EXPOSÉ FÜR EINE DIPLOMARBEIT:
WER SCHREIBT WIKIPEDIA?

SCHREIBEN WIR UNSERE GESCHICHTE SELBST?

DAVID KALTSCHMIDT

betreut durch Dr. Claudia Müller-Birn
Freie Universität Berlin – Fachbereich Informatik

Als Online-Enzyklopädie ist Wikipedia nicht nur Nachschlagewerk sondern auch ein sich stetig wandelndes Geschichtsbuch. Eine global verteilte Nutzerschaft liest und schreibt über lokale Ereignisse noch während sie sich entwickeln. Diese Arbeit soll Möglichkeiten untersuchen, inwiefern man die Herkunft der Autoren bestimmen und damit Einflusssphären auf politische Ereignisse sichtbar machen kann. Vorhandene Analysemethoden und Visualisierungen sollen auf Eignung untersucht, gegebenenfalls weiterentwickelt und als Proof of Concept in eine Software integriert werden.

INHALTSVERZEICHNIS

Als Ende Januar 2011 die Welle des öffentlichen Protestes von Tunesien nach Ägypten überschwappte, rief eine kleine Gruppe von Oppositionsparteien und politischer Aktivisten über die Website Facebook zu einem *Tag des Zornes* auf. Am 25. Januar hatte die Facebook-Gruppe über 80.000 Unterstützer. In den landesweit organisierten Protesten gingen zehntausende auf die Straße. Aufgrund der andauernden Proteste schränkte die Regierung erst den Zugang zu sozialen Netzwerken wie Twitter ein, bevor sie am 28. Januar Ägypten vollständig vom Internet trennte.[1, 2]

Die Nutzung dieser Informationsnetzwerke hatte direkten Einfluss auf die politischen Entwicklungen. Facebook diente zur Planung und Organisation der Proteste, wohingegen Twitter als Informationsmedium während der Proteste eingesetzt wurde. Parallel dazu wurden auf der Online-Enzyklopädie Wikipedia die Ereignisse minutiös festgehalten[3], so dass diese Website als Sammelbecken für Informationen genutzt werden konnte. In der Diplomarbeit soll die Herkunft dieser Informationsbeiträge untersucht werden, um mithilfe der Ergebnisse eine Aussage über die Nutzung von Wikipedia als politisches Werkzeug machen zu können.

Das freie Online-Lexikon, an dem jeder mitschreiben kann, zeichnet sich nicht nur durch eine hohe Qualität aus,[4] sondern erfreut sich auch an stetiger Popularität[5]. Gleichzeitig ist, Dank der von Wikipedia eingesetzten Software MediaWiki[1], der Aufwand sehr niedrig, an einem Artikel mitzuarbeiten. Einen Internetzugang vorausgesetzt, kann jede Person die Entwicklungen von aktuellen Ereignissen im zugehörigen Artikel zeitnah beschreiben und innerhalb von Sekunden publizieren. Für die globale Leserschaft der Wikipedia werden diese Artikel dann zu einer Quelle für Hintergrundinformationen. Im Gegensatz zum globalen Zugang zu Wikipedia, beginnt ein politisches Ereignis häufig lokal. Gebunden an einen bestimmten Ort, ist es ersteinmal auf ein Land begrenzt. Dies spiegelt sich auch in den Artikeln wider: es gibt einen Artikel über die Revolution in Ägypten[2] und einen über die Revolution in Tunesien[3].

In einer Welt, in der Wissen mit Macht gleichgesetzt wird, könnte die Autorschaft einer solchen Online-Referenz von erheblicher strategischer Bedeutung sein. Politische Umbrüche sind sehr empfindliche Prozesse, die leicht durch die gefühlte Einflussnahme von *äußeren* Kräften beeinflusst werden können, zum Beispiel wenn die politische Opposition als Handlanger einer ausländischen Macht diskreditiert wird. Gleichzeitig können

1 http://www.mediawiki.org
2 http://en.wikipedia.org/wiki/Egyptian_Revolution_of_2011
3 http://en.wikipedia.org/wiki/Tunisian_Revolution

durch Beiträge von *innen* die tatsächlichen Zustände ohne Zensur publik gemacht werden. Das kontinuierliche und zeitnahe Mitschreiben an einem Artikel über das entsprechende Ereignis erweitertet das Nachschlagewerk zu einem Nachrichtenmedium, das ständig sowohl korrigiert als auch aktualisiert wird. Die Geschichte eines Landes wird also nicht mehr nur in der Retrospektive von den Gewinnern, d.h. den aktuellen Herrschern, geschrieben, sondern täglich, sogar stündlich, von seinen Bürgern.

