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

Über den Mikroblogging-Dienst Twitter lassen sich in Echtzeit 140 Zeichen lange Textnachrichten veröffentlichen. Seit dem Start des Mikroblogging-Dienstes im Jahr 2006 sind die Nutzerzahlen kontinuierlich angestiegen. 2010 konnte Twitter 75 Millionen aktive Nutzer verzeichnen [CCL10]. Im Jahr 2013 wird Twitter täglich von zirka 100 Millionen Menschen weltweit aktiv genutzt. Dies berichtete Twitter 2013 in seinem Prospekt zum Börsengang [ti13]. Zur Gesamtanzahl der Nutzer-Konten gibt es von Twitter keine Informationen. Dies kann mitunter damit begründet werden, dass die Gesamtanzahl der Nutzer-Konten auch inaktive Nutzer einschliesst und somit keine Informationen über die tatsächliche Aktivität im Netzwerk liefert. Auch andere soziale Netzwerke ziehen die aktiven Nutzer als Metrik heran, des weiteren wird die Metrik vom Interactive Advertising Bureau (IAB) empfohlen. [IAB]

Die Twitter-Nutzer verfassen täglich mehr als 500 Millionen Nachrichten, sogenannte Tweets [ti13]. [1] Die meisten dieser Tweets sind öffentlich zugänglich und können von allen Twitter-Nutzern uneingeschränkt betrachtet werden. Twitter bietet zusätzlich eine sogenannte Streaming-API an, welche es ermöglicht Tweets programmatisch zu empfangen. [2] Die Streaming-API stellt ein Echtzeit-Sample der aktuell versendeten Tweets bereit und liefert maximal 1% aller Tweets die zum aktuellen Zeitpunkt verfasst wurden [MPLC13]. Über die sogenannte Filter-API lassen sich die Tweets nach bestimmten Kriterien wie Nutzer-ID, geografischer Region oder Schlüsselwörtern filtern. [3]

[1] Im Abschnitt Grundlagen wird der Begriff Tweet genauer untersucht, für den Moment sollen darunter die Nachrichten verstanden werden, welche von den Twitter-Nutzern verfasst werden

[2] API: Application Programming Interface oder auch Programmierschnittstelle

[3] https://dev.twitter.com/docs/streaming-apis

Ein Tweet besteht aus einer Reihe von Informationen. Neben dem Verfasser, ist der Tweet-Text die wichtigste Information die in einem Tweet enthalten ist. Der Tweet-Text wird vom Nutzer verfasst und abgesendet, er beinhaltet die zentrale Information eines Tweets. In den 140 Zeichen des Tweet-Textes teilen Twitter-Nutzer Informationen unterschiedlicher Ausprägung aus. Unter anderem wird über privates, Sportergebnisse, Großereignisse, persönliche Erfahrungen oder persönliche Meinungen berichtet. Auch Bilder und Web-Links können in einem Tweet-Text enthalten sein.

Mit Hilfe der Streaming-API ist es erstmals möglich, große Mengen nutzergenerierter Informationen unterschiedlichster Ausprägung direkt zu erhalten. Durch die Möglichkeiten die Twitter bietet kann theoretisch jeder Mensch Nachrichten und Informationen über das Twitter-Netzwerk verbreiten und weitergeben. Diese Masse an nutzergenerierten Informationen bietet Wissenschaftlern in verschiedenen Bereichen zahlreiche neue Möglichkeiten.

Sakaki et al interpretieren die Twitter-Kurznachrichten beispielsweise als Sensor-Daten [SOM10]. Der Twitter-Nutzer fungiert dabei als Sensor, der ein beliebiges Ereignis erfährt oder erlebt. Möglicherweise berichtet der Twitter-Nutzer im Tweet-Text über dieses Ereignis. Damit kann der Text als Sensor-Datum interpretiert werden, wenn auch erhebliches Rauschen in der Gesamtheit der Tweets zu erwarten ist. Sakaki et al zeigen aber, dass mit diesem Vorgehen, Erdbebenzentren lokalisiert oder die Trajektorie eines Typhoons vorhergesagt werden können.

Bild Twitter-Nutzer als sensor

Auch die Sozialwissenschaften und die Meinungsforschung profitieren von dem enormen Informationsfundus der durch Twitter geboten wird. Tumasjan et al. untersuchen in [TSSW11] wie sich die politische Landschaft im Twitter-Netzwerk wiederspiegelt. Die Wissenschaftler haben zur Bundestagswahl 2009 100.000 Tweets analysiert und stellten fest, dass die Erwähnungen von Parteien und Politikern in Twitter, den Wahlausgang sehr genau wiederspiegelten.

Die Kommunikation innerhalb des Twitter-Netzwerks kann aber auch neue Einsichten über die globale Kommunikation oder die Ausbreitung von Nachrichten liefern. Garcia-Gavilanes et al. erforschen in [GGMQ14] die Kommunikation zwischen Ländern. Es wird gezeigt, dass die globale Kommunikation innerhalb des Twitter-Netzwerks nicht nur von der geografischen Distanz abhängig ist, sondern auch von sozialen, ökonomischen und kulturellen Attributen eines Landes.

Selbst die Epidemieforschung kann von den Daten des Twitter-Netzwerks profitieren. So zeigten Szomsor et al. in [SKD11], dass die Vorhersage der Schweingrippe im Jahr 2009 durch die Analyse von Tweets eine Woche früher möglich gewesen wäre als dies mit konventionellen Frühwarnsystemen der Fall war.

Diese Erkenntnisse und Informationen sind allerdings nur gewinnbringend einzusetzen, wenn der Standort des Twitter-Nutzers bekannt ist. Die Information, dass eine Krankheit ausgebrochen ist, ist mit einer exakten Georeferenz wertvoller als ohne diese. Auch die Arbeit von Sakaki et al. ist auf eine Georeferenz angewiesen, wobei die Wissenschaftler angeben, dass die ungefähre Position für ihre Anwendung ausreichend ist. Bei der Untersuchung internationaler Kommunikation wiederum, ist es wichtig zu Wissen in welchem Land ein Tweet verfasst wurde. In diesem Fall kann die Georefrenz eine größere Region umfassen und muss nicht GPS-Genauigkeit aufweisen. Wohingegen eine detaillierte Untersuchung des politischen Klimas innerhalb Deutschlands eine Auflösung auf Bundesländer-Ebene erforderlich machen würde.

Twitter bietet seinen Nutzern die Möglichkeit ihren Standort im Nutzerprofil anzugeben. Hecht et al. stellen in [HHSC11] eine erste ausführliche Analyse der eingegebenen Standort-Daten bereit. Ab 2009 ermöglichte Twitter ein "per-tweet geo-tagging" [CCL10]. Dadurch können Anwendungen, auf Endgeräten mit GPS, Längen- und Breitengrad des aktuellen Standorts als Georeferenz an den Tweet anhängen. Nur ca. 1,7% der Twitter-Kurznachrichten enthalten allerdings eine konkrete Georeferenz in dieser Form. [4]

1.2 Problembeschreibung

Um gewinnbringende Informationen aus den Tweets erzeugen zu können, ist es wichtig Tweets eine Georeferenz zuordenen zu können. Die Anzahl der Twitter-Kurznachrichten die mit Hilfe von Längen- und Breitengrad unmittelbar einem geografischen Ort zugeordnet werden können ist sehr gering.

Es ist also wichtig ein Verfahren zu finden um Twitter-Nutzer oder Tweets eine Georeferenz zuzuordnen. Mit Hilfe der in einem Tweet vorhandenen Daten sollte eine möglichst

[4] Prüfung durch Datensatz XYZ was sich mit den Ergebnissen von [PCV13] und [SHP+13]

genaue Position bestimmt werden. Dies soll auch möglich sein, wenn keine konkrete geografische Angabe in Form von Längen- und Breitengrad vorliegt.

1.3 Fragestellungen und Anforderungen

Die folgenden Fragestellungen sollen beantwortet:

Q1 Wie kann Twitter-Nutzern eine Georeferenz zugeordnet werden?

1.3.1 Anforderungen

Das erarbeitete verfahren soll folgende Anforderungen erfüllen.

R1 Zuordnung einer Georefrenz zu einem Twitter-Nutzer. (R1)

R2 Unabhängig von kommerziellen Anbietern geografischer Informationen, oder sonstiger benötigter Daten. (R2)

R3 Das Ergebnis ist eine Georeferenz welche einer geografischen Hierarcheiebene enstpricht. Folgende Hierarcheiebenen werden angeboten (R3):

a) Land oder Staat

b) Administartionsebene erster Ordnung [5]

c) Administartionsebene zweiter Ordnung [6]

d) Stadt

R4 Es soll möglich sein eine Mindestanforderung für die Konfidenz, mit welcher die Georeferenz bestimmt wurde, anzugeben.

R5 Verfahren unabhängig von Sprache und Schriftzeichen weltweit einsetzbar.

[5] in D Bundesländer, bspsw. Baden-Württemberg, Bayern usw.
[6] in D Regierungsbezirke bspsw. Regierungsbezirk Stuttgart, Regierungsbezirk Karlsruhe usw.

1.4 Gliederung der Arbeit

Abschnitt 2: Grundlagen

In diesem Abschnitt sollen die Grundlagen für die entwickelte Methode vermittelt werden. Es wird auf den Mikroblogging-Dienst Twitter eingegangen und es werden grundsätzliche Methoden und Verfahren vorgestellt welche zum Verständnis der entwickelten Methode benötigt werden. Ebenso werden häufig genutzte geografische Grundbegriffe vermittelt.

Abschnitt 3: Stand der Technik

Es werden aktuelle Ansätze betrachtet, eingeordnet und in Bezug auf die angegebenen Anforderungen untersucht. Es werden sowohl die Verfahren zur 'Änalyse'ünd Zuordnung als auch die Verfahren zum abbilden der geografischen Einheiten untersucht und eingeordnet.

Abschnitt 4: Lösungsansatz

In diesem Kapitel wird, unter Berücksichtigung der gegebenen Anforderungen, ein Verfahren zur Lösung der Fragestellungen entwickelt. Um einen Überblick zu gewährleisten, wird das Verfahren zunächst allgemein betrachtet, danach wird jeder Verfahrensschritt dargelegt. Es wird gezeigt wie aus Tweet-Daten der Standort eines Twitter-Nutzers bestimmen werden kann. Dabei werden Methoden der Sprachverarbeitung, Statistik und geografische Hierarchien eingesetzt.

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 5: Referenzimplementierung der entwickelten Methode

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

Abschnitt 6: Leistungsbewertung der entwickelten Methode

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

Abschnitt 7: Schlussfolgerungen

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

Abschnitt 8: Zusammenfassung und Ausblick

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

2 Grundlagen

In den folgenden Abschnitten werden eine Reihe von Begriffen und Verfahren genutzt, die hier eingeführt werden sollen. Dies ermöglicht dem Leser die Beantwortung der Fragestellungen aus Abschnitt 4 nachzuvollziehen.

Zunächst werden einige geografische Grundbegriffe eingeführt. Danach wird auf Twitter eingegangen, es werden grundsätzliche Funktionen und Begriffe von Twitter eingeführt.

Zum Schluss wird die genutzte Datenbasis und der Einfluss der von Twitter genutzten Sampling-Strategie vorgestellt und erläutert.

2.1 Geografische Grundlagen und Begriffe

In diesem Kapitel sollen geografische Grundbegiffe erläutert werden. Einige geografische Begriffe werden in verschiedenen wissenschaftlichen Bereichen unterschiedlich genutzt und teilweise widersprüchlich definiert. Um Missverständnissen vorzubeugen wird hier definiert was in der vorliegenden Arbeit unter den einzelnen Begriffen zu verstehen ist. Eine Reihe von Begriffen wird selbst definiert um bestimmte Sachverhalte im Kontext dieser Arbeit klarer ausdrücken zu können.

2.1.1 Geografische Koordinaten

Geografische Koordinaten bestehen aus zwei Werten, einem Wert für den Längengrad und einem für den Breitengrad. Mit diesen zwei Werten kann eine Position auf dem Globus exakt bestimmt werden.

Die Längen- und Breitengrade beschreiben ein imaginäres Netz auf dem Globus. Dabei ziehen sich die Breitengrade wie ein Gürtel um den Globus. Der Breitengrad mit dem Wert 0, also 0 Grad Breite, verläuft entlang des Äquators. Die Längegrade hingegen verlaufen vom Nord- zum Südpol. Der Längegrad mit dem Wert 0 wurde 1884 durch ein Konsortium festgelegt, da vertikal des Globus keine natürliche Marke wie der Äquator, verläuft. Die Werte liegen im IT-Umfeld meist als Fliesskommazahlen vor und beschreiben jeweils einen Winkel.

2.1.2 Geodätisches Referenzsystem

Ein geodätisches Referenzsystem dient als einheitliche Grundlage zur Angabe einer Position auf dem Globus mit Hilfe geografischer Koordinaten. Dazu wird ein kartesisches Rechtssystem mit definierter Lage und Ausrichtung festgelegt. Die Lage und Ausrichtung erfolgt relativ zur Erde. Der Ursprung des Koordinatensystems liegt im Zentrum des Globus, meist im Masseschwerpunkt der Erde. Die Z-Achse zeigt meist in Richtung Nordpol und die X-Achse in Richtung 0 Grad Länge und 0 Grad Breite. Mit diesen zwei Werten ist die Lage eines kartesischen Rechtssystem eindeutig definiert. In diesem Koordinatensystem sind Referenzpunkte festgelegt. Diese Referenzpunkte werden benötigt um einen Referenzellipsoid zu verankern. Der Referenzellipsoid soll eine möglichst

genaue Approximation der Erde darstellen und diese im geodätischen Referenzsystem repräsentieren. Ein Punkt auf diesem Ellipsoid enstpricht damit einem Punkt auf der Erde.

Mit diesen Komponenten kann nun ein Punkt auf dem Ellipsoid eindeutig bestimmt werden. Der Längen und Breitengrad eines Punktes auf dem Ellipsoid lässt sich folgendermaßen bestimmen: Durch den Punkt auf dem Ellipsoid und dem Ursprung des Koordinatensystems kann ein Gerade gezogen werden. Der Breitengrad ist nun der Winkel zwischen dieser Linie und der Äquatorebene. Der Längengrad ist der Winkel zwischen der X-Achse, also dem Null Meridian und demjenigen Median welcher durch den Punkt geht. Durch eine Projektion der Punkte auf de Ellipsoid können diese Punkte beispielsweise auf einer Karte dargestellt werden.

Heutzutage ist das Referenzsystem WGS84 weit verbreitet.

[1]

2.1.3 Georeferenz

Eine Georeferenz (engl. Spatial Reference) wird auch als Raumbezug bezeichnet. Ist einem Datensatz eine Lage beziehungsweise Position zugeordnet so wird diese als Georeferenz bezeichnet. Die Art wie die Georeferenz angegeben wird und deren Genauigkeit hängen von den Anforderungen ab, die an die Georeferenz gestellt werden. Beispielsweise stellen die in Kapitel 1 erwähnten Anwendungen unterschiedliche Anforderungen an die Genauigkeit der Georeferenz.

Die Georeferenz lässt sich weiter unterteilen in: [1]

Direkte Georeferenz (direkter Raumbezug) Unter direktem Raumbezug versteht man die Angabe einer konkreten Koordinate bezüglich eines geeigneten geodätischen Referenzsystems.

