Karlsruhe Institute of Technology

Dezentrale Systeme und Netzdienste
Institut für Telematik

Lehrstuhl
Prof. Dr. Hannes Hartenstein

Fakultät für Informatik

Diplomarbeit
2014

Analyse internationaler Nachrichtenflüsse im Twitter-Netzwerk

Peter Michael Bolch

Mat.Nr.: 1345211

Referent:
Betreuer: Matthias Keller

Ich erkläre hiermit, dass ich die vorliegende Diplomarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Karlsruhe, 2014 Peter Michael Bolch

Inhaltsverzeichnis

Todo list

1 Einleitung

1.1 Motivation und Hintergründe

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

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

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

[2] API: Application Programming Interface oder auch Programmierschnittstelle

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

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

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

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

Bild Twitter-Nutzer als sensor

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

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

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

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

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

1.2 Problembeschreibung

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

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

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

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

1.3 Fragestellungen und Anforderungen

Die folgenden Fragestellungen sollen beantwortet:

Q1 Wie kann Twitter-Nutzern eine Georeferenz zugeordnet werden?

1.3.1 Anforderungen

Das erarbeitete verfahren soll folgende Anforderungen erfüllen.

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

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

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

a) Land oder Staat

b) Verwaltungsebene erster Ordnung [5]

c) Verwaltungsebene zweiter Ordnung [6]

d) Stadt

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

R5 Verfahren unabhängig von Sprache und Schriftzeichen weltweit einsetzbar.

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

1.4 Gliederung der Arbeit

Abschnitt 2: Grundlagen

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

Abschnitt 3: Stand der Technik

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

Abschnitt 4: Lösungsansatz

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

Bottom-Up:

1. NGramme aus Indikatoren erzeugen
2. Geomapping
3. Datenstruktur
4. Treffer zählen (NGramm + Geoid gleich usw.)

5. Geografische Hierarchiebene
6. Unsicherheit bei Lokalisierung messen (neuer Daten)
7. Justierung der Lokalisierungsunsicherheit auf geografischen Hierarchiebenen

Abschnitt 5: Referenzimplementierung der entwickelten Methode

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

Abschnitt 6: Leistungsbewertung der entwickelten Methode

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

Abschnitt 7: Schlussfolgerungen

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

Abschnitt 8: Zusammenfassung und Ausblick

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

2 Grundlagen

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

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

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

2.1 Geografische Grundlagen und Begriffe

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

Geodätisches Referenzsystem Ein geodätisches Referenzsystem dient als einheitliche Grundlage zur Angabe einer Position auf der Erde. In diesem Referenzsystem werden unter anderem Referenzpunkte und ein geeignetes Koordinatensystem festgelegt. [1]

Georeferenz Eine Georeferenz (engl. Spatial Reference) wird auch als Raumbezug bezeichnet. Unter dem Begriff der Georeferenz versteht man die zugeordnete der Lage beziehungsweise Position zu einem Datensatz. Die konkreten Angaben zum Raumbezug und deren Genauigkeit hängen von den Anforderungen ab, die an die Georeferenz gestellt wird. Auch die in Kapitel 1 erwähnten Anwendungen stellen unterschiedliche Anforderungen an die Genauigkeit der Georeferenz. Die Georeferenz lässt sich weiter unterteilen in: [1]

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

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

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

auch Telefonvorwahlen. Alle diese Angaben, mit Ausnahme der Adresse, definieren eine geografische Fläche. Diese Fläche ist nicht zwingenderweise klar abzugrenzen.

geografische Objekt Ein geografisches Objekt, ist ein Objekt der Realwelt, dessen Position durch eine Georeferenz bestimmt werden kann. Die EN ISO 19110 Norm beschreibt ein geografisches Objekt folgendermaßen: "Geographische Objekte sind Erscheinungen der realen Welt, die einen Bezug zur Erde (Raumbezug) haben..."

Toponyme Ein Toponym ist ein konkreter Name für ein geografisches Objekts. Beispiele hierfür sind: Städtenamen, Ländernamen oder Landschaftsnamen.

Georeferenzierung Unter Georeferenzierung versteht man die Zuordnung einer Georeferenz zu einem Datensatz. Also den Vorgang einem Datensatz, zum Beispiel einem Twitter-Nutzer eine Georeferenz zuzuordnen. [1]

Aufrgund der einfacheren Verständlichkeit werden die folgenden zwei Begriffe definiert, welche in der restlichen Arbeit verwendet werden.

Geografische Position Unter geografischer Position wird hier ein konkreter Ort, unter Angabe geografischer Koordinaten verstanden. Eine geografische Position entspricht somit einer direkten Georeferenz (direktem Raumbezug).

Geografische Region Unter einer geografischen Region werden Flächen verstanden welche nicht mit einem Punkt, in Form von Längen- und Breitengrad, beschrieben werden können. Hierbei kann es sich beispielsweise um Bundesländer oder Länder. Somit entspricht der Begriff geografische Region einer indirekten Georeferenz (indirektem Raumbezug).

Geografische Hierarchie In der vorliegenden Arbeit wird eine geografische Hierarchie verwendet um eine Einteilung der Erde in geografische Regionen umzusetzen. Dabei können geografische Regionen wiederum geografische Regionen oder geografische Positionen enthalten, wodurch sich eine hierarchische Gliederung ergibt. Diese Einteilung spiegelt

Bild mit Hierarchie von Deutschland

im wesentlichen die Einteilung der Erde in Staaten und deren individuellen Verwaltungseinheiten wieder. Im vorliegenden Fall ist das Staatsgebiet, also die Fläche über die sich der Staat erstreckt, von Interesse. [2]

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

Eine Aufteilung der Erde in geografische Regionen lässt sich auf oberster Ebene mit Hilfe von Ländern und deren Grenzen umsetzen. Daraus resultiert eine Einteilung, welche direkt intuitiv verständlich ist und vielen Anforderungen an geografische Informationen gerecht wird. Die meisten Länder sind in weitere administrative Einheiten aufgeteilt. Diese geografischen Regionen werden hier als Administartionsebenen bezeichnet. Es wird zwischen Administrationsebenen erster und zweiter Ordnung unterschieden. Des weiteren werden Städte in der untersten Ebene der Hierarchie dargestellt. Ausnahmen sind besipielsweise Stadtstaaten wie der Vatikan-Staat oder das Fürstentum Monaco, welche aufgrund ihrer Größe nicht in Verwaltungsbezirke unterteilt werden und keine Städte. Wenn man als Beispiel Deutschland heranzieht, ergibt sich eine Einteilung wie in Bild 2.1 dargestellt wird. Die oberste Ebene beschreibt das Land worauf die zweite Ebene die Bundesländer darstellt. Auf der dritten Ebene werden die Regierungsbezirke abgebildet, worauf die Städte in der letzten Ebene folgen. Analog kann die Einteilung für die USA vorgenommen werden woraus sich die Hierarchie Country->State->County->City ergibt.

