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

Die Verbreitung von Nachrichten und Informationen erfolgt immer stärker durch nutzergenerierte Inhalte. Eine Plattform hierfür bietet der Microblogging-Dienst Twitter. Die Nutzer können 140 Zeichen lange Nachrichten, sogenannte “Tweets“, erstellen und veröffentlichen können. Längst ist Twitter zu einem Massenphänomen geworden und übernimmt die Rolle eines Nachrichtenmediums [POM+13]. Die Twitter-Nutzer verfassen täglich mehr als 500 Millionen Tweets [ti13]. Durch die Möglichkeiten die Twitter bietet kann theoretisch jeder Mensch Nachrichten und Informationen über das Twitter-Netzwerk verbreiten und weitergeben. In den Tweets wird unter anderem über Großereignisse, persönliche Erfahrungen oder Erlebnisse berichtet. Die Tweets sind zum Größtenteil öffentlich zugänglich. Dieses enorme Potenzial an Informationen sollte genutzt und verarbeitet werden.

Es lassen sich daraus Erkenntnisse ableiten, mit denen politische und wirtschaftliche Entscheidungsprozesse unterstützt und verbessert werden können. Aber auch Gefahrenpotentiale können rechtzeitig erkannt werden, um gezielte Gegenmaßnahmen einzuleiten. Dies setzt aber voraus, dass die Informationen aus den Tweets möglichst exakten geografischen Koordinaten zugeordnet und dadurch die Ereignisse lokalisiert werden können.

Sakaki et al zeigen, dass mit Hilfe von Tweets mit geografischen Koordinaten Erdbebenzentren lokalisiert oder die Trajektorie eines Typhoons vorhergesagt werden können [SOM10]. Tumasjan et al. untersuchen in [TSSW11] wie sich die politische Landschaft im Twitter-Netzwerk widerspiegelt. Die Wissenschaftler haben zur Bundestagswahl 2009 100.000 Tweets analysiert und stellten fest, dass die Anzahl der Erwähnungen von Parteien und Politikern in Twitter, den Wahlausgang sehr genau abbildeten.

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 Schweinegrippe im Jahr 2009 durch die Analyse von Tweets eine Woche früher möglich gewesen wäre als dies mit konventionellen Frühwarnsystemen der Fall war.

Diese wichtigen Erkenntnisse und Vorhersagen konnten nur aufgrund von Tweets mit geografischen Koordinaten ermittelt werden. Allerdings weisen nur ca. 1% der Twitter-Kurznachrichten geografische Koordinaten auf [SHP+13]. Dies ist ein sehr geringer Wert, wenn mehr Tweets mit einer zugeordneten geografischen Position zur Verfügung stehen würden, könnten diese Verfahren effizienter genutzt werden.

1.2 Problembeschreibung

Wie kann es ermöglicht werden, dass Tweets ohne geografische Koordinaten eine geografische Position zugeordnet werden kann?

Twitter bietet seinen Nutzern diverse Möglichkeiten persönliche Angaben zu machen. Unter anderem kann im Nutzerprofil ein Standort angegeben werden. Bei der Eingabe des Nutzer-Standortes wird vom Twitter-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. Dieses Feld eignet sich deshalb gut um daraus eine geografische Position abzuleiten. Betrachtet man den Nutzer-Standort jedoch genauer fällt auf, dass dieser nicht ohne weiteres zur Bestimmung einer geografischen Position verwendet werden kann.

Naheliegend ist, dass der Nutzer im Nutzer-Standort einen Ortsnamen (Toponym) verwendet um seinen Standort anzugeben. Der Nutzer-Standort wird jedoch über ein Freitext-Feld eingegeben und direkt abgespeichert. Durch die freie Eingabe werden unter ande-

rem Abkürzungen, größere geografische Regionen oder spezielle Bei- und Spitznamen im Nutzer-Standort eingegeben. Wenn der Nutzer-Standort ein Toponym darstellt, können zudem Probleme wie Mehr- und Doppeldeutigkeiten auftreten.

Der Nutzer-Standort muss aber nicht zwangsweise Werte mit geografischem Bezug enthalten.

Zudem gibt der Nutzer-Standort nicht unbedingt den Ort an, von dem der Tweet versendet wurde. Tweets mit geografischen Koordinaten werden zumeist von mobilen Endgeräten versendet. Der Nutzer muss sich also zum Zeitpunkt des Absendens eines Tweets nicht an dem im Nutzer-Standort angegeben Ort aufhalten.

Der Nutzer-Standort ist eine freiwillige Angabe des Twitter-Nutzers im Nutzer-Profil. Von Hecht et al. [HHSC11] wird der Inhalt der Nutzer-Standorte von 100.000 Nutzer-Profilen manuell analysiert. Ca. 66% aller analysierten Nutzer-Standorte enthalten demnach einen Wert mit geografischem Bezug. Es könnten somit zusätzlich 66% der Twitter-Nutzer eine geografische Position zugeordnet.

Der Nutzer-Standort bietet somit ein großes Potenzial um weiteren Twitter-Nutzern eine geografische Position zuzuordnen. Um eine Geolokalisierung von Tweets mit Hilfe des Nutzer-Standortes realisieren zu können, müssen die oben genannten Probleme bezüglich der angegebenen Werte im Nutzer-Standort gelöst werden. Zusätzlich muss verifiziert werden, dass der Nutzer-Standort den Ort des Absendens eines Tweets hinreichend genau beschreibt.

Abhängig von der Anwendung sind zudem die Anforderungen bezüglich der Genauigkeit, Trefferquote und der gewünschten Zuordnung des Ergebnisses zu einer geografischen Region unterschiedlich.

1.3 Fragestellungen

Aus der Problembeschreibung ergeben sich folgende Fragestellungen:

- Wie können die oben genannten Probleme, der eingegebenen Werte im Nutzer-Standort, weitestgehend eliminiert werden?
- Wie genau kann aus dem Nutzer-Standort die Position, von der ein Tweet abgesendet wurde, bestimmt werden?
- Ist es möglich die Ergebnisse bezüglich Genauigkeit und Trefferquote zu justieren?

1.4 Zielsetzung

Das übergeordnete Ziel dieser Arbeit besteht darin, Tweets mit Hilfe des angegebenen Nutzer-Standortes einer geografischen Position zuzuordnen. Dadurch soll die Position von der ein Tweet versendet wurde möglichst genau bestimmt werden.

Es soll dazu ein Verfahren zur Geolokalisierung von Tweets entwickelt werden, welches die Probleme der angegebenen Werte im Nutzer-Standort so weit wie möglich eliminiert. Das Verfahren soll es ermöglichen, durch eine Vorgabe die Genauigkeit und die Trefferquote der Ergebnisse zu bestimmen. Twitter wird weltweit genutzt, das Verfahren soll deshalb auch unabhängig von unterschiedlichen Sprachen und Schriftzeichen funktionieren.

1.5 Gliederung der Arbeit

In Kapitel 2 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 Abschnitt 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.

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 Abschnitt werden die Ergebnisse der Referenzimplementierung bewertet und, soweit sinnvoll, gegenüber bestehenden Ansätze einer kritischen Betrachtung unterzogen.

Abschnitt 7: Zusammenfassung und Ausblick

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

2 Grundlagen

Im ersten Abschnitt wird der Microblogging-Dienst Twitter und seine Funktionen vorgestellt. Danach werden geografische Begriffe und Verfahren, sowie Verfahren z

2.1 Der Microblogging-Dienst Twitter

Zunächst wird definiert welchen Dienst Twitter anbietet. Danach wird ein geschichtlicher Überblick gegeben und einige Funktionen von Twitter erläutert. Zum Schluss wird aufgezeigt welche Informationen in einem Tweet übermittelt werden.

2.1.1 Was ist Twitter?

Twitter wird als Kurznachrichten-Dienst, Microblogging-Dienst oder auch als Soziales-Netzwerk bezeichnet.

Twitter Geschäftsführer Kevin Thau hat 2010 auf dem Nokia-World-Congress 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 verbreitet werden geändert hat. Thaus Meinung nach kann praktisch jeder durch Twitter 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 weiterer Beispiele derselben Art finden.

In [POM+13] wird ein Vergleich zwischen sogenannten Newswire anbietern und Twitter gezogen. [1] Petrovic et al. fanden heraus, das nahezu über 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.

[1] 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.

Die Bezeichnung Kurznachrichten-Dienst ist irreführend, da dieser mit sms [2] 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. Beim sms wird eine Nachricht direkt an einen Empfänger gesendet und nicht öffentlich verbreitet. Im sms steht die Konversation zweier Nutzer im Vordergrund, wohingegen Nachrichten im Twitter-Netzwerk einen Broadcast 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 von den Dächern“ auch im Twitter-Netzwerk zu einer passenden Redewendung wird. Aufgrund der Erkenntnisse von [KLPM10] und der schlüssigen Argumentation von Kevin Thau ist es naheliegend Twitter als Microblogging-Dienst zu bezeichnen.

2.1.2 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. Die Gründer sind davon überzeugt sind, dass anhand der reinen Nutzer-Zahlen und gesendeten Twitter-Kurznachrichten nicht die “Gesundheit“ des Twitter-Netzwerks nachvollzogen werden. Andererseits werden durch diese Maßnahme auch strategische Ziele verfolgt. [3] 2013 ging Twitter an die Börse und vermeldete 100 Millionen täglich aktive Nutzer

[2] small messeneger service
[3] http://www.pbs.org/mediashift/2007/05/twitter-founders-thrive-on-micro-blogging-constraints137

und über 500 Millionen Twitter-Kurznachrichten, die täglich über den Dienst versendet werden.

2.1.3 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". Danach werden Funktionen wie das weitergeben eines Tweets, Favorisieren und Antworten erklärt. Zum Schluss wird auf den gesendeten Tweet-Inhalt eingegangen.

Abbildung 2.1: Die Tweet-Timeline

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 Infor-

mationen. In Abbildung 2.1 ist in der Mitte die Timeline des Benutzers dargestellt in der die Tweets zu sehen sind. Folgende Bereiche können unterschieden werden.

1. Nutzername und Informationen über den Nutzer.
2. Profilbild
3. Allgemeine Informationen über den Benutzer und dessen Netzwerk
4. Nutzer-Timeline: Tweets des Nutzers in umgekehrter chronologischer Reihenfolge
5. Button zum Folgen

Unter dem Profilbild links sind Informationen des Nutzers aufgelistet. Diese Informationen kann der Nutzer selbst einstellen.

Folgen (Following/Follower) 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 auch im deutschsprachigen Raum etabliert. Auch auf der Twitter Website wird ’"Follower“ nicht ins deutsche übersetzt. In der vorliegenden Arbeit wird ebenfalls auf eine Übersetzung verzichtet.

In Abbildung 2.1 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“. In 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 ein Twitte-Nutzer Tweets anderer Nutzer empfangen und lesen kann. 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. Genauer gesagt wird der Tweet übernommen und ein Hinweis hinzugefügt, dass es sich um einen sogenannten "Retweet" handelt. Es wird damit gekennzeichnet, dass dieser Tweet nicht von dem Nutzer selbst verfasst wurde. 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. Ein Nutzer gibt nur diejenigen Tweets weiter denen er eine Gewisse Relevanz beimisst. Oder von denen er erwartet, dass sie seine Follower interessieren könnten. 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. Nachrichten in Twitter stellen keinesfalls gesicherte Fakten dar. Auch eine große Verbreitung durch viele Retweets ändert daran nichts. Durch diese Funktion können Nutzer zu Tweet-Aggregatoren werden. Der Nutzer abonniert viele Nutzer, aber gibt nur relevante oder themenspezifische Tweets an seine Follower weiter.

2.1.4 Die Twitter Streaming API

Die Twitter Streaming API [4] ermöglicht es Tweets programmatisch abzurufen. Dabei wird von Twitter ein Sample aller aktuell erstellten Tweets zur Verfügung gestellt. Das Sample beinhaltet maximal 1% aller aktuell erstellten Tweets. Die Streaming API ermöglicht es ebenfalls nach bestimmten Eigenschaften der Tweets zu suchen. Beispielsweise lassen sich bestimmte Stichworte angeben anhand derer die Ergebnisse gefiltert werden sollen. Es können auch ausschließlich Tweets abgefragt werden welche eine Georeferenz in Form von Längen- und Breitengrad aufweisen.

Durch eine programmatische Abfrage an die Streaming Api können automatisiert große Datensätze zu Analysezwecken erstellt werden.

[4] Application Programming Interface

2.1.5 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 erscheint dieser wie in Abbildung 2.2. Folgende Bereiche sind zu sehen:

1. Profilbild des Verfassers
2. Name, Benutzername und Zeit
3. Der Tweet-Text

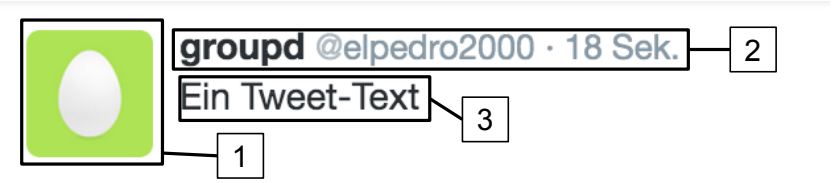

Abbildung 2.2: Ein Tweet

Erfasst man einen Tweet allerdings über die Streaming API wird eine Reihe weiterer Daten mitgeliefert. Mit jedem Tweet werden auch Daten aus dem Twitter-Profil des Verfassers versendet. Es werden im folgenden der Nutzer-Standort, die Nutzer-Zeitzone und die geografischen Koordinaten betrachtet.

Geografische 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 erfolgt über ein GPS-Modul in einem Smartphone oder mit Hilfe von GeoIp an einem PC. Das Anhängen der Koordinaten an einen Tweet wird vollautomatisch durch das Programm übernommen, mit welchem der Tweet verfasst wurde.

Abbildung 2.3: Eingabe des Nutzer-Standortes

Nutzer-Standort Der Nutzer-Standort wird vom Benutzer in den Profil Einstellungen eingegeben. In Abbildung 2.3 ist der Dialog zur Nutzer-Standort Eingabe in den Twitter-Profil Einstellungen abgebildet.

Der Nutzer-Standort wird direkt übernommen und unterliegt keiner weiteren Verarbeitung seitens Twitter. Dies bedeutet der Nutzer-Standort wird übernommen wie er vom Nutzer eingegeben wird. Das Dialogfeld ist lediglich auf ein Maximum von 30 Zeichen beschränkt.

Nutzer-Zeitzone Die Nutzer-Zeitzone kann vom Benutzer in den Profil Einstellungen aus einer Liste gewählt werden. Der Datenwert der Nutzer-Zeitzone stellt also garantiert eine Zeitzone dar. In Abbildung 2.4 ist der Auswahldialog in den Twitter-Profil Einstellungen zu sehen.

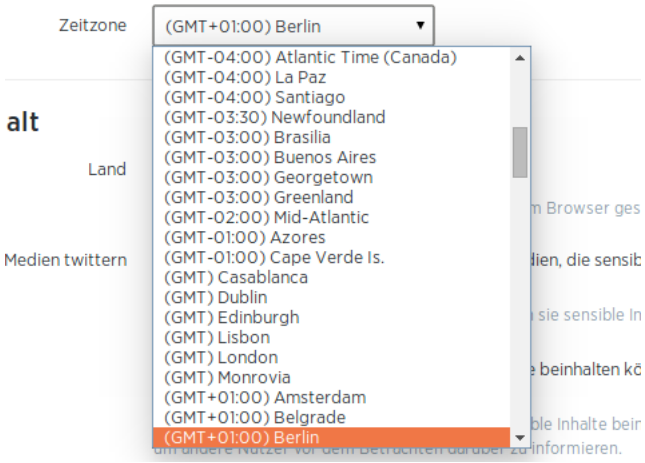

Abbildung 2.4: Zeitzonen Auswahldialog

Es wird nicht geprüft ob der Nutzer sich auch in dieser Zeitzone befindet. Der Nutzer könnte eine bewusste Fehleingabe machen oder aber die Zeitzone nicht wählen. Der Standardwert der Nutzer-Zeitzone ist dann “Pacific Time (US and Canada)“.

2.2 Geografische Grundlagen und Begriffe

In diesem Kapitel sollen geografische Grundbegriffe 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.

2.2.1 Geografische Koordinaten

Eine Position auf dem Globus wird mit zwei Werten beschrieben, dem sogenannten Längengrad und dem sogenannten Breitengrad. Diese beiden Werte werden als geografische Koordinaten bezeichnet. Mit diesen zwei Werten und einem geodätischen Referenzsystemsystem kann die Position auf dem Globus exakt bestimmt werden. Die geografischen Angaben in Twitter liegen bezüglich des geodätischen Referenzsystems WGS84 vor und sind nur für dieses gültig. Werden diese bezüglich eines anderen geodätischen Referenzsystems ausgewertet ist die Position auf dem Globus nicht korrekt. Es ist dann eine Transformation der Werte erforderlich. Heutzutage ist das Referenzsystem WGS84 weit verbreitet.