Indirekte Georeferenz (indirekter Raumbezug) Unter indirektem Raumbezug werden alle Angaben verstanden die eine ungenaue Position bezüglich eines beliebigen

[1] Vergleiche Geoinformatik Lexikon der Universität Roctock: http://www.geoinformatik.uni-rostock.de/lexikon.asp und Vorlesungen zur Geo-Informatik von Prof. Dr.-Ing. Ralf Bill : http://www.geoinformatik.uni-rostock.de/vorlesungsthem.asp

Referenzysystems bestimmen. Ungenau ist in dem Sinne zu verstehen, dass die Angabe der Position auch eine Fläche beschreiben kann. Zusätzlich muss das gewählte Referenzsystem nicht zwingenderweise unveränderlich sein. Beispiele für die Angabe eines indirekten Raumbezugs wären Länder, Adressen, Postleitzahlen oder auch Telefonvorwahlen. Alle diese Angaben, mit Ausnahme der Adresse, definieren eine geografische Fläche. Diese Fläche ist nicht zwingenderweise klar abzugrenzen.

2.1.4 geografische Objekte

Ein geografisches Objekt, ist ein Objekt der Realwelt, dessen Position durch eine Georeferenz bestimmt werden kann. Die EN ISO 19110 Norm beschreibt ein geografisches Objekt folgendermaßen: "Geographische Objekte sind Erscheinungen der realen Welt, die einen Bezug zur Erde (Raumbezug) haben..." Es wird insbesondere nicht festegelegt ob es sich dabei um einen direkten oder einen indirekten Raumbezug handeln muss. Beispiele für geografische Objekte sind: Städte, Länder, Häuser oder auch Fahrzeuge.

2.1.5 Toponyme

Ein Toponym ist ein Name für ein geografisches Objekts. Beispiele hierfür sind: Städtenamen, Ländernamen oder Landschaftsnamen. Ein Toponym muss nicht eindeutig sein. Zu einem geografischen Objekt können mehrere unterschiedliche Toponyme existieren. Aus einem Toponym kann, eine Georeferenz abgeleitet werden.

2.1.6 Georeferenzierung

Unter Georeferenzierung versteht man die Zuordnung einer Georeferenz zu einem Datensatz. Also den Vorgang einem Datensatz, zum Beispiel einem Twitter-Nutzer eine Georeferenz zuzuordnen. [1] In diesem Sinne stellen die Datensätze geografische Objekte dar, da diese einen Bezug zur Erde haben können.

2.1.7 Ortsverzeichnis

Sehr umfangreich mal schauen. Ein Orstverzeichnis beinhaltet eine Liste von Werten, meist Zeichenketten, welchen eine

2.1.8 Geografische Position

Unter einer geografischen Position soll hier eine Position auf dem Globus verstanden werden deren Wert durch geografische Koordinaten angegeben wird.

2.1.9 Geografische Region

Unter einer geografischen Region werden hier Flächen auf dem Globus verstanden. Diese können nicht durch einen einzelnen Punkt beschrieben werden. Flächen werden üblicherweise durch eine Menge von Punkten beschrieben. Die Punkte liegen als geografische Koordinaten vor und werden durch Geraden verbunden. Die so eingeschlossene Fläche innerhalb des Polygons beschreibt dann die geografische Region. Bundesländer oder Länder sind Beispiele für geografische Regionen. Somit entspricht der Begriff geografische Region einer indirekten Georeferenz (indirektem Raumbezug).

2.1.10 Geografische Hierarchie

In der vorliegenden Arbeit wird eine geografische Hierarchie verwendet um eine Einteilung der Erde in geografische Regionen umzusetzen. Dabei können geografische Regionen wiederum geografische Regionen oder geografische Positionen enthalten, wodurch sich eine hierarchische Gliederung ergibt. Diese Einteilung spiegelt im wesentlichen die Einteilung der Erde in Staaten und deren individuellen Verwaltungseinheiten wieder. Im vorliegenden Fall ist das Staatsgebiet, also die Fläche über die sich der Staat erstreckt, von Interesse. [2]

[2] Zur genauen Definition eines Staatsgebietes vergleiche [JJ21]

Im Gegensatz dazu könnte die Erde auch in ein Gitternetz eingeteilt werden. Die einzelnen Zellen würden dann als Referenz für eine geografische Region verwendet werden. Dieses vorgehen wird unter anderem in [SMvZ09] angewendet.

Eine Aufteilung der Erde in geografische Regionen lässt sich auf oberster Ebene mit Hilfe von Ländern und deren Grenzen umsetzen. Daraus resultiert eine Einteilung, welche direkt intuitiv verständlich ist und vielen Anforderungen an geografische Informationen gerecht wird. Die meisten Länder sind in weitere administrative Einheiten aufgeteilt. Diese geografischen Regionen werden hier als Administartionsebenen bezeichnet. Es wird zwischen Administrationsebenen erster und zweiter Ordnung unterschieden. Ausnahmen sind besipielsweise Stadtstaaten wie der Vatikan-Staat oder das Fürstentum Monaco, welche aufgrund ihrer Größe nicht in Verwaltungsbezirke unterteilt werden und keine Städte. Des weiteren werden in der untersten Ebene der Hierarchie Städte dargestellt.

Wenn man als Beispiel Deutschland heranzieht, ergibt sich eine Einteilung wie in Bild 2.1 dargestellt wird. [3] Die oberste Ebene beschreibt das Land worauf die zweite Ebene die Bundesländer darstellt. Auf der dritten Ebene werden die Regierungsbezirke abgebildet, worauf die Städte in der letzten Ebene folgen. Analog kann die Einteilung für die USA vorgenommen werden, woraus sich die Hierarchie Country->State->County->City ergibt.

Bis auf die letzte Ebene wird den Objekten in der Hierarchie eine geografische Region zugeordnet. Lediglich die unterste Ebene, die der Städte, wird durch eine geografische Position exakt beschrieben. Die Ausdehnung einer Stadt wird in der gegebenen Hierarchie also nicht berücksichtigt. Jede Stadt wird als konkrete geografische Position mit Koordinaten repräsentiert.

Ortsverezichnis Oder auch Gazetteer genannt. Auch auf Geocoding API eingehen.

2.2 Twitter

In diesem Kapitel werden grundlegende Begriffe rund um das Twitter-Netzwerk erläutert. Weiter werden die Mechanismen in Twitter erläutert und an praktischen Beispielen

[3] Aus Platzgründen sind im Bild pro Ebene nur einige wenige geografische Objekte aufgezählt.

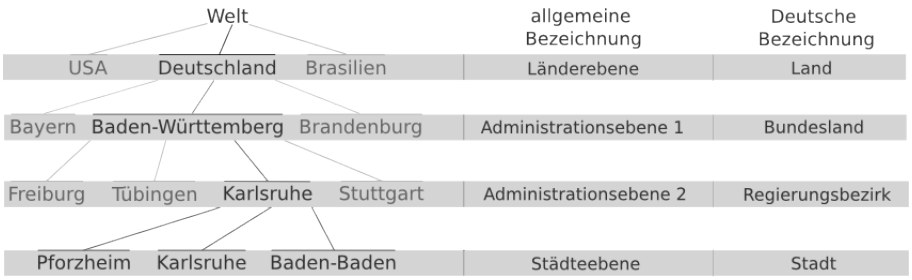

Abbildung 2.1: Die Hierarchiebenen exemplarisch. Bis in die Städtebene wird nur der Pfad Welt -> Deutschland -> Baden-Württemberg -> Karlsruhe ->K arlsruhe,Pforzheim,Baden-Baden dargestellt.

erklärt. Zum Schluss wird aufgezeigt welche Informationen pro Tweet übermittelt werden und welche Daten zur Lokalisierung verwendet werden können.

2.2.1 Geschichtliches

Twitter wurde 2006 von Jack Dorsey, Biz Stone, Noah Glass und Evan Williams gegründet. Ursprünglich war Twitter zur internen Kommunikation innerhalb der Firma Odeo geplant. Schnell wurde allerdings klar, dass in dem Dienst mehr potenzial steckt und so wurde Twitter öffentlich gemacht. Seitdem erfreut sich der Dienst einer wachsenden Nutzer-Gemeinde. Die Twitter-Gründer haben von Anfang an keine exakten Nutzer-Zahlen oder die Anzahl der versendeten Twitter-Kurznachrichten bekanntgegeben. Dies geschah einerseits, weil die Gründer davon überzeugt sind, dass anhand der reinen Nutzer-Zahlen und gesendeten Twitter-Kurznachrichten nicht die '"Gesundheit“ des Twitter-Netzwerks nachvollzogen werden kann, andererseits werden durch diese Massnahme auch strategische Ziele verfolgt. [4] 2013 ging Twitter an die Börse und vermeldete 100 Millionen täglich aktive Nutzer und über 500 Millionen Twitter-Kurznachrichten, die täglich über den Dienst versendet werden.

2.2.2 Was ist Twitter?

Twitter wird als Kurznachrichten-Dienst, Mikroblogging-Dienst oder auch als soziales-Netzwerk bezeichnet. Twitter Geschäftsführer Kevin Thau hat 2010 auf dem Nokia-

[4]http://www.pbs.org/mediashift/2007/05/twitter-founders-thrive-on-micro-blogging-constraints137

World Kongress öffentlich bestritten, dass Twitter ein Soziales-Netzwerk ist. Laut Thau handelt es sich um ein Nachrichten-, Inhalts- und Informations-Netzwerk. Er begründete dies damit, dass Twitter die Art und Weise wie Nachrichten verteilt werden geändert hat und praktisch jeder zum Journalisten werden kann. Als Beispiel nennt er die Landung des Fluges 1549 auf dem Hudson River. Die Augenzeugen hätten damals keine Mails versendet um die Nachricht zu verbreiten, sondern die Nachricht via Twitter weitergegeben. Es lassen sich eine Reihe weitere Beispiele derselben Art finden. In [POM+13] wird ein Vergleich zwischen sogenannten Newswire anbietern und Twitter gezogen. [5] Es stellte sich heraus, dass über nahezu alle Nachrichten, welche in den Newswires verbreitet wurden auch im Twitter-Netzwerk berichtet wird. Nachrichten zu bestimmten, vermutlich sehr speziellen Themen oder Auslandsnachrichten wurden ausschliesslich in Twitter gefunden. Diese Erkenntnisse decken sich mit der Einschätzung von Kevin Thau. In [KLPM10] wird die Einschätzung, bei Twitter handele es sich nicht um ein soziales Netzwerk, wissenschaftlich bestätigt. Kwak et al überprüfen die in [NP03] beschriebenen Eigenschaften sozialer Netzwerke und kommen zu dem Schluss, dass Twitter diese Eigenschaften nicht erfüllt.

Die Bezeichnung Kurznachrichten-Dienst ist irreführend, da dieser mit sms (small messeneger service) in Verbindung gebracht werden kann. Tatsächlich galt der sms in der Anfangsphase von Twitter als Vorbild für den Dienst. In Twitter werden Nachrichten allerdings standardmäßig allen Benutzern zur Verfügung gestellt und können eingesehen werden. Des weiteren wird eine Liste der Nachrichten, welche von einem Nutzer verfasst wurden, als Liste in umgekehrter chronologischer Reihenfolge auf dessen Profil dargestellt. Damit ähnelt das Twitter-Profil einem Blog mit Einträgen deren Länge 140 Zeichen nicht überschreiten darf. Die Darstellung als Liste, und die Funktion einen Tweet standardmäßig allen Nutzern freizugeben unterscheidet sich grundlegend von der Funktion des sms, bei dem eine Nachricht direkt an einen Empfänger gesendet wird und nicht öffentlich ist. Im sms steht die Konversation zweier Nutzer im Vordergrund, wohingegen Nachrichten im Twitter-Netzwerk einen Brodcast an alle Nutzer darstellen.

Die 140 Zeichen langen Nachrichten in Twitter werden als Tweets bezeichnet. Tweet bedeutet übersetzt Zwitschern, womit die Redenwendung '"Die Spatzen zwitschern es

[5] Newswire stellt eine Art Nachrichtenaggregator dar, über welchen Nachrichten aus verscheidenen Quellen aggregiert und weitergegeben werden. In deutschland kommt die Deutsche Presse agentur diesem Konzept am nächsten.

von den Dächern“ auch im Twitter-Netzwerk zu einer passenden Redenwendung wird. In der vorliegenden Arbeit wird Twitter deshalb als Mikroblogging-Dienst bezeichnet.

2.2.3 Funktionen von Twitter

Der Mikroblogging-Dienst Twitter bietet neben dem Profil, auf dem die Tweets des Nutzers angezeigt werden, noch eine Reihe weiterer Funktionen. Im folgenden soll das Twitter-Profil und die Timeline kurz erläutert werden. Eine der zentralen Funktionen von Twitter ist das sogennante Folgen, womit sich Nutzer win Netzwerk aufbauen können aus dem sie Twitter Nachrichten erhalten. Danach werden Funktionen wie das weitergeben eines Tweets, Favorisieren und Antworten erklärt. Zum Schluss wird auf den gesendeten Tweet Inhalt eingenagen und der Netzwerk-Charakter von Twitter untersucht.

Abbildung 2.2: Die Twitter-Timeline auf einem Twitter Profil. 1: Nutzername und Informationen über den Nutzer. 2: Profilbild 3: Allgemeine Informationen über den Benutzer und dessen Netzwerk 4: Nutzer-Timeline: Tweets des Nutzer in umgekehrter chronologischer Reihenfolge 5: Button zum Folgen

Das Nutzer-Profil und die Nutzer-Timeline Das Nutzer-Profil kann über die Url http://twitter.com/BENUTZERNAME abgerufen werden und bietet neben der Nutzer-Timeline, in der die Tweets des Nutzers angezeigt werden, eine Reihe an weiteren Informationen. In Abbildung 2.2 ist in der mitte die Timeline des Benutzers dargestellt in der dei Tweets zu sehen sind. Unter dem Profilbild links sind Informationen des Nutzers aufgelistet. Diese Informationen kann der Nutzer selbst einstellen und entscheiden welche er angeben möchte.

Folgen (Following/Follower/Tweeps) Diese Funktion erlaubt es Tweets eines bestimmten Nutzers zu abonnieren. Im Twitter-Umfeld spricht man von '"following“ oder '"folgen“, wenn man die Tweets eines bestimmten Nutzers abonniert. Hat man Tweets eines bestimmten Nutzers abonniert so wird man als dessen '"Follower“ bezeichnet. Das englische Wort '"Follower“ hat sich im Twitter-Umfeld und darüber hinaus eingebürgert und wird selten übersetzt. Auch auf der Twitter Website wird '"Follower“ nicht ins deutsche übersetzt. In der vorliegenden Arbeit wird deshalb auch auf eine Übersetzung verzichtet.

In Abbildung 2.2 an Position 3 wird unter '"Folge ich“ die Anzahl der Twitter-Nutzer angezeigt denen der Beispielnutzer folgt. Neben dem Feld '"Folge ich“ wird unter '"Follower“ angezeigt wieviele Nutzer dem Beispielnutzer folgen.

Persönliche Timeline Jeder Twitter-Nutzer hat seine persönliche Timeline, auf dieser werden die Tweets derjenigen Nutzer angezeigt, denen er folgt. Die Timeline kann als Aggregation von Tweets betrachtet werden. Diese Timeline ist die zentrale Stelle, an der die Nutzer Tweets anderer Nutzer empfangen und lesen. Auch hier werden die Tweets in umgekehrter chronologischer Reihenfolge angezeigt.

Weiterleiten eines Tweets (Retweet) Unter einem Retweet versteht man das weiterleiten eines Tweets den man nicht selbst verfasst hat an die eigenen Follower. Genauer gesagt wird der Tweet übernommen und ein Hinweis hinzugefügt, dass es sich um einen sogenannten Retweet handelt, und nicht einen vom Nutzer selbst verfassten Tweet. Diese Funktion wird hauptsächlich genutzt um Nachrichten schnell zu verbreiten ohne diese

neu eingeben zu müssen. Die Weitergabe an die eigenen Follower impliziert einen gewissen Grad an Kontrolle und Filterfunktion. Der weitergebende Nutzer kontrolliert und filtert die Nachrichten die er erhält und gibt diejenigen weiter, denen er eine Gewisse Relevanz beimisst, oder von denen er erwartet, dass sie seine Follower interessieren. Mit dieser Funktion können einzelne Nutzer eine Art Filterfunktion übernehmen, welche früher Journalisten vorbehalten war. Es darf jedoch nicht vergessen werden, dass der Nutzer nur im Rahmen seiner eigenen Möglichkeiten einen Tweet verifizieren kann und Nachrichten in Twitter keinesfalls gesicherte Fakten darstellen. Auch können Nutzer durch diese Funktion zu Tweet-Aggregatoren werden, welche Tweets von mehreren Nutzern erhalten oder sammeln, aber nur relevante oder themenspezifische Tweets weitergeben.

