Karlsruhe Institute of Technology

Hier steht der Titel der Diplom- oder Studienarbeit

Diplomarbeit
von

Vorname Nachname

an der Fakultät für Informatik, Institut für Telematik
Forschungsbereich Dezentrale Systeme und Netzdienste

Erstgutachter:	Prof. Dr. Hannes Hartenstein
Zweitgutachter:	??
Betreuer:	Dipl.-Inform. Vorname Nachname
Bearbeitungszeit:	01. Januar 2010 – 30. Juni 2010

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

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 [26]. Die Twitter-Nutzer verfassen täglich mehr als 500 Millionen Tweets [33]. 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 [28]. Tumasjan et al. untersuchen in [32] 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 [7] 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 [31], 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 [29]. 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 (Toponyme) verwendet, um seinen Standort anzugeben. Der Nutzer-Standort wird jedoch über ein Freitext-Feld eingegeben und direkt abgespeichert. Durch die freie Eingabe werden unter anderem 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. [11] 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 Tweets 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 der Qulität der Ergebnisse zu justieren?
- Ausnutzen einer geografischen Hierarchie zur besseren Bestimmung des geografischen Standortes?

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. Es sollen Vorgaben bezüglich der Güte und der Trefferquote der Ergebnisse gemacht werden können. Das Verfahren soll abhängig von dieser Vorgabe justiert werden können.

Die Zuordnung von geografischen Positionen erfolgt in der Regel durch geografische Koordinaten. Exakte geografische Positionen in Form von Lägen- und Breitengrad sind aber nicht immer erforderlich. Für einige Anwendungen sind zwar exakte geografische Zuordnungen notwendig, vielfach ist jedoch eine flächige Zuordnung ausreichend. Es soll dazu eine geeignete Einteilung des Globus in geografische Regionen mit unterschiedlichen Hierarchieebenen vorgenommen werden. Für jede Hierarchieebene soll eine separate Evaluierung der Ergebnisse möglich sein.

Twitter wird weltweit genutzt, das Verfahren soll deshalb auch unabhängig von unterschiedlichen Sprachen und Schriftzeichen funktionieren.

1.5 Gliederung der Arbeit

In Kapitel 2 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 des entwickelten Verfahrens benötigt werden. Ebenso werden häufig genutzte geografische Grundbegriffe eingeführt. In Kapitel 3 werden verwandte Arbeiten betrachtet, eingeordnet und in Bezug auf die

angegebenen Anforderungen untersucht.. Kapitel 4 stellt ein Verfahren zur Lösung der Fragestellungen vor. Um einen Überblick zu gewährleisten, wird das Verfahren zunächst allgemein betrachtet, danach wird jeder Verfahrensschritt dargelegt. Das Verfahren läuft in zwei Schritten, einer Lernphase und der eigentlichen Geolokalisierung. Im Kapitel 4, Evaluierung, werden die Ergebnisse der Referenzimplementierung bewertet. Es werden die besten erreichten Ergebnisse dargestellt und interpretiert. Des weiteren werden verschiedene Einstellungen für Schwellwerte zur Justierung des Verfahrens bezüglich der Qualität getestete. In Kapitel 5 wird das Verfahren und die Ergebnisse zusammengefasst. Im Ausblick werden mögliche Verbesserungen und Ideen zur Weiterentwicklung gegeben.

Literatur

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

[2] Swarup Chandra, Latifur Khan, and Fahad Bin Muhaya. Estimating twitter user location using social interactions–a content based approach. *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*, pages 838–843, 2011.

[3] 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.

[4] 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.

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

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

[7] 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.

[8] Judith Gelernter and Nikolai Mushegian. Geo-parsing messages from microtext. *Transactions in GIS*, 15:753–773, 2011.

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

[10] 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.

[11] 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.

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

[13] Yohei Ikawa, M Enoki, and M Tatsubori. Location inference using microblog messages. *Proceedings of the 21st international conference companion on World Wide Web*, pages 687–690, 2012.

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

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

[16] 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.

[17] Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. Ï'M Eating a Sandwich in Glasgow": Modeling Locations with Tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents - SMUC '11*, page 61, 2011.

[18] 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.

[19] 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.

[20] Alan M. MacEachren, Anuj Jaiswal, Anthony C. Robinson, Scott Pezanowski, Alexander Savelyev, Prasenjit Mitra, Xiao Zhang, and Justine Blanford. Senseplace2: Geotwitter analytics support for situational awareness. In *VAST 2011 - IEEE Conference on Visual Analytics Science and Technology 2011, Proceedings*, pages 181–190, 2011.

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

[22] 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.

[23] 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.

[24] 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.

[25] Sharon Myrtle Paradesi. Geotagging tweets using their content. In *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference*, pages 355–356, 2011.

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

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

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

[34] Chong Wang, Jinggang Wang, Xing Xie, and Wei-Ying Ma. Mining geographic knowledge using location aware topic model. In *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval*, GIR '07, pages 65–70, New York, NY, USA, 2007. ACM.

[35] Allison Gyle Woodruff and Christian Plaunt. Gipsy: Automated geographic indexing of text documents. *Journal of the American Society for Information Science*, 45:1–21, 1994.

Abbildungsverzeichnis