Ein geodätisches Referenzsystem besteht aus einem kartesischen Rechtssystem mit definierter Lage und Ausrichtung, einem Referenzellipsoid und einer Festlegung zur Messung der Winkel.

Die Lage und Ausrichtung des kartesische Rechtssystems erfolgt relativ zur Erde. Der Ursprung des Koordinatensystems liegt im Zentrum des Globus. Die Z-Achse zeigt dabei in Richtung Nordpol und die X-Achse in Richtung 0 Grad Länge (Nullmeridian) und 0

Grad Breite (Äquator). Mit diesen zwei Werten ist die Lage eines kartesischen Rechtssystem eindeutig definiert. In Abbildung 2.5 sind sowohl die Lage des Nullmeridian und des Äquators sowie die Lage des kartesischen Referenzsystems dargestellt.

In diesem Koordinatensystem sind zusätzlich Referenzpunkte festgelegt. Diese Referenzpunkte werden benötigt um einen Referenzellipsoid zu verankern. Auf diesem Ellipsoid sind ebenfalls definierte Referenzpunkte festgelegt, die mit den Referenzpunkten im Koordinatensystem zur Deckung gebracht werden. Der Referenzellipsoid soll eine möglichst genaue Approximation der Erde darstellen und diese im geodätischen Referenzsystem repräsentieren. Ein Punkt auf diesem Ellipsoid entspricht 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 P auf dem Ellipsoid lässt sich folgendermaßen bestimmen:

Durch den Punkt P auf dem Ellipsoid und den Ursprung z des Koordinatensystems wird eine Gerade g gezogen. Der Wert für den Breitengrad ist nun der Winkel ϕ zwischen g und der Äquatorebene. Nun wird der Punkt P auf einen Punkt Q auf der Äquatorebene projiziert. Zwischen z und dem projizierten Punkt Q kann nun wiederum eine Gerade h gezogen werden. Der Wert für den Längengrad ist der Winkel λ zwischen der X-Achse und der Gerade h (Abbildung 2.5). [5]

Referenzen Bild

2.2.2 Georeferenz

Ist einem Datensatz, einem Datum oder einem Objekt eine geografische Lage oder Position zugeordnet so wird diese als Georeferenz (auch Raumbezug genannt) bezeichnet. Eine Georeferenz kann auf verschiedene Arten und mit unterschiedlicher Genauigkeit angegeben werden. Dies hängt 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. [5]

[5] Vergleiche Geoinformatik Lexikon der Universität Rostock (abgerufen Juli 2014): http://www.geoinformatik.uni-rostock.de/lexikon.asp
Vorlesungen zur Geo-Informatik von Prof. Dr.-Ing. Ralf Bill (abgerufen Juli 2014): http://www.geoinformatik.uni-rostock.de/vorlesungsthem.asp Juli 2014

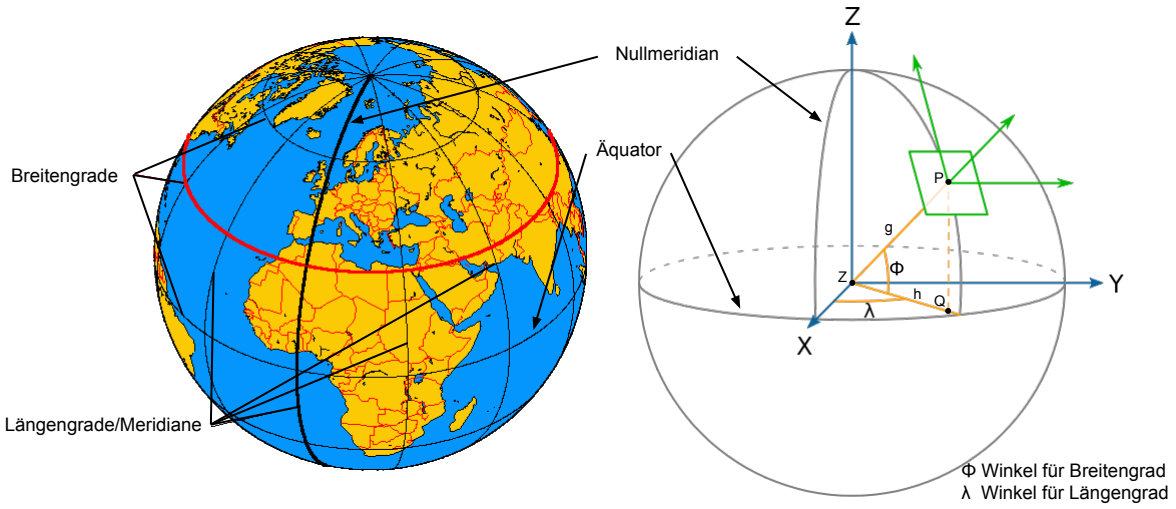

Abbildung 2.5: Nullmeridian und Äquator (linke Bildseiten);Lage des kartesischen Rechtssystems(rechte Bildseite)

Die Georeferenz lässt sich bezüglich der Genauigkeit weiter unterteilen in:

Direkte Georeferenz (Direkter Raumbezug) Unter einer direkten Georeferenz versteht man die Angabe einer konkreten Koordinate bezüglich eines geeigneten geodätischen Referenzsystems. [5]

Indirekte Georeferenz (Indirekter Raumbezug) Unter indirektem Raumbezug werden alle Angaben verstanden die eine ungenaue Position bezüglich eines beliebigen 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. [5]

2.2.3 Geografische Objekte

Ein geografisches Objekt ist ein Objekt der Realwelt dessen Position durch eine Georeferenz bestimmt werden kann. Die EN ISO 19110:2005 Norm beschreibt ein geografisches

Objekt folgendermaßen:

“Geografische Objekte sind Erscheinungen der realen Welt, die einen Bezug zur Erde (Raumbezug) haben...“ [ISO].

Es wird insbesondere nicht festgelegt ob es sich dabei um eine direkte oder eine indirekte Georeferenz handelt. Beispiele für geografische Objekte sind: Städte, Länder, Häuser oder auch Fahrzeuge. Insbesondere sind auch Menschen geographische Objekte, da sie zu jeder Zeit einen Bezug zur Erde haben.

2.2.4 Geografische Position

Unter einer geografischen Position wird in dieser Arbeit eine Position auf dem Globus verstanden die durch geografische Koordinaten bestimmt.

2.2.5 Toponyme

Toponyme werden in der vorliegenden Arbeit definiert als, Namen für geografische Objekte mit unveränderlicher geografischer Position sein. Beispiele für Toponyme sind Städtenamen, Ländernamen oder Landschaftsnamen.

2.2.6 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 ein Polygone beschrieben. Das Polygone wird durch eine Menge geografischer Positionen bestimmt. Diese werden in einer festgelegten Reihenfolge durch eine Linie verbunden.

Bundesländer oder Länder sind Beispiele für geografische Regionen.

2.2.7 Geografischer Bezug

Kann einem Datenwert in irgendeiner Weise eine Georeferenz zugeordnet werden hat dieser Datenwert geografischen Bezug.

Datenwerte mit geografischem Bezug können weiter unterteilt werden in Datenwerte mit unmittelbarem geografischen Bezug oder mittelbarem geografischen Bezug.

Datenwerte mit unmittelbarem geografischem Bezug Einem Datenwert mit unmittelbarem geografischen Bezug lässt sich durch die in ihm enthaltene Information eine Georeferenz zuweisen.

Beispielsweise haben Zeitzonen unmittelbaren geografischen Bezug, da die in ihnen enthaltene Information unmittelbar einer Georeferenz zugewiesen werden kann. Auch Toponyme die eindeutig sind haben unmittelbaren geografischen Bezug.

Werte mit mittelbarem geografischem Bezug Ein Wert hat genau dann mittelbaren geografischen Bezug, wenn die in ihm enthaltene Information nicht direkt auf ein geografisches Objekt verweist, ihm aber trotzdem eine Georeferenz zugeordnet werden kann. Dabei ist die eigentliche Information des Wertes unerheblich. Der geografische Bezug erfolgt beispielsweise durch die geografisch begrenzte Verwendung des Wertes.

Dies soll an einem Beispiel erläutert werden:
Auf einer Website sollen die Nutzer alternative Begriffe eingeben. Die Eingabe erfolgt über ein Freitext-Feld. Die folgenden drei Datenwerte werden von Nutzern eingegeben.

1. Äbierra
2. Grumbeer
3. Tüfte

Die drei Begriffe bezeichnen ein Gemüse, genauer Kartoffeln. Die Datenwerte bezeichnen also insbesondere kein geografisches Objekt und haben somit keinen unmittelbaren geografischen Bezug. Jeder dieser Bezeichnungen stammt aber aus unterschiedlichen Regionen Deutschlands, denn es handelt sich um dialektische

Begriffe. Äbierra wird in Baden-Württemberg, Grumbeer in der Pfalz und Tüfte in Norddeutschland verwendet. Durch ihre geografisch begrenzte Verwendung können sie damit einer geografischen Region zugeordnet werden. Durch diese Datenwerte kann also auf eine Region Deutschlands geschlossen werden und somit auf ein geografisches Objekt.

2.2.8 Geografische Indikatoren

Liefert ein Datensatz Informationen zu einem Objekt, dessen Georeferenz unbekannt ist, kann aus diesen Daten möglicherweise eine Georeferenz abgeleitet werden. Dies ist genau dann der Fall, wenn in dem Datensatz Datenwerte enthalten sind, die mittelbaren oder unmittelbaren geografischen Bezug aufweisen. Diese Datenwerte können als Hinweis auf die Georeferenz des Datensatzes genutzt werden. Diese werden hier als geografische Indikatoren bezeichnet.

In Abbildung 2.6 ist der Zusammenhang zwischen geografischen Indikatoren, geografischem Bezug, geografischem Objekt und einer Georeferenz dargestellt. Das geografische Objekt A hat eine Georeferenz G. Der Datensatz A liefert Informationen zum geografischen Objekt A. Information a und Information b haben einen geografischen Bezug zu G und sind somit geografische Indikatoren. Ist zum geografischen Objekt A keine Georeferenz bekannt, so lässt sich durch die geografischen Indikatoren a und b die Georeferenz G ableiten.

2.2.9 Geografische Hierarchie

In der vorliegenden Arbeit wird eine geografische Hierarchie verwendet um eine Einteilung der Erde in geografische Regionen umzusetzen.

Eine Aufteilung der Erde in geografische Regionen lässt sich auf oberster Ebene mit Hilfe von Ländern und deren Grenzen umsetzen. Die meisten Länder sind in weitere administrative Einheiten aufgeteilt. Diese geografischen Regionen werden hier als Verwaltungseinheiten bezeichnet. Es wird zwischen Verwaltungseinheiten erster und zweiter Ordnung unterschieden. Der Vatikan-Staat und das Fürstentum Monaco sind dabei Ausnahmen. Beide werden aufgrund ihrer Größe nicht in weitere Verwaltungseinheiten

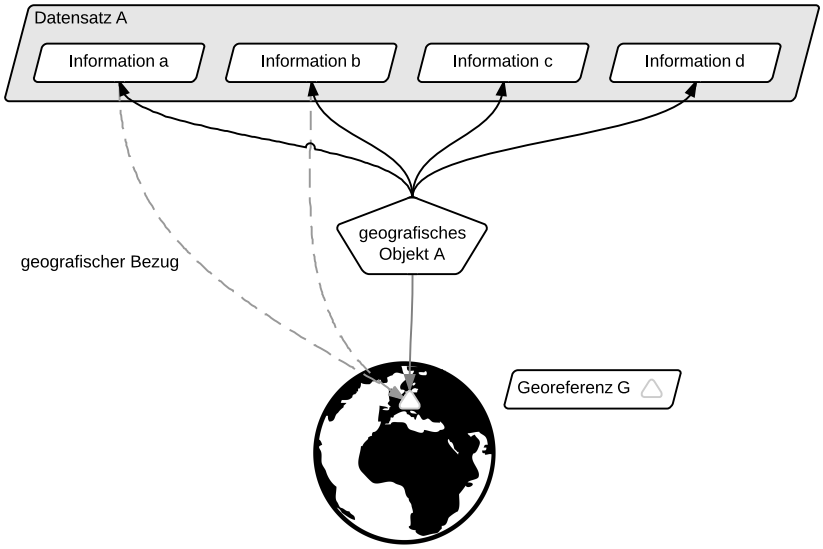

Abbildung 2.6: Geografische Indikatoren

unterteilt. In der untersten Ebene der Hierarchie werden schließlich Städte dargestellt. In der vorliegenden Arbeit wird das Ortsverzeichnis von geonames.org verwendet. Dieses bildet die hier vorgestellte geografische Hierarchie ab.

Wenn man als Beispiel Deutschland heranzieht, ergibt sich eine Einteilung wie in Abbildung 2.7 dargestellt. [6] 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. Jedes Objekt einer Hierarchieebene beschreibt eine geografische Region. Insbesondere besteht in einer solchen Hierarchie eine Teilmengenbeziehung zwischen den Ebenen. Ein Objekt in einer Ebene liegt immer innerhalb der ihm übergeordneten Objekte. Insbesondere liegt die geografische Region die ein Objekt beschreibt komplett innerhalb des ihm übergeordneten Objekts.

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

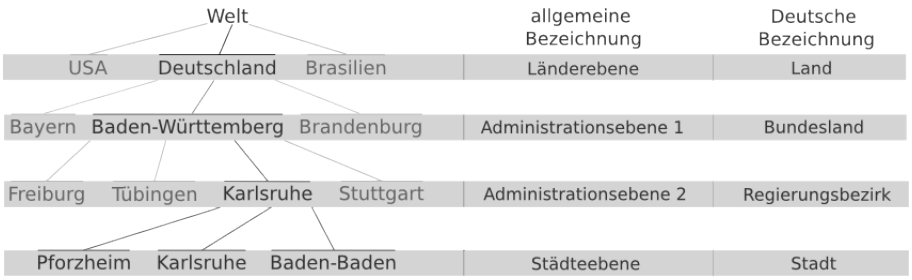

Abbildung 2.7: Beispiel für geografische Hierarchieebenen

In Abbildung 2.8 ist die Einteilung des Globus in Länder und Verwaltungseinheiten dargestellt.

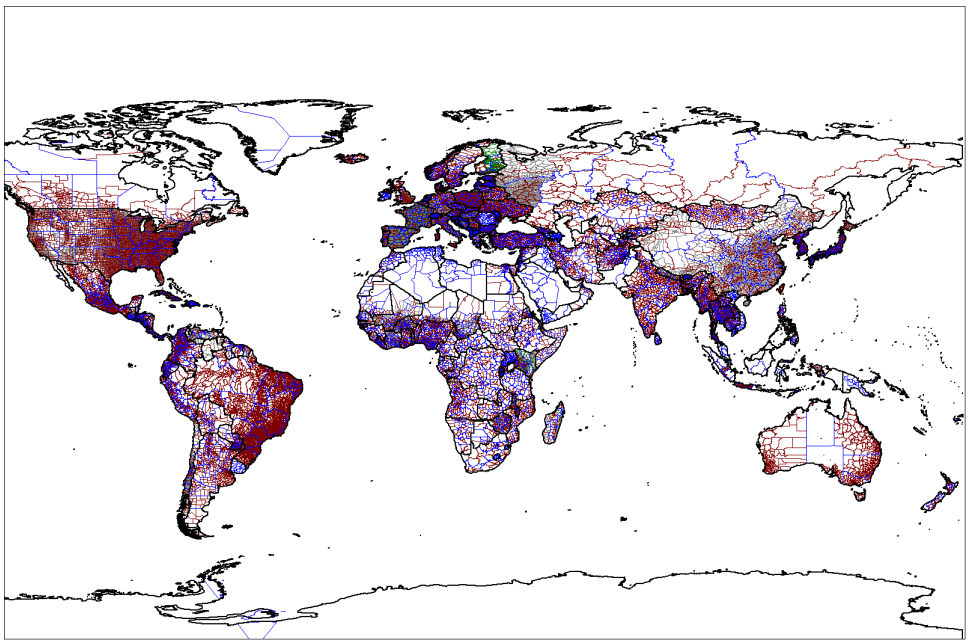

Abbildung 2.8: Aufteilung der Welt in Verwaltungseinheiten

2.2.10 Ortsverzeichnisse

In einem Ortsverzeichnis [7] sind Toponyme und deren zugehörige Georeferenzen gespeichert. Ortsverzeichnisse können sehr umfangreich sein und werden in Form einer Datenbank bereitgestellt. Meistens werden zu den Toponymen noch weitere Informationen

[7] Englisch: Gazetteer

hinterlegt. So bilden die Ortsverzeichnisse beispielsweise oft geografische Hierarchiebeziehungen ab. Aber auch Angaben zur Bevölkerungszahl oder die Angabe von alternativen Toponymen sind oft hinterlegt.