1 retweet Reichweite 1000 Nutzer

Diagramm Retweet, Filterfunktion

Hashtags Hashtags werden genutzt um Tweet Nachrichten zu kategorisieren oder Metag Informationen zu liefern. Ein Hashtag kann vom verfasser selbst als solches ausgezeichnet werden indem ein # vor das gewünschte Wort, welches als Hashtag fungieren soll, gesetzt wird. Hahstags ermöglichen es Tweets nach Stichworten zu filtern. Anhand der Hashtags werden auch die Twitter-Trends analysiert. Twitter Trends

Antworten und direktes ansprechen eines Nutzers Twitter bietet die Möglichkeit einzelne Nutzer direkt anzusprechen. Mit Hilfe des @-Symbols kann ein Nutzer referenziert werden. Der referenzierte Nutzer, besipielsweise @alfred, wird dann benachrichtigt, dass er in einem Tweet erwähnt wurde. Der erwähnte Nutzer muss dabei nicht Follower des Verfassers sein. Eine weitere Funktion im Twitter-Netzwerk ist das Antworten auf einen Tweet. Über eine Schaltfläche wird es ermöglicht auf einen Tweet zu Antworten. Das @-Symbol und der Nutzername des Verfassers werden automatisch eingetragen, womit eine Benachrichtigung an den Verfasser des Ursprungstweets erfolgt. Es ist möglich, das auf einen Antwort-Tweet wiederum geantwortet wird, wodurch ein sogenannter Thread oder Konversation entsteht. Auch ist es möglich, dass an einer solchen Konversation mehrere Twitter-Nutzer beteiligt sind. Dies ist dann der Fall, wenn im ursprünglichen Tweet, auf weitere User referenziert wurde. Aber auch wenn ein Nutzer auf eine bestehende Konversation antwortet, werden alle beteiligtien Nutzer referenziert.

siehe Bild ref1

siehe Bild ref2

siehe Bild ref3

siehe Bild ref4

Diagramm Antwort, Antwort Thread, Bild Antworten Button, Referenzieren

Favorisieren Mit dieser Funktion lässt sich ausdrücken, dass man einen Tweet interessant oder gut findet. Auch Zustimmung wird durch favorisieren ausgedrückt. Einen Tweet zu favorisieren kann aber auch bedeuten ”ich habe deine Reaktion registriert“, oft um einen Antwort-Thread nicht abrupt abzubrechen sondern eine zustimmende Rückmeldung zu geben ohne extra einen Tweet zu verfassen.

2.2.4 Daten einer Twitter-Nachricht

Neben den direkt sichtbaren Informationen enthält ein Tweet eine Reihe weiterer Daten. Betrachtet man einen einzelnen Tweet, beispielsweise auf twitter.com, wird der Tweet-Text, der Verfasser und die Zeit, wann der Tweet verfasst wurde, mitgeteilt. Die Gesamtheit der Daten die in einem Tweet enthalten sind werden hier allgemein als Tweet-Daten bezeichnet.

siehe Bild

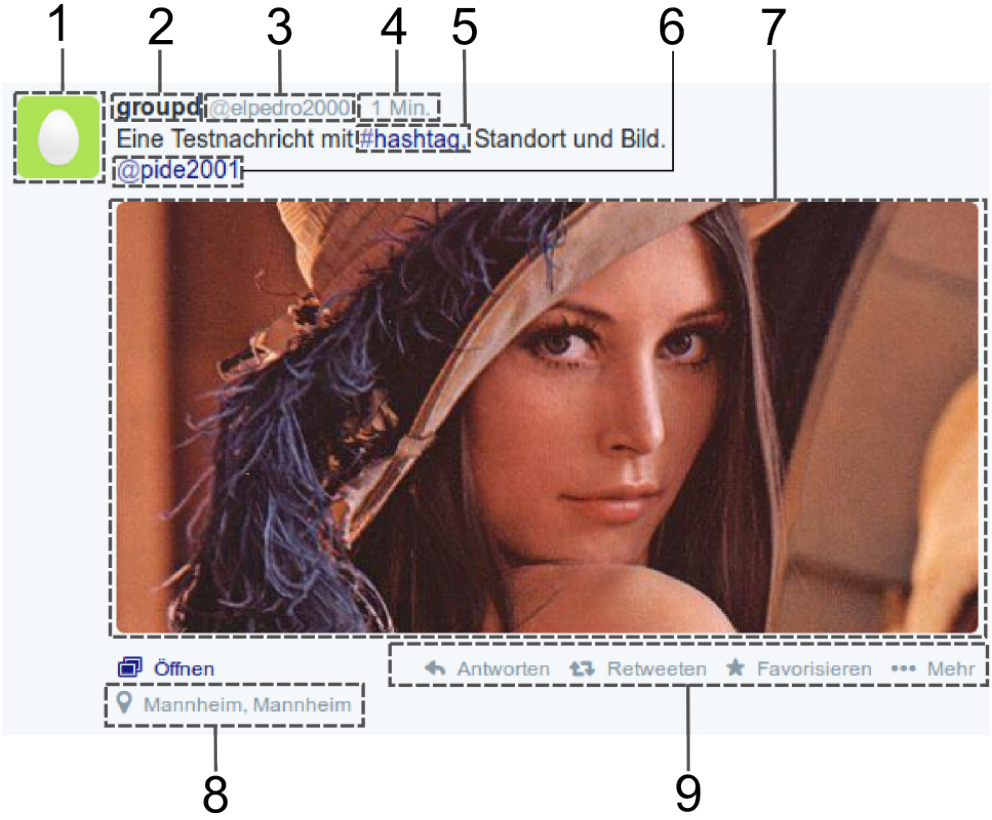

Abbildung 2.3: Was ist zu sehen?

Koordinaten In den Tweet-Daten können geografische Koordinaten in Form von Längen- und Breitengrad angegeben sein. Diese Koordinaten zeigen an wo sich der Verfasser befand als er den Tweet abgesetzt hat. Wenn diese Koordinaten angegeben sind hat der Nutzer explizit zugestimmt, dass die Koordinaten seines aktuellen Aufenthaltsortes dem Tweet angehängt werden. Die Bestimmung der Koordinaten und das anhängen der Koordinaten an einen Tweet werden vollautomatisch durch das Programm übernommen mit welchem der Tweet verfasst wurde. Auf Smartphones wird meist das integrierte GPS-Modul genutzt um die Koordinaten zu bestimmen. Bei der Nutzung an einem PC wird der Tweet häufig über den Browser verfasst und die Position mit Hilfe von GeoIp ermittelt.

Daten Neben den sichtbaren Daten, welche in der Timeline angezeigt werden, enthält ein Tweet eine Reihe weiterer interessanter Informationen.

2.2.5 geografischer Indikator

Unter einem geografischen Indikator wird eine Angabe verstanden, welche direkt einem Nutzer zugeordnet werden kann und die Auskunft über die geografische Position oder Region des Nutzers geben kann. Im Zuge dieser Arbeit wurden potentielle geografische Indikatoren untersucht und eine Reihe von Eigenschaften identifiziert anhand derer sich geografische Indikatoren kategorisieren lassen. Diese Eigenschaften haben Einfluss darauf, wie und ob eine Georeferenz aus dem Indikator abgeleitet werden kann. Dabei ist zu unterscheiden ob sich die Eigenschaft auf den, durch den Nutzer eingegebenen, Wert bezieht oder auf die Information die durch die Angabe geliefert werden soll.

Objektivität der Werte geografischer Indikatoren

Der Wert eines geografischen Indikators ist genau dann objektiv wenn zwei Nutzer für denselben geografischen Ort oder dieselbe geografische Region immer den selben Wert eingeben. Ein Beispiel für einen objektiven geografischen Indikator wäre eine Liste von Ländern aus der ein Nutzer wählen kann. Der Nutzer hat dabei eine Wahl, kann aber nur aus einem begrenzten Anzahl an Möglichkeiten wählen.

Geben zwei Nutzer unterschiedliche Werte ein, obwohl sie denselben Ort oder dieselbe Region beschreiben wollen, ist dieser Wert nicht objektiv.

Zuverlässigkeit der Werte geografischer Indikatoren

Ein Wert ist genau dann zuverlässig wenn er in jedem Fall die Information enthält, welche durch das Feld repräsentiert werden soll. Unzuverlässig ist der Wert, wenn er nicht in jedem Fall die Information enthält welche druch das Feld repräsentiert werden soll.

Gesicherte Werte geografischer Indikatoren

Als gesichert gilt ein Wert genau dann wenn die enthaltene Information in jedem Fall dem tatsächlichen Wert entspricht. Im Umfeld von Twitter ist diese Eigenschaft nicht stichhaltig nachzuprüfen. Alle Angaben die ein Benutzer eingibt werden nicht verfiiziert und können dementsprechen auch nicht gesichert sein. Es besteht die Möglichkeit das ein Nutzer in jedem Feld Falschangaben macht. Dies gilt es im erarbeiteten Verfahren zu beachten.

unmittelbare geografische Indikatoren in Tweet-Daten

Unmittelbare geografische Indikatoren sind solche aus denen direkt eine geografische Position abgeleitet werden kann.

bzw. die Intention des Nutzers bei der Eingabe darauf abzielt eine geografische Position zu beschreiben. Mit einer Gewissen Sicherheit kann ein geografischer Bezug abgeleitet werden, da die Intention des Feldes einen Standort angibt.

2.2.6 Geoinformationen in Twitter Daten

Überarbeiten sub-sec:Geoninformati in Twitter Daten

Welche Tweet-Daten können zur Georeferenzierung herangezogen werden

Nur optional

Um diese Frage zu beantworten, müssen die Tweet-Daten eingehend untersucht werden. Dabei spielt nicht nur die reine Information die den Daten entnommen werden kann eine Rolle, sondern auch wie die Daten generietr oder eingegeben wurden. Beispielsweise kann bei einem Tweet, dem ein Längen- und Breitengrad mit einer Genauigkeit von 14 Nachkommastellen zugeordnet ist, davon ausgegangen werden, dass die geografische Position der tatsächlichen geografischen Position, von welcher der Tweet abgesetzt wurde, entspricht. Es liegt hier die Vermutung nahe, dass diese Werte durch ein mobiles GPS [6] erfasst worden sind. Anders verhält sich dies beispielsweise beim Tweet-Text, eine Erwähnung der Stadt New York, muss nicht bedeuten, dass der Tweet aus dieser Stadt stammt. Es impliziert nicht einmal, dass der Verfasser jemals in dieser Stadt war. Im folgenden werden einige Datenfelder, welche mit jedem Tweet versandt werden, untersucht. Dabei wird die Eignung dieser Daten als geografischer Indikatoren bewertet. Währenddessen werden anhand geeigneter Beispiele die Bergiffe gesicherter -, ungesicherter -, mittelbarer - und unmittelbarer geografischer Indikator eingeführt.

mögliche geografische Indikatoren

Nutzer-Standort Der Nutzer-Standort ist ein unmittelbarer geografischer Indikator. Als Nutzer-Standort kann der Twitter-Nutzer eine beliebige Zeichenfolge eingeben. Es handelt sich beim Nutzer-Standort deshalb um einen ungesicherten geografischen Indikator, es ist deshalb damit zu rechnen, dass unter Umständen keine geografische Position angegeben ist und andererseits keine einheitliche Angabe bezüglich des selben Standorts erwartet werden kann. Beispielsweise beschreiben die Zeichenketten “Karlsruhe, Deutschland“ und “Baden-Württemberg, Karlsruhe“ den selben Ort. Noch deutlicher wird dieser Umstand, wenn man alternative Namen oder umgangssprachliche Namen für Städte betrachtet. Mit “The Big Apple“ und “New York, USA“ oder mit “Motown“ und “Detroit, MI“ sind dieselben Orte gemeint. Auch die Genauigkeit bezüglich der geografischen Position ist nicht zuverlässig vorhersagbar, sehr konkrete geografische Positionen, wie die

[6] Global Positioning System

Angabe einer Stadt oder eines Stadtteils, oder aber eine geografische Region wie beispielsweise ein Land oder ein Kontinent, sind möglich.

Nutzer-Zeitzone Die Nutzer-Zeitzone stellt dagegen einen gesicherten, unmittelbaren geografischen Indikator dar. Bei der Nutzer-Zeitzone kann aus einer Liste möglicher Werte gewählt werden, womit keine Ungenauigkeiten bezüglich der Eingabe besteht und eine definierte Zeichenkette erwartete werden kann, deren geografische Region klar definiert ist. Die Nutzer-Zeitzone beschreibt allerdings in jedem Fall eine größere geografische Region, die nicht immer mit den konventionellen Ländergrenzen korrespondiert und somit eine Bestimmung der geografischen Position nahezu unmöglich macht.

Bei beiden Indikatoren besteht natürlich die Möglichkeit der Falscheingabe durch den Benutzer. Dieser Umstand wird jedoch durch die Analyse der Daten ausgemerzt.

Irgendwo auf den Umstand eingehen, dass Timezone nicht angegeben werden wird und dann der Standard gewählt wird der us central pacific time ist? .....

Umschreiben und woanders darauf eingehen!

3 Stand der Technik

Die Georeferenzierung von Tweets oder Twitter-Nutzern ist ein Feld an dem nach wie vor aktiv geforscht wird. Nicht zuletzt trägt auch die große Verfügbarkeit an Twitter-Daten zu dem Umstand bei, dass Twitter in den letzten Jahren Forschungsgegenstand zahlreicher Publikationen war.

In diesem Abschnitt sollen bestehende Ansätze zur Georeferenzierung im Twitter-Umfeld untersucht werden. Es werden Kriterien zur Einordnung der bestehenden Ansätze erarbeitet und erläutert. Die Arbeiten werden mit Hilfe der Kriterien schematisch eingeordnet um einen Überblick zu erhalten. Zum Schluss wird untersucht ob die Arbeiten die bereits formulierten Anforderungen aus 1.3.1 erfüllen, und wie sich die vorliegende Arbeit von den bestehenden Ansätzen abgrenzt.

3.1 Kategorisierung bestehender Ansätze

In früheren Arbeiten wurde bereits versucht, eine Einordnung der bestehenden Verfahren vorzunehmen. Es ist interessant die Kategoriesierungsansätze und die verwandten Arbeiten einiger Autoren zu studieren. Es lässt sich dadurch die Entwicklung zum Thema Lokalisierung im Twitter-Umfeld beobachten. Einige Kategorisierungsansätze werden im folgenden aufgelistet und erläutert.

Sowohl in [HHSC11] als in [CCL10] beschränken sich die verwandten Arbeiten nicht auf die Lokalisierung im Twitter-Umfeld, es werden Arbeiten zur Lokalisierung von Web-Inhalten im Allgemeinen aufgelistet. Dies lässt darauf schliessen, dass sich vor den Jahren 2010/2011 nur wenige Arbeiten mit der Lokalisierung im Twitter-Umfeld beschäftigt haben.

Kategorisierung über die untersuchte Ressource

[HHSC11] nimmt deshalb eine Kategorisierung anhand der untersuchten Ressource vor. Es wird unterschieden zwischen Forschungen zur "Lokalisierung von Microblogging-Seiten und deren Inhalten" und der "Lokalisierung von Nutzern, welche Inhalte zu Web 2.0 Seiten beisteuern". Zusätzlich wird in dieser Arbeit das "Verhalten der Nutzer im Umgang mit der Veröffentlichung ihres aktuellen Standorts" und die "Vorhersage privater Informationen" betrachtet. Darauf soll hier allerdings nicht weiter eingegangen werden.

