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

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, oder selbst Nachrichten und Informationen für andere Nutzer bereit zu stellen. Die direkte Einflussnahme der Nutzer auf die Verbreitung der Nachrichten kann gemessen und analysiert werden, dadurch lässt sich beispielsweise ein Interesse der Nutzer an Nachrichten anderer Länder ableiten. Des weiteren können die Nachrichten, welche von Nutzern selbst generiert werden, interessante Informationen, beispielsweise über Katastrophen oder Krankheiten enthalten, womit sich neue Möglichkeiten für den Katastrophen- oder Seuchenschutz ergeben. Damit aus diesen Informationen gewinnbringende Erkenntnisse gezogen werden können muss bekannt sein, wo der Tweet abgesetzt wurde. Die Erkentniss, dass beispielsweise eine Krankheit ausgebrochen ist, ist ohne konkrete Angaben über den Ort nicht hilfreich. Auch bei der Untersuchung der Verbreitung von Nachrichten ist es wichtig den Ursprung der Twitter-Kurznachricht zu kennen. Oftmals existieren keine konkreten geografischen Angaben zu einer Twitter-Kurznachricht. Nur ca. 1,6% der Twitter Kurznachrichten enthalten eine konkrete geografische Angabe in Form von Längen- und Breitengrad.

1.2 Problembeschreibung

Die Lokalisierung von Twitter-Kurznachrichten hilf

1.3 Fragestellungen und Zielsetzungen

Wie können Twitter Nachrichten ohne konkrete Ortsangaben, in Form von Längen- und Breitengrad, zuverlässig lokalisiert und direkt auf geographische Einheiten abgebildet werden?

1.4 Anforderungen

1. Unabhängig von kommerziellen Anbietern geografischer Informationen, oder sonstiger benötigter Daten, einsetzbar sein.

2. Bestimmung des Ortes von welchem die Twitter-Kurznachricht abgesetzt wurde möglichst exakt.

3. Ergebniss der Lokalisierung ist ein konkreter geografischer Ort. Zur Auswahl sollen die folgenden Hierarchieebenen stehen:
   a) Land
   b) Verwaltungsebene erster Ordnung [1]
   c) Verwaltungsebene zweiter Ordnung [2]
   d) Stadt

4. Lokalisierung unter Angabe einer maximalen Unsicherheit.

5. Verfahren unabhängig von Sprache und Schriftzeichen.

6. Minimaler Aufwand zur Lokalisierung einer Twitter-Kurznachricht mit unbekanntem Ort.

1.5 Gliederung der Arbeit

Abschnitt 1: Grundlagen und Stand der Technik

In diesem Kapitel sollen die Grundlagen für die entwickelte Methode vermittelt werden. Es werden aktuelle Ansätze untersucht, die verschiedenen Verfahren untersucht und die Probleme der aktuellen Lösungen in Bezug auf die postulierten anforderungen diskutiert.

Abschnitt 3: Lösungsansatz

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 [3]

2. Lokalisierung von Social Media Daten (Lokalisierungsprozess)

3. Geografische Hierarchiebenen [4]

4. Sicherheit anhand der Verteilungswahrscheinlichkeiten

5. Einsatz der geografischen Hierarchiebenen zur Justierung der Sicherheit

[1] in D bspsw. Länder, Baden-Württemberg, Bayern usw.
[2] in D bspsw. Regierungsbezirke, Regierungsbezirk Stuttgart, Regierungsbezirk Karlsruhe usw.
[3] Datenbankschema oder Informationsschema
[4] 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.

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 4: Referenzimplementierung der entwickelten Methode

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

Abschnitt 5: Leistungsbewertung der entwickelten Methode

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

Abschnitt 6: Schlussfolgerungen

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

Abschnitt 7: 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 bspsw. 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 Geoniformationen 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 Bereichenünd 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

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

[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.