Ortsverzeichnisse können genutzt werden um Toponymen eine Georeferenz zuzuweisen. Dabei wird ein gegebenes Toponym im Ortsverzeichnis nachgeschlagen um die zugehörige Georeferenz zu erhalten.

Eines der bekanntesten Ortsverzeichnisse ist das frei erhältliche Ortsverzeichnis von geonames.org. Dieses Ortsverzeichnis kann als CSV Datei[8] heruntergeladen werden. Neben umfangreichen Informationen wie Bevölkerungszahl, Sprache, Längen- und Breitengrad wird die geografische Hierarchie abgebildet.

Die Daten von geonames.org werden aus diversen Quellen automatisch zusammengetragen. [9] Es werden allerdings auch Einträge von Nutzern erstellt. Mittlerweile hat sich eine aktive Community rund um das Projekt entwickelt.

Das Ortsverzeichnis beinhaltet ca. 8,8 Millionen Toponyme und zugehörige Informationen. [10] Hinzu kommen ca. 8 Millionen alternative Toponyme. Geonames.org ist damit eine der umfangreichsten frei erhältlichen Ortsverzeichnisse.

Jedem Toponym ist eine geografische Position in Form von Längen- und Breitengrad zugeordnet. Es wird allerdings kein Polygon zur Beschreibung der geografischen Region angegeben. Durch die Teilmengen Beziehung die in der geografischen Hierarchie abgebildet ist, können jedoch immer alle geografischen Hierarchieebenen bestimmt werden.

In der vorliegenden Arbeit wird die geonames.org Datenbank mit Stand Dezember 2013 als Basis für geografische Informationen genutzt. Auch die verwendete geografische Hierarchie lässt sich aus der Datenbank gewinnen.

[8] Comma Sperated Value

[9] Die genutzten Datenquellen können unter http://www.geonames.org/data-sources.html eingesehen werden

[10] Stand Dez. 2013, die Daten unterliegen ständiger Bearbeitung, die Anzahl der Einträge kann deshalb variieren

2.2.11 Toponyme in geografischen Indikatoren

In einem Datensatz zu einem geografischen Objekt können als Datenwerte Toponyme auftauchen. Toponyme eignen sich grundsätzlich gut als geografische Indikatoren. Sie geben den Namen eines geografischen Objekts an und haben unmittelbaren geografischen Bezug. Mit Hilfe von Ortsverzeichnissen kann Toponymen eine Georeferenz zugeordnet werden.

Ein Problem kann allerdings darstellen, dass Toponyme nicht standardisiert sind. Im folgenden werden die Probleme genauer betrachtet.

Alternative Toponyme

In Ortsverzeichnissen kann eine große Anzahl an Toponymen hinterlegt werden. Aufgrund der immensen Vielfalt an Toponymen ist es aber nahezu unmöglich alle existierenden Toponyme abzudecken.

Neben den offiziellen Namen für Städte, Länder usw. existieren eine Reihe von alternativen Toponymen. Auch Toponyme die spezielle geografische Objekte bezeichnen sind möglich.

Ein Beispiel für alternative Toponyme sind Bei- und Spitznamen für Städte. In Wikipedia sind für die Stadt Detroit, im US-Bundesstaat Michigan, folgende Beinamen angegeben:

“The Motor City“, “Motown“, “Hockeytown“, “Rock City“, “The D“.

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 Bei- und Spitznamen. Durch eine Abfrage an dieses Ortsverzeichnis könnte somit keine Georeferenz bestimmt werden. Eine Abfrage in Google Maps hingegen bietet bei Eingabe der oben genannten Bei- und Spitznamen Detroit als Vorschlag an. [11]

[11] http://de.wikipedia.org/wiki/Detroit (abgerufen Juli 2014)

Mehrdeutigkeiten von Toponymen

Toponyme sind oft Doppel- oder Mehrdeutig und verweisen somit auf mehrere Georeferenzen.

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 2.2 sind einige Städte-Namen und die Vorkommen in den USA aufgelistet.

Tabelle 2.1: Häufige deutsche Städtenamen in den USA

Name	Anzahl in den USA
Hannover	40
Berlin	39
Hamburg	30

Bei einer Abfrage des Toponyms “Hamburg“ auf ein Ortsverzeichnis, wie geonames.org, würden somit mindestens 31 Georeferenzen als Ergebnis erscheinen. Tatsächlich liefert geonames.org folgendes Ergebnis:

Tabelle 2.2: Abfrage an geonames.org für Hamburg

Land	Anzahl
USA	29
Süd Afrika	2
Deutschland	1
Kanada	1
Surinam	1

Diese Mehrdeutigkeit stellt ein Problem dar. Es kann durch eine Abfrage an ein Ortsverzeichnis keine eindeutige Entscheidung getroffen werden welche Georeferenz dem Toponym zugewiesen werden soll.

3 Verwandte Arbeiten

Durch das aufkommen der nutzergenerierte Die Geolokalisierung von Tweets ist ein Feld an dem nach wie vor aktiv geforscht wird.

An die Geolokalsierung werden dabei gewisse Mindestanforderungen bezüglich der Genauigkeit und der Trefferquote gestellt. Die entwickelten Verfahren unterscheiden sich sowohl durch die verwendete Methode als auch durch den Untersuchungsgegenstand.

In diesem Abschnitt sollen bestehende Ansätze zur Geolokalisierung 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.

3.1 Kategorisierung bestehender Ansätze

In früheren Arbeiten wurde bereits versucht, eine Einordnung der bestehenden Verfahren vorzunehmen. Es ist interessant die Kategorisierungsansätze 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 Erweiterter Kategorisierungsansatz

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)

3.2

Durch die Kategorisierung wird deutlich, dass auf dem Nutzer-Standort in Kombination mit der Zeitzone bisher keine NLP ausgeführt wurde. Es ist sinnvoll NLP auf den vorgenannten feldern auszuführen, da die Inhalte ähnlich dynamisch sein können wie im Text Feld. Der Nutzer-Standort bietet den Vorteil, dass die Intention ist, dass ein Toponym eingegeben.

Es wird selten eine geografische Hierarchie zum abbilden der Positionen verwendet.

4 Lösungsansatz

Es wird ein Verfahren zur Geolokalisierung von Tweets vorgestellt.

In Kapitel 3 wurden bestehende Arbeiten zur Geolokalisierung von Tweets untersucht. Die Arbeiten wurden anhand der verwendeten Verfahren zur Geolokalisierung und anhand der untersuchten geografischen Indikatoren aus dem Tweet klassifiziert.

Das vorgestellte Verfahren besteht aus zwei Abschnitten. Zunächst soll aus einer Menge von Tweets eine Datenbasis eingelernt werden. Die in dieser Datenbasis enthaltenen Informationen sollen dann zur Geolokalsierung von Tweets genutzt werden können.

Die Idee besteht darin durch Training ein probabilistisches Sprachmodell einzulernen.

Zum einlernen sind verfügbar: Nutzer-Standort, geografische Koordinaten.

Im ersten Kapitel wird ganz allgemein auf die Geolokalisierung eingegangen. Es werden die benötigten Komponenten für eine Geolokalisierung identifiziert und benannt. Hierzu wird betrachtet was eine Geolokalisierung leisten muss. Danach wird darauf eingegangen wie dies mit Hilfe einer geeigneten Datenbasis umgesetzt werden kann. Zum Schluss wird ein generelles Vorgehen zur Geolokalisierung, unter Zuhilfenahme einer Datenbasis, an einem Beispiel erläutert.

Für die Geolokalisierung eines Twitter-Nutzers werden die Datenwerte des Nutzer-Standortes und der Nutzer-Zeitzone verwendet. Diese werden eingehend untersucht. Auf Basis dieser Untersuchung wird ein Lernverfahren entwickelt um eine Datenbasis einzulernen.

Zum Schluss wird ein Verfahren zur Geolokalisierung vorgestellt welches die wahrscheinlichste Georeferenz zu einem gegebenen Twitter-Nutzer ermittelt. Dabei wird die eingelernte Datenbasis als Grundlage für die Bestimmung der Georeferenz verwendet.

4.1 Verwendete Datenbasis

4.2 Verfahren zum Einlernen der Georeferenz-Basis

BILD MIT Ablauf

Vorverarbeitung des Nutzer-Standortes

Quantitative Betrachtung der Referenzwerte bezüglich der geografischen Position

Vorverarbeitung der geografischen Koordinaten Als Referenzwerte sollen die Informationen aus dem Nutzer-Standort verwendet werden Bei der Verwendung des Nutzer-Stadort als Referenzwert können Probleme auftreten. Die Um die Informationen aus dem Nutzer-Standort als Referenzwert nutzen zu können

muss deshalb eine Vorverarbeitung stattfinden.

Der Nutzer-Standort soll zunächst eingehender betrachtet werden.

4.2.1 Untersuchung der Werte im Nutzer-Standort

4.2.2 Vorverarbeitungsschritte des Nutzer-Standortes

Bereinigung des Nutzer-Standortes

Zusammenfassen von Toponymen mit Hilfe von a priori Wissen aus einem Ortsverzeichnis

Angleichung durch alphanumerische Sortierung

Fragmentierung zur Extraktion potenziell geografischer Indikatoren

Umwandlung aller Buchstaben in Kleinbuchstaben

Eliminierung von Mehr- und Doppeldeutigkeiten

4.2.3 Quantitative Betrachtung der Referenzwerte bezüglich der geografischen Position

Zählen

Problem der Genauigkeit geografischen Koordinaten in Tweets

4.2.4 Vorverarbeitung der geografischen Koordinaten

Im nächsten Schritt sollen die Vorkommen der Referenzwerte quantitativ erfasst werden. Dazu werden die Vorkommen eines referenzwertes an einer geografischen Position gezählt.

4.3 Auflösen des Nutzer-Standortes mit Hilfe der Georeferenz-Basis

Grundsätzliche Idee des auflösens mit absoluten Häufigkeiten

Hier Idee skizzieren.

Schwellwert für den absoluten Häufigkeitswert zur Justierung der Konfidenzen.

4.3.1 Auflösen des Nutzer-Standortes basierend auf absoluten Häufigkeiten

Ablauf

Auswertung

4.3.2 Probleme bei der ausschließlichen Betrachtung der absoluten Häufigkeiten

Problem am Beispiel und mit Daten.

Einbeziehung der relativen Häufigkeiten. Aufgrund von Beispielbetrachtung The und La Plata

4.3.3 Auflösen des Nutzer-Standortes basierend auf absoluten und relativen Häufigkeiten

Idee skizzieren

Ablauf

Schwellwert für die absolute Häufigkeit und relativen Häufigkeiten

4.4 Ausnutzen der impliziten geografischen Hierarchie

Idee skizzieren

4.4.1 Hochziehen und summieren

4.4.2 Anpassung der Schwellwerte

4.5 Geolokalisierung

Enthält ein Datensatz Informationen zu einem geografischen Objekt, so können in diesem geografische Indikatoren enthalten sein. Durch diese geografischen Indikatoren kann dem Datensatz, und damit dem geografischen Objekt, eine Georeferenz zugewiesen werden.

Die Zuordnung einer Georeferenz mit Hilfe von geografischen Indikatoren soll Geolokalisierung genannt werden.

In Abbildung 4.1 wird dies dargestellt. Im Gegensatz zu Abbildung 2.6 ist kein direkter Verweis vom geografischen Objekt zu einer Georeferenz vorhanden. Stattdessen wird mit Hilfe der geografischen Indikatoren a und b durch die Geolokalisierung eine Georeferenz zugeordnet.

Um aus den geografischen Indikatoren eine Georeferenz ableiten zu können soll eine Datenbasis verwendet werden. Diese ordnet den geografischen Indikatoren eine Georeferenz zu. Die gespeicherte Georeferenz ist bekannt und kann genau bestimmt werden. Das bedeutet: Ein Hinweis auf eine geografische Position wird auf eine konkrete, bekannte geografische Position abgebildet. Eine solche Datenbasis soll Georeferenz-Basis genannt werden.

Im einfachsten Fall liegt ein geografischer Indikator vor, dem direkt eine Georeferenz zugewiesen werden kann. Dies kann Beispielsweise ein eindeutiges Toponym sein. Ein Ortsverzeichnis könnte hier als Georeferenz-Basis verwendet werden. Liegt der Datenwert “Karlsruhe“ vor so würde die Geolokalisierung durch eine Abfrage an die Georeferenz-Basis die Stadt Karlsruhe als Georeferenz zuweisen.

Wie in Kapitel 2.2.11 bereits erläutert sind Toponyme allerdings nicht immer eindeutig oder bekannt. Die Abfrage an ein Ortsverzeichnis liefert potenziell mehrere Ergebnisse was eine weitere Verarbeitung der Ergebnisse erfordert. Ist das Toponym nicht bekannt so kann kein Ergebnis geliefert.

Des weiteren sind die Datenwerte eines Datensatzes nicht immer direkt zu verwenden. Dies kommt ganz darauf an wie die Daten erhoben wurden. Nutzereingaben auf Webseiten können beispielsweise aus einer Liste gewählt, oder in ein Freitext-Feld eingegeben werden. Werden die Datenwerte durch eine Liste erhoben liegt eine klar definierte Menge an möglichen Datenwerten vor. Haben die Datenwerte in der Liste geografischen Bezug

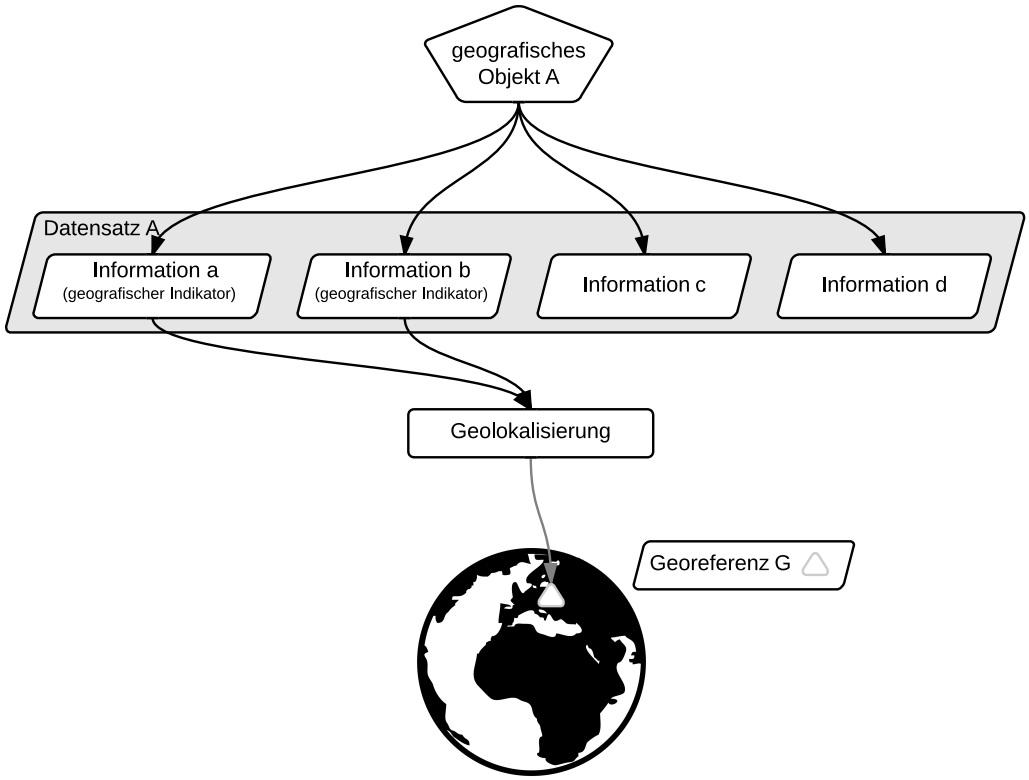

Abbildung 4.1: Geografische Hierarchieebenen

so können ihnen Georeferenzen zugeordnet werden. Die so entstandenen Paare aus Datenwert und Georeferenz können in der Georeferenz-Basis abgespeichert werden.

Soll ein Nutzer in ein Freitext-Feld seinen aktuellen Standort eingeben, und der Datenwert wird direkt übernommen muss dieser vorverarbeitet werden. Es ist zwar wahrscheinlich, dass der Nutzer ein Toponym angibt, aber es kann durch die direkte Übernahme der Eingabe zu Problemen kommen. Zunächst kann nicht einmal entschieden werden ob der Datenwert überhaupt einen geografischen Indikator darstellt. Zudem können in einem Freitext-Feld mehrere geografische Indikatoren auftauchen. Diese können widersprüchlich sein oder aber eine geografische Position genauer spezifizieren. Durch die direkte

Übernahme des Wertes können alle in Kapitel 2.2.11 aufgeführten Probleme auftreten. Dies macht die Zuordnung einer Georeferenz durch eine Ortsverzeichnis schwierig. Soll trotzdem mit Hilfe eines Ortsverzeichnisses eine Geolokalisierung durchgeführt werden ist zumindest eine umfangreiche Vor- und Nachverarbeitung nötig.