Kategorisierung über die verwendete Methode

[CCL10] klassifiziert die vorgestellten Arbeiten anhand der verwendeteten Methodik. Es wird auf Arbeiten zur Lokalisierung von Webseiten, Web-Logs, Suchanfragen und Web-Nutzern verwiesen. Diese werden in die folgenden drei Kategorien eingeteilt.

"Inhaltsanalyse mit Begriffen in einem geografischen Verzeichnis (Content analysis with terms in a gazetteer)" Es wird darunter eine einfache Datenbanksuche verstanden. Es werden einzelne Wörter in einer Datenbank nachgeschlagen um diese einem konkreten geografischen Ort zuweisen zu können. Dabei kann sowohl lokal auf eine Geo-Datenbank als auch auf Internet Ressourcen zurückgegriffen werden. In der Regel durchläuft der untersuchte Text eine manuelle oder automatische Vorverarbeitung um potenziell geografische Begriffe, sogenannte Toponyme, herauszufiltern.

"Inhaltsanalyse mit probabilistischen Sprachmodellen (Content analysis with probabilistic language models)" Dabei werden Texte oder Textteile einer Twitter-Kurznachricht zu vordefinierten geografischen Regionen wie Ländern oder Städten zugeordnet. Nach einer Vorverarbeitung des Textes erfolgt eine statistische Auswertung, um danach den Text oder einzelne Textteile, wie beispielsweise Wörter, einer geografischen Region zuzuordnen. Eine unbekannter Text kann dann mit Hilfe der zuvor gelernten Zuordnung einer geografischen Region zugeordnet werden.

“Schlussfolgerungen durch soziale Verbindungen (Inference via social relations)“ es werden soziale Verbindungen, die in Netzwerken abgebildet sind, herangezogen um Rückschlüsse auf den geografischen Ort des untersuchten Inhaltes oder einer Person ziehen zu können.

Preidhorsky et al. schlagen in [PCV13] eine weitere Einteilung anhand der Methodik vor. Allerdings werden hier ausschließlich Arbeiten im Twitter-Umfeld betrachtet.

“Geocoding“ Im wesentlichen entspricht dies der “Inhaltsanalyse mit Begriffen in einem geografischen Verzeichnis“ aus [CCL10]. “Geocoding“ wird als Begriff in vielen Fachrichtungen unterschiedlich definiert, was zu Missverständnissen führen kann. In [Gol08] wird genauer auf den Begriff des Geocoding und die Poblematik eingegangen und eine Definition des Begriffs vorgeschlagen. Im vorliegenden Kontext ist es präziser und weniger missverständlich die Methodik als “Inhaltsanalyse mit Begriffen in einem geografischen Verzeichnis“ zu bezeichnen, anstatt den Begriff “Geocoding“ einzusetzen.

“Geografische Themenmodelle (geografic Topic Modeling)“ wird definiert als die Verbindung von “Themenmodellierung“ und “Standorterkennung (Location Awareness)“. Durch klassisches “Themenmodellierung“ lässt sich aus aus Texten eine Menge von Themen extrahieren. Durch eine Lernphase werden Wörterbücher zu den Themen erstellt. Mit Hilfe dieser Themen-Wörterbücher kann später das Thema eines Textes bestimmt werden. [BNJ12] Unter “Standorterkennung“ wird hier verstanden, dass nicht nur das Thema sondern auch eine bestimmte Region extrahiert werden kann. Dies kann durch geografischen Koordinaten in Twitter-Kurznachrichten realisiert werden. Im Unterschied zur Kategorie “Inhaltsanalyse mit probabilsitischen Sprachmodellen“ aus [CCL10] wird hier jedoch keine vorgegebene geografische Region gefordert. Vielmehr ergeben sich die geografischen Regionen aus den Themenmodellen und den zugehörigen geografischen Koordinaten. Es wird damit eine kontinuierliche Region beschrieben, welche nicht zwangsweise durch Stadt-, Staaten- oder Ländergrenzen beschränkt ist.

“Statistische Klassifizierung (Statistical classifiers)“ Diese Kategorie entspricht der “Inhaltsanalyse mit probabilsitischen Sprachmodellen“ wobei in [CCL10] nur eine Arbeit in dieser Kategorie betrachtet wird. [PCV13] listet mehrere Arbeiten auf, die sich in diese Kategorie einordnen lassen.

"Informationen aus sozialen Verbindungen (Social Network Information)" analog zu "Schlussfolgerungen durch soziale Verbindungen" aus [CCL10] werden soziale Verbindungen herangezogen um den Standort zu bestimmen.

Priedhorsky et al. wählen eine ähnliche Einteilung wie vormals Cheng et al. in 2010, die verwandten Arbeiten stammen allerdings aus dem Twitter-Umfeld. Dabei ist zu bemerken, dass sich die verwendeten Methoden zur Lokalisierung im Twitter-Umfeld nicht wesentlich von denen in anderen Bereichen unterscheiden. Um die Arbeiten im Twitter-Umfeld sinnvoll voneinander abgrenzen zu können muss die Kategorisierung mehr Dimensionen umfassen. Es müssen mehr Kriterien zur Kategorisierung herangezogen werden als die reine Methodik.

Mahmud et al. betrachten in [MND12] hauptsächlich Arbeiten im Twitter-Umfeld. Diese werden in die folgenden Kategorien unterteilt.

1. "Inhaltsbasierte Standortschätzung von Tweets (Content-based Location Estimation from Tweets)"
2. "Inhaltsbasierte Standortextrahierung von Tweets (Conetnt-based Location Extraction from Tweets"
3. "Standortschätzung ohne den Tweet Inhalt zu nutzen (Location Estimation without using Tweets Content)"

"Inhaltsbasierte Standort-Schätzung von Tweets (Content-based Location Estimation from Tweets)" hier wird die geografische Position durch eine Inhaltsanalyse der Twitter-Kurznachricht geschätzt. Die Schätzung erfolgt dabei durch probabilistische Modelle. Diese Kategorie vereint damit "Geografische Themenmodelle", "Statistische Klassifizierung" aus [PCV13] mit "Inhaltsanalyse mit probabilsitischen Sprachmodellen" aus [CCL10] und ist damit als genereller anzusehen, als die vorgenannten Kategorien.

"Inhaltsbasierte Standort-Extrahierung von Tweets (Content-based Location Extraction from Tweets" die verwandten Arbeiten in dieser Kategorie versuchen direkte Hinweise auf einen geografischen Ort aus einer Twitter-Kurznachricht zu extrahieren. Diese Kategorie ähnelt dem "Geocoding" beziehungsweise der "Inhaltsanalyse mit Begriffen in einem geografischen Verzeichnis".

"Standortschätzung ohne den Tweet Inhalt zu nutzen (Location Estimation without using Tweets Content)" hierunter versteht der Autor alle Informationen die nicht unmittelbar im Tweet-Text enthalten sind. Dazu zählen Informationen aus dem Nutzerprofil oder Informationen über die sozialen Verbindungen des Nutzers.

[MND12] nutzt ebenfalls die Methodik um die Arbeiten zu kategorisieren. Allerdings wird hier eine generellere Einteilung vorgenommen. So wird unterteilt, ob der Standort geschätzt oder extrahiert wurde. Mahmud et al. bringen aber auch eine weitere Dimension ein. Es wird hier zusätzlich unterschieden ob das angewendete Verfahren den Tweet-Inhalt nutzt oder andere Informationen.

Dies ist sinnvoll, denn die genannten Methoden lassen sich sowohl auf den Tweet-Inhalt als auch auf andere Informationen, beispielsweise aus dem Nutzerprofil, anwenden.

Frühere Arbeiten verweisen auf ein weiteres Spektrum an Arbeiten aus anderen Bereichen, wie Lokalisierung von Flickr Bildern oder Web-Log Einträgen. Arbeiten zur Lokalisierung im Twitter-Umfeld werden hier seltener erwähnt. In späteren Arbeiten, wie in [PCV13], wird hingegen fast ausschließlich auf Arbeiten aus dem Twitter-Umfeld verwiesen. Dies spiegelt die steigende Anzahl der Arbeiten zur Lokalisierung im Twitter-Umfeld wieder. Betrachtet man die Ausarbeitungen zur Lokalisierung im Twitter-Umfeld genauer, wird allerdings schnell klar, dass die Kategorisierung der Arbeiten anhand der verwendeten Methodik, dem Umfang nicht mehr gerecht wird.

Bei genauerer Betrachtung der Arbeiten stellt man allerdings fest, dass diese Klassifizierungen dem Umfang der Arbeiten nicht gerecht wird. [HHSC11] verweist auf ähnliche Ansätze mit einem anderen Untersuchungsgegenstand. [CCL10] kategorisiert die Arbeiten anhand der Methodik, und verweist ebenso auf andere Untersuchungsgegenstände. [PCV13] verweist ausschliesslich auf Arbeiten im Twitter-Umfeld und kategorisiert diese anhand der verwendeten Methodik. Die Methodeneinteilung ist aufgrund der Begriffswahl missverständlich und kann somit zu Problemen führen.

3.1.1 ttt<sss

In [SHP+13] werden die folgenden Dimensionen zur Abgrenzung herangezogen.

Indikatoren aus [SHP+13]

Allerdings lassen sich noch andere Dimensionen zur Klassifizierung der Arbeiten heranziehen. Wird besipielsweise der Text einer Twitter-Kurznachricht durch eine einfache Geokodierung untersucht wird dies andere Ergebnisse liefern als eine Untersuchung auf Basis eines geografischen Themenmodells.

[HHSC11] nutzen diese Methode um eine Ground-Truth zu bestimmen indem das Userlocation-Feld in Wikipedia nachschlagen wird. Wikipedia bietet zu vielen Artikeln eine grografische Position in Form von Längen- und Breitengrad an, diese werden dann der untersuchten Twitter-Kurznachricht zugeordnet. [HGG12] nutzen die Yahoo und die Google Geocoding Api um das Userlocation-Feld eingehender zu untersuchen.

Eine weitere zu betrachtende Dimension stellt daher der konkrete Untersuchungsgegenstand in Form des Indikators dar.

Betrachtet man die Gesamtheit an arbeiten im Bereich der Lokalisierung im Twitter Netzwerk drängen sich noch mehr Dimensionen zur Klassifizeirung der arbeiten auf.

1. Räumliche Indikatoren
2. Techniken
3. Fokus der Lokalisierung

Tabelle einfügen, bereits fertig, nur noch Format anpassen (Lesbarkeit)

Requirements Tabelle einfügen

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) Verfahren mit Inhaltsanalysen
    b) Verfahren mit Indikatoren einzelne oder mehrere

c) Welche Verfahren kommen beim mapping auf geografische Entitäten zum Einsatz

geografische Entität definieren

3.1.2 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

In diesem Kapitel wird ein Verfahren zur Georeferenzierung von Twitter-Nutzern vorgestellt. Die Fragestellungen aus Kapitel 1.3 werden, unter Berücksichtigung der Anforderungen aus Kapitel 1.3.1, beantwortet.

Zunächst soll ein Überblick über die Funktion und den Ablauf des Verfahrens gegeben werden ohne detailliert auf die einzelnen Verfahrensschritte einzugehen. Danach wird das Verfahen detaillierter betrachtet und die einzelnen Verfahrensschritte eingehender erläutert.

4.1 Überblick

Das erarbeitete Verfahren soll es ermöglichen Twitter-Nutzern eine Georeferenz zuzuordnen. Dabei sollen als Eingabe für die Georeferenzierung lediglich der Nutzer-Standort und die Nutzer-Zeitzone aus dem Profil eines Twitter-Nutzers verwendet werden. Als Ergebnis soll eine Georeferenz mit einem Konfidenzwert zurückgeliefert werden. Dabei hat der Anwender die Möglichkeit, sowohl die Genauigkeit bezüglich der geografischen Position, als auch einen Schwellwert für die gewünschte Konfidenz anzugeben. In Abbildung 4.1 ist der generelle Ablauf dieser Georeferenzierung dargestellt.

Die Georeferenzierung nach dem Schema aus Abbildung **??** setzt voraus, dass aus den eingegebenen Indikatoren eine Georeferenz abgeleitet werden kann. Da es sich beim Nutzer-Standort und der Nutzer-Zeitzone um unmittelbare geografische Indikatoren handelt, lässt sich ein geografischer Bezug aus diesen Indikatoren ableiten. Der Wert des Nutzer-Standortes kann vom Nutzer frei eingegeben werden. Es wird keinerlei Kontrolle oder Verifizierung durch Twitter durchgeführt. Wie in 2 bereits analyisert wurde ist der Wert des Nutzer-Standortes nicht objektiv, nicht zuverlässig und nicht gesichert. Dies

evtl. allgemeine und diesen Teil nach unten ziehen wenn konkret auf Twitter eingegangen wird.

Schema Darstellung Eingabe->Ausgabe Sketchbook A1: Unterschrift Die Eingabe besteht aus dem Nutzer-Standort sowie der Nutzer-Zeitzone. Als Rahmenbedingungen wird die gewünschte Hierarchie-Ebene sowie der Schwellwert für die Konfidenz angeben. Als Ausgabe erhält man eine Georeferenz

Neues Schema mit Objekt welchem mehrere geografische Indikatoren zugeordnet wer-

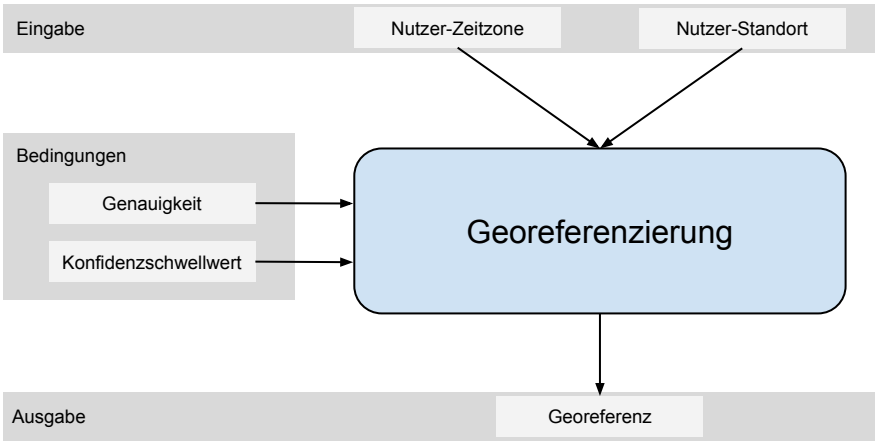

Abbildung 4.1: Ablauf Georeferenzierung

führt zu Problemen bei der Auswertung des Nutzer-Standortes. Um diesen Problemen zu begegnen soll aus einer umfangreichen Tweet-Datensammlung die Zuordnung der eingegebenen Indikatoren zu geografischen Positionen gelernt werden. Diese gelernten Zuordnungen sollen in einer Datenbasis gespeicert werden. Die Datenbasis wird Georeferenz-Basis genannt.

Zur Erstellung der Georeferenz-Basis werden zunächst einige Schritte zur Vorverarbeitung der Indikatoren durchgeführt. Zusätzlich wird, die in jedem Tweet anegegebene Georeferenz, verarbeitet. In der Georeferenz-Basis werden nun die so erstellen Indikator-Werte mit den entsprechenden Georeferenzen gespeichert. Dadurch entsteht ein Wörterbuch mit Indikatoren und zugehörigen geografischen Objekten, die Georeferenz-Basis. Der gesamte Ablauf zur Erstellung der Georeferenz-Basis wird in Abbildung **??** dargestellt.

