkte Einflussnahme der Nutzer kann gemessen und analysiert werden, dadurch lässt sich beispielsweise ein Interesse der Nutzer an Nachrichten anderer Länder ableiten. Daher ist es interessant Nachrichtenflüsse im Twitter Netzwerk zu untersuchen.

Um diese Nachrichtenflüsse untersuchen zu können, muss bekannt sein von welchem Standort ein Tweet abgesetzt wurde. Die Genauigkeit der Ortsauflösung, also beispielsweise auf Städte-, Länder- oder Bundesebene ¹, hängt davon ab, ob internationale oder aber nationale Nachrichtenflüsse von Interesse sind.

Beim absetzen eines Tweet werden allerdings nicht immer konkrete Daten über den aktuellen Standort des Senders angehängt. Vergleichsweise selten erhält man geographische Koordinaten oder andere Daten, welche mit Hilfe von IT Systemen unmittelbar auf einen konkreten Ort aufgelöst werden können.

1.1.1 alte Motivation

Die Auslandsnachrichten in Fernsehen und Zeitungen bestimmen das Weltbild der Menschen. Für viele Menschen ist es die einzige Möglichkeit, sich ein Bild von der Welt zu machen. In den Kommunikationswissenschaften wird im Teilgebiet der Nachrichtengeographie untersucht, welche Nachrichtenflüsse zwischen Ländern bestehen. Es wird betrachtet, über welche Länder in den klassischen Nachrichtenmedien, wie Fernsehen oder Zeitungen, berichtet wird. Die Nachrichtengeographie versucht, diese Nachrichtenflüsse durch bestimmte Faktoren, wie beispielsweise Wirtschaftsmacht, zu erklären.

¹Hier werden als Begrifflichkeiten nur die deutschen Administrationsebenen genannt, andere Länder haben hierfür andere Begriffe die nicht vollständig aufgelistet werden können

Die Verbreitung von Nachrichten und Informationen findet immer stärker auch in sozialen Netzwerken wie Twitter statt. Längst ist Twitter zu einem Massenphänomen geworden und übernimmt die Rolle eines Nachrichtenmediums [POM⁺13]. Eine interessante Forschungsfrage ist daher, welche länderübergreifenden Nachrichtenflüsse im Twitter-Netzwerk bestehen und wie diese in Bezug zur klassischen Nachrichtengeographie zu bewerten sind. Twitter bietet seinen Nutzern, im Gegensatz zu klassischen Nachrichtenmedien, die Möglichkeit, direkt Einfluss auf die Verbreitung von Informationen zu nehmen. Diese direkte Einflussnahme der Nutzer kann gemessen und analysiert werden, wodurch sich das Interesse der Nutzer für Nachrichten aus anderen Ländern ableiten lässt.

1.2 Problembeschreibung

Allgemein die Problematik der Lokalisierung von Social Media Daten betrachten und erläutern. Danach insbesondere auf Twitter und die Problematik der Informationsflüsse eingehen.

1.3 Fragestellungen und Zielsetzungen

Wie können Interaktionen, Benutzer, oder Daten aus Sozialen Netzwerken lokalisiert werden, auch wenn keine geografischen Koordinaten angegeben sind? Lokalisierung anhand von Indikatoren bzw. Sekundärinformationen.² Wie können diese auf konkrete geografische Entitäten³ abgebildet werden.

1.4 Gliederung der Arbeit

KAPITEL1: Grundlagen und Stand der Technik

In diesem Kapitel sollen die Grundlagen für die entwickelte "Methode zur Ortsbestimmung von Social Media Daten in Abwesenheit geografischer Koordinaten"⁴ vermittelt werden. Des weiteren werden aktuelle Ansätze bezüglich der Lokalisierung von Social Media Daten untersucht, die verschiedenen Verfahren untersucht und die Probleme der aktuellen Lösungen diskutiert.

KAPITEL2: Technologien und Standards

Fussnote beachten!⁵

²hier konkrete, mittelbare und unmittelbare geografische Indikatoren umschreiben um diese später zu definieren-> keine Vorwärtsverweise

³geografische Entität noch nicht definiert, allgemeine geografische Begriffe verwenden

⁴Eventuell als "Universelle Methode zur ..."

⁵Hier bin ich mir unsicher ob dies Sinn macht. Theoretisch könnte man hier die geografischen Standards und Grundbegriffe definieren sowie die genutzten Komponenten der Implementierung.

