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 änderndes Geschichtsbuch. Eine global verteilte Nutzerschaft liest und schreibt über lokale Ereignisse 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. Eine Web-Anwendung soll diese Analyse für einen beliebigen Wikipedia-Artikel ermöglichen.

INHALTSVERZEICHNIS

1 MOTIVATION

Die Rolle von Twitter und Facebook als demokratisierende Instrumente in den politischen Revolutionen der arabischen Welt zu Beginn des Jahres 2011 ist umstritten. Die Nutzung dieser Informationsnetzwerke hatte direkten Einfluss auf die politischen Entwicklungen. Facebook diente in der Regel zur Planung und Organisation der Proteste, wohingegen Twitter als Informationsmedium während der Proteste eingesetzt wurde. Die Nutzung von Wikipedia als politisches Werkzeug ist weniger verstanden und soll in dieser Arbeit untersucht werden.

Die Online-Enzyklopädie, an der jeder mitschreiben kann, genießt einen hohen Stand sowohl in der Qualität als auch der Aktualität der Informationen. 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 einer *äußeren* Macht zerstört werden können, zum Beispiel wenn die politische Opposition als Marionette Amerikas gebrandmarkt wird. Gleichzeitig können durch Beiträge von *innen* die tatsächlichen Zustände ohne Zensur publik gemacht werden. 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 die Verfasser des Artikels aus dem Land das Schauplatz des Umbruches ist? Oder kommen die Beiträge aus den angrenzenden Staaten oder einer anderen Macht mit eigenen politische Zielen? Was sind die Hauptstreitpunkte die zu dem Ereignis führten?

Im Rahmen dieser Arbeit sollen Möglichkeiten untersucht werden, die Artikel der Wikipedien aller Sprachen zu analysieren, um Aussagen über mögliche politische Zusammenhänge treffen zu können. Eine Reihe von Visualisierungen soll dabei helfen, sowohl die Einflussnehmer als auch die Streitpunkte identifizieren zu können.

2 ZIELSTELLUNG

Basierend auf diesen Überlegungen dieser Arbeit wird eine Web-Anwendung erstellt. Diese soll einen gegebenen Artikel automatisch auswerten und die Ergebnisse der Analyse geeignet darstellen, so dass zum Beispiel folgende Informationen erkennbar werden:

A1 Aktivitätslevel der Änderungen des Artikels relativ zur Vergangenheit

A2 Ursprungsländer der Autoren und deren Anteil am Artikel

 Hauptstreitpunkte des Artikels

Bisherige Analysetechniken und Visualisierungen sollen auf Eignung untersucht und gegebenenfalls weiterentwickelt werden.

3 THEORETISCHER HINTERGRUND

- *Nachweise einfügen*
- *zentrale Grundannahmen komplett?*

Laut einer Studie hat ein Großteil der Artikel der Wikipedia eine hohe Qualität. Der Einsatz der Wikimedia-Software hält den Aufwand, an einem Artikel mitzuarbeiten, sehr niedrig. Einen Internetzugang vorausgesetzt, kann jede beliebige Person die Entwicklungen von aktuellen Ereignissen im zugehörigen Artikel zeitnah beschreiben und in Sekunden publizieren. Diese Aktualität ermöglicht es den Bürgern in Ländern, in denen ein Großteil der Medien durch die Regierung kontrolliert wird, Wikipedia als eine alternative Quelle für aktuelle Nachrichten zu nutzen. Für die globale Leserschaft der Wikipedia werden diese Artikel dann zu einer Quelle für Hintergrundinformationen.

Im Gegensatz zum globalen Zugang zu Wikipedia ist ein politisches Ereignis immer lokal. Gebunden an einen bestimmten Ort, ist es häufig auf ein Land begrenzt. Dies spiegelt sich auch in den Artikeln wider: es gibt einen Artikel für die Revolution in Ägypten und einen für die Revolution in Tunesien.

Die politischen Ereignisse werden natürlicherweise von Menschen vorangetrieben, die auch aus einem bestimmten Land kommen. Wikipedia bietet zur Bestimmung der Herkunft eines Autors zwei direkte Ansätze: Für die Beiträge von nicht registrierten Benutzer wird die IP des gespeichert, über die er Zugang zum Internet erlangt. Für einen Großteil der IPs lässt sich daraufhin das Herkunftsland bestimmen. Der zweite Ansatz betrifft die registrierten Benutzer. Ihre IPs sind nicht öffentlich zugänglich. Jedoch erhält jeder Benutzer nach der Registrierung eine *user page* auf der er sich  ofilieren kann. Einige Nutzer haben dort ihr Ursprungsland angegeben. Beide Ansätze zusammen decken jedoch nur einen Teil der Beiträge schreibenden Nutzerschaft ab.

Die englische Wikipedia ist mit derzeit 3,6 Millionen Artikeln mit Abstand die größte.[1] Daneben existieren Wikipedien in mehr als 260 anderen Sprachen, in denen die Artikel desselben Themas untereinander verlinkt sind. Es ist vorstellbar, dass sich

ein bestimmter Artikel in den unterschiedlich Sprachen unterschiedlich entwickelt. Falls ein Land mehrere offizielle Sprachen hat, ließen diese sich entweder gruppiert oder einzeln im direkten Vergleich betrachten.