Bei der Georeferenzierung werden die Indikatoren zunächst derselben Vorverarbeitung wie beim einlernen unterzogen. Die Vorverarbeitungsschritte sind somit generisch, da sie sowohl vor dem erzeugen der Georeferenz-Basis als auch vor der eigentlichen Georeferenzierung durchgeführt werden. Nach der Vorverarbeitung werden die Indikatoren in der Georeferenz-Basis nachgeschlagen und es kann eine Georeferenz zugeordnet werden. Der Ablauf der Georeferenzierung wird in Abbildung **??** dargestellt.

Ablaufplan
A1, A2, A3

In den folgenden Kapiteln soll nun genauer auf die einzelnen Teile des Verfahrens eingegangen werden. Das Verfahren wird dabei in zwei Teilen behandelt.

Einlernen der Georeferenz-Basis In diesem Teil wird erklärt wie mit Hilfe einer Tweet-Sammlung die Georeferenz-Basis durch ein Lern-Verfahren erstellt wird. Zunächst werden die verwendeten Indikatoren untersucht. Basierend auf diesen Erkenntnissen wird ein erster Ansatz zur Umsetzung eines Lern-Verfahrens vorgestellt. Mit Hilfe dieses ersten Ansatzes wird das Verfahren sukzessive weiterentwickelt bis die angestrebten Anforderungen erfüllt werden können. Im Zuge dessen werden die generischen Vorverarbeitungsschritte zur Verarbeitung der Indikatoren entwickelt und erläutert. Auch die Vorverearbeitung, der in jedem Tweet angegeben geografischen Koordinaten, soll hier behandelt werden. Danach soll der gesamte Ablauf des Einlernens noch einmal dargestellt und kurz erklärt werden um einen Gesamtüberblick zu liefern.

Georeferenzierung Die Vorverarbeitungsschritte der Indikatoren wird in diesem Teil nur angeschnitten, da diese bereits im ersten Teil genau erläutert werden. Die Georeferenzierung besteht im Grunde lediglich aus einer Abfrage an die Georeferenz-Basis und die Auswertung der Ergebnisse. Bei der Auswertung wird insbesondere auf die konkrete Berechnung der Konfidenzen eingegangen. Am Ende dieses Teils steht das Ergebnis der Georeferenzierung.

4.2 Generelle Struktur einer Datenbasis zur Georeferenzierung

In diesem Kapitel soll eine generelle Struktur angegeben werden, welche die minimalen Anforderungen einer Georeferenzierung erfüllen kann. Danach wird das Schema und die nötigen Informationen, welche die Datenbasis beinhaltet, genauer betrachtet.

Nach Abbildung **??** wird durch die Georeferenzierung einer Menge geografischer Indikatoren eine Georeferenz zugewiesen. In der einfachsten Variante wird lediglich ein einziger geografischer Indikator an die Georeferenzierung übergeben und genau eine Georeferenz

zurückgegeben. Die Georeferenzierung muss also zu einem gegebenen geografischen Indikator eine Georeferenz bestimmen können. Dies führt zu einer ersten einfachen Struktur für die Datenbasis.

Eine erste einfache Struktur der Datenbasis

Es wird angenommen der geografische Indikator stellt immer ein eindeutiges Toponym dar. Des weiteren sind alle möglichen Toponyme sowie eine zugehörige Georeferenz bekannt. Die Georeferenz liegt als Adresse mit Straße, Hausnummer, Postleitzahl und Ortsname vor.

Jedem möglichen Toponym, welches im gegebenen geografischen Indikator vorkommen kann, soll eine Georeferenz zugeordnet werden können. Daraus ergibt sich die Anforderung, dass die Datenbasis eine Menge von Datensätzen beinhalten muss, die jeweils aus einem Toponym und einer Georeferenz bestehen. Dieser Aufbau entspricht einer Art Wörterbuch in dem Informationen zu einem gegebenen Referenzwert nachgeschlagen werden können. Im vorliegenden Fall kann also zu einem Toponym die enstprechende Georeferenz nachgeschlagen werden. Die Referenzwerte stellen dabei mögliche Werte für die Indikatoren dar. In Abbildung 4.1 ist ein Beispiel für eine sehr simplen Struktur dargestellt.

Tabelle 4.1: Erste einfache Struktur für eine Datenbasis zur Georeferenzierung

Referenzwert	Georeferenz
Zoo-Karlsruhe	Ettlinger Straße 6 - 76137 Karlsruhe
ZKM	Lorenzstraße 19 D - 76135 Karlsruhe
Elbphilharmonie	Dammtorwall 46 - 20355 Hamburg

Wird eine Abfrage auf die Datenbasis mit den Indikatoren “Zoo-Karlsruhe“, “ZKM“ oder “Elbphilharmonie“ durchgeführt, kann nun eine Georeferenz zurückgeliefert werden. Diese simple Struktur reicht grundsätzlich aus um eine Georeferenzierungen durchführen zu können. In dem angeführten Beispiel ist die Menge der möglichen Toponyme sehr begrenzt, aber diese kann beliebig erweitert werden. Damit können sehr mächtige Datenbanken erstellt werden.

Es sollen nun der Referenzwert und die Georeferenz genauer betrachtet werden.

Form der Georeferenz Die Form in der die Georeferenz angegeben wird ist abhängig von der Anwendung. Im Beispiel 4.1 wurden Adressen verwendet. Dazu muss das angegebene Toponym oder die Zeichenkette jedoch eine Adresse besitzen. Ein See in der Wildnis Alaskas wird keine solche Adresse aufweisen. Aber auch die geografische Position einer Stadt oder eines Landes kann nicht durch eine Adresse beschrieben werden. Die Form in der die Georeferenz angegeben wird kommt auf den jeweiligen Anwendungsfall an. Denkbar sind hier unter anderem:

- geografische Koordinaten
- vollständige Adressen
- Länder
- Städte
- adminsitrative Verwaltungsgebiete
- Zeitzonen
- Straßenname und eine Kilometerbezeichnung

Grundsätzlich sind alle Formen, welche eine direkte oder indirekte Georeferenz darstellen, denkbar. Wichtig ist nur, dass die Angabe der Georeferenz in einer Form erfolgt welche für die gegebene Anwendung ausreichend genau ist.

Für den Straßenverkehr ist eine Angabe einer Adresse ausreichend. Für Wanderungen in unerschlossenen Gebieten hingegen sind geografische Koordinaten notwendig.

Form der Referenzwerte Der Referenzwert ergibt sich aus den möglichen Werten der Indikatoren. Die Indikatoren sind dabei zumeist Zeichenketten. Abhängig vom Indikator können diese Zeichenketten verschiedene Informationen enthalten. Im obigen Beispiel sind die Referenzwerte Toponyme. Toponyme sind Namen für geografische Objekte, es kann ihnen also unmittelbar ein geografische Objekt zugeordnet werden. Es muss sich dabei aber nicht um Toponyme handeln. Werte die als Referenzwerte in Frage kommen sollen als geografische Indikatoren bezeichnet werden. Toponyme sind also geografische Indikatoren, da sie direkt mit einem geografischen Objekt in Verbindung gebracht werden können. Geografische Indikatoren können aber auch indirekt mit einem geografischen Objekt in Verbindug gebracht werden. Unbrauchbar sind Werte welche weder direkt

noch indirekt mit einem geografischen Objekt in Verbindung gebracht werden können. Diese Werte stellen keine geografischen Indikatoren dar. Im nächsten Abschnitt werden die geografischen Indikatoren genauer betrachtet.

4.3 Geografische Indikatoren

Geografische Indikatoren sind diejenigen Indikatoren welche einen geografischen Bezug aufweisen. Es sind also genau die Werte, welche in irgendeiner Weise mit einem geografischen Objekt in Verbindung gebracht werden können.

Geografische Indikatoren lassen sich in unmittelbare geografische Indikatoren und mittelbare geografische Indikatoren aufteilen. In diesem Abschnitt soll sowohl auf die unmittelbaren wie auch die mittelbaren geografischen Indikatoren eingegangen werden.

4.3.1 Unmittelbare geografische Indikatoren

Der in einem unmittelbaren geografischen Indikator angegebene Wert beschreibt direkt eine geografische Position oder eine geografische Region. Einem unmittelbaren geografischen Indikator kann durch die in ihm enthaltene Information direkt eine Georeferenz auf ein geografisches Objekt zugewiesen werden.

Im Beispiel in Tabelle 4.1 wurden unmittelbare geografische Indikatoren verwendet. Die Referenzwerte entsprechen Toponymen, sie sind Namen für geografische Objekt. Ihnen kann also unmittelbar ein geografisches Objekt zugeordnet werden.

Ein weiteres Beispiel für einen unmittelbaren geografischen Indikator sind Zeitzonen.

Bei der Zeitzone handelt es sich um einen geografischen Indikator der eine geografische Region beschreibt. Die zugehörige Georeferenz für eine Zeitzone könnten die Länder sein welche die Zeitzone umfasst. Aber auch die Kontinente oder Städte welche in dieser Zeitzone liegen könnten, je nach Anwendugsfall, von Interesse sein.

4.3.2 Mittelbare geografische Indikatoren

Der Indikator muss aber nicht unmittelbar einer Georeferenz zuzuordnen sein. Dies ist genau dann der Fall, wenn der in eiem geografischen Indikator angegebene Wert in erster Linie nicht einem geografischen Objekt zugeordnet werden kann.

Die Möglichkeiten was einen mittelbaren geografischen Indikator darstellen kann sind nahezu unbegrenzt. An einem Beispiel soll ein gezeigt werden was einen mittelbaren geografischen Indikator ausmacht.

Beispiel eines mittelbaren geografischen Indikators Es liegen drei Begriffe vor.

1. Äbierra
2. Grumbeer
3. Tüfte

Nach einer kurzen Recherche kann festgestellt werden, dass es sich bei allen drei Begriffen um Kartoffeln handelt. Die Information die hinter jedem dieser Begriffe steckt ist also die Bezeichnung eines Gemüses. Die Begriffe bezeichnen also insbesondere kein geografisches Objekt und sind somit keine unmittelbaren geografischen Indikatoren. Jede dieser Bezeichnungen stammt aber aus unterschiedlichen Regionen Deutschlands, denn es handelt sich um dialektiche Begriffe. Durch ihre geografisch begrenzte Verwendung können sie damit einer geografischen Region zugeordnet werden. Durch die Verwendung eines dieser Worte kann also auf eine Region Deutschlands geschlossen werden. Dadurch kann jedem Begriff eine Georeferenz zugewiesen werden.

Mit folgender Datenbasis kann eine Georeferenzierung der obigen Begriffe durchgeführt werden:

Tabelle 4.2: Kartoffeln in verschiedenen Dialekten

Referenzwert	Georeferenz
Äbbiera	Württemberg
Grumbeer	Pfalz
Tüfte	Norddeutschland

In diesem Beispiel wurde die Zuordnung zu einem geografischen Objekt über die Verwendung der Begriffe in bestimmten Regionen ermöglicht. Aber auch andere Informationen können dazu dienen einem Indikator eine Georeferenz zuzuordnen. Es kommt grundsätzlich nicht darauf an was genau mit dem Begriff bezeichnet wird. Lediglich die geografisch begrenzte Verwendung kann hier als Hinweis auf eine Georeferenz dienen.

Ein solcher geografischer Indikator wird als "mittelbar geografischer Indikator" bezeichnet.

4.4 Probleme bei der Verwendung von Toponymen als geografische Indikatoren

Toponyme sind die am häufigsten auftretende Art geografischer Indikatoren. Es handelt sich dabei um unmittelbar geografische Indikatoren.

Die Georeferenzierung von Toponymen kann bereits mit der Struktur aus Tabelle 4.1 umgesetzt werden. Jedem Toponym wurde dabei eine Georeferenz zugewiesen. Es wurden dabei allerdings mehrere Annahmen gemacht, die in der Regel nicht gelten.

Im folgenden soll auf diese Annahmen eingegangen werden und die daraus entstehenden Probleme sollen aufgezeigt werden. Des weiteren werden generelle Probleme bei der Verwendung geografischer Toponyme zur Georeferenzierung aufgezeigt.

4.4.1 Alle möglichen Toponyme sind bekannt

Es existieren Datenbanken mit Toponymen welche Millionen von Einträgen beinhalten. Eine dieser Datenbanken ist das frei erhältliche Ortsverzeichnis von geonames.org. Es beinhaltet ca. 8,9 Millionen Toponyme und zugehörige Georeferenzen. Zusätzlich sind weitere ca. 8 Millionen alternative Toponyme hinterlegt. In den Datensätzen sind neben einer Georeferenz in Form geografischer Koordinaten noch andere Informationen hinterlegt wie beispielsweise Länderkürzel und Einwohnerzahlen. Auch eine geografische Hierarchie wird dabei abgebildet.

Trotz der immensen Anzahl an Toponymen beinhalten solche Datenbanken nicht alle möglichen Toponyme. Aufgrund der immensen Vielfalt ist es nahezu unmöglich alle Toponyme abzudecken.

Eine vollständige, weltweite Abdeckung von Toponymen ist nahezu unmöglich. Im folgenden soll an einigen Beispielen die Vielfalt der Toponyme belegt werden.

Vielfalt der Toponyme

Da Toponyme nicht definiert sind und neben den offiziellen Namen für Städte, Länder usw. weitere Namen vorkommen it die Vielfalt nahezu unbegrenzt.

Spitznamen oder alternative Namen für geografische Objekte In Wikipedia sind für die Stadt Detroit, im US-Bundesstaat Michigan, folgende Spitznamen angegeben:

- The Motor City
- Motown
- Hockeytown

Die ersten zwei dürften weltweit einen Gewissen Bekanntheitsgrad haben. Hockeytown, Rock City und The D dürften allerdings weniger bekannt sein. Tatsächlich beinhaltet die geonames.org Datenbank keinen dieser Spitznamen. Google Maps hingegen bietet bei Eingabe der oben genannten Spitznamen Detroit als Vorschlag an.

Das Problem wird noch gravierender wenn man lokal begrenzte Spitznamen einbezieht. Es ist durchaus denkbar, dass Spitznamen für einen Stadtteil nur lokal verwendet werden und über die Stadt- oder Landesgrenze hinaus nicht bekannt ist. Diese Spitznamen können nur schwer erfasst werden. Solche Spitznamen existieren auch für Länder und administrative Verwaltungsebenen.

Weitere Beispiele die zu einer größeren Vielfalt der Toponyme führen:

- Historische Toponyme
- Kulturell bedingte Toponyme
- Toponyme für landschaftliche Besonderheiten

- Toponyme für bestimmte Landschaften

4.4.2 Die Toponyme sind eindeutig geografischen Objekten zuzuweisen und umgekehrt

Auch dies ist im allgemeinen nicht der Fall. Toponyme sind oft Doppel- oder Mehrdeutig und verweisen somit auf mehrere Georeferenzen. Allerdings können zu einer Georeferenz auch mehrere eindeutige Toponyme existieren, welche nur auf diese Georeferenz verweisen. Doppel- und Mehrdeutigkeiten können also sowohl ausgehend von der Georeferenz, als auch ausgehend vom Referenzwert auftreten.

Zu einem Referenzwert können mehrere Georeferenzen existieren Es gibt zahlreiche Städte-Namen, die in mehreren Ländern verwendet werden. Ein gutes Beispiel hierfür sind US Städte. Da die USA ein Einwanderungsland ist, übernahmen viele Einwanderer bei der Gründung neuer Städte die Namen aus der alten Heimat. So finden sich in den USA zahlreiche Städte deren Namen exakt den deutschen Städtenamen entsprechen. In Tabelle **??** sind einige Städte-Namen und die Vorkommen in den USA aufgelistet.

Tabelle 4.3: Häufige deutsche Städtenamen in den USA

Name	Anzahl in den USA
Hannover	40
Berlin	39
Hamburg	30

Soll nun eine Datenbasis angelegt werden, welche einem Stadt-Namen einen Staat zuweist müssten für Hannover 40, für Berlin 39 und für Hamburg 30 Georeferenzen hinterlegt sein. Als Ergebnis einer Georeferenzierung für den Wert “Hamburg“ würden 30 Georeferenzen in den USA und eine in Deutschland zurückgegeben werden. Die Anzahl der Einträge hat dabei keinerlei Aussage über die Wahrscheinlichekit welches Hannover gemeint ist.

