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

Motivation aus Proposal einfügen und dahingehend abändern, dass Analyse der Tweet-Retweet Paare fehlt

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]

[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.

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 [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.

[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.

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 und Stand der Technik

2.1 Social Media

2.1.1 Geoinformationen in Social Media Daten

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/

2.2 Geografische Grundbegriffe und Geografiedaten

2.2.1 Geografische Grundbegriffe

2.2.2 Geonames.org

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

2.2.3

2.3 ???

2.3.1 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

2.4 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

2.4.1 Probleme früherer Ansätze

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

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

3.1 Indikatoren zur Ortsbestimmung

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

3.1.1 unmittelbar geografische Indikatoren

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

3.1.2 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

3.1.3 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

3.1.4 Encoding

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

3.2 Geolocation Mapping

3.2.1 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

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

3.3.1 Generierung eines Wissendatensatzes

3.3.2 Verknüpfung mit Geodaten

3.3.3 Auflösen auf Administartionsebenen, Länder

3.4 Lokalisieren von Tweets ohne konkrete geografische Daten

3.4.1 Ablauf der Lokalisierung

3.4.2 Lokalisierungssicherheit durch Ausnutzung der geografischen Hierarchiebeziehungen

4 Implementierung

5 Leistungsbewertung

6 Schlussfolgerungen, Ausblick und Fragen

7 Zusammenfassung

8 Ideen und Notizen

8.1 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

8.2 Fragen an Matthias

8.2.1 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. ???

8.2.2 Inhalt

8.3 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.

8.4 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

8.5 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.