Die zentrale Frage der Arbeit ist, inwieweit die kollektive Autorschaft eines gegebenen Wikipedia-Artikels über ein politisches Ereignis das Ereignis selbst widerspiegelt. Kommen zum Beispiel die Verfasser eines Artikels über eine Revolution aus dem Land, das Schauplatz des Umbruches ist? Ändert sich die Verteilung der Herkunft der Beiträge mit der Zeit? Lassen sich innerhalb eines Artikels Kontroversen und deren geographischer Ursprung identifizieren?

Während sich bisherige Studien eher auf das Editierverhalten und Qualitätsmetriken konzentrierten, steht in der Diplomarbeit der geographische Aspekt im Vordergrund. In diesem Rahmen sollen Möglichkeiten untersucht werden, inwieweit der geographische Ursprung der Artikelbeiträge erfasst und aufbereitet werden kann, um etwa Dritte bei einer politischen Analyse eines Artikels zu unterstützen. Eine Reihe von Visualisierungen soll dabei helfen, Aussagen über politische Zusammenhänge ableiten zu können, wie zum Beispiel die Identifikation der Einflussnehmerstaaten oder auch der Streitpunkte.

2 ZIELSTELLUNG

Basierend auf diesen Überlegungen sollen bisherige Analysemethoden und Visualisierungen auf Eignung untersucht, gegebenenfalls weiterentwickelt und als Proof of Concept in eine Software integriert werden. Die Nutzung dieser Software soll für einen gegebenen Artikel eine automatische, quantitative Auswertung durchführen und deren Ergebnisse geeignet darstellen, so dass zum Beispiel folgende Informationen erkennbar werden:

A1 Ursprungsländer der Autoren und deren Anteil am Artikel

A2 Zeitliche Entwicklung der Ursprünge der Autorschaft

A3 Hauptstreitpunkte des Artikels

A4 Vergleich der Ursprünge von dem gleichen Artikel in unterschiedlichen Sprachen, zum Beispiel im englischen und im deutschen Wikipedia

Unter Einsatz der Software soll anhand einer Auswahl von Artikeln über politische Ereignisse eine solche Analyse durchgeführt

werden, um die Kernfrage, ob ein Land seine Geschichte selbst schreibt, beispielhaft zu beantworten. Eine statistische Analyse, die eine allgemeine Aussage für die gesamte Wikipedia erlaubt, schließt die Arbeit ab.

3 THEORETISCHER HINTERGRUND

Die Online-Enzyklopädie Wikipedia gibt es in über 260 Sprachvarianten, von denen die englische mit derzeit 3,6 Millionen Artikeln mit Abstand die größte ist. Die Anzahl der Artikel in den anderen Sprachen sowie die Nutzung der jeweiligen Sprachvariante unterscheiden sich jedoch erheblich.[6] Wenn ein Artikel auch in anderen Sprachen vorhanden ist, sind diese Varianten untereinander verlinkt.