Diese Mehrdeutigkeit stellt ein Problem bei der Georeferenzierung dar. Es kann keine eindeutige Entscheidung getroffen werden welche Georeferenz dem Ort zugewiesen werden soll. Lösung über Konfidenzen

Zu einer Georeferenz können mehrere Referenzwerte existieren Dies wird anhand der Städte-Spitznamen klar. Städte-Spitznamen sind oft eindeutig, allerdings kann eine Stadt mehrere dieser eindeutigen Spitznamen besitzen. Damit ergibt sich eine Mehrdeutigkeit von der Georeferenz zu einem Toponym. Dies ist grundsätzlich auf alle geografischen Objekte übertragbar. Zu jedem geografischen Objekt existieren potenziell meherere gültige Toponyme.

Das Problem hierbei stellt nicht die Beziehung von der Georeferenz zu den Toponymen dar, sondern eher die vielfalt der Toponyme.

4.4.3 Fazit

Es existieren sehr umfangreiche Datenbasen um Toponymen eine Georeferenz zuzuweisen. Es ist allerdings schwer, wenn nicht sogar unmöglich, das Wissen über geografische Objekte und deren zugehörige Toponyme vollständig zu erfassen.

Ein weiteres Problem kann die untersuchte Domäne darstellen. In sozialen Netzwerken können sich eigene Begriffe und Formulierungen etablieren welche im allgemeinen nicht bekannt sind.

Im Twitter-Umfeld haben sich in den letzten Jahren beispielsweise einige spezielle Begriffe und Formulierungen etabliert. Hauptsächlich werden diese aber in den Tweets selbst verwendet. Eine Übertragung auf den Nutzer-Standort kann aber nicht gänzlich ausgeschlossen werden.

Ein Beispiel hierfür ist “Bieberville“ welches in den untersuchten Daten von Hecht et al. öfter vorkommt. “Bieberville“ wird abgeleitet von dem Pop-Star Justin Bieber. Twitter wird oft als “Bieberville“ bezeichnet, da der Pop-Star in Twitter sehr aktiv ist und deshalb viele Fans auch in Twitter aktiv sind. Unter diesem Gesichtspunkt hätte “Bieberville“ keinen geografischen Bezug. Sucht man allerdings im Internet weltweit nach “Bieberville“ stößt man auf einen Imbiß in Groß-Bieberau. “Bieberville“ kann also durchaus einen geografischen Bezug haben, wenngleich es im Twitter-Umfeld nicht als solcher benutzt wird. Ein Nutzer-Stadnort der in einem Land kein Toponym darstellt, kann in einem anderen durchaus ein Toponym sein.

Die Benutzung von “Bieberville“ könnte auch ein temporäres Phänomen darstellen. In den Tweet-Daten welche für diese Arbeit verwendet werden befindet sich kein Eintrag mit dem Nutzer-Stadort “Bieberville“. Dies könnte ein Hinweis auf eine temporäre Verwendung des Begriffes im Nutzer-Standort sein. Leider ist es nicht möglich Nutzer nach ihrem Nutzer-Standort zu suchen und deshalb kann diese Aussage nicht mit letzter Sicherheit getroffen werden. Gibt es Toponyme die tatsächlich nur temporär auftauchen wird das Problem noch größer. Dies ist durchaus denkbar, denn mit wenigen Klicks und einer kurzen EIngabe kann der Nutzer-Stadort geändert werden.

Das Wissen über solche Begriffe und Formulierungen kann nur sehr schlecht erfasst werden, wenn man die enstprechende Domäne nicht untersucht.

Bieberville beispiel in lernen

4.5 Der Nutzer-Standort

Der Nutzer-Standort eines Twitter-Nutzers soll zur Georeferenzierung genutzt werden. Der Nutzer-Standort eignet sich besonders gut zur Erzeugung geografischer Indikatoren. BBei der Eingabe wird vom Nutzer abgefragt, wo dieser sich befindet. Die Intention der Abfrage zielt also darauf ab, dass der Nutzer einen Wert eingibt der auf ein geografisches Objekt verweist welches seinem Standort entspricht. Es ist naheliegend, dass der Nutzer seinen Standort mit Hilfe eines Toponyms angibt. Der Nutzer-Standort wird jedoch vom Nutzer frei eingegeben und keinerlei Kontrolle unterzogen. Dieser muss also nicht zwangsweise ein Toponym enthalten. Des weiteren können alle in Kapitel 4.4 erwähnten Probleme auftreten. Durch die unkontrollierte Eingabe sind tatsächlich alle möglichen Toponyme denkbar.

Hier soll nun der Nutzer-Stadort genauer betrachtet werden und ob dieser überhaupt zur Erzeugung geografischer Indikatoren geeignet ist.

4.5.1 Hat der Nutzer-Standort überhaupt einen geografischen Bezug?

Nach Hecht et al. haben in [HHSC11] den Nutzer-Standort eingehend untersucht und sind zu folgendem Ergebnis gekommen: Wenn der Nutzer überhaupt einen Standort angegegeben konnte in 80% Prozent der Fälle ein geografischer Bezug nachgewiesen werden. In den restlichen 20% der Fälle konnte im Nutzer-Standort kein geografischer Bezug nachgewiesen werden. Die Untersuchung wurde manuell vorgenommen und es durften alle zur Verfügung stehenden Mittel zur Analyse der Daten verwendet werden. Hecht et al. haben deshalb ausschließlich Daten untersucht die nachweislich aus den USA stammten.

Es kann hier also festgehalten werden, dass 80% der Nutzer-Standorte als geografischer Indikator verwendet werden können. Es muss bechtet werden, dass nur Daten aus den USA betrachtet wurden.

Im Rahmen der vorliegenden Arbeit wurde der Nutzer Standort von 1000 Twitter-Nutzern ebenfalls manuell untersucht. [1] Dabei wurde keinerlei Einschränkung zur Herkunft gemacht. Um zu prüfen ob ein geografischer Bezug vorliegt wurden Ortsverzeichnisse von Google-Maps und Geonames.org verwendet. Diese lassen es zu auch in dem Nutzer unbekannten Sprachen und Alphabeten zu suchen. Es konnte dabei in XYZ% der Fälle ein geografischer Bezug nachgewiesen werden.

In den restlichen XYZ% der fälle konnte kein geografischer Bezug mit Hilfe der Datenbanken nachgewiesen werden. Dies bedeutet nicht, dass grundsätzlich kein geografischer Bezug vorhanden ist. Es konnte lediglich anhand der genutzten Quellen kein geografischer Bezug hergeleitet werden.

Der Nutzer-Standort kann also in vielen Fällen einen Hinweis auf die Herkunft des Twitter-Nutzers geben.

4.5.2 Genauigkeit der geografischen Angaben

In 80% der Fälle kann bei einer Angabe des Nutzer-Standortes davon ausgegangen werden, dass dieser geografischen Bezug hat. Wiederum aufgrund dessen, dass der Nutzer

[1] siehe **??**

beliebige Eingaben machen kann, ist nicht sicher wie genau ein solcher geografischer Bezug ist.

Hecht et al. analyiserten ihre Daten auch darauf wie genau die Nutzer ihren Standort angeben. Dabei ist wiederum zu beachten, dass die Daten aus den USA stammten und deshalb die geografischen Verwatungsebenen der USA zugrunde gelegt wurden. Zusätzlich wurden hier auch Dabei wurden folgende Werte festgestellt:

- 64% Stadt
- 20% Staat (Adminstrationsebne erster Ordnung)
- ca. 8% Intrastate (Adminstrationsebene zweiter Ordnung)
- ca. 5% Land

Die restlichen 13% entfallen auf Interstate Regionen, Nachbarschaften und konkrete Adressen. Interstate Regionen sind Regionen die sich über mehrere Staaten hinwegziehen. Beispiele für Interstate Regionen sind "Central United States" oder "West-Coast". Nachbarschaften(Neighbourhoods) sind oft Stadtteile wie "Harlem" oder "Bronx" in New York.

In den eigenen Untersuchungen sind folgende Werte festgestellt worden:

- xyz% Stadt
- zz% Adminstrationsebene erster Ordnung
- tt% Adminstrationsebene zweiter Ordnung
- uu% Land

Es hier gefolgert werden, dass selbst wenn ein geografischer Bezug besteht, nicht mit Sicherheit bestimmt werden kann wie genau dieser ist.

4.5.3 Stimmt der tatsächliche Stadort mit dem angegeben überein?

Diese Frage ist nur schwer zu beantworten.

4.5.4 Partieller geografischer Bezug des Nutzer-Standortes

In einigen der Einträge konnte festgestellt werden, dass nur Teile der Einträge geografischen Bezug haben. Oft wollen die Nutzer noch weitere Informationen geben, diese haben oft keinen Nutzen für eine Georeferenzierung. In der folgenden Liste sind einige Einträge von Nutzer-Standorten aufgelistet.

1. 11th Dimension | California
2. between here and there - Miami

Im ersten Fall kann für Kalifornien ein geografischer Bezug festgestellt werden. 11th Dimension hat hier offensichtlich keinen geografischen Bezug. Im zweiten Fall ist die Aussage "between here and there" nicht zu gebrauchen, Miami kann jedoch als Bezug zur Stadt Miami in Florida, USA gebracht werden.

Es können also auch nur Teile des Nutzer-Stadorts für eine Georeferenzierung von Nutzen sein.

4.5.5 Mehrere wiedersprüchliche geografische Bezüge

Es existieren auch Einträge für Nutzer-Standorte welche mehrere wiedersprüchliche Angaben machen. Das bedeutet es werden zwei oder mehr Werte mit geografischem Bezug angegegeben.

Auch hier sollen einige Beispiele genannt werden:

- Bolton\/Leigh
- Liverpool\/London
- Balikesir \/ Izmir

In diesen Beispielen sind jeweils zwei Städte angegeben. Bolton und Leigh liegen 14km auseinander. Liverpool und London trennen ca. 350km. Balikesir und Izmir ca. 180km.

Es kann nun spekuliert werden wieso der Nutzer zwei Städte angibt. Ist er in einer Stadt aufgewachsen und lebt momentan in der anderen? Pendelt er zwischen den Städten um zu arbeiten?

Wie dem auch sei, es kann nicht eindeutig entschieden werden in welcher Stadt sich der Nutzer aufhält.

4.5.6 Fazit

Ist ein geografischer Bezug des Nutzer-Standortes nachzuweisen, handelt es sich bei dem Eintrag in den meisten Fällen um ein Toponym. Durch die durchgeführte Untersuchung konnte gezeigt werden, dass ca. xyz% der Nutzer-Standorte mit einem geografischen Bezug auf Toponyme zurückzuführen sind. Im Hinblick auf den geografischen Bezug der Einträge im Nutzer-Standort ergab sich kein signifikanter Unterschied zu den Untersuchugen von Hecht et al..

Zusammenfassend kann gesagt werden, dass der Nutzer-Standort häufig einen geografischen Bezug aufweist und es sich in diesen Fällen um Toponyme handelt. Aufgrund der völlig freien Eingabe durch den Nutzer ergeben sich bei der Auswertung der Nutzer-Standorte jedoch einige Probleme die gelöst werden müssen.

4.6 Die Nutzer-Zeitzone

In Kapitel 4.4 wurden die Probleme bei der Verwendung von Toponymen als geografische Indikatoren eingegangen. Dabei wurde die Doppel- und Mehrdeutigkeit von Toponymen betrachtet. Bei der Verwendung des Nutzer-Standortes kann dieses Problem auch auftauchen.

Um diesem Problem zu begegnen soll nun ein weiterer geografischer Indikator hinzugezogen werden, die Nutzer-Zeitzone. Zunächst sollen die generellen Eigenschaften der Nutzer-Zeitzone erläutert werden bevor erklärt wird wie die Nutzer-Zeitzone das Problem der Mehrdeutigkeit von Toponymen beheben kann.

4.6.1 Eigenschaften der Nutzer-Zeitzone

Die Nutzer-Zeitzone stellt einen unmittelbaren geografischen Indikator dar. Sie beschreibt eine eindeutige geografische Region auf dem Globus. Dabei enstprechen die Grenzen der

Region nicht unbedingt den Landesgrenzen oder den Grenzen sonstiger adminstrativer Verwaltungseinheiten. In Abbildung 4.2

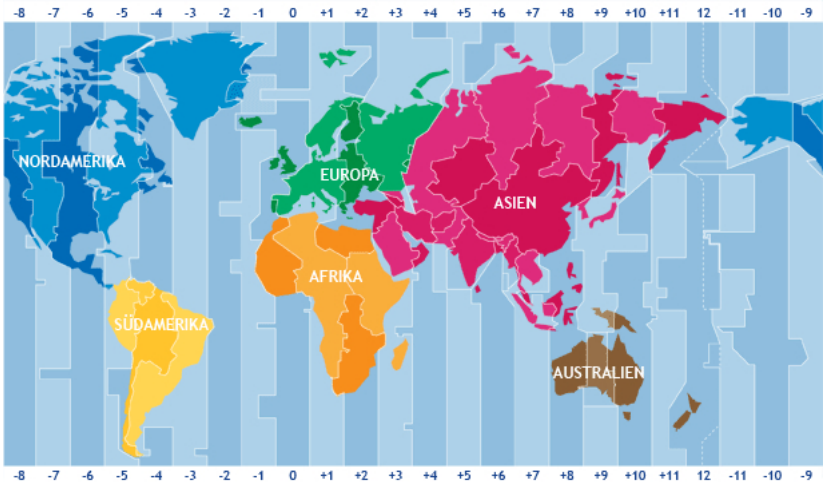

Abbildung 4.2: Zeitzonen der Erde.

Quelle: http://www.zeitzo

Die Nutzer-Zeitzone kann in Twitter über eine Liste gewählt werden. Der Wert der Nutzer-Zeitzone stellt deshalb immer eine Zeitzone dar. Es ist dem Nutzer nicht möglich einen Wert einzugeben der nicht einer Zeitzone entspricht.

Bild Auswahlliste

Allerdings ist nicht gesichert, dass der angegegebene Wert auch der Zeitzone entspricht in der sich der Nutzer aufhält. Der Nutzer könnte eine bewusste Fehleingabe machen oder aber die Zeitzone nicht wählen womit der Stadardwert “Pacific Time (US and Canada)“ gewählt wird. In Abbildung **??** wurden Tweets anhand ihres Längen- und Breitengrades platziert. In der Abbildung sind Tweets aus den USA zu sehen. Die Farben entsprechen den gewählten Zeitzonen.

- Pacific Time Blau
- Eastern Time Rot
- Central Time Grün
- Mountain Time Pink

Die Zeitzonen sind durch die Farben gut zu erkennen. Lediglich die dünne besiedelte Region der Mountain Time kann nur an einige Ballungszentren erkannt werden. Grundsätzlich scheint die Angabe der Zeitzone aber korrekt zu sein.

Bei lernen Zeitzonen Problematik mit standard einbringen

4.6.2 Auflösen von Doppeldeutigkeiten

Mit der Nutzer-Zeitzone lassen sich Doppeldeutigkeiten auflösen.

Besteht zu einem Nutzer-Stadort eine Doppeldeutigkeit, kann nicht enstcheiden werden welches geografische Objekt zugeordnet werden soll. Liegen die beiden geografischen Objekte allerdings in zwei unterschiedlichen Zeitzonen und der Nutzer hat diese angegegeben kann die Doppeldeutigkeit aufgelöst werden. Somit kann dem Nutzer die korrelte Georeferenz zugewiesen werden. Vorausetzung dazu ist natürlich, dass die geografischen Objekte in zwei unterschiedlichen Zeitzonen liegen und die Nutzer-Zeitzone angegegeben ist.

