SCHREIBEN WIR UNSERE GESCHICHTE SELBST?

DAVID KALTSCHMIDT

Diplomarbeit

Betreuer: Dr. Claudia Müller-Birn
2. Prüfer: Prof. Dr. Robert Tolksdorf
Institut für Informatik
Freie Universität Berlin

[6. Juli 2011 at 21:07]

David Kaltschmidt: *Schreiben wir unsere Geschichte selbst?*
Diplomarbeit, © 2011

BETREUER:
Betreuer: Dr. Claudia Müller-Birn
2. Prüfer: Prof. Dr. Robert Tolksdorf

ORT:
Berlin, Germany

JAHR:
2011

ABSTRACT

Short summary of the contents in English...

ZUSAMMENFASSUNG

Kurze Zusammenfassung des Inhaltes in deutscher Sprache, behandle Forschungsfrage, Weg zur Beantwortung und Ergebnis in allgemeinverständlicher Sprache...

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

LISTINGS

ACRONYMS

Teil I

ÜBERLEGUNGEN

1 EINLEITUNG

1.1 MOTIVATION

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]. Dank der von Wikipedia eingesetzten Software MediaWiki[1] ist der Aufwand, an einem Artikel mitzuarbeiten, sehr gering. Einen Internetzugang vorausgesetzt, kann jede Person die Entwicklungen von aktuellen Ereignissen im zugehörigen Artikel zeitnah beschreiben und innerhalb von Sekunden publizieren.

Diese Form der Mitarbeit erweitert das Nachschlagewerk zu einem Nachrichtenmedium, das ständig korrigiert und aktualisiert wird. Das Resultat ist eine einzigartige Quelle des Wissens, in dem sich jedoch die Möglichkeit einer Berichterstattung für jedermann mit der Autorität eines Lexikons vermischt. Eine technologieversierte Öffentlichkeit, die das Internet als effizientes Mittel zur Informationsgewinnung und -verbreitung ansieht, kann Wikipedia zum *fact checking* nutzen und auf dieser Basis handeln.[6, S. 424-427] Die Autorschaft eines solchen Me-

1 http://www.mediawiki.org

diums würde damit unmittelbaren Einfluss auf den politischen Entscheidungsprozess ausüben.

Politische Ereignisse sind häufig auf ein Land oder eine Region begrenzt. Dies spiegelt sich auch in den Artikeln über die Proteste in der arabischen Welt wider: es gibt sowohl einen zusammenfassenden „Mutter-Artikel“[2] als auch einzelne Artikel über die Revolution in Ägypten[3] oder den Aufstand in Libyen[4].

Am libyschen Beispiel ist auch erkennbar, dass solch ein politischer Umbruch ein äußerst empfindlicher Prozess ist. Anfang März 2011 war die Gruppe der Aufständischen klar gespalten in Liberale und Islamisten. Während beide Lager eine Flugverbotszone über Libyen forderten, war sich die Gemeinschaft über einen Einsatz von Bodentruppen uneinig. Durch die Befürwortung eines Bodeneinsatzes liefen die Liberalen Gefahr, sowohl vom Regime als auch von den Islamisten als Handlanger ausländischer Mächte diskreditiert zu werden.[7]

Die kollektive Autorschaft eines Wikipedia-Artikels könnte ähnlich geteilt aussehen und würde damit erste Fragestellungen liefern, deren Analyse am Ende der Diplomarbeit ermöglicht werden soll: Kommen zum Beispiel die Verfasser eines Artikels über eine Revolution aus dem Land, das Schauplatz des Umbruches ist? Werden die Zustände vor Ort tatsächlich von *innen* geschildert? Lassen sich innerhalb eines Artikels Kontroversen und deren geographischer Ursprung identifizieren? Ändert sich die Verteilung der Herkunft der Beiträge mit der Zeit? Wie verändert sich der Artikel nachdem ein Ereignis vorüber ist?

Fragestellung

Geschichte wird von Siegern geschrieben. Ob dieser Aphorismus ausgedient hat, wird die Diplomarbeit nicht beantworten können. Ob die Bürgern eines Landes täglich oder sogar stündlich auf Wikipedia an ihrer Geschichte mitarbeiten, hingegen schon.

1.2 ZIELSETZUNG

- *Ansätze für Lösung des Problems*
- *Warum lösen diese das Problem?*

Während sich bisherige Studien eher auf das Kollaborationsverhalten und Qualitätsmetriken konzentrierten[5], steht in dieser Diplomarbeit der geographische Aspekt im Vordergrund. In diesem Rahmen sollen Möglichkeiten untersucht werden, inwieweit

2 http://en.wikipedia.org/wiki/2010-2011_Middle_East_and_North_Africa_protests
3 http://en.wikipedia.org/wiki/Egyptian_Revolution_of_2011
4 http://en.wikipedia.org/wiki/2011_Libyan_uprising
5 Insbesondere ???