Die Artikel dieser Lexika werden von Freiwilligen auf der ganzen Welt geschrieben, gemeinschaftlich korrigiert und aktualisiert. Jede Änderung eines Artikels erzeugt eine neue Version, die der Versionsgeschichte des Artikels hinzugefügt wird und danach für alle Benutzer einsehbar ist. Jeder Eintrag in der Versionsgeschichte besteht dabei aus der Textänderung, dem Datum der Version, dem Benutzer sowie einem optionalen Kommentar über den Grund der Änderung. Jede Änderung kann mit Hilfe dieser Historie ausführlich begutachtet und bei Missfallen wieder revidiert werden. Dies kann mitunter sogenannte *edit wars* hervorrufen, in denen neue Beiträge von Nutzern mit entgegengesetzten Standpunkten sofort wieder revidiert werden.[7]

Die Mitarbeit an den Artikeln kann mit oder ohne vorherige Registrierung erfolgen. Autoren, die sich registrieren, erlangen sowohl bestimmte Privilegien wie zum Beispiel das Recht, neue Einträge zu erstellen, als auch den Zugang zu Wikipedias sozialem Netzwerk: Jeder Benutzer erhält nach der Registrierung eine *user page* auf der er Informationen über sich veröffentlichen und über die er mit anderen Nutzern Kontakt aufnehmen kann.[8]

Zur Bestimmung der Herkunft eines Autors bietet Wikipedia zwei direkte Ansätze: Für die Beiträge von nicht registrierten Benutzern wird die IP-Adresse gespeichert, über die er Zugang zum Internet erlangt hat. Für einen Großteil der IPs lässt sich daraufhin das Herkunftsland bestimmen. Der zweite Ansatz betrifft die registrierten Benutzer. Ihre IP-Adressen sind maskiert und nicht öffentlich zugänglich.[4] Die registrierten Nutzer können jedoch auf ihrer *user page* Informationen über ihre Person entweder als Freitext oder strukturiert in *user boxes* veröffentlichen. Letztere sind definierte Einheiten mit denen der Nutzer persönliche Eigenschaften wie Herkunftsland, gesprochene Sprachen oder wissenschaftliche Interessen kodifizieren kann. Zusammen decken

4 Eine kleine, von der Wikipedia-Community gewählte Nutzerschaft mit der Berechtigung *checkuser* kann die Adressen demaskieren.

beide Ansätze jedoch nur einen Teil der Beiträge schreibenden Nutzerschaft ab.

Ein zusätzlicher, indirekter Ansatz für die Bestimmung der Herkunft eines Nutzers wird von Lieberman in *You are where you edit: Locating Wikipedia users through edit histories*[9] beschrieben. Er basiert auf der Annahme, dass ein Nutzer mit Vorliebe an Artikeln über Orte in seiner geographischen Nähe mitarbeitet. Diese Artikel sind in der Regel mit geographischen Koordinaten versehen und erlauben so eine sehr grobe Bestimmung des Aufenthaltsortes und dessen Visualisierung auf einer Landkarte.

Eine Analyse der Autorschaft bis auf Satzebene innerhalb eines Artikels wird von Kramer in [10] erforscht. Durch Auswertung der Versionsgeschichte lässt sich zu jedem Satz der Autor bestimmen, der dessen Hauptteil geschrieben hat. Eine automatische Auswertung eines Artikels bis auf Wortebene wird von Adler in [11] vorgestellt. Sie basiert auf dem von Adler selbst entwickelten Reputationssystem [12], das Textstellen eine hohe Vertrauenswürdigkeit zuweist, die von einem vertrauenswürdigen Autor geschrieben oder mindestens einmal bearbeitet worden sind. Basierend auf diesen beiden Arbeiten wurde die Software WikiTrust[5] implementiert, welches die Vertrauenswürdigkeit als weiß-orange *Heatmap* darstellt: zweifelhafte Textstellen werden orange hinterlegt und damit leicht erkennbar.

4 VORGEHENSWEISE

Im Theorieteil der Arbeit sollen diese Ansätze und deren Anwendbarkeit auf die Frage, wer die Geschichte eines Landes schreibt, untersucht werden. Ggfalls wird der aktuelle Forschungsstand daraufhin überprüft, ob es weitere Methoden zur Analyse und Visualisierung gibt und wie diese angepasst beziehungsweise weiterentwickelt und schließlich in eine Software integriert werden können.

4.1 *Theoretischer Teil*

Nach einer Einführung in die Grundlagen und einer Zusammenfassung verwandter Arbeiten, werden Wikipedias Datenstrukturen und die daraus ableitbaren Informationen untersucht, z.B.:

ARTIKEL Ein Artikel hat mindestens einen Autor und ist gegebenenfalls in mehreren Sprachen vorhanden.

