Karlsruhe Institute of Technology

Dezentrale Systeme und Netzdienste
Institut für Telematik

Lehrstuhl
Prof. Dr. Hannes Hartenstein

Fakultät für Informatik

Diplomarbeit
2014

Mein Titel

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 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 und kann die Rolle eines Nachrichtenmediums übernehmen [POM+13]. Daher ist es interessant Nachrichtenflüsse in sozialen Netzwerken wie Twitter zu untersuchen.

Um die Nachrichtenflüsse jedoch untersuchen zu können muss bekannt sein von welchem Standort ein Tweet abgesetzt wurde. Die Genauigkeit der Ortsauflösung, also beispielsweise auf Städtebene oder Länderbene, hängt vom Untersuchungsgegenstand ab. Dabei können beispielsweise internationale oder aber nationale Nachrichtenflüsse von Interesse sein.

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

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.

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 Nachrich-

tenmedien, 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. [1] Wie können diese auf konkrete geografische Entitäten [2] 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" [3] 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! [4]

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.

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

[2] geografische Entität noch nicht definiert, allgemeine geografische Begriffe verwenden

[3] Eventuell als "Universelle Methode zur ..."

[4] 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.

Top-Down:

1. Genereller Aufbau der Wissensbasis [5]
2. Lokalisierung von Social Media Daten (Lokalisierungsprozess)
3. Geografische Hierarchiebenen [6]
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

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 Refernzimplementierung bewertet und, soweit sinnvoll, gegenüber bestehenden Ansätze einer kritischen Betrachtung unterzogen.

KAPITEL6: Schlussfolgerungen

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

[5] Datenbankschema oder Informationsschema

[6] 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.

KAPITEL7: Zusammenfassung und Ausblick

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

2 Entwurf

2.1 Indikatoren zur Bestimmung der geografischen Lokation

2.1.1 unmittelbar geografische Indikatoren

1. Mögliche Alternativen
2. Begründung warum Userlocation und Timezone
3. Beispiele und Auswertungen (manuell getaggter Datensatz)
4. Verweis auf ïn justin biebers heart"

Eventuell in Einleitung

2.1.2 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

2.1.3 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
5. Zeitzone als ßchärfenden Indikator für doppeldeutige Namen"

Checken wie oft das vorkommt und wie groß der Nutzen ist

2.1.4 Encoding

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

2.2 Geolocation Mapping

2.2.1 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 ich hier anwenden(Recherch

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

2.3.1 Generierung eines Wissendatensatzes

2.3.2 Verknüpfung mit Geodaten

2.3.3 Auflösen auf Administartionsebenen, Länder

2.4 Lokalisieren von Tweets ohne konkrete geografische Daten

2.4.1 Ablauf der Lokalisierung

2.4.2 Lokalisierungssicherheit durch Ausnutzung der geografischen Hierarchiebeziehungen

3 Implementierung

4 Leistungsbewertung

5 Schlussfolgerungen, Ausblick und Fragen

6 Zusammenfassung

7 Ideen und Notizen

7.1 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

7.2 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.

Literaturverzeichnis

[FVMF13] Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. Traveling trends: Social butterflies or frequent fliers? *CoRR*, abs/1310.2671, 2013.

[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.