In Abbildung 4.2 ist ein Beispiel zur Verwendung einer Georeferenz-Basis dargestellt. Der Datenwert "Karlsruhe" wird dabei auf eine Georeferenz aufgelöst indem eine Abfrage an die Georeferenz-Basis durchgeführt wird. Die zurückgelieferte Georeferenz wird dem Datensatz zugeordnet.

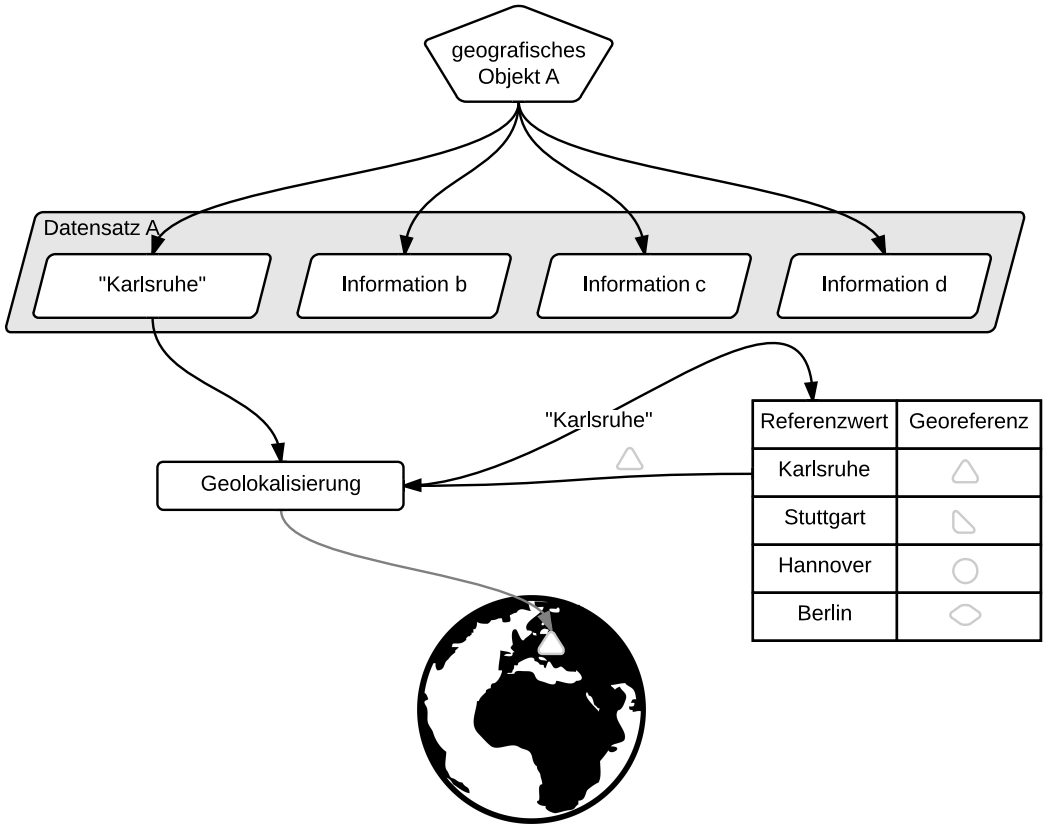

Abbildung 4.2: Geolokalisierung mit Referenz-Basis

Im folgenden Kapitel soll eine erste Struktur für eine Georeferenz-Basis vorgestellt wer-

den. Es sollen dabei die minimalen Anforderungen an eine solche Datenbasis erfüllt werden.

4.6 Minimale Struktur einer Georeferenz-Basis zur Geolokalisierung

Nach Abbildung 4.1 wird bei der Geolokalisierung einem oder mehreren geografischen Indikatoren eine Georeferenz zugewiesen. In der einfachsten Variante wird lediglich ein einziger geografischer Indikator an die Geolokalisierung übergeben, und genau eine Georeferenz zurückgegeben. Die Geolokalisierung muss also zu einem gegebenen geografischen Indikator eine Georeferenz bestimmen können. Dies führt zu einer ersten einfachen Struktur für die Georeferenz-Basis.

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 soll eine Georeferenz zugeordnet werden können. Die Georeferenz-Basis muss also eine Menge von Toponymen und zugehörigen Georeferenzen beinhalten 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 entsprechende Georeferenz nachgeschlagen werden. Die Referenzwerte stellen dabei mögliche Werte für den geografischen Indikator dar. In Abbildung 4.1 ist ein Beispiel für eine sehr simple Struktur dargestellt.

Tabelle 4.1: Beispiel für eine Georeferenz-Basis

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 Georeferenz-Basis mit den geografischen Indikatoren “Zoo-Karlsruhe“, “ZKM“ oder “Elbphilharmonie“ durchgeführt, kann nun eine Georeferenz

zurückgeliefert werden. Diese simple Struktur reicht grundsätzlich aus um eine Geolokalisierung 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.

Die meisten Ortsverzeichnisse sind nach dieser Struktur aufgebaut, wenngleich sie neben der Georeferenz noch andere Informationen zu einem Toponym liefern.

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.

Mögliche Angaben für die Georeferenz

- geografische Koordinaten
- vollständige Adressen
- Ländernamen
- Städtename
- Namen für Verwaltungseinheiten
- Zeitzonen
- Straßenname und Kilometerangabe
- ...

Grundsätzlich sind alle Formen, welche eine direkte oder indirekte Georeferenz darstellen, denkbar. Die Angabe muss lediglich die gegebenen Anforderungen an die Geolokalisierung erfüllen.

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

Wie in der Liste zu erkennen ist kann die Georeferenz auch als Toponym angegeben werden. Wenn nun ein Toponym abgefragt wird, wird als Georeferenz ein Toponym zurückgeliefert. Dies macht auf den ersten Blick wenig Sinn. Die geografischen Indikatoren sind jedoch vorerst nur Hinweise auf eine Georeferenz. Liefert eine Georeferenz-Basis ein Ergebnis zurück, ist der geografische Bezug bestätigt und dem Datensatz kann eine bekannte Georeferenz zugeordnet werden.

4.7 Der Nutzer-Standort und die Nutzer-Zeitzone in Twitter

In diesem Kapitel sollen die genutzten Informationen aus dem Twitter-Profil eingehender untersucht werden. Dabei werden zum Nutzer-Standort quantitative Daten erhoben um die Eignung des Nutzer-Standorts zur Geolokalisierung zu überprüfen. Zusätzlich wird untersucht welche Datenwerte im Nutzer-Standort vorkommen können.

Bei der Nutzer-Zeitzone wird geprüft ob diese einen geografischen Indikator darstellt.

4.7.1 Der Nutzer-Standort

Der Nutzer-Standort eines Twitter-Nutzers soll als geografischer Indikator verwendet werden. Bei der Eingabe des Nutzer-Standortes 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. Es ist naheliegend, dass der Nutzer seinen Standort mit Hilfe eines Toponyms angibt. Der Nutzer-Standort wird jedoch über ein Freitext-Feld abgefragt und direkt abgespeichert. Dieser muss also nicht zwangsweise Werte mit geografischem Bezug enthalten.

Sollte es sich bei dem eingegebenen Wert um Toponyme handeln, sind alle in Kapitel 2.2.11 erwähnten Probleme zu erwarten. Durch die unkontrollierte Eingabe sind tatsächlich alle möglichen Toponyme denkbar. Des weiteren können auch geografische Indikatoren mit mittelbarem geografischen Bezug auftauchen. Aber auch Werte die keinen geografischen Bezug aufweisen sind möglich.

Zuerst sollen einige allgemeine Kennzahlen zum Nutzer-Standort betrachtet werden. Hecht et al. haben in [HHSC11] den Nutzer-Standort von 10000 Twitter-Nutzern manuell untersucht. Um zu bestimmen ob ein Nutzer-Standort geografischen Bezug hat, wurden alle zur Verfügung stehenden Hilfsmittel verwendet. Dabei wurden allerdings nur Nutzer aus den USA betrachtet.

Im Zuge der vorliegenden Arbeit wurde eine eigene Untersuchung von 1000 Nutzer-Standorten verschiedener Nutzer vorgenommen. Dazu wurden 1000 Tweets untersucht. Die Tweets haben eine zusätzliche Georeferenz in Form von geografischen Koordinaten. Diese geben die Position an, von der eine Tweet versendet wurde. Für jeden Tweet wurde bestimmt ob der Nutzer-Standort einen geografischen Bezug hat, und darauf basierend eine Georeferenz zugewiesen. Als Hilfsmittel hierfür wurden die Ortsverzeichnisse von Google-Map und Geonames.org verwendet. Es wurde keinerlei Einschränkung bezüglich der Herkunft des Twitter-Nutzers oder der verwendeten Sprache und des verwendeten Alphabets gemacht.

Zum Schluss werden die Datenwerte im Nutzer-Standort anhand von Beispielen betrachtet.

Geografischer Bezug des Nutzer-Standorts

Hecht et al. konnten den Datenwerten in den Nutzer-Standorten in 80% der Fälle einen geografischen Bezug feststellen. In den restlichen 20% der Fälle konnte im Nutzer-Standort kein geografischer Bezug festgestellt werden.

In den eigenen Untersuchungen konnten 76% der Nutzer-Standorte ein geografischer Bezug nachgewiesen werden.

In den restlichen 24% der Fälle konnte kein geografischer Bezug mit Hilfe der Ortsverzeichnisse nachgewiesen werden. Dies bedeutet nicht, dass grundsätzlich kein geografischer Bezug vorhanden ist. Es konnte lediglich anhand der genutzten Quellen kein geografischer Bezug hergeleitet werden. Beispielsweise wurde “Swag City“ nicht zugewiesen, denn der Spitzname für die Stadt “Ann Arbor“ war in den Datenbanken nicht hinterlegt.

Genauigkeit der geografischen Angaben

Hecht et al. analysierten ihre Daten darauf wie genau die Nutzer ihren Standort angeben. Dabei ist zu beachten, dass die Daten aus den USA stammen und deshalb die Verwaltungseinheiten der USA zugrunde gelegt wurden.

Die Genauigkeiten der Standortangabe nach Hecht et al. sind in 4.2 angegeben.

Tabelle 4.2: Genauigkeit Standortangabe Hecht et al.

Anteil in % gerundet	geografische Hierarchieebene
64%	Stadt
20%	Staat
ca. 8%	Intrastate
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.

Bei den eigenen Untersuchungen wurden die geografischen Hierarchieebenen aus Unterkapitel 2.2.9 zugrunde gelegt. Die Genauigkeiten der Standortangabe aus den eigenen Untersuchungen sind in Tabelle 4.3 angegeben.

Tabelle 4.3: Genauigkeit Standortangabe eigene Untersuchungen

Anteil in % gerundet	geografische Hierarchieebene
77%	Stadt
8%	Verwaltungseinheit erster Ordnung
5%	Verwaltungseinheit zweiter Ordnung
10 %	Land

Es ist nicht sicher welche geografische Hierarchieebene im Nutzer-Standort angegeben wird.

Die unterschiedlichen Ergebnisse zwischen der eigenen Untersuchung und den Ergebnissen von Hecht et al. können dadurch Zustande kommen, dass in der vorliegenden Arbeit Twitter-Nutzer aus der ganzen Welt untersucht wurden. Aber auch der zeitliche Abstand zwischen den beiden Untersuchungen kann dafür verantwortlich sein.

Partieller geografischer Bezug des Datenwertes im Nutzer-Standort

In einigen Nutzer-Standorten konnte festgestellt werden, dass nur Teile des eingegebenen Datenwertes geografischen Bezug haben. Es werden oft weitere Informationen angegeben die keinen geografischen Bezug haben. Als Beispiel hierfür sollen “11th Dimension | California“ und “between here and there - Miami“ betrachtet werden. Im ersten Fall kann für “California“ ein unmittelbarer geografischer Bezug festgestellt werden. “11th Dimension“ hat hier offensichtlich keinen geografischen Bezug. Im zweiten Fall ist die Aussage “between here and there“ ohne geografischen Bezug. “Miami“ kann jedoch als Bezug zur Stadt Miami in Florida gebracht werden.

Es können also auch nur Teile des Nutzer-Standorts für eine Geolokalisierung von Nutzen sein.

Widersprüchliche Datenwerte im Nutzer-Standort

Es existieren auch Datenwerte in denen mehrere widersprüchliche Angaben gemacht werden. Dies bedeutet es werden zwei oder mehr Datenwerte mit geografischem Bezug angegeben, die auf unterschiedliche geografische Objekte verweisen.

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.

Geografische Hierarchien im Nutzer-Standort

Es ist auch möglich, dass im Nutzer-Standort teile einer geografischen Hierarchie angegeben sind. Beispielsweise die Angabe einer Stadt in Kombination mit einem Land.

In den USA wird beispielsweise oft die Stadt und der zugehörige Bundesstaat angegeben. In Brasilien hingegen wird oft ein Bundesstaat und das Land angegeben. Hier einige Beispiele:

- Los Angeles, California
- Mato Grosso, Brazil
- Bronx, New York

Im ersten Beispiel handelt es sich um die Stadt Los Angeles und den US-Bundesstaat Kaliforniern. Im zweiten Beispiel wird zuerst ein brasilianischer Bundesstaat angegeben und danach das Land. Im dritten Beispiel wird ein Stadtteil von New York angegeben und dann die Stadt oder der US-Bundesstaat. Da der US-Bundesstaat denselben Namen hat wie die Stadt, und beide in einer Hierarchiebeziehung mit Bronx stehen, kann hier nicht unterschieden werden was gemeint ist.

Domänenspezifische Toponyme im Nutzer-Standort

In sozialen Netzwerken können sich eigene Begriffe und Formulierungen etablieren. Diese sind im allgemeinen nicht bekannt.

Im Twitter-Umfeld haben sich in den letzten Jahren einige spezielle Begriffe und Formulierungen zur Verwendung in Tweet-Texten etabliert. Das im Twitter-Umfeld auch spezielle Toponyme verwendet werden, kann 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 seiner Fans auch in Twitter aktiv sind hat sich dieser Name etabliert. Unter diesem Gesichtspunkt hätte “Bieberville“ keinen geografischen Bezug. Sucht man allerdings im Internet 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-Standort der in einem Land kein Toponym darstellt, kann in einem anderen durchaus ein Toponym sein. Ist in einem Ortsverzeichnis beispielsweise “Bierberville“ als Bezeichnung für den Imbiss in Groß-Bieberau hinterlegt, würde dieser als Georeferenz zugeordnet werden, was im Umfeld von Twitter in einem Großteil der Fälle falsch wäre.

Solche Begriffe und Formulierungen können nur sehr schlecht mit einem Ortsverzeichnis zu Georeferenzen zugeordnet werden.

4.7.2 Die Nutzer-Zeitzone

In Kapitel 2.2.11 wurde auf 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 als geografischen Indikator kann dieses Problem auch auftauchen.

Um diesem Problem zu begegnen soll die Nutzer-Zeitzone als weiterer geografischer Indikator hinzugezogen werden. 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.

Eigenschaften der Nutzer-Zeitzone

Die Nutzer-Zeitzone stellt einen geografischen Indikator dar. Sie beschreibt eine eindeutige geografische Region auf dem Globus und somit hat direkten geografischen Bezug. Dabei entsprechen die Grenzen der Region nicht unbedingt den Landesgrenzen oder den Grenzen sonstiger Verwaltungseinheiten. In Abbildung 4.3 sind die Zeitzonen der Erde dargestellt.

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

In Abbildung 4.4 wurden Tweets anhand ihres Längen- und Breitengrades platziert. Jeder Punkt in der Abbildung entspricht einem Tweet. Es wurden nur Tweets aus den USA

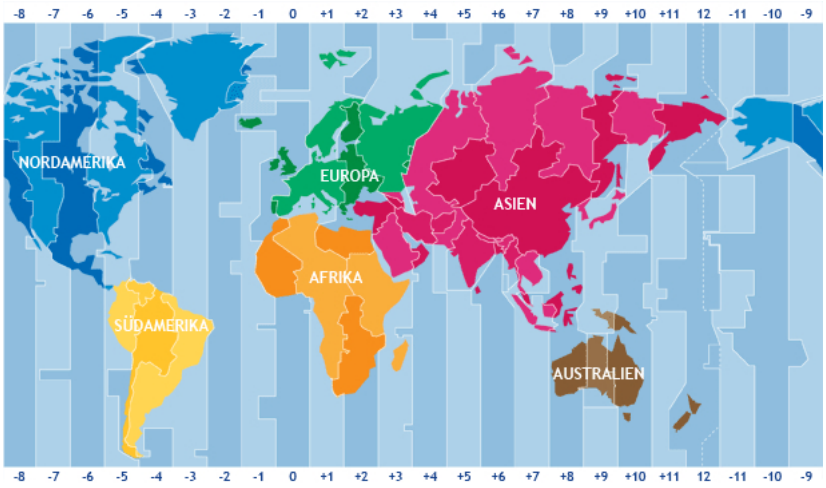

Abbildung 4.3: Zeitzonen der Erde.

ausgewählt. Anhand der Nutzer-Zeitzone wurde jedem Tweet eine Farbe zugeordnet. In Tabelle 4.4 sind die Farbzuordnungen aufgelistet.