4.7 Einlernen geografischer Indikatoren am Beispiel von Twitter

Um die oben genannten Probleme beheben zu können soll ein Lernverfahren entwickelt werden. Dabei sollen aus einer Tweet-Sammlung die genutzten geogarfischen Indikatoren und deren Zuordnung zu einer Georeferenz automatisch gelernt werden. Es soll eine Datenbasis erzeugt werden, welche es ermöglicht den genutzten geografischen Indikatoren eine Georeferenz zuzuweisen.

Der Vorteil besteht darin, dass eine domänenspezifische Datenbasis geschaffen wird welche potenziell mehr geografische Indikatoren zuordenen kann als ein normales Ortsverzeichnis. Es können dadurch domänenspezifische Eigenheiten bei der Eingabe berücksichtigt werden. Auch domäneninterne Begriffe sollen hierdurch gelernt werden können.

Zunächst soll ein genereller Ablauf zum einlernen geografischer Indikatoren vorgestellt werden. Danach werden die nötigen Vorverarbeitungsschritte für den Nutzer-Standort

und die Nutzer-Zeitzone vorgestellt. Anhand desssen wird die in Abschnitt 4.2 vorgestellte Struktur der Datenbasis angepasst und erweitert. Dabei wird am grundsätzlichen Prinzip nichts geändert, es wird nach wie vor einem Referenzwert eine Georeferenz zugewiesen. Die Referenzwerte leiten sich dabei aus den potenziellen geografischen Indikatoren ab.

Zum Schluss wird erläutert wie aus den gelernten Daten einem Twitter-Nutzer eine Georeferenz zugewiesen werden kann.

4.7.1 Generelles Verfahren zum einlernen von geografischen Indikatoren

Um die Datenbasis einzulernen soll nun ein generelles Verfahren vorgeschlagen werden. Es muss mindestens ein Wert vorhanden sein aus welchem unmittelbare oder mittelbare geografische Indikatoren extrahiert werden können. Um den Wert für die Georeferenz zu bestimmen muss zu jedem dieser Werte eine geografische Koordinate vorhanden sein.

Die Tweet-Sammlung beinhaltet pro Datensatz zwei Werte aus denen geografische Koordinaten extrahiert werden können. Jedem solchen Paar ist zusätzlich eine geografische Koordinate zugewiesen. Womit die Tweet-Sammlung die Anforderungen erfüllt.

In Abbildung **??** ist der generelle Ablauf des Lernverfahrens am Beispiel der Twitter-Daten dargestellt. Es wird der Nutzer-Standort und die Nutzer-Zeitzone einer Vorverarbeitung unterzogen. Daraus resultieren die geografischen Indikatoren. Diese wiederum werden mit der zugehörigen geografischen Koordinate in der Datenbasis gespeichert. So kann jedem geografischen Indikator eine Georeferenz zugeordnet werden.

4.7.2 Vorverarbeitung des Nutzer-Standortes und der Nutzer-Zeitzone

Aus dem Nutzer-Stadort und der Nutzer-Zeitzone werden durch die Vorverarbeitung mehrere Werte erzeugt. Diese Werte müssen nicht zwangsweise einen gegrafischen Indikator darstellen. Sie werden deshalb potenzielle geografische Indikatoren genannt. Bei

unmittelbaren geografischen Indikatoren kann zunächst nicht entschieden werden ob diese einen geografischen Indikator darstellen oder nicht. Es ist also sinnvoll keine Werte zu verwerfen welche einen mittelbaren geografischen Indikator darstellen könnten.

Durch die Vorverarbeitung sollen nun potenzielle geografische Indikatoren erzeugt werden.

Ziel ist es aus dem Nutzer-Standort möglichst viele Informationen zu extrahieren und etwaige Probleme zu beseitigen.

In der folgenden Liste sind einige Nutzer-Standorte angegeben. Anhand dieser Liste sollen die Vorverarbeitungsschritte motiviert und demonstriert werden.

1. Bélem-PA
2. West Sussex, England
3. South Florida
4. Pitmedden, Scotland, UK
5. Mato Grosso & Rio de Janeiro
6. -****-
7. USA \/ Los Angeles
8. Nottingham\/London
9. Los Angeles, USA
10. I ♡ New York
11. †˜ Los Angeles˜†
12. earth-sea
13. In front of the computer
14. 11th Dimension | California
15. York
16. York

Eliminierung von Sonder- und Satzzeichen

Es fällt auf, dass oft Sonder- und Satzzeichen verwendet werden. Beispielsweise als Trenner zwischen Toponymen unterschiedlicher geografischer Hierarchieebenen wie bei “West Sussex, England“, “USA \/ Los Angeles“ oder “Bélem-PA“. Das Trennzeichen wird nicht einheitlich verwendet. Es kann deshalb nicht entschieden werden ob ein Satzzeichen als Trenner zweier Hierarchiebenen fungiert oder nicht. Bei “USA \/ Los Angeles“ wird \/ als Trenner für Hierarchieebenen verwendet bei “Nottingham\/London“ werden zwei Städte angegeben. Es ist also insbesondere nicht klar welcher Zusammenhang zwischen den Toponymen, die durch ein Zeichen getrennt sind, besteht.

Bei “I ♡ New York “ werden Sonderzeichen zum ausdrücken von Emotionen genutzt. In “†˜ Los Angeles˜†“ werden Sonderzeichen als Dekoration genutzt. Einige Nutzer-Standorte bestehen ausschließlich aus Sonder- und Satzzeichen.

Zur Ableitung eines geografsichen Bezugs bringen die Sonder- und Stazzeichen keinen Mehrwert. Es gibt Fälle in denen Sonder- oder Satzzeichen Bestandteil des Toponyms sein. Ein Beispiel hierfür wäre “3. Arrondisement“ in Paris, das weglassen des Punktes würde hier aber zu keinen Problemen führen.

Grundsätzlich ist es aufgrund der Vielfalt schwierig definitiv zu entscheiden ob Satz- und Sonderzeichen einen Mehrwert bieten bei der Georeferenzierung. Das hinzufügen von Satz- und Sonderzeichen kann allerdings mehr Probleme mit sich bringen als es Vorteile bringt. Es sollen deshalb in einem ersten Vorverarbeitungsschritt alle Sonder- und Satzzeichen entfernt werden.

Liste nach dem entfernen von Sonder- und Satzzeichen:

1. Bélem PA
2. West Sussex England
3. South Florida
4. Pitmedden Scotland UK
5. Mato Grosso Rio de Janeiro
6.
7. USA Los Angeles
8. Nottingham London
9. Los Angeles USA
10. I New York
11. Los Angeles
12. earth sea
13. In front of the computer
14. 11th Dimension California
15. York
16. York

Der Wert 6 existiert nun nicht mehr, der Wert ist leer und wird somit nicht weiter betrachtet.

Zusammenfassen von Toponymen

In diesem Schritt sollen Toponyme, welche aus zwei oder mehr Wörtern bestehen mit Hilfe eines Ortsverzeichnisses zusammengefasst werden. Dies kann nur für bekannte To-

ponyme durchgeführt werden. Damit bilden beispielsweise “Los“ und “Angeles“ eine Einheit. Dies soll mit einem + Zeichen gekennzeichnet werden.

Daraus resultiert:

1. Bélem PA
2. West+Sussex England
3. South Florida
4. Pitmedden Scotland UK
5. Mato+Grosso Rio+de+Janeiro
6. USA Los+Angeles
7. Nottingham London
8. Los+Angeles USA
9. I New+York
10. Los+Angeles
11. earth sea
12. In front of the computer
13. 11th Dimension California
14. York
15. York

In diesem Schritt wird insbesondere keine Georeferenzierung vorgenommen. Dieser Schritt dient dazu, möglichst früh vorhandenes Wissen über Toponyme einzubeziehen.

Im wesentlichen sollen die Werte hiemrit auf die nächsten Verarbeitungsschritte vorbereitet werden um eine unnötige Fragmentierung der Werte zu vermeiden.

Alphanumerische Sortierung

In diesem Schritt sollen die Werte alphanumerisch sortiert werden. Dadurch ist die Reihenfolge in der die Werte angegeben werden unerheblich.

Nach der Sortierung liegt folgende Tabelle vor:

1. Bélem PA
2. England West+Sussex
3. Florida South
4. Pitmedden Scotland UK
5. Mato+Grosso Rio+de+Janeiro
6. Los+Angeles USA
7. London Nottingham
8. Los+Angeles USA
9. I New+York
10. Los+Angeles

11. earth sea
12. computer front In of the
13. 11th California Dimension
14. York
15. York

Die Werte 6 und 8 sind nun gleich. Durch das zusammenfassen der Toponyme im vorherigen Schritt werden bekannte Toponyme nicht getrennt.

Dieser Schritt stellt einen Kompromiss dar. Sortiert man die Werte werden Werte mit gleichem Inhalt aber unterschiedlicher Reihenfolge angeglichen. Aber es werden auch potenzielle Toponyme die aus mehreren Werten bestehen auseinandergezogen.

Aus “Motor City Michigan USA“ würde “City Michigan Motor USA“ entstehen. Der Zusammenhang zwischen Motor und City wäre nicht mehr vorhanden.

Erzeugung von N-Grammen

Es sollen nun N-Gramme bis zum Grad 3 erzeugt werden. Aus einem Nutzer-Standort werden mehrere potenzielle geografische Indikatoren erzeugt. Dieses Vorgehene löst gleich mehrere Probleme.

Zum ersten sind sowohl Toponyme als auch Werte die kein Toponym darstellen in den Nutzer-Standorten vorhanden. Durch die Erzeugung von N-Grammen können diese getrennt voneinander betrachtet werden.

Zum zweiten können in einem Nutzer-Stadort meherere Toponyme enthalten sein. Der Wert “Pitmedden Scotland UK“ enthält mit UK das Land, mit Scottland den Landesteil und mit Pitemedden eine Stadt. Allen drei Werten kann mit Hilfe der geografischen Koordinaten aus dem Tweet somit eine Georeferenz zugewiesen werden.

Besteht eine Beziehung der Werte zueinander, kann dies durch Bi- und Trigramme abgebildet werden. Aus “Pitmedden Scotland UK“ wird “Pitmedden Scotland“,“Scotland UK“ und “Pitmedden Scotland UK“ erzeugt.

Dieser Schritt soll nur an einigen ausgewählten Beispielen aus der Liste erfolgen die Bestandteile der N-Gramme werden mit <>gekennzeichnet:

1. <England><West+Sussex>
2. <England>
3. <West+Sussex>
4. <Mato+Grosso><Rio+de+Janeiro>
5. <Rio+de+Janeiro>
6. <Mato+Grosso>
7. <11th><California><Dimension>
8. <11th><California>
9. <California><Dimension>
10. <11th>
11. <California>
12. <Dimension>
13. <York>
14. <York>

Der Wert 7 ist ein Trigramm. Bei den Werten 1,4,8 und 9 handelt es sich um Bigramme. Die Werte von 2,3,5,6,10,11 und 12 sind Unigramme.

Auch hier entsteht aufgrund des zusammenfassens bekannter Toponyme keine Fragmentierung der Werte, da zusammengehörige Toponyme gemeinsam betrachtet werden.

Hinzufügen der Zeitzone

Es soll an dieser Stelle die Zeitzone hinzugefügt werden. Da die Zeitzone eine begrenzte Anzahl an Werten darstellt und diese vom Nutzer nicht frei eingegeben werden können muss hier keine weitere Vorverarbeitung vorgenommen werden.

Jeder Wert der in der aktuellen Liste vorkommt soll einmal mit und einmal ohne Zeitzone existieren. Damit wird garantiert, das eingelrnete Werte die eine falsche Zeitzone aufweisen trotzdem berücksichtigt werden bei einer Anfrage.

Auch hier wird die Liste weiter eingeschränkt, da diese sonst den Rahmen sprengt:

1. England West+Sussex
2. <England>
3. <West+Sussex>
4. <England><West+Sussex>London
5. <England>London
6. <West+Sussex>London
7. <11th><California><Dimension>
8. <11th><California>
9. <California><Dimension>
10. <11th>

11. <California>
12. <Dimension>
13. <11th><California><Dimension>Pacific Time (US & Canada)
14. <11th><California>Pacific Time (US & Canada)
15. <California><Dimension>Pacific Time (US & Canada)
16. <11th>Pacific Time (US & Canada)
17. <California>Pacific Time (US & Canada)
18. <Dimension>Pacific Time (US & Canada)
19. York
20. <York>London
21. York
22. <York>Eastern Time

Mit Hilfe der Zeitzone können nun auch die beiden letzten Einträge unterschieden werden. Zum einen York in England zum anderen York in den USA.

Fazit

Nach der Vorverarbeitung liegen eine Menge von potenziellen geografischen Indikatoren mit zugehörigen geografischen Koordinaten vor. Diese Indikatoren und Referenzwerte können nun in der Datenbasis gespeichert werden.

Mit der nun erzeugten Datenbasis ist es grundsätzlich möglich eine Georeferenzierung durchzuführen. Allerings wird jedem potenziellen geografischen Indikator eine Georeferenz zugeordnet. Das ist grundsätzlich problematisch, da einige der Indikatoren keinen geografischen Bezug aufweisen. Es sollte entschieden werden können ob es sich um einen geografischen Indikator handelt oder nicht.

Des weiteren können zu einem Referenzwert beliebig viele Datensätze existieren, selbst wenn diese auf denselben Ort verweisen. Bei einer Abfrage an die Datenbank werden potenziell große Mengen an Datensätzen zurückgeliefert die zunächst keine eindeutige Georeferenz liefern.

4.7.3 Häufigkeistwert

Es soll nun ein Häufigkeitswert in der Datenbasis eingeführt werden. Der Häufigkeitswert soll angeben wie oft eine Kombination aus Referenzwert und Georeferenz in der Datenbasis vorhanden ist. Dadurch werden Duplikate in der Datenbasis vermieden.

Beim abspeichern eines neuen Datensatzes soll nun zunächst geprüft werden ob dieser bereits in der Datenbasis vorhanden ist. Ist dies der Fall, so wird der Häufigkeitswert des entsprechenden Eintrags um 1 erhöht. Ist der Datensatz noch nicht gespeichert so wird ein neuer Datnesatz angelegt und der Häufigkeitswert mit 1 initialisiert. In Abbildung **??** ist der Ablauf und die dargestellt.

Der Häufigkeitswert misst damit wie oft ein potenzieller geografischer Indikator an einer geografischen Position vorkommt. Es ist zu erwarten, dass ein geografischer Indikator gehäuft an einer bestimmten geografischen Position oder Region auftritt. Indikatoren die keinen geografischen Bezug haben treten hingegen sehr verteilt auf oder nur sehr selten. Durch die Struktur der Datenbasis kann dies analysiert werden.

Die Georeferenz besteht momentan aus den geografischen Koordinaten des zugehörigen Tweets. Die geografische Position wird zumeist mit Hilfe von GPS-Modulen in Smartphones bestimmt. Diese können eine Position auf wenige Meter genau bestimmen. Das bedeutet, zwei Tweets die wenige Meter voneinander abgesetzt wurden haben unterschiedliche Werte für den Längen- und Breitengrad. Dies kann für die Bestimmung der Häufigkeiten problematisch sein, da die Werte in der Regel nicht exakt übereinstimmen. Um dies zu umgehen sollen die geografischen Koodinaten auf Städte abgebildet werden. Dazu müssen die geografischen Koordinaten einer Vorverarbeitung unterzogen werden.

4.7.4 Vorverarbeitung der geografischen Koordinaten

Die geografischen Koordinaten sollen auf Stödte aufgelöst werden. Dies enstpricht der untersten Ebene der geografischen Hierarchie. Insbesondere werden damit auch alle anderen geografischen Hierarchieebenen bestimmt.