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.

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 Sprachvarianten eines Artikels anhand einfacher Metriken wie Artikellänge, Anzahl der Autoren und Aktivitätslevel (Anzahl der Revisionen in einem festen Zeitintervall).

1.3 GLIEDERUNG DER ARBEIT

Das Kapitel GRUNDLAGEN beginnt mit einer Übersicht über bisherige Ergebnisse in den Gebieten Autorenanalyse, Georeferenzierung und Visualisierung. Entlang dieser Überlegungen sollen bisherige Analysemethoden und Visualisierungen auf Eignung untersucht, gegebenenfalls weiterentwickelt und als Proof of Concept in einer Software umgesetzt werden.

Unter Einsatz der Software wird im Kapitel EXPERIMENTELLES 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 deskriptive, statistische Analyse einer Gruppe von politischen Artikeln schließt die Arbeit ab.

2 GRUNDLAGEN

weave together important concepts for this thesis and split prior research in areas:

- *wikipedia production*
- *contribution*
- *geoanalysis*
- *visualization*

2.1 WIKIPEDIA

this section should cover the basics to understand components of wikipedia

- *Artikel*
- *Sprachen*
- *Versionshistorie*

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.[8] Wenn ein Artikel zum selben Thema in Wikipedias unterschiedlicher Sprachen vorhanden ist, sind diese Varianten in der Regel über sogenannte Interwiki-Links 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.[9]

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.[10]

2.2 AUTORENANALYSE

introduce collective authorship and name some important concepts. prior research in:

- *text-longevity*
- *attribution*

Eine Analyse der Autorschaft bis auf Satzebene innerhalb eines Artikels wird von Kramer in [11] 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 [12] vorgestellt. Sie basiert auf dem von Adler selbst entwickelten Reputationssystem [13], das Textstellen eine hohe Vertrauenswürdigkeit zuweist, die von einem vertrauenswürdigen Autor geschrieben oder mindestens einmal bearbeitet worden sind.[1]

Für eine Analyse der Artikel bis auf Satzebene werden Algorithmen wie in [11] auf ihre Anwendbarkeit untersucht.

2.3 GEOREFERENZIERUNG

explain this intermediate step to assign a location to a contribution

- *pick up where he left off: D. Hardy. „Volunteered geographic information in Wikipedia". Diss. UNIVERSITY OF CALIFORNIA, SANTA BARBARA, 2011*
- *WikiScanner*
- *Erik Zachte's: Wikipedia edits visualized*[2]
- *Indirect approach M.D. Lieberman und J. Lin. „You are where you edit: Locating Wikipedia users through edit histories". In:* ICWSM'09 *(2009), S. 106–113*

1 Basierend auf diesen beiden Arbeiten wurde die Software WikiTrust implementiert, welches die Vertrauenswürdigkeit als weiß-orange *Heatmap* darstellt: zweifelhafte Textstellen werden orange hinterlegt und damit leicht erkennbar. Über ein API ist eine mit Vertrauenspunkten annotierte Version eines Artikels abrufbar: http://www.wikitrust.net/vandalism-api

2 http://infodisiac.com/blog/2011/05/wikipedia-edits-visualized/

Zur Bestimmung der Herkunft eines Autors bietet Wikipedia zwei direkte Ansätze: Für jeden Beitrag eines nicht registrierten Benutzers wird die IP-Adresse gespeichert, über die er Zugang zum Internet erlangt hat. Die registrierten Nutzer können jedoch auf ihrer *user page* Informationen über ihre Person entweder als Freitext oder strukturiert in *user boxes* veröffentlichen.

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*[15] 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.

2.4 VISUALISIERUNG

Write about prior works of visualizing the aspects of attribution and georeference

- *Erik Zachte's: Wikipedia edits visualized*[3]
- *Wikitrust 1*

3 http://infodisiac.com/blog/2011/05/wikipedia-edits-visualized/

Teil II

EXPERIMENTELLES

3 APPARATUS

Welche Instrumente stehen zur Verfügung und wie können diese weiterentwickelt werden?

3.1 WIKIPEDIAS DATENSTRUKTUREN

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.

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 Im Internet existieren zahlreiche Dienste, die Schnittstellen anbieten, um Informationen über Nutzer und deren Beiträge zu erhalten, z.B.: WikiTrust[1] oder WikiWatcher[1]

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[2]. Eine solche Kopie enthält alle Artikel inklusive Versionsgeschichte und ist damit jedoch sehr groß[3].

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.

1 Das WikiWatcher-Teilprojekt *Poor Man's Check User* erlaubt eine Auflösung des Benutzernamens in eine IP-Adresse, wenn dieser Nutzer in der Vergangenheit beim Ändern eines Artikels das Session-Limit überschritten hatte. Inzwischen wurde diese Sicherheitslücke in der WikiMedia-Software jedoch behoben. http://wikiwatcher.virgil.gr/pmcu

2 http://dumps.wikimedia.org

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

TOOLSERVER Die Wikimedia Deutschland e.V. stellt Server bereit,[4] 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.