Tabelle 4.4: In Abbildung 4.4 werden folgende Farben verwendet

Zeitzone	Farbe
Pacific Time	Rot
Eastern Time	Grün/Gelb
Central Time	Blau
Mountain Time	Pink

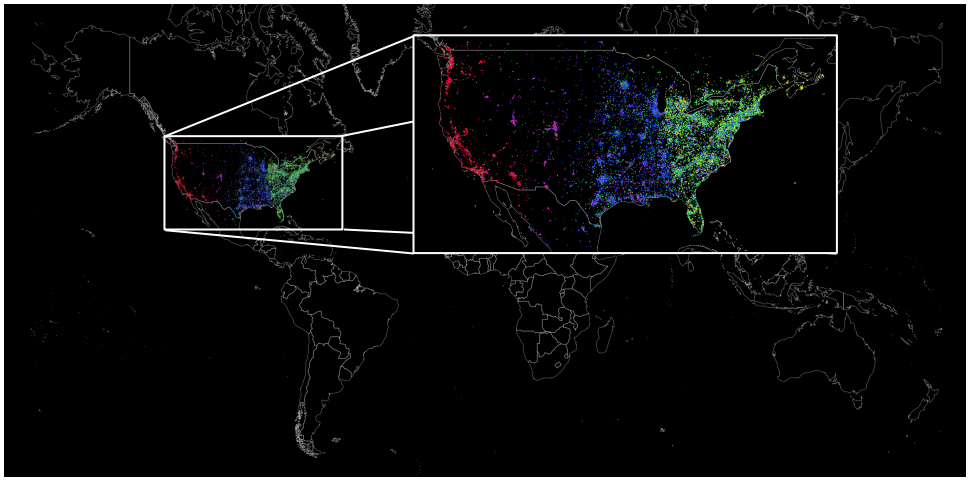

Abbildung 4.4: Tweets, abhängig der Zeitzone eingefärbt

Die Zeitzonen sind an den Farben gut zu erkennen. Lediglich die dünn besiedelte Region der Mountain Time kann nur an einigen Ballungszentren erkannt werden. Grundsätzlich scheint der Großteil der Angaben aber korrekt zu sein.

Auflösen von Doppeldeutigkeiten

Ist ein Toponym Doppel- oder Mehrdeutig kann nicht entschieden werden welches geografische Objekt zugeordnet werden soll. Liegen die beiden geografischen Objekte allerdings in zwei unterschiedlichen Zeitzonen und der Nutzer hat eine Nutzer-Zeitzone angegeben kann die Doppeldeutigkeit aufgelöst werden. Es muss lediglich die Nutzer-Zeitzone betrachtet werden. Somit kann dem Nutzer eine korrekte Georeferenz zugewiesen werden. Voraussetzung hierfür ist natürlich, dass die geografischen Objekte in zwei unterschiedlichen Zeitzonen liegen und die Nutzer-Zeitzone angegegeben ist.

4.7.3 Fazit

In ca. 80% der Fälle kann der Nutzer-Standort tatsächlich einen Hinweis auf eine Georeferenz liefern. Die Eingaben sind allerdings nicht standardisiert und können somit nicht ohne Vorverarbeitung verwendet werden. Wie oben dargestellt kann der Nutzer-Standort nur partiellen geografischen Bezug haben. Es können mehrere Informationen verschiedener geografischer Hierarchieebenen im Datenwert auftauchen. Und die Datenwerte im Nutzer-Standort können keinerlei geografischen Bezug aufweisen. Dies macht es grundsätzlich schwierig die Datenwerte im Nutzer-Standort direkt als geografischen Indikator zu verwenden.

Ist ein geografischer Bezug des Nutzer-Standortes nachzuweisen, handelt es sich bei dem Eintrag in den meisten Fällen um ein Toponym. Bei der durchgeführten Untersuchung wurden nur Ortsverzeichnisse zur Bestimmung eines geografischen Bezugs verwendet. In Ortsverzeichnissen sind nur Toponyme hinterlegt es kann deshalb davon ausgegangen werden, dass ca. 76% der Nutzer-Standorte auf Toponyme zurückzuführen sind.

Widersprüchliche Angaben, geografische Hierarchien und partieller geografischer Bezug können bei der Geolokalisierung zu Problemen führen. Diese entstehen meist dadurch, dass der Datenwert nicht nur eine konkrete, sondern mehrere Angaben beinhaltet.

Um den Problemen zu begegnen sollen die Datenwerte im Nutzer-Standort einer Vorverarbeitung unterzogen werden. Dadurch sollen möglichst viele einzelne Informationen aus dem Nutzer-Standort extrahiert werden.

Es können Beispielsweise Datenwerte mit geografischem Bezug von Datenwerten ohne geografischen Bezug getrennt untersucht werden.

Sollten Mehrdeutigkeiten im Nutzer-Standort auftauchen können diese teilweise durch die Hinzunahme der Nutzer-Zeitzone aufgelöst werden.

4.8 Verfahren zum einlernen geografischer Indikatoren am Beispiel von Twitter

Die Probleme bei der Nutzung von Ortsverzeichnissen sind meist auf deren Unvollständigkeit zurückzuführen. Um dies zu umgehen soll die Georeferenz-Basis automatisch eingelernt werden.

Der Vorteil besteht darin, dass eine domänenspezifische Georeferenz-Basis geschaffen wird. Diese kann potenziell mehr geografische Indikatoren zuordnen als ein normales Ortsverzeichnis. Denn es können dadurch domänenspezifische Eigenheiten bei der Verwendung von Toponymen berücksichtigt werden. Auch domäneninterne Begriffe sollen hierdurch gelernt werden können.

Im ersten Unterkapitel soll der Ablauf und die nötigen Daten zur Umsetzung eines solchen Lernverfahrens erläutert werden.

Danach wird auf die Vorverarbeitung des Nutzer-Standortes eingegangen. Es werden basierend auf den erzeugten Referenzwerten und den zugeordneten Georeferenzen absolute Häufigkeiten bestimmt. Des weiteren wird die in Abschnitt 4.6 vorgestellte Struktur der Georeferenz-Basis angepasst und erweitert. Dabei wird am grundsätzlichen Prinzip nichts geändert, es wird nach wie vor einem Referenzwert eine Georeferenz zugewiesen.

4.8.1 Allgemeiner Ablauf des Lernverfahrens

Die eingelernte Georeferenz-Basis soll es ermöglichen durch die Datenwerte im Nutzer-Standort und der Nutzer-Zeitzone eine Georeferenz zu bestimmen. Nach der Struktur aus Kapitel 4.6 besteht eine Georeferenz-Basis zumindest aus Referenzwerten und einer zugeordneten Georeferenz. Die Referenzwerte sollen aus dem Nutzer-Standort und der Nutzer-Zeitzone erzeugt werden. Um eine Georeferenz zuordnen zu können muss zu jedem Tupel aus Nutzer-Standort und Nutzer-Zeitzone eine Georeferenz vorliegen. Als Basis für das Lernverfahren werden die Tweet-Lerndaten genutzt. Jeder Tweet enthält dabei neben dem Nutzer-Standort und der Nutzer-Zeitzone eine Georeferenz in Form von Längen- und Breitengrad. Die Tweets sollen mit Hilfe des Längen- und Breitengrades der nächstgelegenen Stadt zugeordnet werden. Dadurch kann jedem Referenzwert in der Georeferenz-Basis später eine Stadt als Georeferenz zugeordnet werden. Es entstehen somit Tupel aus einem Referenzwert und einer zugehörigen Georeferenz. Diese sollen in der Georeferenz-Basis gespeichert werden.

Zusätzlich soll ein neues Feld in der Georeferenz-Basis eingeführt werden. Dieses soll angeben, wie oft die Kombination aus einem Referenzwert und einer Stadt vorgekommen ist. Das Feld gibt also die absolute Häufigkeit der Vorkommen an. Nachdem die Georeferenz-Basis eingelernt ist liegen somit für jeden Referenzwert eine Georeferenz und die absolute Häufigkeit vor. Basierend darauf können dann die relativen Häufigkeiten für jedes paar aus Referenzwert und Stadt berechnet werden. Bei der Geolokalisierung werden diese Werte genutzt um die wahrscheinlichste Georeferenz für einen gegebenen Nutzer-Standort und Nutzer-Zeitzone zu bestimmen.

In Abbildung 4.5 ist das Verfahren zum einlernen einer Georeferenz-Basis an einem Beispiel dargestellt.

In den nächsten Unterkapiteln sollen nun die einzelnen Schritte des Verfahrens genauer betrachtet werden. Zudem wird die Erweiterung der Datenbasis erläutert.

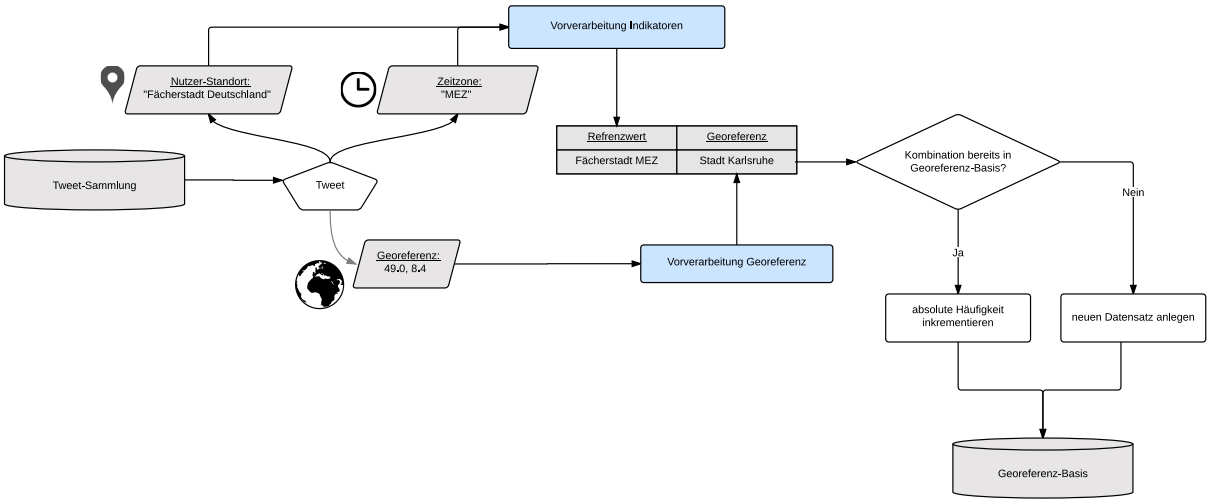

Abbildung 4.5: Verfahren zum einlernen einer Georeferenz-Basis

4.8.2 Vorverarbeitung des Nutzer-Standortes und der Nutzer-Zeitzone

Es ist zunächst nötig die genutzten geografischen Indikatoren einer Vorverarbeitung zu unterziehen. 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 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 werden oft Sonder- und Satzzeichen im Nutzer-Standort verwendet. Beispielsweise als Trenner zwischen Toponymen unterschiedlicher geografischer Hierarchieebenen. Beispiele hierfür sind “West Sussex, England“, “USA \/ Los Angeles“ oder “Bélem-PA“. Das Trennzeichen wird dabei nicht einheitlich verwendet. Es kann deshalb nicht entschieden werden ob ein Satzzeichen als Trenner zweier Hierarchieebenen genutzt wird oder nicht. Bei “USA \/ Los Angeles“ wird \/ als Trenner für Hierarchieebenen verwendet. Bei “Nottingham\/London“ hingegen werden zwei Städte angegeben. Es ist also insbesondere nicht klar welcher Zusammenhang zwischen den Datenwerten, die durch ein Sonder- oder Satzzeichen 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.

Grundsätzlich ist es aufgrund schwierig zu entscheiden ob Satz- und Sonderzeichen einen Mehrwert bieten um einen Nutzer-Stadort genauer zu bestimmen. Dies ist immer dann der Fall, wenn die Sonder- und Satzzeichen Bestandteil eines Toponyms sind. Es gibt tatsächlich Fälle in denen Sonder- oder Satzzeichen Bestandteil eines Toponyms sind. Ein Beispiel hierfür wäre “3. Arrondisement Paris“. Das weggelassen des Punktes hätte allerdings keinen Einfluss auf die gelieferte Information.

In den oben genannten Fällen bringen Sonder- und Satzzeichen keine zusätzlichen Informationen. 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.

Unnötige Sonder- und Satzzeichen wurden entfernt und die entstandenen Werte können nun einfacher weiterverarbeitet werden.

Zusammenfassen von Toponymen

Oft bestehen Toponyme aus zwei oder mehr Worten. Diese sollen mit Hilfe eines Ortsverzeichnisses zusammengefasst werden. Dies kann selbstverständlich nur für bekannte Toponyme durchgeführt werden.

“Los“ und “Angeles“ bilden gemeinsam “Los Angeles“ und sollen in der weiteren Verarbeitung gemeinsam betrachtet werden. Dies soll zunächst 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 Geolokalisierung vorgenommen. Er dient dazu möglichst früh vorhandenes Wissen über Toponyme einzubeziehen und somit eine unnötige Fragmentierung zu vermeiden. Im wesentlichen sollen die Werte hiermit für die nächsten Verarbeitungsschritte vorbereitet werden.

Alphanumerische Sortierung

In diesem Schritt sollen die Werte alphanumerisch sortiert werden.

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 die Sortierung werden Werte mit gleichem Inhalt aber unterschiedlicher Reihenfolge angeglichen. Durch das zusammenfassen der Toponyme im vorherigen Schritt werden bekannte Toponyme nicht getrennt.

Dieser Schritt stellt allerdings einen Kompromiss dar. Es werden zwar Werte mit gleichem Inhalt und unterschiedlicher Reihenfolge angeglichen. Aber es werden auch potenzielle Toponyme, die aus mehreren Teilen bestehen, auseinandergezogen. Aus “Motor City Michigan USA“ würde “City Michigan Motor USA“ entstehen. Der Zusammenhang zwischen Motor und City wäre nicht mehr vorhanden und könnte auch nicht wiedergewonnen werden.

Erzeugung von N-Grammen

Es sollen nun N-Gramme bis zum Grad 3 erzeugt werden. Dieses Vorgehen löst gleich mehrere Probleme.

Zum ersten können sowohl geografische Indikatoren als auch Werte die kein geografischen Indikator darstellen in den Nutzer-Standorten vorhanden sein. Durch die Erzeugung von N-Grammen können diese getrennt voneinander betrachtet werden. Es löst das Problem des partiellen geografischen Bezugs des Nutzer-Standortes aus 4.7.1 in der Hinsicht, dass die Werte für die weitere Verarbeitung getrennt betrachtet werden können.

Zum zweiten können in einem Nutzer-Standort mehrere geografische Indikatoren enthalten sein. Der Wert "Pitmedden Scotland UK" enthält mit "UK" das Land, mit "Scottland" den Landesteil und mit "Pitemedden" eine Stadt. Mit dem Längen- und Breitengrad aus dem zugehörigen Tweet kann nun allen drei Werten eine Georeferenz zugeordnet werden.

Aber auch zwei verschiedene Städte wie "Nottingham\/London" können in einem Wert vorkommen. Diese werden zu "Nottingham\/London", "Nottingham" und "London". Wird hier wiederum beiden Werten die geografische Koordinate zugeordnet können drei Fälle auftreten. Erstens: Der Nutzer war in keiner der beiden Städte als er den Tweet abgesetzt hat. Damit sind beide Zuordnungen unbrauchbar. Zweitens: Er war in einer der beiden Städte, dann ist zumindest einer der entstandenen Datensätze brauchbar. Dies löst das Problem aus 4.7.1 in der Hinsicht, dass die Werte für die weitere Verarbeitung getrennt betrachtet werden können.

Besteht allerdings eine Beziehung der Werte zueinander, kann diese 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 zur Verdeutlichung mit <>gekennzeichnet in späteren Beispielen wird dies weggelassen, da durch das zusammenfassen von Werten mit einem plus klar ist welches die Elemente des NGramms darstellen

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.

In diesem Schritt werden aus dem Nutzer-Standort mehrere potenzielle geografische Indikatoren erzeugt die als Referenzwerte genutzt werden können. Bevor diese als Referenzwerte genutzt werden soll aber noch die Nutzer-Zeitzone einbezogen werden.

Hinzufügen der Nutzer-Zeitzone

Die Zeitzone stellt eine begrenzte Anzahl an Werten dar. Diese werden vom Nutzer nicht frei eingegeben. Es wird hier hier deshalb keine weitere Vorverarbeitung vorgenommen.