Bis auf die letzte Ebene wird den Objekten in der Hierarchie eine geografische Region zugeordnet. Lediglich die unterste Ebene, die der Städte, wird durch eine geografische Position genau beschrieben. Die Ausdehnung einer Stadt wird in der gegebenen Hierarchie also nicht berücksichtigt. Jeder Stadt werden als konkrete geografische Koordinate in Form von Längen- und Breitengrad repräsentiert.

[2] Die genaue Definition eines Staatsgebietes kann in [JJ21] nachgelesen werden.

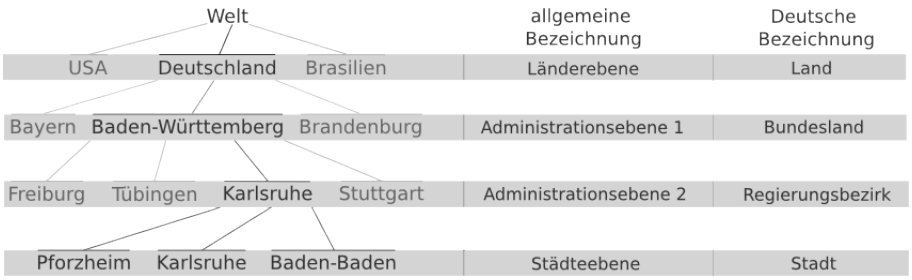

Abbildung 2.1: Die Hierarchiebenen exemplarisch. Bis in die Städtebene wird nur der Pfad Welt -> Deutschland -> Baden-Württemberg -> Karlsruhe ->K arlsruhe,Pforzheim,Baden-Baden dargestellt.

2.2 Twitter

In diesem Kapitel werden grundlegende Begriffe rund um das Twitter-Netzwerk erläutert. Weiter werden die Mechanismen in Twitter erläutert und an praktischen Beispielen erklärt. Zum Schluss wird aufgezeigt welche Informationen pro Tweet übermittelt werden und welche Daten zur Lokalisierung verwendet werden können.

2.2.1 Geschichtliches

Twitter wurde 2006 von Jack Dorsey, Biz Stone, Noah Glass und Evan Williams gegründet. Ursprünglich war Twitter zur internen Kommunikation innerhalb der Firma Odeo geplant. Schnell wurde allerdings klar, dass in dem Dienst mehr potenzial steckt und so wurde Twitter öffentlich gemacht. Seitdem erfreut sich der Dienst einer wachsenden Nutzer-Gemeinde. Die Twitter-Gründer haben von Anfang an keine exakten Nutzer-Zahlen oder die Anzahl der versendeten Twitter-Kurznachrichten bekanntgegeben. Dies geschah einerseits, weil die Gründer davon überzeugt sind, dass anhand der reinen Nutzer-Zahlen und gesendeten Twitter-Kurznachrichten nicht die '"Gesundheit“ des Twitter-Netzwerks nachvollzogen werden kann, andererseits werden durch diese Massnahme auch strategische Ziele verfolgt. [3] 2013 ging Twitter an die Börse und vermeldete 100 Millionen täglich aktive Nutzer und über 500 Millionen Twitter-Kurznachrichten, die täglich über den Dienst versendet werden.

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

2.2.2 Was ist Twitter?

Twitter wird als Kurznachrichten-Dienst, Mikroblogging-Dienst oder auch als soziales-Netzwerk bezeichnet. Twitter Geschäftsführer Kevin Thau hat 2010 auf dem Nokia-World Kongress öffentlich bestritten, dass Twitter ein Soziales-Netzwerk ist. Laut Thau handelt es sich um ein Nachrichten-, Inhalts- und Informations-Netzwerk. Er begründete dies damit, dass Twitter die Art und Weise wie Nachrichten verteilt werden geändert hat und praktisch jeder zum Journalisten werden kann. Als Beispiel nennt er die Landung des Fluges 1549 auf dem Hudson River. Die Augenzeugen hätten damals keine Mails versendet um die Nachricht zu verbreiten, sondern die Nachricht via Twitter weitergegeben. Es lassen sich eine Reihe weitere Beispiele derselben Art finden. In [POM+13] wird ein Vergleich zwischen sogenannten Newswire anbietern und Twitter gezogen. [4] Es stellte sich heraus, dass über nahezu alle Nachrichten, welche in den Newswires verbreitet wurden auch im Twitter-Netzwerk berichtet wird. Nachrichten zu bestimmten, vermutlich sehr speziellen Themen oder Auslandsnachrichten wurden ausschliesslich in Twitter gefunden. Diese Erkenntnisse decken sich mit der Einschätzung von Kevin Thau. In [KLPM10] wird die Einschätzung, bei Twitter handele es sich nicht um ein soziales Netzwerk, wissenschaftlich bestätigt. Kwak et al überprüfen die in [NP03] beschriebenen Eigenschaften sozialer Netzwerke und kommen zu dem Schluss, dass Twitter diese Eigenschaften nicht erfüllt.

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

[4] 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.

Nachrichten im Twitter-Netzwerk einen Brodcast an alle Nutzer darstellen.

Die 140 Zeichen langen Nachrichten in Twitter werden als Tweets bezeichnet. Tweet bedeutet übersetzt Zwitschern, womit die Redenwendung '"Die Spatzen zwitschern es von den Dächern“ auch im Twitter-Netzwerk zu einer passenden Redenwendung wird. In der vorliegenden Arbeit wird Twitter deshalb als Mikroblogging-Dienst bezeichnet.