3.2 AUTORENSCHAFT

Introduce types of authors (roles) as well as methods to determine contribution/attribution

- *Autoren*
- *Bots*
- *Wer überlebt?*
- *Algorithmen, welche Unterschiede?*

3.3 GEOREFERENZIERUNG

- *registered vs. unregistered vs. bots vs. admins*
- *IPs of unregistered users: Geo lookup*
- *Autoren-Profile: Information Extraction*
- *Geographische Zuordnung vom user profile*

Zur Bestimmung der Herkunft eines Autors bietet Wikipedia zwei direkte Ansätze: Für jeden Beitrag eines nicht registrierten Benutzers wird die IP-Adresse gespeichert, über die er Zugang zum Internet erlangt hat. Der zweite Ansatz betrifft die registrierten Benutzer. Ihre IP-Adressen sind maskiert und nicht öffentlich zugänglich.[5] 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 beide Ansätze jedoch nur einen Teil der Beiträge schreibenden Nutzerschaft ab.

4 http://toolserver.org

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

3.3.1 *IP lookup*

- *Dienste*
- *Genauigkeit*

Mit frei verfügbaren[6] Online-Diensten wie *Quova*[7] oder *geoplugin*[8] lässt sich für einen Großteil der IPs daraufhin das Herkunftsland bestimmen.

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:

IP ⇒Geolocation-Dienst ⇒Koordinaten und Land

3.3.2 *Information Extraction*

- *IE approach with Machine Learning L. Xiao u. a. „Information extraction from the web: System and techniques". In:* Applied Intelligence *21.2 (2004), S. 195–224*
- *unsupervised IE: O. Etzioni u. a. „Unsupervised named-entity extraction from the web: An experimental study". In:* Artificial Intelligence *165.1 (2005), S. 91–134*

3.4 VISUALISIERUNG

- *Darstellung der geographischen Analyse*
- *per Wort, Satz, Artikel, Wort*

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 Dynamisches Blasendiagramm[9] über die Entwicklung unterschiedlicher Sprachvarianten

6 Die vorgestellten Dienste haben ein tägliches Kontingent an Anfragen. Hilfstechniken wie Caching können diese Einschränkungen jedoch mindern.

7 http://developer.quova.com

8 http://www.geoplugin.com/webservices

9 http://en.wikipedia.org/wiki/Motion_chart

V5 *Heatmap* des Artikels mit Stellen höchster Aktivität

V6 Landeskürzel für eine gegebene Textstelle

4 EXPERIMENTELLES

4.1 DATENSATZ

- *Wahl einer Kategorie/Artikel*
- *Wieso repräsentativ für politische Ereignisse?*

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*[1] angegeben werden. Als Ergebnis erhält man eine Sammlung von Artikeln über politische Ereignisse.

4.2 DURCHFÜHRUNG

- *Beispielhafte Durchführung*
- *Sammlung der Ergebnisse*

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.

1 http://en.wikipedia.org/wiki/Category:Revolutions_by_country

Teil III

ERGEBNISSE

5 ERGEBNISSE

- *Statistische Auswertung*

Anhand eines ausgewählten Datensatzes von politischen Ereignissen wie *Revolutions by country* soll eine statistische Auswertung erfolgen, um die Frage zu beantworten, wer die Geschichte eines Landes schreibt.

6

SCHLUSSFOLGERUNGEN

- *Interpretation der Ergebnisse*
- *Vermutungen bestätigt*

6.1 AUSBLICK

Teil IV

APPENDIX

LITERATUR

[1] The Economist. *Protest in Egypt: Another Arab regime under threat*. 2011. URL: http://www.economist.com/node/18013760.

[2] sueddeutsche.de. *Krise in Ägypten - Die Kinder des 6. April und der Tag der Entscheidung - Politik*. 2011. 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] A. Chadwick. *Routledge handbook of Internet politics*. Taylor & Francis, 2009. ISBN: 0203962540.

[7] The Economist. *Libya: A civil war beckons*. 2011. URL: http://www.economist.com/node/18290470.

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

[9] 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.

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

[11] 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.

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

[13] 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.

[14] D. Hardy. „Volunteered geographic information in Wikipedia“. Diss. UNIVERSITY OF CALIFORNIA, SANTA BARBARA, 2011.

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

[16] L. Xiao u. a. „Information extraction from the web: System and techniques“. In: *Applied Intelligence* 21.2 (2004), S. 195–224.

[17] O. Etzioni u. a. „Unsupervised named-entity extraction from the web: An experimental study“. In: *Artificial Intelligence* 165.1 (2005), S. 91–134.