VERSIONSGESCHICHTE Diese Historie liefert Informationen wie Benutzername oder IP-Adresse, Datum der Version sowie die inkrementelle Textänderung.

5 http://www.wikitrust.net

USER PAGES & USER BOXES Auf den *user pages* kann ein registrierter Benutzer Informationen über sich veröffentlichen, die Aufschluss über seine Herkunft geben könnten.

EXTERNE QUELLEN Webservices wie WikiTrust bieten Schnittstellen an, die Informationen über Nutzer und deren Beiträge bereithalten.

Daraufhin werden Wege gesucht, die Extraktion der relevanten Daten zu automatisieren und deren Speicherung zur weiteren Verarbeitung zu vereinheitlichen. Im Bezug auf die Herkunft sind sowohl das Land als auch die Geo-Koordinaten interessant. Basierend auf der Versionsgeschichte würde für nicht registrierte Benutzer eine Gewinnung von Daten dann beispielsweise folgende Schritte durchlaufen:

$$\text{IP} \Rightarrow \text{Geo-Location-Service} \Rightarrow \text{Koordinaten} \Rightarrow \text{Land}$$

Für eine Analyse der Artikel bis auf Satzebene werden Algorithmen wie in [10] auf ihre Anwendbarkeit untersucht. Auf Basis der strukturierten Daten in Form von Artikeln, Sätzen, Ländern, Koordinaten und Sprachen sollen nun Visualisierungen gefunden werden, welche die Fülle an Informationen zugänglich machen. Mögliche Visualisierungen wären etwa:

V1 Revisionshistogramm à la Google Finance

V2 Heatmap einer Landkarte mit Ursprüngen der Revisionen

V3 Netzwerkgrafik, die Metriken desselben Artikels in verschiedenen Sprachvarianten anzeigt

V4 Heatmap des Artikels mit Stellen höchster Aktivität

V5 Landeskürzel für eine gegebene Textstelle

4.2 *Praktischer Teil*

Die Methoden zur Datenextraktion und Visualisierung werden anschließend in eine Software integriert. Die Gewinnung der von dieser Anwendung zu verarbeitenden Daten kann aus einer der folgenden Quellen erfolgen:

DB-KOPIE Monatlich angefertigte Moment-Aufnahmen der gesamten Wikipedia-Datenbank sind öffentlich verfügbar[6]. Eine solche Kopie enthält alle Artikel inklusive Versionsgeschichte und ist damit jedoch sehr groß[7].

6 `http://dumps.wikimedia.org`

7 Eine Kopie der englischen Wikipedia-Datenbank umfasst derzeit 5,4 Terabyte.

ARTIKELEXPORT Jeder einzelne oder mehrere Artikel der Wikipedia kann auch separat exportiert werden. Diese Daten umfassen ebenfalls die Versionsgeschichte und sind im Umfang bedeutend kleiner.

TOOLSERVER Die Wikimedia Deutschland e.V. stellt Server bereit,[8] welche einen direkten Zugang zu einer replizierten, schreibgeschützten Wikipedia-Datenbank ermöglichen. Die Nutzung eines solchen Servers vermeidet es zwar, eine eigene komplette Kopie der gesamten Wikipedia-Datenbank halten zu müssen, bedarf jedoch einer Anmeldung.

Für die Entwicklung der Software wird jedoch keine komplette Datenbank-Kopie benötigt. Mithilfe der Export-Funktion von Artikeln lässt sich ein kleiner Datensatz generieren, an dem die Anwendung getestet werden kann. Über dieselbe Export-Funktion kann auch eine Kategorie wie zum Beispiel *Revolutions by country*[9] angegeben werden. Als Ergebnis erhält man eine Sammlung von Artikeln über politische Ereignisse. Mit der fertigen Anwendung können diese analysiert und die Daten entsprechend der gewählten Visualisierung aufbereitet werden.