An die Referenzwerte, die aus dem Nutzer-Standort erzeugt wurden soll nun die Zeitzone angehängt werden. Jeder Referenzwert soll einmal mit und einmal ohne Zeitzone existieren. Damit wird garantiert, dass eingelernte Referenzwerte die eine falsche Zeitzone aufweisen, trotzdem berücksichtigt werden können. Beispielsweise ist die Nutzer-Zeitzone “Pacific Time (US & Canada)“ für den Nutzer-Standort “Jakarta“ nicht korrekt. Aus dieser Kombination würden die Referenzwerte “Jakarta Pacific Time (US & Canada)“ und “Jakarta“ entstehen. Bei einer Abfrage von “Jakarta“ kann dann trotzdem der Referenzwert “Jakarta“ genutzt werden. Würde allerdings nur die Kombination “Jakarta Pacific Time (US & Canada)“ als Referenzwert existieren würde keine Georeferenz zurückgeliefert werden können.

Die Elemente der Zeitzonen werden wiederum mit einem Plus zusammengefasst um deutlich zu machen, dass es sich um ein Element eines NGramms handelt. Um die Nutzer-Zeitzone von den aus dem Nutzer-Standort generierten Elementen unterscheiden zu können, wird die Nutzer-Zeitzone kursiv geschrieben. Auch hier wird die Liste weiter eingeschränkt und es werden lediglich noch zwei Beispiele betrachtet.

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. Dies löst das Problem der Mehrdeutgkeiten von Toponymen. Mit Hilfe der zusätzlichen Zeitzone können geografische Objekte, zumindest wenn sie in zwei verschiedenen Zeitzonen liegen, unterschieden werden.

Fazit

Nach der Vorverarbeitung liegen eine Menge von potenziellen geografischen Indikatoren mit zugehörigen geografischen Koordinaten vor. Bei der Vorverarbeitung wurden einige Probleme des Nutzer-Standortes beseitigt. Insbesondere können Datenwerte separat voneinander betrachtet werden.

Die so erzeugten potenziellen geografischen Indikatoren können nun in der Georeferenz-Basis als Referenzwerte gespeichert werden. Mit der so erzeugten Georeferenz-Basis ist es grundsätzlich möglich eine Geolokalisierung durchzuführen. Allerdings wird jedem potenziellen geografischen Indikator eine Georeferenz zugeordnet. Das ist grundsätzlich

problematisch, da Referenzwerten die keinen geografischen Bezug aufweisen eine Georeferenz zugeordnet wird. Dies kann bei der Geolokalisierung zu Fehlern führen. Ein Referenzwert mit geografischem Bezug wird gleich behandelt wie ein Referenzwert ohne geografischen Bezug. Es sollte entschieden werden können ob der Referenzwert einen geografischen Bezug hat oder nicht. Bisher existiert noch kein Hinweis auf den geografischen Bezug. Es existiert lediglich eine Datenbasis die Referenzwerten eine Georeferenz zuweist.

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 alle auf dieselbe Georeferenz verweisen.

4.8.3 Absolute Häufigkeiten

Es sollen nun absolute Häufigkeiten in der Georeferenz-Basis eingeführt werden. Die absoluten Häufigkeiten sollen angeben wie oft ein Tupel aus Referenzwert und Georeferenz in der Georeferenz-Basis vorhanden ist. Dadurch werden Duplikate in der Georeferenz-Basis vermieden. Zusätzlich kann ermittelt werden ob ein Referenzwert an einer bestimmten Position gehäuft auftritt.

Vor dem abspeichern eines neuen Tupels soll nun zunächst geprüft werden ob dieses bereits in der Georeferenz-Basis vorhanden ist. Ist dies der Fall, so wird die absolute Häufigkeit des entsprechenden Eintrags um 1 erhöht. Ist das Tupel noch nicht gespeichert so wird ein neuer Datensatz angelegt und die absolute Häufigkeit mit 1 initialisiert.

Die absolute Häufigkeit misst damit wie oft ein Referenzwert an einer geografischen Position vorkommt. Es ist zu erwarten, dass ein Referenzwert mit geografischem Bezug gehäuft an einer bestimmten geografischen Position oder in einer Region auftritt. Während Referenzwerte die keinen geografischen Bezug haben sehr verteilt oder nur sehr selten auftreten.

Betrachtet man die Tweet-Lerndaten kann beim Vergleich der Georeferenzen ein Problem entstehen. Die Georeferenz aus den Tweet-Lerndaten besteht aus den geografischen Koordinaten des zugehörigen Tweets. Die geografische Position wird zumeist mit Hilfe von GPS-Modulen mobiler Endgeräte wie Smartphones bestimmt. Diese können eine

Position oft auf wenige Meter genau bestimmen. Das bedeutet, zwei Tweets die wenige Meter voneinander abgesetzt wurden haben unter Umständen 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 Koordinaten auf Städte abgebildet werden.

4.8.4 Vorverarbeitung der geografischen Koordinaten

Die geografischen Koordinaten sollen auf Städte aufgelöst werden. Dies entspricht der untersten Ebene der geografischen Hierarchie. Bei der Auflösung auf eine Stadt werden durch die geografischen Hierarchieebenen implizit auch die Verwaltungseinheiten und das Land bestimmt.

Jedem Tweet soll mit Hilfe von Voronoi-Regionen 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. Wird ein Tweet innerhalb einer Voronoi-Region abgesetzt wird er der enstprechenden Stadt zugeordnet.

Es bleibt zu definieren was unter einer Stadt zu verstehen ist. In dicht besiedelten Gebieten können viele kleine Städte vorhanden sein. Damit wäre die Positionsangabe wiederum zu genau. Deshalb soll die Definition einer Stadt hier über die Einwohnerzahl stattfinden. Es werden nur Städte verwendet deren Einwohnerzahl 15000 überschreitet. Der Tweet wird also der nächstgelegenen Stadt mit mehr als 15000 Einwohnern zugeordnet. In Ballungsräumen sind damit mehr Städte zu erwarten, womit die Genauigkeit zunimmt. Es ist allerdings auch zu erwarten das dort tendenziell mehr Tweets abgesetzt werden als in ländlichen Gebieten. Womit in jeder Voronoi-Region zu einer Stadt ausreichend Tweets vorhanden sind. Damit können potenziell mehr Tweets zu einer Stadt zugeordnet werden. Wohingegen in ländlichen Gebieten ein größeres Einzugsgebiet pro Stadt zustande kommt. Es sind allerdings auch weniger Tweets zu erwarten. Durch das größere Einzugsgebiet werden dennoch ausreichend viele Tweets auf eine Stadt zugeordnet.

Die Distanzen, welche zwischen der tatsächlichen Position des Tweets und der zugeordneten Stadt liegen wurden protokolliert. Dies spiegelt den Fehler der bei einer solchen Zuordnung entsteht wieder.

In Tabelle 4.5 sind die Ergebnisse dargestellt.

Tabelle 4.5: Fehlerdistanzen zwischen Tweet Ursprung und zugeordneter Stadt (in km)

Durchschnitt	7
Median	3.5
0.25 Quantil	1.7
0.75 Quantil	6.9
0.85 Quantil	10
0.95 Quantil	24.1
0.98 Quantil	44.2
Größte Distanz	3424.5
Kleinste Distanz	0

Im Median liegt die Fehlerdistanz zwischen der tatsächlichen Position und der zugeordneten Stadt bei 3,5 Kilometern. Über die Quantile können die Fehlerdistanzen noch genauer untersucht werden. Das 0.25 Quantil sagt aus, dass 25% aller Fehlerdistanzen unter 1,7 Kilometern liegen. 95% der Fehlerdistanzen liegen unter 24,1 Kilometer. Dies ist ausreichend genau.

4.8.5 Erweiterte Struktur der Georeferenz-Basis

Hier soll nun die erweiterte Struktur der Georeferenz-Basis vorgestellt werden. Mit der absoluten Häufigkeit ist ein neuer Wert pro Datensatz hinzugekommen. Des weiteren wurde die Georeferenz auf eine Stadt aufgelöst und damit implizit die Verwaltungseinheit erster Ordnung (Adm1), die Verwaltungseinheit zweiter Ordnung (Adm2) und das Land mitbestimmt. Diese sollen nun zusätzlich in der Struktur hinterlegt werden. Die daraus resultierende Struktur der Georeferenz-Basis ist in 4.6 inklusive einiger Beispieleinträge dargestellt.

Tabelle 4.6: Struktur der Georeferenz-Basis mit geografischer Hierarchie

Referenzwert	Stadt	Adm2	Adm1	Land	abs. Häufigkeit
Los+Angeles USA	Los Angeles	LA County	CA	USA	30
Los+Angeles USA	San Francisco	SF County	CA	USA	3
Los+Angeles	Los Angeles	LA County	CA	USA	70
USA	Los Angeles	LA County	CA	USA	80
Heilbronn	Heilbronn	Regierungsbezik Stuttgart	BaWü	BRD	90

4.8.6 Überblick

Hier soll nun ein Überblick über das gesamte Verfahren zum einlernen der Georeferenz-Basis gegeben werden.

In Abbildung 4.6 ist der Gesamtablauf des Einlernens dargestellt. Zunächst werden die einzelnen Vorverarbeitungsschritte für den Nutzer-Standort und die Nutzer-Zeitzone durchgeführt. Parallel kann der Längen- und Breitengrad auf eine Stadt aufgelöst werden. Die dadurch entstandenen potenziellen geografischen Indikatoren und die zugehörige Georeferenz wird nun in die Georeferenz-Basis gespeichert. Dabei wird überprüft ob dieser Datensatz bereits vorhanden ist. Ist dies der Fall, wird der Häufigkeitswert des entsprechenden Datensatzes inkrementiert. Ist der Datensatz noch nicht vorhanden wird er angelegt und die absolute Häufigkeiten mit 1 initialisiert.

Die Vorverarbeitung extrahiert dabei zusätzliche Informationen aus dem Nutzer-Standort. Durch die Bereiningung der Werte im Nutzer-Standort, die alphanumerische Sortierung, das identifzieren von Toponymen mit mehreren Worten und die darauffolgende Erzeugung von N-Grammen werden zusätzliche Informationen aus jedem Nutzer-Standort gewonnen. Die Nutzer-Zeitzone wird einbezogen um Doppeldeutigkeiten auflösen zu können.

Durch die neue Datenstruktur, mit den absoluten Häufigkeiten und der abgebildeten geografischen Hierarchie, lassen sich nun tiefergehende Analysen durchführen um eine robuste Geolokalisierung zu ermöglichen. Die absoluten Häufigkeiten geben dabei an wie oft ein Referenzwert in einer bestimmten Region vorkommt.

Durch dieses Verfahren lässt sich eine Datenbasis erzeugen die domänenspezifische Eigenheiten, in Bezug auf die Verwendung spezieller Begriffe oder Formulierungen, berücksichtigt. Des weiteren werden Toponyme, die in Ortsverzeichnissen unter Umständen nicht hinterlegt sind, berücksichtigt. Auch geografische Indikatoren mit mittelbarem geografischen Bezug, zum Beispiel die Verwendung spezieller Begriffe in einer geografischen

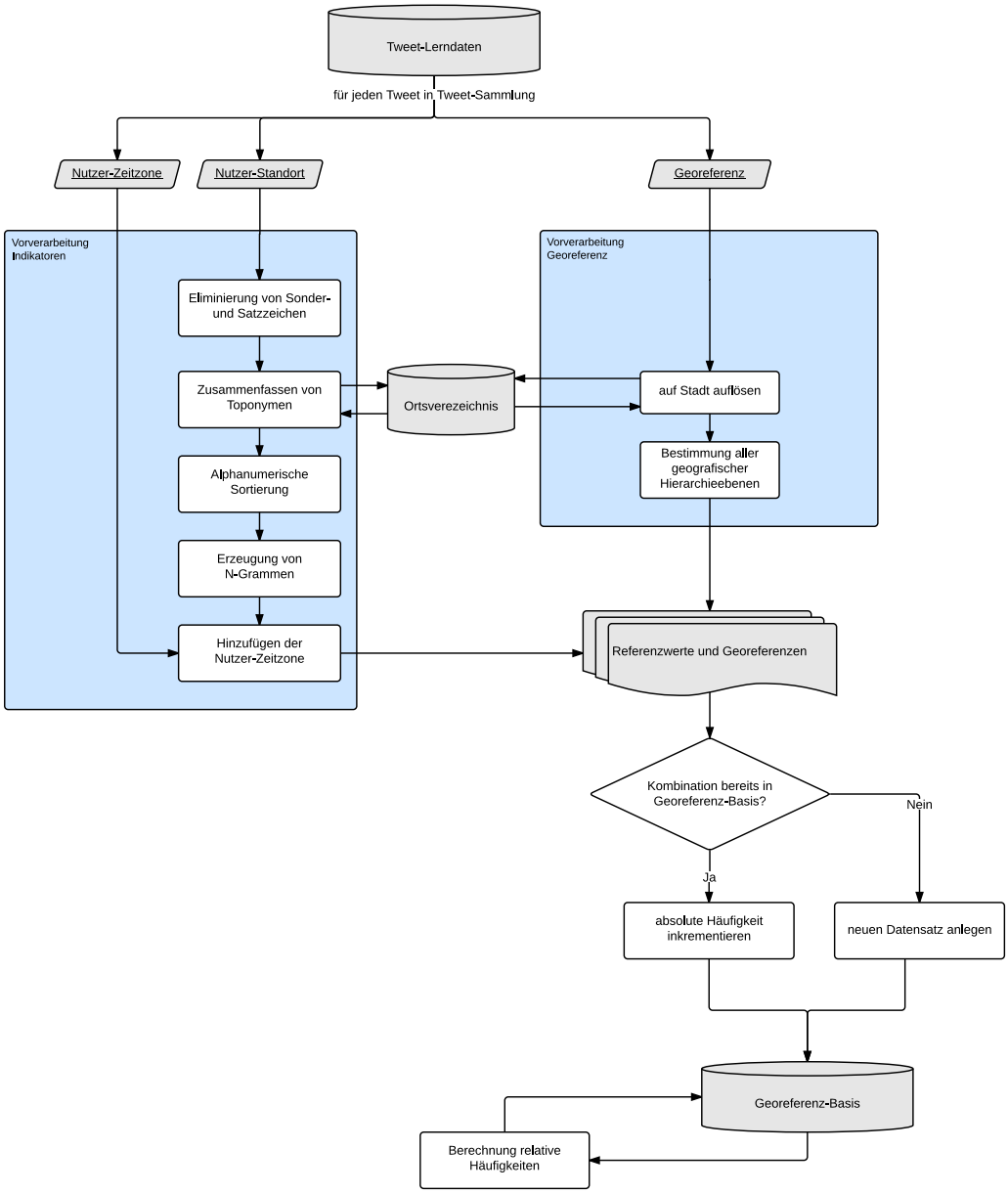

Abbildung 4.6: Ablaufplan einlernen

Region, können einbezogen werden.

4.9 Geografischer Bezug der eingelernten Referenzwerte

Die eingelernten Referenzwerte beinhalten alle Werte aus den Nutzer-Standorten der Tweet-Lerndaten. Es sind also auch Referenzwerte vorhanden die keinen geografischen Bezug haben. Es ist die Frage zu beantworten: Wie kann bestimmt werden ob ein Referenzwert geografischen Bezug hat oder nicht? Oder: Wie kann vermieden werden, dass ein Referenzwert, der keinen geografischen Bezug hat, zur Geolokalisierung genutzt wird?

Dies ist wichtig, denn durch die Referenzwerte wird in der eigentlichen Geolokalisierung einem geografischen Indikator eine Georeferenz zugewiesen. Wird einem geografischen Indikator durch einen Referenzwert ohne geografischen Bezug eine Georeferenz zugewiesen ist diese mit hoher Wahrscheinlichkeit fehlerhaft. Dies wiederum führt zu schlechten und unzuverlässigen Ergebnissen. Es muss also ein Verfahren gefunden werden um zu entscheiden ob die Referenzwerte einen geografischen Bezug haben oder nicht.

Es ist zu beachten, dass die Vorverarbeitung keine Aussage zum geografischen Bezug macht, sondern vielmehr die Referenzwerte aus den Nutzer-Standorten extrahiert. Dies soll sicherstellen, das möglichst viele Informationen aus den Nutzer-Standorten gezogen werden können und insbesondere keine Informationen verloren gehen.

Um zu entscheiden ob ein Referenzwert geografischen Bezug hat oder nicht wird die absolute Häufigkeit verwendet. Die absoluten Häufigkeiten geben an wie oft ein Referenzwert in einer bestimmten Region, zunächst in der Voronoi-Region der entsprechenden Stadt, vorkommt. Daraus kann nun ein geografischer Bezug abgeleitet werden.