2.2.3 Funktionen von Twitter

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

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

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

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

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

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

Weiterleiten eines Tweets (Retweet) Unter einem Retweet versteht man das weiterleiten eines Tweets den man nicht selbst verfasst hat an die eigenen Follower. Genauer

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

1 retweet Reichweite 1000 Nutzer

Diagramm Retweet, Filterfunktion

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

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

siehe Bild ref1

siehe Bild ref2

siehe Bild ref3

lichen Tweet, auf weitere User referenziert wurde. Aber auch wenn ein Nutzer auf eine bestehende Konversation antwortet, werden alle beteiligtien Nutzer referenziert.

siehe Bild ref4

Diagramm Antwort, Antwort Thread, Bild Antworten Button, Referenzieren

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

2.2.4 Daten einer Twitter-Nachricht

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

siehe Bild

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

2.2.5 Geoinformationen in Twitter Daten

Welche Tweet-Daten können zur Georeferenzierung herangezogen werden

Nur optional

Um diese Frage zu beantworten, müssen die Tweet-Daten eingehend untersucht werden. Dabei spielt nicht nur die reine Information die den Daten entnommen werden kann eine Rolle, sondern auch wie die Daten generietr oder eingegeben wurden. Beispielsweise kann bei einem Tweet, dem ein Längen- und Breitengrad mit einer Genauigkeit von 14 Nachkommastellen zugeordnet ist, davon ausgegangen werden, dass die geografische Position der tatsächlichen geografischen Position, von welcher der Tweet abgesetzt wurde, entspricht. Es liegt hier die Vermutung nahe, dass diese Werte durch ein mobiles

Abbildung 2.3: Was ist zu sehen?

GPS [5] erfasst worden sind. Anders verhält sich dies beispielsweise beim Tweet-Text, eine Erwähnung der Stadt New York, muss nicht bedeuten, dass der Tweet aus dieser Stadt stammt. Es impliziert nicht einmal, dass der Verfasser jemals in dieser Stadt war. Im folgenden werden einige Datenfelder, welche mit jedem Tweet versandt werden, untersucht. Dabei wird die Eignung dieser Daten als geografischer Indikatoren bewertet. Währenddessen werden anhand geeigneter Beispiele die Bergiffe gesicherter -, ungesicherter -, mittelbarer - und unmittelbarer geografischer Indikator eingeführt.

2.2.6 geografischer Indikator

Unter einem geografischen Indikator wird eine Angabe verstanden, welche direkt einem Nutzer zugeordnet werden kann und die Auskunft über die geografische Position oder Region des Nutzers geben kann. Im Zuge dieser Arbeit wurden potentielle geografische

[5] Global Positioning System

Indikatoren untersucht und eine Reihe von Eigenschaften identifiziert anhand derer sich geografische Indikatoren kategorisieren lassen. Diese Eigenschaften haben Einfluss darauf, wie und ob eine Georeferenz aus dem Indikator abgeleitet werden kann. Dabei ist zu unterscheiden ob sich die Eigenschaft auf den, durch den Nutzer eingegebenen, Wert bezieht oder auf die Information die durch die Angabe geliefert werden soll.

Objektivität der Werte geografischer Indikatoren

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

Geben zwei Nutzer unterschiedliche Werte ein, obwohl sie denselben Ort oder dieselbe Region beschreiben wollen, ist dieser Wert nicht objektiv. Eingaben in Freitext Felder, welche ohne weitere Verarbeitung übernommen werden fallen in diese Kategorie.

Zuverlässigkeit der Werte geografischer Indikatoren

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

Gesicherte Werte geografischer Indikatoren

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

mittelbare und unmittelbare geografische Indikatoren

Diese Eigenschaft bezieht sich, im Gegensatz zu den vorgenannten auf die Information die durch das Feld oder die Eingabe repräsentiert werden soll.

Hierbei handelt es sich um Indikatoren welche unter Umständen einen geogarfischen Bezug haben können, aber die Intention bei der Eingabe war nicht eine geografische Position mitzuteilen.

Beispielsweise die Verwendung bestimmter umgangssprachlicher oder dialektischer Wörter die einen Hinweis auf eine geografische Region geben können. Die Intention des Nutzers war nicht, dadurch einen Hinweis auf die geografische Position oder seinen Aufenthalstort zu geben, allerdings lässt sich mittelbar eine geografische Region daraus ableiten.

unmittelbare geografische Indikatoren in Tweet-Daten

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

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

3 Stand der Technik

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

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

3.1 Kategorisierung bestehender Ansätze

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

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

Kategorisierung über die untersuchte Ressource

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

Kategorisierung über die verwendete Methode

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.1.1 ttt<sss

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

Indikatoren aus [SHP+13]

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

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

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

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

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

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

Requirements Tabelle einfügen

1. Naiver Ansatz -> Geocoding mit Google Maps API V3, nur Indikatoren die geografische Namen enthalten. Prinzipiell einfache Datenbankabfrage mit ein wenig semantik. Keine Jargon Namen wie Big Apple etc.
    a) Funktion der GMaps Api V3
    b) Einschränkungen der GMaps Api V3
    c) zurückgelieferte Daten der GMaps Api V3
    d) Kurze Beschreibung wie ich die API genutzt habe
2. aktuelle Ansätze
    a) Verfahren mit Inhaltsanalysen
    b) Verfahren mit Indikatoren einzelne oder mehrere

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

geografische Entität definieren

3.1.2 Probleme früherer Ansätze

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

4 Lösungsansatz

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

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

4.1 Überblick

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

Die Georeferenzierung nach dem Schema aus Abbildung 4.1 setzt voraus, dass aus den eingegebenen Indikatoren eine Georeferenz abgeleitet werden kann. Da es sich beim Nutzer-Standort und der Nutzer-Zeitzone um unmittelbare geografische Indikatoren handelt, lässt sich ein geografischer Bezug aus diesen Indikatoren ableiten. Die Idee besteht nun darin aus einer umfangreichen Tweet-Datensammlung die Zuordnung der eingegebenen Indikatoren zu geografischen Positionen zu lernen und diese in einer Datenbasis zu