Jeder geografischen Koordinate soll die am nächsten gelegene Stadt zugeordnet werden. Als Ergebnis liegt nun pro Referenzwert statt einer geografischen Koordinate eine

Stadt vor. Dies kann als Übergang einer kontinuierlichen Darstellung durch geografische Koordinaten zu einer diskreten Darstellung durch Städte angesehen werden.

Durch die Nearest-Neighbour Zuordnung der geografischen Koordinaten zu Städten wird der Globus in Voronoi-Regionen eingeteilt. Die Punkte zur Erzeugung der Voronoi-Regionen bilden dabei die Städte. Jeder Voronoi-Region kann also eine Stadt zugewiesen werden. Wird ein Tweet innerhalb einer Voronoi-Region abgesetzt wird er der enstprechenden Stadt zugeordnet.

In Abbildung **??** ist die Zuordnung schematisch dargestellt.

nearestNeighbourI

4.7.5 Finale Struktur der Datenbasis

Hier soll nun die Finale Struktur der Datenbasis vorgestellt werden. Mit dem Häufigkeitswert ist ein neuer Wert pro Datensatz hinzugekommen. Des weiteren wurde die Georeferenz auf eine Stadt aufgelöst und damit implizit die Administartionsebene erster Ordnung (Adm1), die Adminstrationsebene zweiter Ordnung (Adm2) und das Land. Neben der Stadt sollen auch die anderen Ebenen pro Datensatz hinterlegt sein. Die daraus resultiernde Struktur der Datenbasis ist in 4.4 inklusive einiger Beispieleinträge dargestellt.

Tabelle 4.4: Struktur der Datenbasis mit geografischer Hierarchie

Referenzwert	Stadt	Adm2	Adm1	
<Los+Angeles><USA>	Los Angeles	Los Angeles County	California	Unite
<Los+Angeles>	Los Angeles	Los Angeles County	California	Unite
<USA>	Los Angeles	Los Angeles County	California	Unite
<Heilbronn>	Heilbronn	Regierungsbezik Stuttgart	Baden-Württemberg	Bundes

Eine andere Darstellung für die Datenbasis kann in **??** betrachtet werden. Der Baum stellt die geografische Hierarchie dar, an die Blätter wird der Referenzwert und die zugehörige Häufigkeit angehängt.

Hierarchiebaum mit Referenzwerten

4.8 Identifizierung geografischer Indikatoren

Die Frage ob es sich bei dem Referenzwert tatsächlich um einen geografischen Indikator handelt kann mit Häufigkeitswerten beantwortet werden. Kommt in einer Voronoi-Region einer Stadt besonders häufig ein bestimmter Referenzwert vor ist dieser sehr wahrscheinlich ein gegografischer Indikator.

Handelt es sich nicht um einen geografischen Indikator wird der Wert entweder sehr verteilt oder sehr selten vorkommen. Im ersten Fall existieren zwar viele Einträge in der Datenbank, aber der Häufigkeitswert wird sehr niedrig sein. Im zweiten Fall wird lediglich der Häufigkeitswert sehr gering sein.

Fazit

Der Referenzwert muss nun nicht zusätzlich auf ein geografisches Objekt der gewünschten Hierarchieebene aufgelöst werden. Aus der Datenbasis kann der Referenzwert für alle Hierarcheieebenen direkt bestimmt werden.

NAch wie vor ist nicht bekannt was für ein geografisches Objekt der Referenzwert beschreibt, ist der Referenzwert ein Land wird

Hierarchiebenen Soll nun ein gegrafischer Indikator

Dadurch wird zunächst das Problem gelöst, dass meherere Einträge zu einem Referenzwert existieren. Denn diese werden nun, vorausgesetzt die Georeferenz stimmt überein

4.9 Einlernen einer Datenbasis

Das Problem besteht grundsätzlich darin, dass nicht alle verwendeten Toponyme bekannt sein können. Zum anderen ist nicht bekannt welche Werte einen mittelbaren geografischen Indikator darstellen.

Die oben genannten Datenbasen

Hinzu soll das Wort Guatsle und Filzl kommen. Beide Wörter beschreiben keine Kartoffel, haben aber aufgrund des Dialekts einen Ein Guatsle ist ein BonBon oder eine Süßigkeit und wird im württembergischen verwendet. Filzl bedeutet Mett und wird in der Pfalz verwendet.

Damit sieht die Liste folgendermaßen aus.

Tabelle 4.5: verschiedene dialektische Begriffe

Referenzwert	Georeferenz
Äbbiera	Württemberg
Guatsle	Württemberg
Grumbeer	Württemberg
Grumbeer	Pfalz
Filzl	Pfalz
Tüfte	Norddeutschland

Mit dieser Liste ergeben sich nun einige Möglichkeiten zur Georeferenzierung.

Beispiel für Lernen Es liegen drei Briefe vor deren Absender unbekannt ist. In den Briefen beschreiben die Autoren ihr Mittagessen. Mit Hilfe der Datenbasis aus 4.5 soll nun entschieden werden woher die Alle Briefe sind in deutscher Sprache verfasst, es soll genügen eine geografische Region innerhalb Deutschlands zu bestimmen. [2]

Die Aufgabe besteht also darin dem Brief eine Georeferenz zuzuordnen. Dies kann durch eine Analyse seines Inhaltes realisiert werden. Der Inhalt eines Briefes sind in der Regel Wörter. Womit jedes Wort einen potenziellen Indikator darstellt.

Nach einer Analyse der Wörter fallen folgende drei Begriffe auf:

Hat man eine solche Datenbasis erstellt können die Briefe automatisch untersucht werden. Für jedes Wort in den Briefen wird geprüft ob es sich um einen dieser Referenzwerte handelt. Wenn ein Treffer vorliegt wird dem Brief die enstprechende Georeferenz zugewiesen.

In diesem Beispiel werden mehrere Probleme der Georeferenzierung klar.

Anhand von Indikatoren wird einem übergeordneten Objekt eine Georeferenz zugewiesen. Die Indikatoren können der Inhalt des Objekts sein oder diese sind auf irgendeine

[2]Die restlichen deutschssprachigen Länder sollen ignoriert werden

Art mit dem Objekt verknüpft. Im Beispiel eines Briefes sind die Worte im Brief potenzielle Indikatoren.

Auch ist nicht garantiert, dass in einem Brief das Wort Kartoffel in dialektischer From vorkommt. In einem Geschäftsbrief werden dialektische Worte nicht vorkommen und diesen kann somit keine Georeferenz zugewiesen werden.

Die Zuordnung ist hier mit einer Gewissen Unsicherheit verbunden. Der Autor könnte unter Umständen in der Region aufgewachsen sein aber nun in einer anderen Region wohnen und dieses Wort aus Gewohnheit benutzen. Dies soll hier aber zunächst ignoriert werden.

Es gibt dabei allerdings zwei grundsätzliche Probleme. Zunächst muss bekannt sein, dass die Referenzwerte auch in den untersuchten Indikatoren vorkommen. In einem Geschäftsbrief werden keine dialektischen Begriffe genutzt werden und somit bringt die

1. Wie kann einem solchen Wert eine Georeferenz zugeordnet werden?
2. Wie kann sichergestellt werden, dass dieser Wert auch genutzt wird und somit eine Georeferenzierung erst möglich wird?

Zu jeder Zeichenkette ist eine Georeferenz bekannt Im Beispiel in 4.1 ist für jedes mögliche Toponym eine Georeferenz in Form einer Adresse bekannt. Dadurch wird angenommen, dass alle Anfragen an die Datenbasis ausschliesslich die drei angegbenen Werte beinhalten. Dabei ist nicht die Anzahl der möglichen Werte das Problem. Die Tabelle könnte nach beliebig erweitert werden. Das Problem sind Anfragen die nicht in der Datenbasis gespeichert sind.

Es gibt mehrere Möglichkeiten wieso die Werte nicht in der Datenbasis vorhanden sind.

1. Die

Die geografischen Indikatoren sind Toponyme Dies muss nicht unbedingt der Fall sein. Vorstellbar ist hier

Diese Struktur erfüllt grundsätzlich alle Eigenschaften die benötigt werden um eine Georeferenzierung durchzuführen. In der Praxis ist eine Georeferenzierung auf diese Weise allerdings nur schwer umzusetzen.

Vollständigkeit der Datenbasis

Im Minimalbeispiel in 4.1 wird davon ausgegangen, dass alle Toponyme und die zugehörigen Georeferenzen bekannt sind.

Doppel- und Mehrdeutigkeiten bei geografischen Indikatoren

Einheit der Georeferenz

Im obigen Beispiel werden lediglich drei Werte angegeben für die eine Georeferenz

Erstellung einer Datenbasis zur Georeferenzierung

Wie kann eine solche Datenbasis nun erstellt werden. Im einfachsten Fall sind alle verwendeten geografischen Indikatoren und deren Georeferenz bekannt und können so manuell eingepflegt werden. Dies ist vorstellbar für ein verzeichnis von Ladengeschäften,

Schema zum einlernen der Datenbasis

Um die Georeferenz-Basis einlernen zu können wird zunächst eine umfangreiche Tweet-Sammlung benötigt. Jeder Tweet in dieser Sammlung muss den Nutzer-Standort, die Nutzer-Zeitzone sowie eine zugehörige Georeferenz enthalten. Die Georeferenz liegt dabei in Form von Längen- und Breitengrad vor.

Pro Tweet werden, entsprechend dem Ablauf aus Abbildung **??**, folgende Verarbeitungsschritte ausgeführt:

1. Vorverarbeitung der Indikatoren
2. Vorverarbeitung der geografischen Koordinaten
3. Speichern der Daten in die Georeferenz-Basis

Der letzte Schritt, das Speichern der Daten in die Georeferenz-Basis, läuft dabei immer nach dem gleichen Prinzip ab und wird deshalb kurz erläutert.

Aus den Schritten 1 und 2 resultiert mindestens ein Tupel bestehend aus einem Toponym und einer Georeferenz. Es wird zunächst überprüft ob dieses Tupel bereits existiert. Ist exakt dieses Tupel in der Georeferenz-Basis vorhanden wird der Häufigkeitswert dieses Datensatzes um einen Zähler erhöht. Ist das Tupel nicht in der Georeferenz-Basis vorhanden wird ein neuer Datensatz angelegt und der Häufigkeistwert mit 1 initialisiert.

Die Schritte 1 und 2 können potenziell mehrere Verarbeitungsschritte beinhalten. Diese werden in den folgenden Kapiteln erarbeitet.

4.9.1 Einlernen der Datenbasis

4.10 Elbauen in Verwendung der Nutzer-Zeitzone

Bei der Nutzer-Zeitzone hingegen kann der Nutzer nur aus einer List von Werten wählen. Damit ist gesichert, dass der EIntrag einen geografischen Bezug hat. Aber auch die Nutzer-Zeitzone wird nicht geprüft, sodass der Nutzer eine beliebige Zeitzone wählen kann.

4.11 Was über Geocoding API's und Datenbanken

Es existieren auch freie Web-Services über welche eine Georeferenzierung auf diese Art durchgeführt werden kann. Die bekanntesten Anbieter sind Google, Yahoo, Microsoft, Map Quest, geonames.org und Cloud Made. Bei allen Anbietern bestehen jedoch gewisse Einschränkungen um die Anzahl der Anfragen zu reglementieren. Google beispielsweise beschränkt die Anzahl der Abfragen pro Tag auf 2500. Es gibt allerdings bei allen Anbietern die Möglichkeit gegen ein Entgelt mehr Anfragen von dem Service auswerten zu lassen. Es besteht aber auch die Möglichkeit freie Datenbanken zu nutzen. Ein Beispiel hierfür ist die Datenbank von geonames.org.

Des weiteren wird in den obigen Beispielen davon ausgegangen, dass die Indikatoren in einem bestimmten Alphabet sowie einer bestimmten Sprache vorliegen. Es drfte nahezu unmöglich sein mittelbare geografische Indikatoren in einer fremden Sprache zu identifizieren. Toponyme hingegen sind auch in verschiedenen Sprachen und Alphabeten verfügbar.

Wo muss jetzt Identifizierung eines Objekts anhand geografischer Indikatoren hin? Wichtig da mehr text und so in einlernen?

5 Implementierung

Im Rahmen dieser Diplomarbeit ist eine Referenzimplementierung des vorgestellten Verfahrens entstanden. In Auszügen soll die Referenzimplementierung hier vorgestellt werden. Hierbei sollen insbesondere Probleme bei der Umsetzung betrachtet werden, und wie diese gelöst wurden. Damit soll die Möglichekit gegeben werden, in eigenen Implementierungungen die Probleme frühzeitig zu erkennen und zu vermeiden. Des weiteren soll ein Überblick über die genutzten Datensätze und API's gegeben werden.

Datensätze in Grundlagen?

5.1 Komponenten der Referenzimplementierung

5.1.1 Architektur

Allgemeine Architektur der Refernzimplementierung

5.1.2 Präprozessorverarbeitung - Erzeugung der N-Gramme

Warum Präprozessoren -> schnelleres ändern der Vorverarbeitung.

5.2 Datenbank

5.2.1

Eventuell was über die Geo Indexe in der Datenbank und die Nearest Neighbour Berechnungen.

5.3 Geografie Daten

in Implementierung verschieben

5.4 Data Sample

Beschreibung wie Daten erzeugt wurden, Zeiträume, Analysen

5.5 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

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

[BNJ12] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2012.

[CCL10] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, page 759, 2010.

[EOSX10] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, 2010.

[FVMF13a] Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. Traveling trends: social butterflies or frequent fliers? *... conference on Online social ...*, 2013.

[FVMF13b] Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. Traveling trends: Social butterflies or frequent fliers? *CoRR*, abs/1310.2671, 2013.

[GGMQ14] R Garcia-Gavilanes, Y Mejova, and D Quercia. Twitter ain't without frontiers: Economic, social, and cultural boundaries in international communication. *... cooperative work & social ...*, pages 1511–1522, 2014.

[Gol08] Daniel W Goldberg. *A Geocoding Best Practices Guide*. North American Association of Central Cancer Registries (NAACCR), 2008.

[HGG12] S Hale, D Gaffney, and M Graham. Where in the world are you? geolocation and language identification in twitter. *Proceedings of ICWSM'12*, (2013), 2012.

[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.

[IAB] Interactive Advertising Bureau IAB. Social media ad metrics definitions. Internet.

[JJ21] G. Jellinek and W. Jellinek. *Allgemeine Staatslehre.* J. Springer, 1921.

[KA08] Balachander Krishnamurthy and Martin Arlitt. A few chirps about twitter. In *Proceedings of the first workshop on Online social networks (WOSP '08)*, pages 19–24, 2008.

[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.

[KLPM10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter , a social network or a news media? In *The International World Wide Web Conference Committee (IW3C2)*, pages 1–10, 2010.

[MND12] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Where is this tweet from? inferring home locations of twitter users. *ICWSM*, pages 511–514, 2012.

[MPLC13] Fred Morstatter, J Pfeffer, H Liu, and KM Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. *Proceedings of ICWSM*, pages 400–408, 2013.

[NP03] M E J Newman and Juyong Park. Why social networks are different from other types of networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 68:036122, 2003.

[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.

[SHP+13] Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Mühlhäuser. A multi-indicator approach for geolocalization of tweets. *Seventh International AAAI Conference on Weblogs and Social Media*, pages 573–582, 2013.

[SKD11] Martin Szomszor, Patty Kostkova, and Ed De Quincey. #swineflu: Twitter predicts swine flu outbreak in 2009. In *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering*, volume 69 LNICST, pages 18–26, 2011.

[SMvZ09] Pavel Serdyukov, Vanessa Murdock, and Roelof van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*, page 484, 2009.

[SOM10] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. *Proceedings of the 19th international conference on World wide web*, pages 851–860, 2010.

[ti13] twitter inc. Final initial public offering(ipo) prospectus, 11 2013.

[TSSW11] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Election forecasts with twitter: How 140 characters reflect the political landscape, 2011.