Neben der beispielhaften Präsentation der Software soll im Schlussteil der Arbeit auch eine statistische Auswertung erfolgen. Der dazu benötigte Datensatz kann ebenfalls über die Export-Funktion durch Angabe von geschichtsbezogenen Kategorien geliefert werden. Die Software sollte diese Daten so verarbeiten können, dass sie statistisch ausgewertet werden können.

4.3 *Auswertung*

Im Schlussteil der Arbeit werden die Visualisierungen der Anwendung beispielhaft für eine Reihe von Artikeln über politische Ereignisse als Analyse-Werkzeug gebraucht. Dabei könnte zum Beispiel sichtbar werden, dass sich ein bestimmter Artikel in verschiedenen Sprachvarianten unterschiedlich entwickelt. Falls ein Land mehrere offizielle Sprachen hat, könnte man diese entweder gruppiert oder einzeln im direkten Vergleich betrachten. Ebenso könnten sich in Anlehnung an die *edit wars* Streitpunkte anhand von Textstellen herauskristallisieren, die besonders umkämpft sind. Anhand eines ausgewählten Datensatzes von politischen Ereignissen soll eine statistische Auswertung erfolgen, um die Frage zu beantworten, wer die Geschichte eines Landes schreibt. Als Ausblick werden Ansätze zur Erweiterung aufgezeigt, etwa eine Mustererkennung, die anhand der Herkunfts- und Streitmuster Vorhersagen über politische Unruhen erlauben könnte.

8 http://toolserver.org
9 http://en.wikipedia.org/wiki/Category:Revolutions_by_country

5 ZEITPLAN

Der ungefähre zeitliche Ablauf lässt sich in folgende Phasen einteilen (Umfang in Wochen, insgesamt 22 Wochen):

2 Recherche, Literatur, Forschungsstand

2 Theoretische Basis

6 Ansätze zur Analyse und Visualisierung

6 Implementierung der Konzepte

2 Auswertung mit Beispielen und Statistik

4 Abschlussphase

[1] The Economist. *Protest in Egypt: Another Arab regime under threat*. URL: http://www.economist.com/node/18013760.

[2] sueddeutsche.de. *Krise in Ägypten - Die Kinder des 6. April und der Tag der Entscheidung - Politik*. URL: http://www.sueddeutsche.de/politik/krise-in-aegypten-die-kinder-des-april-rufen-zum-protest-1.1053426.

[3] Wikipedia. *Revision history of 2011 Egyptian revolution*. URL: http://en.wikipedia.org/w/index.php?title=2011_Egyptian_revolution&dir=prev&action=history.

[4] J. Giles. „Internet encyclopaedias go head to head". In: *Nature* 438.7070 (2005), S. 900–901. ISSN: 0028-0836.

[5] Wikipedia. *Wikipedia Page Views*. URL: http://stats.wikimedia.org/EN/TablesPageViewsMonthly.htm.

[6] Wikipedia. *Statistics*. URL: http://stats.wikimedia.org/EN/Sitemap.htm.

[7] B. Suh u. a. „Us vs. them: Understanding social dynamics in Wikipedia with revert graph visualizations". In: *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*. IEEE. 2007, S. 163–170.

[8] Wikipedia. *Why create an account?* URL: http://en.wikipedia.org/wiki/Wikipedia:Why_create_an_account%3F.

[9] M.D. Lieberman und J. Lin. „You are where you edit: Locating Wikipedia users through edit histories". In: *ICWSM'09* (2009), S. 106–113.

[10] M. Kramer, A. Gregorowicz und B. Iyer. „Wiki trust metrics based on phrasal analysis". In: *Proceedings of the 4th International Symposium on Wikis*. ACM. 2008, S. 1–10.

[11] B.T. Adler u. a. „Assigning trust to wikipedia content". In: *Proceedings of the 4th International Symposium on Wikis*. ACM. 2008, S. 1–12.

[12] B.T. Adler und L. De Alfaro. „A content-driven reputation system for the Wikipedia". In: *Proceedings of the 16th international conference on World Wide Web*. ACM. 2007, S. 261–270.