Schema Darstellung Eingabe->Ausgabe Sketchbook A1: Unterschrift Die Eingabe besteht aus dem Nutzer-Standort sowie der Nutzer-Zeitzone. Als Rahmenbedingungen wird die gewünschte Hierarchie-Ebene sowie der Schwellwert für die Konfidenz angeben. Als Ausgabe erhält man eine Georeferenz und einen Konfidenzwert, der angibt wie sicher das Ergebnis ist.

speichern. Diese Datenbasis soll im folgenden als Georeferenz-Basis bezeichnet werden. Damit dies realsiert werden kann, muss die Tweet-Datensammlung pro Datensatz den Nutzer-Standort, die Nutzer-Zeitzone und eine Georeferenz in Form von Längen- und Breitengrad beinhalten. [1]

Der Wert des Nutzer-Standortes ist nicht objektiv und unzuverlässig, dies macht eine Analyse und eine entsprechende Vorverarbeitung dieses Indikators nötig. Zur Erstellung der Georeferenz-Basis werden deshalb zunächst einige Schritte zur Vorverarbeitung der Indikatoren durchgeführt. Durch die Vorverarbeitung werden die eingegebenen Indikatoren genauer analysiert, zusätzliche Informationen extrahiert und in eine einheitliche Form gebracht. Die Indikatoren werden durch die Vorverarbeitung derart verändert, dass aus einem Indikator-Paar mehrere Datensätze enstehen können, welche Informationen aus den ursprünglichen Indikatoren enthalten. Danach wird mit Hilfe des Längen- und Breitengrades jedem Tweet eine konkrete Instanz der untersten Ebene der geografischen Hierarchie zugeordnet. Legt man die in Abschnitt 2 vorgestellte geografische Hierarchie zugrunde, wird dadurch jedem Tweet eine Stadt zugeordnet. Durch diese Zuordnung werden jedem Tweet implizit auch die höheren Ebenen der geografischen Hierarchie zugeordnet. Dieser Schritt markiert den Übergang von einer kontinuierlichen Darstellung der Georeferenz, durch Längen- und Breitengrad, in eine diskrete Darstellung, durch eine Stadt. Die so verarbeiteten Datensätze werden nun einer simplen statistischen Auswertung unterzogen indem ermittelt wird wie häufig einer Stadt ein bestimmter Indikator zugeordnet wurde. Die Georeferenz-Basis beinhaltet nun die verarbeiteten Indikatoren und die zuvor zugeordnete geografische Position, in Form einer Stadt, sowie die Häufigkeit in der die Kombination Indikator/Stadt ermittelt wurde. Der Ablauf zur Erstellung der Georeferenz-Basis wird in Abbildung **??** dargestellt.

Bei der Georeferenzierung werden die Indikatoren zunächst derselben Vorverarbeitung wie beim einlernen unterzogen. Die Vorverarbeitungsschritte sind somit generisch, da sie sowohl vor dem erzeugen der Georeferenz-Basis als auch vor der eigentlichen Georeferenzierung durchgeführt werden. Nach der Vorverarbeitung wird in der Georeferenz-Basis nach diesen vorverarbeitenden Indikatoren gesucht. Das Ergebnis dieser Suche wird unter Berücksichtigung des Konfidenzschwellwertes und der geografischen Hierarchie ausgewertet und die zugeordnete geografische Position wird ausgegeben. Der Ablauf der Georeferenzierung wird in Abbildung **??** dargestellt.

[1] Siehe Kapitel 1 genutzte Daten

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

Im ersten Teil wird das Einlernen der Georeferenz-Basis behandelt. Im Zuge des Einlernens der Georeferenz-Basis sollen auch die generischen Vorverarbeitungsschritte erläutert werden. Die Vorverarbeitungsschritte lassen sich in zwei Klassen einteilen. Die erste Klasse beinhaltet Verarbeitungsschritte um die Indikatoren für die weitere Verarbeitung zu optimieren und vorzubereiten, diese extrahieren keine zusätzlichen Informationen aus den Indikatoren. Die zweite Klasse beinhaltet Vorverarbeitungsschritte, durch die versucht wird zusätzliche Informationen aus den Indikatoren zu extrahieren. Diese Verarbeitungsschritte hängen stark zusammen mit der verwendeten geografischen Hierarchie und der späteren Auswertung. Deshalb werden diese Vorverarbeitungsschritte in Bezug auf die verwendete geografische Hierarchie und das Auswertungs-Verfahren betrachtet. Danach soll der gesamte Ablauf des Einlernens noch einmal dargestellt und kurz erklärt werden um einen Gesamtüberblick zu liefern.

Zuletzt soll auf die Georeferenzierung eingegangen werden. Die Vorverarbeitungsschritte der Indikatoren wird in diesem Teil nur angeschnitten, da diese bereits im ersten Teil genau erläutert werden. Nachdem die Indikatoren vorverarbeitet wurden, wird in der Georeferenz-Basis nach korrespondierenden Indikatoren gesucht. Die Ergebnisse dieser Suche werden unter Berücksichtigung des Konfidenzschwellwertes und der gewünschten georafischen Genauigkeit, in Form der geografischen Hierarchieebene, ausgewertet. Das Ergebnis dieser Auswertung ist die gewünschte Georeferenz.

4.2 Einlernen der Georeferenz-Basis

Zunächst sollen die Vorverarbeitungsschritte der ersten Klasse erklärt werden. Danach soll der Nutzer-Standort als geografischer Indikator genauer betrachtet werden und die Extrahierung tiefergehender Informationen aus den Indikatoren werden erläutert. Dabei wird insbesondere auf den Zusammenhang zwischen den Vorverarbeitungsschritten der zweiten Klasse, der geografischen Hierarchie und der darauffolgenden Auswertung der Ergebnisse eingegangen. Die Tatsache, dass der Wert des Nutzer-Standortes nicht objektiv und unzuverlässig ist muss dabei beachtet werden.