4.9.1 Die absolute Häufigkeit als Hinweis auf geografischen Bezug zu Städten

Eine hohe absolute Häufigkeit kann ein Hinweis auf den geografischen Bezug eines Referenzwertes darstellen. Aufgrund der Eigenschaften des Nutzer-Standorts ist anzunehmen, dass in der Voronoi-Region einer Stadt der Name der zugehörigen Stadt häufig vorkommt. Dadurch kann eine Relevanz des Referenzwertes zu einer Stadt abgeleitet werden. Tritt der Referenzwert nicht häufig auf, so ist er von nur wenigen Nutzern als

Nutzer-Standort in einer Stadt angegeben worden und somit für die Stadt nicht relevant. In Abbildung 4.7 sind die Tweets in denen “Istanbul“ im Nutzer-Standort vorkommt aufgetragen. Es ist deutlich eine Häufung um die Stadt Istanbul zu erkennen. Werden die Tweets rund um Istanbul nun auf die Stadt Istanbul abgebildet, wird die Kombination aus dem Referenzwert “Istanbul“ und der Georeferenz Istanbul sehr häufig vorkommen. In den Nutzer-Standorten der Tweets rund um Istanbul taucht “Istanbul“ tatsächlich 972 mal auf. Damit kann eine Gewisse Relevanz für den Referenzwert “Istanbul“ zur Stadt Istanbul abgeleitet werden.

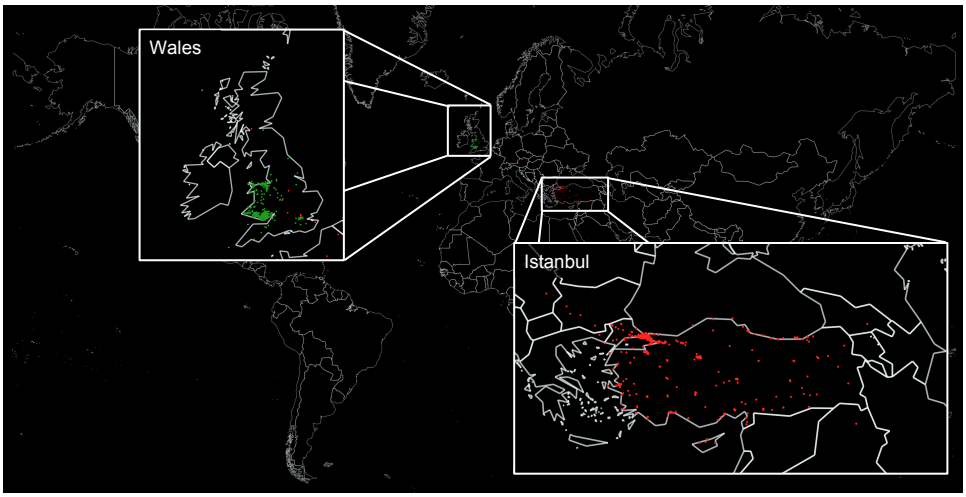

Abbildung 4.7:

Es kann also ein Schwellwert für die Häufigkeit eingeführt werden um Referenzwerte mit geografischem Bezug zu identifizieren. Allerdings garantiert die absolute Häufigkeit noch nicht, dass ein Referenzwert einen geografischen Bezug hat. Es können auch Werte an einem bestimmten Ort häufig vorkommen, die keinen geografischen Bezug haben. Um dies erkennen zu können muss ein weiterer Wert berechnet werden.

4.9.2 Relative Häufigkeiten als Hinweis auf geografischen Bezug zu Städten

Betrachtet man die absoluten Häufigkeiten isoliert voneinander wird die Verteilung des Referenzwertes außer acht gelassen. Ein Referenzwert kann eine gleichmäßige Verteilung über mehrere Städte aufweisen. Das bedeutet, der Referenzwert wird in vielen

unterschiedlichen Städten benutzt. Es können trotzdem Häufungen in Städten auftreten. Diese können sogar über einem gewählten Schwellwert für die absolute Häufigkeit liegen. Die Häufung kann jedoch relativ gesehen sehr gering sein. Dies ist ein Hinweis darauf das der Referenzwert keinen geografischen Bezug hat. Die relative Häufung sagt hier aus, das ein Referenzwert sehr verteilt auftritt. Es ist also wichtig nicht nur die absoluten Häufigkeiten, sondern auch die relativen Häufigkeiten der Referenzwerte zu berücksichtigen.

Berechnung der relativen Häufigkeiten

Um die relativen Häufigkeiten zu berechnen soll das Vorkommen eines Referenzwertes in einer Stadt, durch die Gesamtanzahl der Vorkommen des Referenzwertes geteilt werden. Damit erhält man den prozentualen Anteil der auf eine Stadt entfallenden Vorkommen eines Referenzwertes. Als Basis für diese Berechnung dienen die absoluten Häufigkeiten.

Sei (r_i, c_j) ein Datensatz der Georeferenz-Basis mit Referenzwert r_i und Georeferenz c_i. Des weiteren liefert $H(r_i, c_j)$ die absolute Häufigkeit zu einem Referenzwert r_i und einer Georefrenz c_i.

Damit kann die relative Häufigkeit $rel_{(r_i,c_j)}$ für jede Kombination (r_i, c_j) durch die folgende Formel berechnet werden. n_c ist dabei die Anzahl aller Georeferenzen.

$$d_{r_i,c_j} = \frac{H(r_i, c_j)}{\sum_{j=0}^{n_c} H(r_i, c_j)} \tag{4.1}$$

Beispiel

“La Plata“ tritt rund um die Stadt La Plata in Argentien 91 mal auf. Der Referenzwert “La Plata“ hat offensichtlich einen geografischen Bezug zu einer Stadt.

Obwohl “the“ keinen offensichtlichen geografischen Bezug hat ist die absolute Häufigkeit von 91 Vorkommen in Jakarta hoch. Der Schwellwert könnte nun aufgrund der Erfahrung mit dem Referenzwert “La Plata“ auf 90 angesetzt werden. Dann würde davon ausgegangen werden, dass der Referenzwert “the“ einen geografischen Bezug hat. Betrachtet

Abbildung 4.8: Tweets mit Nutzer-Standort “The“

man allerdings die Abbildung 4.8 fällt auf, dass Tweets mit dem Wert “the“ im Nutzer-Standort sehr verteilt auf dem Globus auftreten. Im Gegensatz dazu tritt “La Plata“ in den Nutzer-Standorten sehr konzentriert auf. In Abbildung 4.9 wird die Verteilung von Tweets deren Verfasser “La Plata“ im Nutzer-Standort enthalten dargestellt. Es ist deutlich eine Häufung um die Stadt La Plata in Argentien zu erkennen, weltweit tritt der Referenzwert aber sehr selten auf.

Berechnet man nun die relativen Häufigkeiten kann dies abgebildet werden. In Tabelle 4.7 sind die zugeordneten Städte, die absoluten Häufigkeiten und die berechneten relativen Häufigkeiten für die Vorkommen des Wortes “the“ aufgetragen. Die Einträge sind absteigend nach dem Wert der relativen Häufigkeit sortiert. In Tabelle 4.7 werden nur die vier ersten Einträge dargestellt. Insgesamt kam das Wort “the“ in 2824 verschiedenen Städten vor. Dabei wurde es insgesamt 5764 mal verwendet. Trotz der hohen absoluten Häufigkeit in Jakarta liegt die relative Häufigkeit bei lediglich 1,6%. Aus den relativen Häufigkeiten lässt sich nun die, in Abbildung 4.8 vermutete, globale Verteilung ablesen.

Die Verteilung erklärt sich dadurch, dass das Wort “the“ im englischen sehr häufig auftritt. Englisch ist die internationale Verkehrssprache und wird dementsprechend global und sehr häufig verwendet. Dies spiegelt sich in der Verteilung der vorkommen auf dem Globus wieder. Die Häufung um Jakarta kann teilweise damit erklärt werden, dass sehr

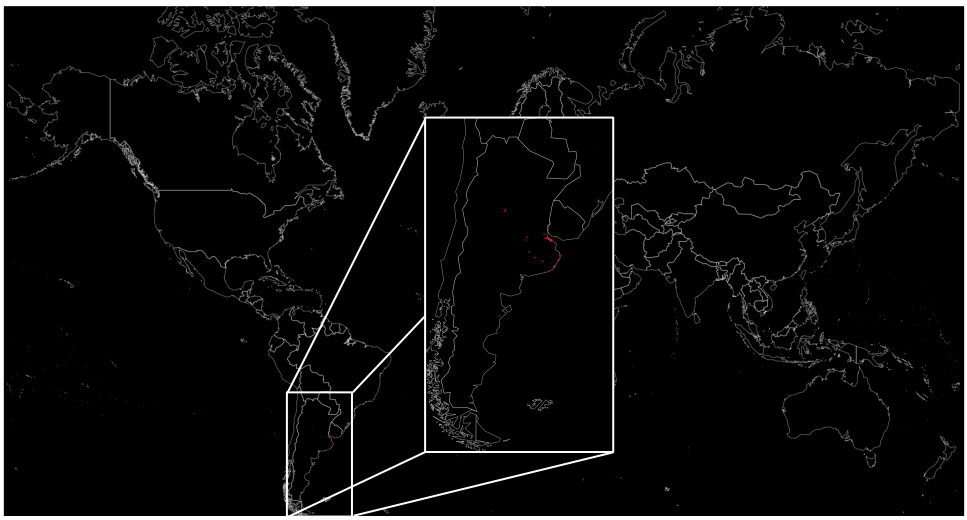

Abbildung 4.9: Tweets mit Nutzer-Standort “La Plata“

Tabelle 4.7: “the“

Stadt	abs. Häufigkeit	rel. Häufigkeit in %
Jakarta	91	1,6
Singapore	27	0,5
Bekasi	25	0,4
Philadelphia	23	0,4
...	...	...

viele Tweets die in und um Jakarta abgesetzt werden eine Angabe von Längen- und Breitengrad aufweisen.

In Tabelle 4.8 wird dieselbe Auswertung für “La Plata“ dargestellt. “La Plata“ ist insgesamt 129 mal in 23 verschiedenen Städten aufgetaucht. In der Stadt La Plata, in Argentien, kam es 91 mal vor. Dies entspricht einer relativen Häufigkeit von 70,5%.

“La Plata“ taucht erwartungsgemäß am häufigsten rund um die Stadt La Plata auf.

Es lässt sich also anhand der relativen Häufigkeiten ein geografischer Bezug des Referenzwertes nachweisen. Damit kann man zunächst bestimmen ob ein Referenzwert geografischen Bezug hat oder nicht. Die relativen Häufigkeiten können nach dem einlernen der Referenzwerte berechnet und in der Georeferenz-Basis hinterlegt werden. Mit Hilfe eines Schwellwertes für die relativen Häufigkeiten, kann nun bestimmt werden wann ein Referenzwert geografischen Bezug hat.

Tabelle 4.8: “La Plata“

Stadt	abs. Häufigkeit	rel. Häufigkeit in %
La Plata	91	70,5
Villa Gesell	9	7,0
Mar del Plata	5	3,9
Quilmes	3	2,3
...	...	...

Aber auch die ausschließliche Betrachtung der relativen Häufigkeiten reicht nicht aus um die geografische Relevanz nachzuweisen. Da die relativen Häufigkeiten basierend auf den Vorkommen des Referenzwertes berechnet werden sagen diese wiederum nichts über die absoluten Häufigkeit aus. Kommt ein Referenzwert zwei mal in zwei verschiedenen Städten vor, liegen die jeweiligen relativen Häufigkeiten bei 50%. Dies ist ein hoher Wert. Da der Referenzwert aber nur einmal vorkam, ist es sehr unwahrscheinlich das er einen geografischen Bezug hat.

4.9.3 Geografischer Bezug zu Verwaltungseinheiten und Ländern

Beim einlernen der Georeferenz-Basis werden durch die absoluten Häufigkeiten die Vorkommen pro Stadt gespeichert. Mit den daraus errechneten relativen Häufigkeiten kann nicht entschieden werden, ob für den Referenzwert eine globale Verteilung vorliegt oder ob der Referenzwert unter Umständen nur regional begrenzt, zum Beispiel in einem Land, auftritt. Für Referenzwerte die ein Land oder eine Verwaltungseinheit bezeichnen ist auf Städteebene eine geringe relative Häufigkeit zu erwarten. Diese Referenzwerte treten in einer größeren geografischen Region als der Voronoi-Region einer Stadt auf. Sie sind somit über mehrere Städte verteilt und werden auf Stadtebene einer geringe relative Häufigkeit aufweisen. Es ist beispielsweise zu erwarten das ein Ländername in den Nutzer-Standorten von Tweets aus dem gesamten Land auftritt. Durch die Unterteilung des Landes in Stadtgebiete wird der Wert sehr verteilt auf die Städte des Landes auftreten.

Soll nun statt einer Stadt eine Verwaltungseinheit oder das Land als Georeferenz bestimmt werden, kann mit diesen absoluten Häufigkeiten auf Stadtebene keine Aussage über den geografischen Bezug gemacht werden.

Am folgenden Beispiel soll dieser Sachverhalt erläutert werden.

In Abbildung 4.7 sind in grün Tweets dargestellt, welche im Nutzer-Standort “Wales“ enthalten. Wales entspricht einer Verwaltungseinheit erster Ordnung und gehört zu Großbritannien. Die Tweets sind in der gesamten geografischen Region, über die sich Wales erstreckt, verteilt. Außerhalb von Wales tritt “Wales“ im Nutzer-Standort sehr selten auf. Die relativen Häufigkeiten auf Stadtebene werden in Tabelle 4.9 dargestellt. Diese bestätigen eine Verteilung über eine größere geografische Region. Anhand der relativen Häufigkeiten kann aber nicht entschieden werden ob der Referenzwert global verteilt ist, oder in einer bestimmten geografischen Region, wie einem Land, auftritt. “Wales“ taucht in insgesamt 78 Städten 346 mal in Nutzer-Standorten auf.

Tabelle 4.9: “Wales“

Stadt	abs. Häufigkeit	rel. Häufigkeit in %
Cardiff	44	12,7
Newport	32	9,2
Carmarthen	24	6,9
Swansea	18	5,2
...	...	...

Die relative Häufigkeit von 12,7% deutete eher darauf hin, dass “Wales“ keinen geografischen Bezug aufweist. Auf Städteebene ist dies auch durchaus korrekt. Allerdings kann aus diesen Ergebnissen kein geografischer Bezug zu einer der anderen geografischen Hierarchieebenen abgeleitet werden. Das Problem ist, dass Wales keinen geografischen Bezug zu einer Stadt aufweist, wohl aber zu einer Verwaltungseinheit erster Ordnung und daher zu einer geografischen Region.

Um dieses Problem lösen zu können müssen zu einem Referenzwert die relativen Häufigkeiten für die anderen geografischen Hierarchieebenen berechnet werden. Damit kann dann die Verteilung der Referenzwerte auf diese Hierarchieebenen betrachtet werden.

Berechnung der relativen Häufigkeiten zu Verwaltungseinheiten und Ländern

Da zu jedem Referenzwert die zugehörigen Verwaltungseinheiten und Länder bekannt sind können die absoluten und relativen Häufigkeiten direkt aus der Georeferenz-Basis berechnet werden.

Es müssen lediglich die absoluten und relativen Häufigkeiten aufsummiert werden, bei denen der Wert der jeweiligen geografischen Hierarchieebenen übereinstimmen. Im Beispiel aus Tabelle 4.9 müssen alle absoluten und relativen Häufigkeiten derjenigen Städte aufsummiert werden, die in derselben Verwaltungseinheit erster Ordnung liegen. Betrachtet man die Verwaltungseinheiten zu allen Städten in denen "Wales" im Nutzer Standort vorkommt ergibt sich folgende Liste.

1. Wales 35
2. England 30
3. unterschiedliche Verwaltungseinheiten 13

Aus dieser Betrachtung alleine lässt sich noch nicht entscheiden ob der Referenzwert "Wales" einen geografischen Bezug zu einer Verwaltungseinheit hat. Denn der Referenzwert kam sowohl in 30 Städten in England als auch in 30 Städten in Wales vor, was keinen signifikanten Unterschied darstellt. Summiert man allerdings die Vorkommen und relativen Häufigkeiten pro Stadt auf ergibt sich daraus Tabelle 4.10.

Tabelle 4.10: "Wales"

Adm1	abs. Häufigkeit	rel. Häufigkeit in %
Wales	298	86,1
England	38	11,0
National Capital Region	1	0,3
Stockholm	1	0,3
...	...	...

Die relative Häufigkeit von 86,1% weißt nun deutlich darauf hin, dass der Referenzwert einen geografischen Bezug zu Wales hat. Mit diesem Vorgehen, kann der geografische Bezug eines Referenzwertes auf jeder der geografischen Hierarchieebenen untersucht werden.

Analog können die Werte für die Verwaltungseinheit zweiter Ordnung und dem Land berechnet werden. Dieses Vorgehen ermöglicht es einen geografischen Bezug eines Referenzwertes auf jeder der geografischen Hierarchieebenen zu prüfen.