Auf den Seiten der Wikipedia werden regelmäßig sogenannte *edit wars* ausgetragen.[2] Dabei werden neue Beiträge von Nutzern mit entgegengesetztem Standpunkt sofort wieder revidiert. Bei politischen Umbrüchen gibt es in der Regel ein ähnliches Muster[1] khack. Die Annahme ist, dass in den Artikeln der Wikipedia bestimmte Textstellen besonders "umkämpft" sind. Ungeachtet der Brisanz solcher Stellen findet in der gesamten Arbeit keine inhaltliche Analyse statt, da für einen gegebenen Artikel alle Auswertungen automatisiert erfolgen sollen.

Im Rahmen dieser Arbeit wird die Analyse der Herkunft und der Vergleich desgleichen Artikels mit unterschiedlichen Sprachen kombiniert. Zusätzlich soll bis auf die Ebene einer Textstelle, zum Beispiel eines Absatzes mit hoher Änderungsfrequenz, nachvollziehbar sein, woher der Beitrag stammt. Im Theorieteil der Arbeit sollen diese Ansätze deren Anwendbarkeit auf die Frage, wer die Geschichte eines Landes schreibt, untersucht werden. Ebenfalls wird der aktuelle Forschungsstand daraufhin überprüft, ob es ähnliche Methoden zur Analyse und Visualisierung gibt und wie diese angepasst beziehungsweise weiterentwickelt werden können.

4 VORGEHENSWEISE

Die Überlegungen aus dem theoretischen Teil werden in einer Web-Anwendung praktisch umgesetzt und deren Verwendung an einigen Beispielen am Schluss der Arbeit erläutert.

4.1 *Theoretischer Teil*

Zu Beginn werden Wikipedias (Meta-)Datenstrukturen und die daraus ableitbaren Informationen untersucht, z.B.:

D1 Revisionshistorie

D2 Geographischer Ursprung der Revisionen eines Artikels

D3 Vergleich der Aktivitäten desselben Artikels in Wikipedien anderer Sprachen

Die Revisionshistorie (D1) hilft dabei sowohl das Aktivitätslevel (A1) als auch die Stellen der höchsten Aktivität (A3) zu identifizieren.

Auf Basis der Daten sollen nun Visualisierungen gefunden werden, die die Fülle an Informationen so zugänglich machen, dass

sich die oben genannten Aufgaben erledigen lassen. Mögliche Visualisierungen wären etwa:

V1 Revisionshistogramm à la Google Finance

V2 Heatmap einer Landkarte mit Ursprüngen der Revisionen

V3 Netzwerkgrafik, die Aktivitäten desselben Artikels in verschiedenen Wikipedias anzeigt

V4 Heatmap des Artikels mit Stellen höchster Aktivität

V5 Landeskürzel für eine gegebene Textstelle

- *Helfen die Ansätze, das Problem zu lösen?*
- *Was ist machbar?*

4.2 *Praktischer Teil*

Es wird eine Web-Anwendung erstellt, mit deren Hilfe man die oben genannten Aufgaben erledigen kann. Die Anwendung fragt dazu die nötigen Daten bei Wikipedia ab, analysiert sie und bereitet sie entsprechend der gewählten Visualisierung auf.

Die Art der Datenquelle, die automatischen Analysemethoden sowie die geeigneten Visualisierungen werden aus dem theoretischen Teil übernommen. Die fertige Anwendung wird dann im letzten Teil der Arbeit an einigen Beispielen erläutert.

4.3 *Auswertung*

Die aktuellen politischen Umbrüche in der arabischen Welt genießen ein nie da gewesenes Potential an digitaler Integration. In einer Fallstudie wird die Applikation mit seinen Visualisierungen für eine Reihe von Artikeln über politische Ereignisse als Analyse-Werkzeug benutzt. Anhand einer Liste von politischen Ereignissen soll eine quantitative Auswertung erfolgen, um die Frage zu beantworten, wer die Geschichte eines Landes schreibt. Am Ende der Arbeit werden Ansätze zur Erweiterung aufgezeigt, etwa eine Mustererkennung, die anhand der Herkunfts- und Streitmuster Vorhersagen über politische Unruhen erlauben könnte.

5 ZEITPLAN

Erstellung eines Zeitplans mit Meilensteinen über folgende Phasen:

- *Recherche, Literatur*

- *Theoretische Basis*
- *Ansätze zur Analyse*
- *Ansätze zur Visualisierung*
- *Implementierung der Konzepte*
- *Auswertung der Anwendung/Beispiele*
- *Abschlussphase*

[1] *Statistics - Wikipedia, the free encyclopedia*. URL: http://en.wikipedia.org/wiki/Special:Statistics.

[2] Bongwon 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* (2007), S. 163–170. DOI: 10.1109/VAST.2007.4389010.