Der Nutzer-Standort als geografischer Indikator

Zunächst soll der Nutzer-Standort als geografischer Indikator eingehender betrachtet werden. In Abschnitt 2 wurde der Nutzer Standort bezüglich seiner Eigenschaften untersucht. Demnach kann der Nutzer-Standort als unmittelbarer geografischer Indikator angesehen werden, da als Eingabe eine geografische Position gefordert wird. In zahlreichen anderen Arbeiten wurde deshalb versucht den Nutzer-Standort durch eine Anfrage an Geografie-Datenbanken oder Geocoding-APIs auf eine konkrete geografische Position aufzulösen. Dieser naive Ansatz ignoriert allerdings die Eigenschaften des Wertes der im Nutzer-Standort gespeichert ist und die folgenden Probleme verursacht.

Referenzen auf frühere Ansätze mit Gazetteer Abfrage

Nach Abschnitt 2 ist der Wert des Nutzer-Standortes nicht objektiv, nicht zuverlässig und nicht gesichert.

Nicht zuverlässig Diese Eigenschaft sagt aus, dass der eingegebene Wert nicht zwingend ein Toponym darstellt. Vom Nutzer kann jeder beliebige Wert eingegeben werden, unabhängig davon ob dieser ein Toponym ist oder nicht. Beim naiven Ansatz, der Abfrage an eine Geocoding-Api oder eine Geografie-Datenbank wird allerdings implizit angenommen das es sich beim eingegebenen Wert um ein Toponym handelt. Dadurch kann es zu Fehlern bei der Auflösung kommen. Dadurch das der Wert nicht zuverlässig ist wird die Eigenschaft, dass der Nutzer-Standort ein unmittelbarer geografischer Indikator ist, in gewisser Weise relativiert. Denn theoretisch könnte jeder Nutzer einen Wert eingeben, der kein Toponym darstellt, womit die Eigenschaft des unmittelbaren geografischen Indikators hinfällig wäre. Es muss allerdings beachtet werden, dass der Nutzer-Standorts von einem Großteil der Nutzer zu seinem vorgesehenen Zweck genutzt wird. Nach Hecht et al in [HHSC11] wird in der Fälle der Nutzer-Standort verwendet um ein Toponym anzugeben. [2] Im Zuge der vorliegenden Arbeit wurde eine manuelle Untersuchung an xyz Tausend Twitter-Profilen vorgenommen bei denen xyz Prozent der Nutzer ein Toponym im Nutzer-Standort gespeichert hatten.

xyz Prozent

kleines Beispiel mit Hilfe der Geocoding Api erstellen

[2] Hecht et al untersuchen allerdings nur Nutzer-Standorte die in englischer Sprache eingegeben wurden.

Nicht gesichert Dadurch besteht keine Garantie, dass der eingegebene Wert, sollte er ein Toponym darstellen, dem tatsächlichen Standort des Nutzers entspricht. Dieser Umstand wird bei einer Auflösung mit Hilfe einer Geocoding-API oder Geografie-Datenbank nicht berücksichtigt.

Nicht Objektiv Dies bedeutet, dass dieselbe georafische Position von jeweils zwei Nutzern mit unterschiedlich Toponymen bezeichnet werden kann. Dies wäre an sich kein Problem, wenn Toponyme über die Geocoding-API oder eine Geografie-Datenbank korrekt aufgelöst werden würden. Allerdings ist in Kombination mit der Unzuverlässigkeit des Wertes zu beachten, dass auch Werte die nicht aufgelöst werden können und somit nicht direkt als Toponym identifiziert werden, einen geografischen Ort beschreiben können und damit ein Toponym darstellen. Dies kann der Fall sein wenn ein Ort beispielsweise eine lokale Bezeichnung besitzt welche in Datenbanken nicht hinterlegt ist. Als Beispiel hierfür sollen Spitznamen für Städte dienen. Bei Wikipedia sind für die Stadt Detroit im US-Bundesstaat Michigan folgende Spitznamen angegeben: The Motor City, Motown, Hockeytown, Rock City und The D. Die ersten zwei dürften weltweit einen Gewissen Bekanntheitsgrad haben, wohingegen Hockeytown, Rock City und The D weniger bekannt sein dürften. Es existieren Datenbanken mit solchen Städte-Spitznamen, aber für diese kann keine Vollständigkeit garantiert werden. Auch besteht die Möglichkeit das in Teilen der Welt Spitznamen für Städte existieren die kaum über die Grenzen des Landes bekannt sind, oder aber die Schreibweise ist landesintern anders als diese in der Datenbank hinterlegt ist. Eine weitere Fehlerquelle sind netzwerkinterne Schreibweisen die sich etabliert haben und ausschließlich innerhalb eines Netzwerks verwendet werden. Beachtet man nun noch, dass das Verfahren unabhängig von der verwendeten Sprache und dem verwendeten Alphabet funktionieren soll, ist eine Auflösung mit Hilfe von Geocoding-APIs oder einer Geografie-Datenbank sehr schwer umzusetzen und vor allem schwer zu kontrollieren. Zudem ist es fraglich in wie weit und ob diese Spitznamen im Nutzer-Standort überhaupt verwendet werden. Neben Spitznamen von Städten sind aber auch Toponyme denkbar die einen bestimmten Landstrich oder eine landschaftliche Besonderheit beschreiben. Beispieslweise beschreibt 'Än der Förde'eine geografische Region an der Ostsee, meist ist damit die Kieler Förde gemeint, es wird allerdings nicht explizit daraf hingewiesen. Diese geografischen Bezeichnungen sind unter Umständen nur in speziellen Datenbanken zu finden und können so nur schlecht aufgelöst werden.

Diese sehr speziellen Fälle sind im Twitter-Netzwerk nicht unbedingt zu erwarten, es kann allerdings auch nicht gänzlich ausgeschlossen werden, dass solche Bezeichnungen verwendet werden. Des weiteren kann aufgrund der nicht Objektivität und der Unzuverlässigkeit nicht garantiert werden, dass der aufgelöste Ort tatsächlich der Standort ist den der Nutzer angeben wollte. Es gibt zahlreiche Städte, die in mehreren Ländern vorkommen. 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. So gibt es in den USA 30 Städte mit dem Namen Hamburg, 40 mit dem Namen Hannover, 39 mit dem Namen Berlin und 25 mit dem Namen Frankfurt um nur einige beispiele zu nennen. Es gibt auch Orte die in einem Land mehrfach vorkommen. Innerhalb Deutschlands gibt es beispielsweise 12 Gemeinden mit dem Namen Hausen und zahlreiche Stadtteile die diesen Namen tragen.

