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

Motivation aus Proposal (abändern).

1.2 Problembeschreibung

Lokalisierung von Tweets ohne konkrete geografische Angaben. Problem das sehr wenige Tweets Geotags haben.

2 Grundlagen und Stand der Technik

2.1 Geoinformationen in Social Media Daten

definieren!

1. gesicherte Geoinformationen vs. ungesicherte Geoinformationen
2. konkrete Geolocations (bsp. Städte <–> Länder) Referenz: From Justin Biebers Heart
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

2.2 Twitter

Allgemeine Informationen zu Twitter.

1. Was ist Twitter -> Tweets/Mechanismen/"Wie wird Twitter genutzt"

Paper raussuchen

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.

2.3 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

2.4 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

2.5 N-Gramme

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

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

2.6 Stand der Technik

Zwei vorgehen bei aktuellen Ansätzen Zum ersten, dass zusammenfassen zu Gruppen von Hinweisen auf einen bestimmten Standort und das abbilden dieser Gruppen auf geografische Entitäten oder direktes abbilden der Indikatoren auf den Globus und das darauffolgende Gruppieren nach Indikatoren um Häufungen festzustellen.

1. Naiver Ansatz -> Geotagging 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

in allen ander Arbeiten glei- ches Prinzip?

geografische Entität definieren

2.6.1 Probleme früherer Ansätze

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

2.6.2 Vorteile neuer Ansatz bei Mapping auf Geografische Daten

Notwendigkeit/Vorteile von Hierarchiebeziehungen im Mapping auf Geograohie Daten

3 Entwurf

3.1 Indikatoren zur Bestimmung der geografischen Lokation

3.1.1 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

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
5. Zeitzone als ßchärfenden Indikator für doppeldeutige Namen"

Checken wie oft das vorkommt und wie groß der Nutzen ist

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 ich hier anwenden(Recherch

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 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 geogetaggten und nicht geogetaggten tweets + Mentalität in bestimmten Ländern)
   c) Indikatoren mit geografischem Bezug, oder hinreichendem geografischen Bezug, Mittelbar oder unmittelbar

   In Einleitung
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.