KAPITEL3: Entwicklung einer Methode zur konkreten Ortsbestimmung von Social Media Daten in Abwesenheit geografischer Koordinaten oder anderer konkreter Ortsangaben

In diesem Kapitel 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 ⁶
2. Lokalisierung von Social Media Daten (Lokalisierungsprozess)
3. Geografische Hierarchieebenen ⁷
4. Sicherheit anhand der Verteilungswahrscheinlichkeiten
5. Einsatz der geografischen Hierarchieebenen 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 Hierarchieebene
6. Unsicherheit bei Lokalisierung messen (neuer Daten)
7. Justierung der Lokalisierungsunsicherheit auf geografischen Hierarchieebenen

KAPITEL4: Referenzimplementierung der entwickelten Methode

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

KAPITEL5: Leistungsbewertung der entwickelten Methode

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

⁶Datenbankschema oder Informationsschema

⁷In Grundlagen und Stand der Technik behandelt bei Geografie, hier nur erklären wie verwendet 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 Implementierung.

KAPITEL6: Schlussfolgerungen

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

KAPITEL7: Zusammenfassung und Ausblick

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

2 Grundlagen

2.1 Soziale Medien

2.1.1 Geoinformationen in Daten sozialer Medien

1. gesicherte Geoinformationen vs. ungesicherte Geoinformationen
2. konkrete Geolocations (bsp. Städte <-> Länder) [HHSC11]
3. unmittelbar geografische Indikatoren
4. mittelbar geografische Daten bspw. Hashtags, Inhaltsanalysen ohne spezielle geografische Hinweise
5. Lokalisierung von Social-Media Elementen (Videos, User, Nachrichten, Bilder) kleine Übersicht
6. Hinleitung zu Twitter

gesichert und ungesicherte Geoinformationen definieren!

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.

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

3 Stand der Technik

3.1 Stand der Technik

Zweistufiger Prozess bei den meisten, mir bekannten Ansätzen. Untersuchung auf Häufungen von Informationen bzw. Indikatoren anhand der konkreten geografischen Angaben. Meistens Cluster Verfahren auf geografischen Daten in Verbindung mit Indikatoren/Informationen die vorverarbeitet wurden.

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 Bereichen und daraus lernen.
- b) Verfahren mit Inhaltsanalysen
- c) Verfahren mit Indikatoren einzelne oder mehrere
- d) Welche Verfahren kommen beim mapping auf geografische Entitäten zum Einsatz

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

geografische Entität definieren

3.1.1 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

4.1.1 unmittelbar geografische Indikatoren

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

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

4.1.2 mittelbar geografische Indikatoren

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

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"

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

4.1.4 Encoding

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

Was bringt die Zeitzone als zusätzlicher Indikator? Verbesserun messen

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 durchschnittl. 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, admin1, Land schlecht möglich mit Distanzen

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

4.3.1 Generierung eines Wissensdatensatzes

4.3.2 Verknüpfung mit Geodaten

4.3.3 Auflösen auf Administrationsebenen, 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

¹eventuell erst in Implementierung darauf eingehen

4.4.4 ???

N-Gramme

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

N-Gramme ->
Nochmal genau
prüfen, Zusam-
menhang zu
Markov Modell
und N-Gram
Statistik her-
ausstellen

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
 - a) Lerndaten mit konkreten geografischen Angaben
 - b) Indikatoren in Lerndaten, welche auch in Datensätzen ohne konkrete geografische Angaben vorkommen (hier eventuelle Diskrepanzen zwischen geogetaggtten und nicht geogetaggtten 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

In Einleitung

4. Wahrscheinlichkeiten für korrekte Lokalisierung kann angegeben und justiert werden
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.

Korrelation zwischen Lokalisierungssicherheit und tatsächlichem Match berechnen

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 eigenen Standort anzugeben. Dabei kann allerdings nicht davon ausgegangen werden, dass der eingetragene Wert nicht doch in einem Zusammenhang mit einer geografischen Lokation steht. Bezeichnungen von Städten in Umgangssprache wie beispielsweise "The Big Apple" für New York City oder Motown für Detroit, sind für einige Personen nicht unmittelbar zuzuordnen, geben allerdings 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 Strategy Impact the Discovery of Information Diffusion in Social Media (De Choudhury, 1)
2. Lerndatensatz
3. Kontrolldatensatz
4. Manuell getaggtter Datensatz
5. Google Maps getaggtter Datensatz

9.6 Vorteile neuer Ansatz bei Mapping auf Geografische Daten

Notwendigkeit/Vorteile von Hierarchiebeziehungen im Mapping auf Geographie Daten

Literaturverzeichnis

- [FVMF13] Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. Traveling trends: Social butterflies or frequent fliers? *CoRR*, abs/1310.2671, 2013.
- [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.
- [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.
- [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.
- [ti13] twitter inc. Final initial public offering(ipo) prospectus, 11 2013.