Diese Umstände machen es schwierig und sehr unsicher eine Auflösung direkt über den angegebenen Wert durchzuführen. Die Idee besteht deshalb darin den Nutzer-Standort so zu behandeln als ob keine Information direkt aus ihm entnommen werden kann. Es soll aus den angegebenen Werten für den Nutzer-Standort eine Zuordnung zu gegrafischen Positionen gelernt werden. Damit werden einige der Probleme die in Bezug auf die Eigenschaften des Nutzer-Standortes bestehen vermieden.

Eliminierung von Satz- und Sonderzeichen

Oft werden im Nutzer-Standort Sonder- und Satzzeichen verwendet. Einige Nutzer haben ausschließlich Sonder- und Satzzeichen als Nutzer-Standort angegeben, andere benutzen diese um Emotionen auszudrücken oder als Dekoration. Beispiele hierfür sind “I ♡ New York“ oder “†˜ Los Angeles˜†“. Es konnte kein Hinweis darauf gefunden werden, dass Sonder- und Satzzeichen zusätzliche Informationen zum angebenen Nutzer-Standort liefern. In den x Millionen verfügbaren Tweets befinden sich y Nutzer welche einen Nutzer-Standort angegeben haben der ausschliesslich aus Sonder- oder Satzzeichen besteht. Es wurden lediglich einige Beispiele gefunden in denen Satzzeichen ein Bestandteil des Standortes sind. Ein Beispiel hierfür ist “Paris, 3. Arrondissement“. Entfernt man die Satzzeichen, liefern diese Werte immernoch ausreichende Informationen über den Standort. In diesem Vorverarbeitungsschritt werden alle Sonder- und Satzzei-

Auswertung von Sonder Stazzeichen Datensätzen in UL

chen entfernt. Dadurch wird gewährleistet, dass keine unnötigen Zeichen in den folgenden Schritten verarbeitet werden müssen.

Tokenisierung

Identifizierung von Toponymen innerhalb des Nutzer-Standortes

Da der Nutzer-Standort ein unmittelbaren geografischen Indikator darstellt kann eine Analyse auf Toponyme Vorteile bringen. In diesem Schritt werden Toponyme, welche aus zwei oder mehr Wörtern bestehen zusammengefasst um in folgenden nicht einzeln behandelt zu werden. Dadurch wird die weitere Verarbeitung effizienrter bezüglich der Anzahl der zu verarbeitenden Worte. Zum Beispiel werden die Token [New] [York] [City] zusammengefasst zu dem Token [New York City] Aber insbesondere soll hier nicht das Toponym analysiert und direkt eine Georeferenz hinzugefügt werden. Die Identifizierung von toponymen dient lediglich dazu die folgenden Verarbeitungsschritte effizienter zu machen.

4.2.1 Der Nutzer-Standort als geografischer Indikator

Der Nutzer-Standort ist ein unmittelbarer geografischer Indikator, es wird vom Benutzer eine geografische Angabe, nämlich der Standort des Nutzers abgefragt. Es kann also zunächst davon ausgegangen werden, dass der Nutzer-Standort eine geografische Angabe darstellt. Allerdings ist der Wert des Nutzer-Standortes nicht objektiv. Diese Eigenschaft resultiert daraus, dass der Nutzer-Standort von jedem Nutzer frei eingegeben werden kann und ohne weitere Verarbeitung durch Twitter gespeichert wird. Aus diesem Umstand kann auch direkt hergeleitet werden, dass der Nutzer-Standort unzuverlässig ist. Damit ist bei der Auswertung mit Unsicherheiten bezüglich des Wertes zu rechnen.

Des weiteren soll das Verfahren unabhängig der verwendeten Sprache und des verwendeten Alphabets durchgeführt werden.

Es kann also davon ausgegangen werden, dass eine geografische Angabe durch den Nutzer gemacht wird, aber man muss aufgrund der nicht Objektivität und der Unzuverlässigkeit

davon ausgehen, dass die Angaben nicht korrekt, nicht überprüfbar, oder überhaupt keine geografische Angabe darstellen.

Dies bedeutet zum einen, dass die Werte die im Nutzer-Standort angegeben werden nicht einheitlich sind, und das die eingegebenen Werte nicht zwingend einen geografischen Ort bezeichnen.

Auch ist der Nutzer-Standort unzuverlässig, womit

Des weiteren soll nach den Anforderungen in Abschnitt 1.3.1 die Auswertung unabhängig von der verwendeten Sprache und den verwendeten Schriftzeichen sein.

In Tabelle **??** sind einige Werte aufgeführt die als Nutzer-Standort angegeben wurden, aber keine geografische Position oder geografische Region bezeichnen. Diese Werte sind natürlcih zur Georeferenzierung unbrauchbar.

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

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

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

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

Umschreiben und woanders darauf eingehen!

Der Nutzer-Standort wird einer mehrstufigen Vorverarbeitung unterzogen. Die Vorverarbeitung des Nutzer-Standortes erfüllt mehrere Aufgaben.

Zeitzone als zusätzlichen geografischen Indikator einführen Unsicherheiten ausräumen Beispiel suchen

Zunächst soll sichergestellt werden, dass keine unnötigen Sonder- oder Satzzeichen in die weitere Verarbeitung einfliessen. Im Nutzer-Standort finden sich oft Sonder- oder Satzzeichen, die für die Bestimmung des Standortes von keinem oder sehr geringem Nutzen sind, diese sollen in der weiteren Verarbeitung nicht berücksichtigt werden. Des weiteren sollen Toponyme welche aus zwei oder mehr Wörtern bestehen erkannt, und als zusammengehörig markiert werden. Die erkannten Toponyme werden nicht zur Georeferenzierung genutzt, sondern ausschließlich um die Wörter zusammenfassen zu können und in weiteren Verarbeitungsschritten als zusammengehörige Einheit zu verarbeiten. Dies bringt den Vorteil, dass in den nächsten Verarbeitungsschritten weniger Wörter verarbeitet werden müssen. Die so verarbeiteten Worte werden zuletzt in Tokens aufgeteilt und alphanumerisch sortiert. Dieser Schritt dient zur Vorverarbeitung für den nächsten Verarbeitugsschritt.