Fazit

Mit der absoluten Häufigkeit besteht ein erster Hinweis darauf ob ein Referenzwert geografischen Bezug hat oder nicht. Die alleinige Betrachtung der absoluten Häufigkeit lässt aber die Verteilung der Werte auf den jeweiligen geografischen Hierarchieebenen außer betracht. Mit den berechneten relativen Häufigkeiten können die Referenzwerte zusätzlich auf ihre Verteilung untersucht werden. Die relativen Häufigkeiten können nach dem Einlernen berechnet und in der Georeferenz-Basis gespeichert werden. Während der Geolokalisierung kann dies absolute und relative Häufigkeit genutzt werden um die geografischen Indikatoren zu bestimmen.

4.10 Verfahren zur Geolokalisierung am Beispiel von Twitter

In Kapitel 4.8 wurde ein Verfahren zum einlernen der Georeferenz-Basis vorgestellt. Im darauffolgenden Kapitel 4.9 wurde eine Möglichkeit vorgestellt wie der geografische Bezug der Referenzwerte untersucht werden kann. Dies soll hier genutzt werden um eine Georeferenz zu bestimmen.

Es soll nun das Verfahren zur Geolokalisierung eines Twitter-Nutzers vorgestellt werden. Dabei dient die Georeferenz-Basis und die in ihr abgelegten Werte als Basis für die Zuweisung einer Georeferenz.

Zunächst werden aus dem Nutzer-Standort und der Nutzer-Zeitzone potenzielle geografische Indikatoren erzeugt. Dies geschieht analog zur Erzeugung von Referenzwerten beim einlernen der Georeferenz-Basis. Es werden also dieselben Vorverarbeitungsschritte für den Nutzer-Standort und die Nutzer-Zeitzone durchgeführt. Daraus resultiert eine Menge potenzieller geografischer Indikatoren. Die Werte der potenziellen geografischen Indikatoren werden nun in der Georeferenz-Basis nachgeschlagen. Dabei werden alle Datensätze deren Referenzwerte mit den potenziellen geografischen Indikatoren korrespondieren zurückgegeben. Auf diesen Datensätzen erfolgt die weitere Verarbeitung und Bestimmung der wahrscheinlichsten Georeferenz.

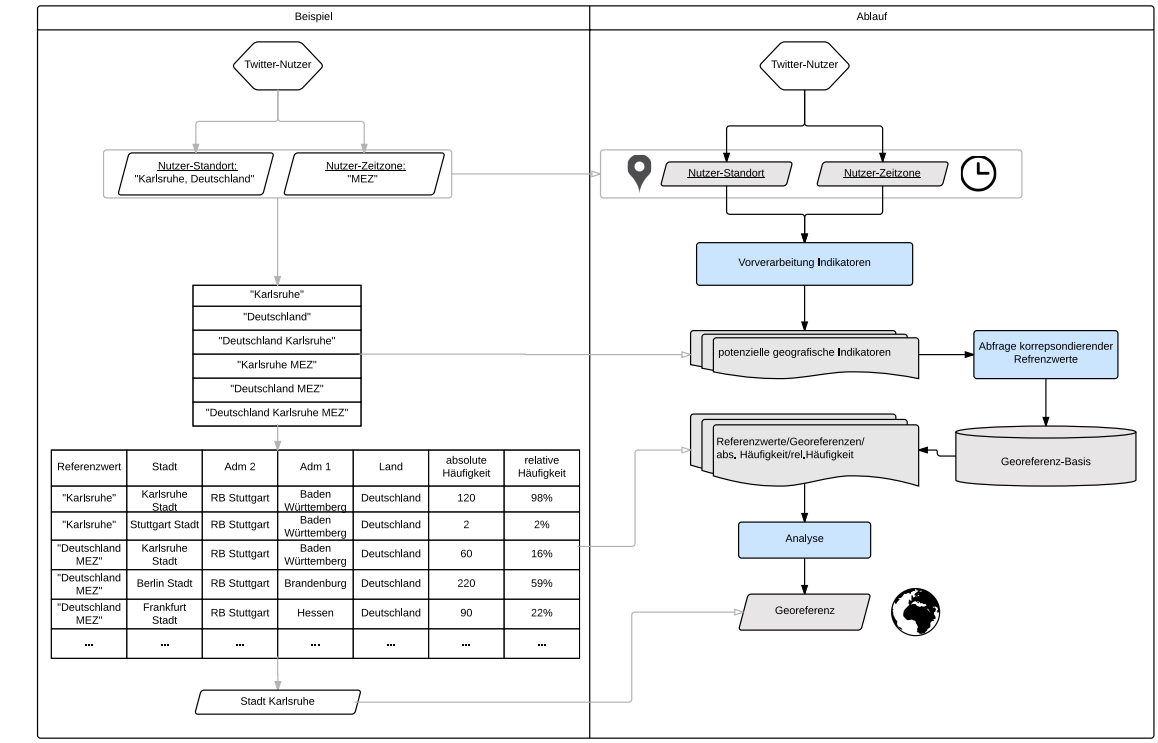

Abbildung 4.10: Ablauf der Geolokalisierung mit Beispiel

Es liegt nun eine Menge an Datensätzen aus der Georeferenz-Basis vor. Die Referenzwerte entsprechen dabei den potenziellen geografischen Indikatoren. Die absoluten und relativen Häufigkeiten werden für jeden Referenzwert separat analysiert. Das Ziel der Analyse ist es, die wahrscheinlichste Georeferenz zu ermitteln.

In Abbildung 4.10 ist der gesamte Ablauf an einem Beispiel dargestellt. In den folgenden Abschnitten soll nun die Analyse genauer betrachtet werden.

4.10.1 Analyse

Die Analyse beinhaltet zwei Schritte. Zuerst müssen diejenigen Referenzwerte gewählt werden, welche am wahrscheinlichsten einen geografischen Bezug haben. Dabei wird jeder

Referenzwert separat betrachtet. In einem nächsten Schritt wird derjenige Referenzwert gewählt, der unter den verbliebenen am wahrscheinlichsten die geografische Position des Nutzers beschreibt.

Auswahl der Referenzwerte mit geografischem Bezug

Es können pro Referenzwert zunächst mehrere Datensätze vorliegen. Aus diesen sollen diejenigen gewählt werden, welche am wahrscheinlichsten einen geografischen Bezug aufweisen. Dazu werden sowohl die absoluten als auch die relativen Häufigkeiten genutzt. Für jeden Referenzwert wird derjenige Datensatz gewählt, der die größte relative Häufigkeit h_{rel} über einem Schwellwert s_{rel} aufweist. Mit diesem Schwellwert lässt sich bestimmen wie verteilt der Referenzwert auftreten kann. Zusätzlich wird geprüft ob die absolute Häufigkeit ebenfalls über einem Schwellwert s_{abs} liegt. Mit diesem Schwellwert lässt sich bestimmen wie häufig der Referenzwert an einer geografischen Position oder in einer geografischen Region auftreten muss.

Bestimmung der wahrscheinlichsten Georeferenz

Nun liegt wiederum eine Menge an Datensätzen vor. Jeder Referenzwert, und damit auch jeder potenzielle geografische Indikator, taucht nur noch ein mal auf.

Aus den verbliebenen Datensätzen soll nun die Georeferenz gewählt werden. Dazu werden die relativen Häufigkeiten verglichen. Es wird der Datensatz mit der höchsten relativen Häufigkeit gewählt. Die Georeferenz dieses Datensatzes wird dann dem Twitter-Nutzer zugewiesen. Damit wird der Referenzwert ausgewählt der die größte relative Häufigkeit aller untersuchten Referenzwerte aufweist.

Die Referenzwerte stellen NGramme dar, wie in der Vorverarbeitung in Unterkapitel 4.8.2 erläutert wird. Es werden hier also insbesondere auch Uni-, Bi- und Trigramme miteinander verglichen. Darauf soll nun eingegangen werden.

Vergleich der relativen Häufigkeiten zu Uni- Bi- und Trigrammen Jedes Element eines Bi- oder Trigrammes kann potenziell einen geografischen Bezug haben. Umso mehr

Elemente ein NGramm beinhaltet umso spezieller kann die Beschreibung des geografischen Objekts sein. Deshalb können NGramme mit einem höheren Grad ein Objekt genauer beschreiben als NGramme mit einem niedrigeren Grad.

Allerdings können NGramme mit einem höheren Grad auch eine schlechtere Beschreibung darstellen. Beispielsweise wenn das zusätzliche Element keinen geografsichen Bezug hat.

Bei NGrammen mit einem Grad größer zwei können also zwei Fälle unterschieden werden.

1. Die Kombination aus den Elementen des NGrammes beschreibt einen Ort genauer
2. Die Kombination aus den Elementen des NGrammes beschreibt einen Ort nicht genauer

Fall1 Ein Beispiel für den ersten Fall ist der Nutzer-Standort "york" mit der Nutzer-Zeitzone "eastern+time+us+canada". Durch die Vorverarbeitung werden folgende potenzielle geografische Indikatoren erzeugt.

1. york
2. york *eastern+time+us+canada*

Eine Stadt Namens York existiert sowohl in Grossbritannien als auch in den USA. Fragt man nun die beiden Referenzwerte in der Georeferenz-Basis ab erhält man folgende Werte:

Tabelle 4.11: "york"

Referenzwert	Stadt	abs. Häufigkeit	rel. Häufigkeit in %
York	York (GB)	97	48,3
york eastern+time+us+canada	York (US)	12	63,2

Die realtive Häufigkeit für york in Kombination mit der Zeitzone ist höher. Die Zeitzone gibt zusätzliche Auskunft darüber welches York gemeint ist. Die Kombination ist spezieller, kommt deshalb seltener vor und potenziell eher dort wo sie zutrifft. In diesem Fall in York in den USA.

In den meisten Fällen beschreibt einer der beiden Indikatoren eine größere geografische Region wie beispielsweise einen Bundesstaat der USA. Wird ein weiterer Wert, beispielsweise ein Städtename hinzugenommen, wird die Angabe des Ortes genauer. Die Wahrscheinlichkeit, dass diese Kombination ausserhalb des Ortes auftritt wird geringer.

Fall 2 Hier können wiederum 2 Fälle unterschieden werden.

1. Beide Elemente beziehen sich auf unterschiedliche geografische Objekte
2. Nur ein Element hat geografischen Bezug das andere nicht

Wenn zu einem Referenzwert mit geografischem Bezug ein Element hinzugefügt wird, welches keinen geografischen Bezug hat, beschreibt dies den Ort nicht genauer. Es ist zu erwarten, dass die Kombination der Elemente sehr selten vorkommt oder sehr verteilt ist. Ist die Kombination verteilter, so ist der relative Wert geringer als der des einzelnen Referenzwertes mit geografischem Bezug. Ist die Kombination seltener kann der Referenzwert bereits durch den Schwellwert s_{abs} aussortiert werden.

Wenn zu einem Referenzwert mit geografischem Bezug ein Element hinzugefügt wird, welches zwar geografischen Bezug hat, aber dieses sich auf ein anderes geografisches Objekt bezieht ist dasselbe Verhalten zu erwarten.

Wahl der Schwellwerte

Die Wahl der beiden Schwellwerte ist abhängig von den Anforderungen. Dabei ist die gewünschte geografische Hierarchie der zurückgegeben Georeferenz ein Faktor. Und die gewünschte Genauigkeit und Trefferquote.

Gewünschte Hierarchieebene der Georeferenz Umso größer die betrachtete geografische Region ist, umso weniger Möglichkeiten zur Einteilung gibt es. Auf Städteebene gibt es 23322 verschiedene geografische Regionen. Für jede Stadt mit mehr als 15000 Einwohnern existiert dabei eine Region. Diese Städte verteilen sich auf 234 verschiedene Länder. Dieselbe Menge an Referenzwerten, verteilt sich auf Länderebene also auf weniger geografische Regionen. Dadurch werden die Werte der relativen Häufigkeit und der

absoluten Häufigkeit insgesamt größer. Relativ zueinander werden allerdings nach wie vor Referenzwerte mit geografischem Bezug größer sein als Referenzwerte ohne geografischen Bezug.

Es müssen also für jeder geografsiche Hierarchieebene geeignete Schwellwerte s_{rel} und s_{abs} gefunden werden.

Genauigkeit und Trefferquote Der zweite Faktor ist die gewünschte Trefferquote und die Genauigkeit.

Umso niedriger der Schwellwert s_{rel} ist, umso größer wird die Wahrscheinlichkeit Referenzwerte zu wählen die keinen geografischen Bezug haben. Daraus resultieren mehr fehlerhafte Zuordnungen einer Georeferenz. Wodurch die Genauigkeit schlechter wird. Dadurch können allerdings mehr Georeferenzen zugeordnet werden, wordurch die Trefferquote verbessert wird. Umso höher der Schwellwert s_{rel} gewählt wird umso mehr Referenzwerte mit geografischem Bezug werden verworfen. Die Wahrscheinlichkeit, dass die gewählten Referenzwerte tatsächlich geografischen Bezug haben ist allerdings höher. Dadurch können weniger Georeferenzen zugeordnet werden. Somit sinkt die Trefferquote. Allerdings sind die zugewiesenen Georeferenzen sicherer womit die Genauigkeit steigt.

Der Schwellwert s_{abs} vermeidet, dass Refrenzwerte gewählt werden die eine hohe realtive Häufigkeit aufweisen aber aufgrund ihrer geringen Vorkommen nicht relevant sind. Die Auswirkungen der Wahl des Schwellwertes s_{abs} verhalten sich Analog zum Schwellwert s_{rel}.

Fazit Die Wahl der Schwellwerte hängt zum einen von der Hierarchieebene und zum anderen von den Anforderungen an die Genauigkeit und die Trefferquote ab. In Bezug auf die geografischen Hierarchieebenen sind lediglich separate Schwellwerte für jede geografischen Hierarchieebenen zu bestimmen, da die relativen und absoluten Häufigkeiten sich insgesamt verändern. Bezüglich der Genauigkeit und der Trefferquote ist ein Kompromiss zwischen den beiden Werten einzugehen. Die Verbesserung der Trefferquote geht mit einer Verschlechterung der Genauigkeit einher und umgekehrt. Es kann also entweder ein Kompromiss gefunden werden der ein Optimum für beide Werte darstellt. Oder einer der Werte wird optimiert.

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 Möglichkeit gegeben werden, in eigenen Implementierungen die Probleme frühzeitig zu erkennen und zu vermeiden.

chap:Implementier

5.1 Verwendete System

CSharp, SQL Server, IIS Server

chap:Implementier sec: Verwendete Systeme

5.2 Architektur

Allgemeine Architektur der Referenzimplementierung. Klassnediagramm Sequenzdiagramm

chap:Implementier sec: Architektur

5.2.1 Präprozessorverarbeitung

Warum Präprozessoren -> schnelleres ändern der Vorverarbeitung. Durchreichen des Tweets und Verarbeitung durch Präprozessoren.

chap:Implementier sec: Architektur subsec: Präprozessorverarbeitung

5.3 Datenbank

Datenbankschema. Geography Datatypes and Methods. Nearest Neighbour search SQL Server

chap:Implementier
sec: Daten-
bank

5.4 Oberfläche zur manuellen Zuordnung von Georeferenzen

Kurz Oberfläche vorstellen plus verwendete Technologien.

chap:Implementier
sec: Ober-
fläche zur
manuellen
Zuordnung
von
Georeferenzen

5.5 Probleme und Fallstricke

chap:Implementier
sec: Probleme
und Fallstricke

6 Leistungsbewertung

chap:Grundlagen
sec:Precision
Recall

chap:Grundlagen
sec:Konfidenzen

6.1 Bestimmung Schwellwerte

Ergebnisse Kontrolldaten. Angabe Schwellwerte über Random 2D Array. Darstellung in Matlab. Maximum finden.

chap:Lesitungsbew
sec:Bestimmung
Schwellwerte

6.2 Naiver Ansatz

Im Vergleich zu einem naiven Ansatz. Google Maps API V3 ohne Vorverarebitung Precision/Recall.

chap:Lesitungsbew
sec:Naiver
Ansatz

6.3 Vergleich zu früheren Ansätzen auf Nutzer-Standort

chap:Lesitungsbew
sec:Vergleich
zu früheren
Ansätzen
auf Nutzer-
Standort

6.4 Vergleich zu früheren Ansätzen NLP

chap:Lesitungsbew
sec:Vergleich
zu früheren
Ansätzen NLP

6.5 Fazit

chap:Lesitungsbew
sec:Fazit

7 Zusammenfassung

chap: Zusammenfassung

8 Ideen und Notizen

8.1 Ideen

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.

[ISO] ISO19110:2005. Geographic information methodology for feature cataloguing (iso 19110:2005).

[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.

[MPLC13a] 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.

[MPLC13b] 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.

Abbildungsverzeichnis