Toponym definieren

alphanumerische Sortierung der Tokens

URL Encoding

4.2.2 Encoding

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

4.2.3 Die Zeitzone als geografischer Indikator

4.3 Einlernen der Georeferenz-Basis

4.3.1 Zuordnung der georeferenzierten Tweets zu geografischen Entitäten

4.3.2 Erzeugung von N-Grammen

einbauen als Beispiel für Ngramisierung Beispiel wenn Karlsruhe Deutschland und Berlin Deutschland und Abstatt Deutschland dann kann durch die geografische Hierarchie und die NGramme Wissen über den Term Deutschland gezogen werden da auf Länderebene Deutschland immer dem gleichen geografischen Ding zugeordnet wird

Vorteil Toponym Identifizierung Verarbeitungsschritte: insbesondere bei der Erzeugung von N-Grammen im Schritt

Erzeugung von N-Grammen Aus den alphanumerisch sortierten Tokens der Vorverarbeitung werden in diesem Schritt N-Gramme erzeugt. Genauer werden N-Gramme der Ordnung 1, 2 und 3 erzeugt. Diese werden auch als Mono-, Bi- und Tri-Gramme bezeichnet.

Jedes dieser so entstandenen N-Gramme wird die Zeitzone angehängt. Das Ergebnis dieser Phase ist eine Datenbasis welche eine Reihe von Zeichenketten aus der Nutzer-Standort und der Nutzer-Zeitzone generiert haben. Jeder dieser Zeicheneketten wird die Häufigkeit ihres Vorkommens und die nächste Stadt mit über 15.000 Einwohnern zugeordnet.

4.3.3 Matching auf N-Gramme und geografische Entität

Counting

1. Datenstruktur

2. Matching
3. Counting

4.4 Georeferenzierung

In der zweiten Phase kann mit Hilfe der Datenbasis die eigentliche Georeferenzierung von Twitter-Nutzern, mit unbekannter geografischer Position, durchgeführt werden. Es wird dem Anwender ermöglicht die Genauigkeit über die geografischen Hierarchien anzugeben und einen Konfidenzschwellwert zu bestimmen.

Bei der Georeferenzierung kann bestimmt werden wie genau die geografische Position oder die geografische Region bestimmt werden soll. Dabei kann zwischen den bereits in Kapitel 2.1 erläuterten Hierarchieebenen gewählt werden. Dies hat den Vorteil, dass die Genauigkeit der Georeferenzierung den jeweiligen Anforderungen angepasst werden kann. Des weiteren kann eine Konfidenz angegeben werden, desto höher die Konfidenzschwelle gewählt wird desto sicherer ist die Georeferenzierung.

4.5 Geolocation Mapping

4.5.1 nearest neighbour mapping

1. Wie genau kann gemappt werden? Fehler Durchschnitt.
2. Mapping auf cities 1000/1000/15000 mit Daten zu durchschnitll. Abstand
3. Hier ist noch Verbesserungspotenzial -> wenn Mapping Distanz zu weit entfernt -> verwerfen!

Welches Fehlermaß kann für das mapping angewandt werden? auf Städteebene gut möglich mit geografischen Distanzen, admin2,amdin1, Land schlecht möglich mit Distanzen

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

4.6.1 Generierung eines Wissendatensatzes

4.6.2 Verknüpfung mit Geodaten

4.6.3 Auflösen auf Administartionsebenen, Länder

4.7 Lokalisieren von Tweets ohne konkrete geografische Daten

4.7.1 Ablauf der Lokalisierung

4.7.2 Lokalisierungssicherheit durch Ausnutzung der geografischen Hierarchiebeziehungen

einbauen!!!

4.7.3 Geografische Grundbegriffe und Geografiedaten

Geografische Grundbegriffe

Geonames.org

[3]

[3] eventuell erst in Implemetierung darauf eingehen

4.7.4 ???

N-Gramme

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

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

4.8 Einbauen in dieses Kapitel

Die folgenden Vorverarbeitungsschritte werden durchgeführt.

1. Eliminierung von Sonderzeichen
2. ausschließlich Kleinschreibung
3. geografische Zuordnung von Teilworten
4. Tokenisierung
5. html Encoding
6. alphanumerische Sortierung

Bessere Umschreibung finden für Teilworte, es ist eig. eine Menge von Worten aus Nutzer-Standort

Diese Vorverarbeitung dient der Bereiningung und Vereinheitlichung des Nutzer-Standorts. Die gegografische Zuordnung dient der

In den einzelnen Stufen der Vorverarbeitung werden Sonderzeichen entfernt, alle Zeichen in Kleinbuchstaben umgewandelt, in der Wortfolge werden nach geografischen Begriffen gesucht die zwei oder mehr Worte beinhalten, die Zeichen werden url-encoded und die einzlenen Wörter werden in Tokens zerlegt welche alphanumerisch sortiert werden. Im nächsten Schritt werden aus den Tokens Uni-Gramme, Bi-Gramme und Tri-Gramme erstellt.

In Bewertung einbauen Legt man zugrunde, dass der Nutzer-Standort als ungesicherter, unmittelbarer geografischer Indikator angesehen werden kann, besteht eine Möglichkeit der Georeferenzierung darin, eine sogenannte Geocoding-API zu nutzen. Die Nutzer-Zeitzone wird in diesem Fall nicht als Indikator herangezogen, da diese von den Geocoding-API's nicht als zusätzliche Information verarbeitet wird. Eine Reihe von bekannten Firmen bietet eine API zur Georeferenzierung an. Die bekanntesten sind Google, Yahoo, Microsoft, Map Quest und Cloud Made. Teilweise sind die Anfragen, welche an die Geocoding-API's gesendet werden können, in ihrer Anzahl begrenzt. Auch die Antwortzeiten der Geocoding-APIs begrenzen die Anzahl möglicher Anfragen pro Zeiteinheit. In Tabelle **??** ist eine Auflistung der Anbieter mit den jeweiligen Begrenzungen dargestellt. Eine detaillierte Analyse der Antwortzeiten wurde im Zuge dieser Arbeit nicht durchgeführt. [4]

Tabelle mit Georef Api Anbietern.

Link in footnote einfügen

Dieses vorgehen wird in einigen Arbeiten angewendet um den Nutzer-Standort zu bestimmen. Dabei wird, mit einer simplen Vorverarbeitung des Nutzer-Standortes, direkt in einer Geografie-Datenbank nach der eingegebenen Zeichenfolge gesucht.

[4] für eine Analyse Vergleiche

5 Implementierung

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

Datensätze in Grundlagen?

5.1 Komponenten der Referenzimplementierung

5.1.1 Architektur

Allgemeine Architektur der Refernzimplementierung

5.1.2 Präprozessorverarbeitung - Erzeugung der N-Gramme

Warum Präprozessoren -> schnelleres ändern der Vorverarbeitung.

5.2 Datenbank

5.2.1

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

5.3 Geografie Daten

in Implementierung verschieben

5.4 Data Sample

Beschreibung wie Daten erzeugt wurden, Zeiträume, Analysen

5.5 geonames.org

Allgemeines zu geonames.org, was ist geonames.org.

1. Woher stammen die Daten?
2. Umfang und Informationen
3. Aktualität
4. Hierarchiebeziehungen im geonames.org Datensatz

6 Leistungsbewertung

7 Schlussfolgerungen, Ausblick und Fragen

8 Zusammenfassung

9 Ideen und Notizen

9.1 Stakeholder analyse

Welche potenziellen Stakeholder profitieren von der Arbeit? Was benötigt jeder dieser Stakeholder? Bedürfnisse analysieren und Begründen.

1. Marketing Professionals
2. Statistiker allgemein
3. Sozialwissenschaftler -> Analyse von Informationsströmen

9.2 Fragen an Matthias

9.2.1 Strukturell

1. Soll ich noch auf die Messung eines Informationsflusses eingehen? Wenn ich keine Informationsflüsse untersuche hängt dieses Thema ein wenig in der Luft.
2. ???

9.2.2 Inhalt

9.3 Ideen

1. Voraussetzungen zur Anwendung des Verfahrens — In Einleitung

a) Lerndaten mit konkreten geografischen Angaben

b) Indikatoren in Lerndaten, welche auch in Datensätzen ohne konkrete geografische Angaben vorkommen (hier eventuelle Diskrepanzen zwischen geogetaggten und nicht geogetaggten tweets + Mentalität in bestimmten Ländern)

c) Indikatoren mit geografischem Bezug, oder hinreichendem geografischen Bezug, Mittelbar oder unmittelbar

2. Auf Jargon Namen für Städte eingehen, wie bspsw. the big apple -> New York City
3. Landesgrenzen-Problematik wird durch meine Lösung obsolet -> auf stakeholder eingehen
4. Wahrscheinlichkeiten für korrekte Lokalisierung kann angegeben und justiert werden

Korrelation zwischen Lokalisierungungssicherheit und tatsächlichem Match berechnen

5. Wenn Wahrscheinlichkeiten auf best. Ebene nicht hoch genug dann verschieben auf Admin2 -> Admin1 -> Länderebene
6. mit vorherigem werden Unsicherheiten bei Lokalisierung abgebildet (Wichtig für Informationsflüsse)
7.

9.4 Formulierungen

9.4.1 unmittelbare ungesicherte geografische Indikatoren

Das “userlocation“ Feld in einem Tweet kann durchaus eine konkrete Lokation beinhalten, jedoch wird auch oft irgendetwas eingetragen. [HHSC11] Es kann sich dabei um beliebige Wörter oder Sätze handeln, die einzige Limitierung ist die Anzahl zur Verfügung stehender Zeichen. Nichtsdestotrotz ist es das Ziel dieses Feldes seinen eigene Standort anzugeben. Dabei kann allerdings nicht davon ausgegangen werde, das der eingetragene Wert nicht doch in einem Zusammenhang mit einer geografischen Lokation steht. Bezeichnungen von Städten in Umgangssprache wie besipielsweise "The Big Apple"für New York City oder Motown für Detroit, sind für einige Personen nicht unmittelbar

zuzuordenen, geben allerdiings eine konkrete Lokation an. Da die Masse an Bei bzw. Spitznamen für Städte nicht überschaubar ist und auch sprachliche Probleme bestehen ist es sinnvoll alle userlocation Einträge gleich zu behandeln und diese in erster Linie als Lokationsangaben zu behandeln. Durch die Einschränkung auf eine Geolocation werden einzelne gleich lautende Einträge, welche aber nicht auf einen konkreten Ort hinweisen in einzelnen Datensätzen abgelegt.

9.5 Datenbasis

1. Welche Datenbasis wurde genutzt
    a) Streaming API
    b) Is the Sample good enough (Morstatter et al 13)
    c) When is it biased? (Morstatter et al)
    d) How does the Data sampling Startegy Impact the Discovery of Information Diffusion in Social Media (De Choudhurry, 1)
2. Lerndatensatz
3. Kontrolldatensatz
4. Manuell getaggter Datensatz
5. Google Maps getaggter Datensatz

9.6 Vorteile neuer Ansatz bei Mapping auf Geografische Daten

Notwendigkeit/Vorteile von Hierarchiebeziehungen im Mapping auf Geograohie Daten

Literaturverzeichnis

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

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

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

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

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

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

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

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

[HHSC11] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. Tweets from justin bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 237–246, New York, NY, USA, 2011. ACM.

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

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

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

[KCLC13] Krishna Y. Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. Spatio-temporal dynamics of online memes: A study of geo-tagged tweets. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 667–678, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

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

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

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

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

[PCV13] Reid Priedhorsky, Aron Culotta, and Sara Y. Del Valle. Inferring the origin locations of tweets with quantitative confidence. *CoRR*, abs/1305.3932, 2013.

[POM+13] S. Petrovic, M. Osborne, R. Mccreadie, C. Macdonald, and I. Ounis. Can twitter replace newswire for breaking news? In *ICWSM - 13*, 2013.

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

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

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

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

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

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