WHERE IS HISTORY BEING WRITTEN? GEOREFERENCING CONTRIBUTIONS TO WIKIPEDIA

DAVID KALTSCHMIDT

Diplomarbeit

Dr. Claudia Müller-Birn
Prof. Dr. Robert Tolksdorf
Institut für Informatik
Freie Universität Berlin

[August 8, 2011 at 17:57]

David Kaltschmidt: *Where is history being written? Georeferencing contributions to Wikipedia*
Diplomarbeit, © 2011

SUPERVISORS:
Dr. Claudia Müller-Birn
Prof. Dr. Robert Tolksdorf

LOCATION:
Berlin, Germany

YEAR:
2011

ABSTRACT

Wikipedia is more than an online encyclopedia. It is also a news channel as well as a self-updating history book. A global readership can follow political events as they unfold, written about by local people and later edited by other volunteers. This thesis describes a method to answer the question to what extent local volunteers write about events in their own country. First, the geographic origin of each individual article contribution is determined. In a second step, a given article is annotated with georeferences on a word level. The properties of these annotations then allow for a statistical geographic analysis of a single article or a category of articles.

ZUSAMMENFASSUNG

Translate english abstract, make sure to cover research question and the method in basic terms...

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 passieren. 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 einer Software umgesetzt werden.

CONTENTS

I THOUGHTS 1

1 INTRODUCTION 2
- 1.1 Context 2
- 1.2 Research Questions 4
- 1.3 Structure 4

2 FOUNDATION 6
- 2.1 Wikipedia 6
- 2.2 Contributions 7
- 2.3 Georeferences 7
- 2.4 Visualization 8

II EXPERIMENTS 9

3 APPARATUS 10
- 3.1 Wikipedia's Data Structures 10
  - 3.1.1 Zugriff 11
- 3.2 Collective Authorship 11
  - 3.2.1 Relevant Edits 11
- 3.3 Georeferences 11
  - 3.3.1 IP Look-up 12
  - 3.3.2 Information Extraction 12
  - 3.3.3 Geographic Profiling 13
  - 3.3.4 Consolidation 13
- 3.4 Visualization 13
  - 3.4.1 Goals 14
  - 3.4.2 Design 14
- 3.5 Data Model and System Overview 14

4 EXPERIMENTS 15
- 4.1 Data Set 15
- 4.2 Application 15

III RESULTS 16

5 ERGEBNISSE 17

6 SCHLUSSFOLGERUNGEN 18
- 6.1 Probleme 18
- 6.2 Ausblick 18

IV APPENDIX 19

Bibliography 20

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

Part I

THOUGHTS

1 INTRODUCTION

1.1 CONTEXT

Start with the arab spring as a thread that will provide examples throughout the thesis

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]

how do these events reflect on social media and wikipedia, see Econ. 'Back to the coffee house'

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.

why is wikipedia being used for this, what makes it special, why is it so easy to publish sth.

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.

why should we care? wikipedia acts as a news channel while the aura of an encyclopedia lends it credibility and authority

1 http://www.mediawiki.org

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 Mediums würde damit unmittelbaren Einfluss auf den politischen Entscheidungsprozess ausüben.

news are about events, events have a place, wikipedia articles cover events at a certain place

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].

political events have different factions, express different views (despite NPOV), foreign meddling

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]

collaborative authorship will reflect this division, where do reports originate, iterations of edits turn it into an historical account, where do the editors post from

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

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

Why is the origin important? A place is linked to meaning, context etc.

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 RESEARCH QUESTIONS

Focus on origin of contributions, how well it can be determined

Im Rahmen dieser Diplomarbeit 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.

can distribution of origins be related to place of event

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

can a statistical analysis be done to answer the main question

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).

scope: only georeferencing, not behavior

1.3 STRUCTURE

name the basic chapters and their function, one part = one paragraph

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

thesis describes a method to help answer the research question

Unter Einsatz der Software wird im Kapitel EXPERIMENTS 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.

results and conclusion

2

FOUNDATION

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

- *F.Å. Nielsen. "Wikipedia research and tools: Review and comments." In: (2011)*
- *why wikipedia?*
- *wikipedia production*
- *contribution/attribution*
- *georeferencing*
- *visualization*

2.1 WIKIPEDIA

this section should cover the basics to understand components of wikipedia

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

language editions, chart

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.

article, draw nice graphic of UI

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.[10]

revision history, changesets

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.[11]

2.2 CONTRIBUTIONS

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 [12] 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 [13] vorgestellt. Sie basiert auf dem von Adler selbst entwickelten Reputationssystem [14], 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 [12] auf ihre Anwendbarkeit untersucht.

2.3 GEOREFERENCES

explain this intermediate step to assign a location to a contribution

- *pick up where he left off: D. Hardy. "Volunteered geographic information in Wikipedia." PhD thesis. UNIVERSITY OF CALIFORNIA, SANTA BARBARA, 2011*
  He's more about the how, I'm about the where. Key findings:
- *"I find that as a group, anonymous contributors write about fewer places than registered contributors, despite outnumbering them five-to-one." [15]*
- *"I find that anonymous contributors are more likely to write about nearby places, and that the geographic effects fit an exponential distance decay function." [15]*

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`

- *"Combined approaches (i.e., where quantitative spatial analysis models are calibrated with surveyed locations) may prove useful." [15, p. 85]*
- *WikiScanner*
- *Erik Zachte's: Wikipedia edits visualized*[2]
- *Indirect approach M.D. Lieberman and J. Lin. "You are where you edit: Locating Wikipedia users through edit histories." In: ICWSM'09 (2009), pp. 106–113*

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*[16] 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 VISUALIZATION

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

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

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

Part II

EXPERIMENTS

3

APPARATUS

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

3.1 WIKIPEDIA'S DATA STRUCTURES

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.

3.1.1 *Zugriff*

Tools and servers to access the articles.

3.2 COLLECTIVE AUTHORSHIP

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

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

3.2.1 *Relevant Edits*

Are all edits relevant? Edit wars? Bots?

3.3 GEOREFERENCES

- *registered vs. unregistered vs. bots vs. admins*
- *incorporate key findings of [15] as laid out in chapter 2.3*
- *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

4 http://toolserver.org

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

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.

3.3.1 *IP Look-up*

- *Services*
- *Accuracy*
- *Active prevention by proxies and anonymizers:*
  J.A. Muir and PC van Oorschot. Internet geolocation and evasion. *Tech. rep. Citeseer, 2006*
  J.A. Muir and P.C.V. Oorschot. "Internet geolocation: Evasion and counterevasion." In: ACM Computing Surveys (CSUR) *42.1 (2009), p. 4*
  M. Duckham and L. Kulik. "A formal model of obfuscation and negotiation for location privacy." In: Pervasive Computing *(2005), pp. 152–170*

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 et al. "Information extraction from the web: System and techniques." In:* Applied Intelligence *21.2 (2004), pp. 195–224*
- *unsupervised IE: O. Etzioni et al. "Unsupervised named-entity extraction from the web: An experimental study." In:* Artificial Intelligence *165.1 (2005), pp. 91–134*
- *if city is mentioned, determine country (needs disambiguation, e.g. Berlin)*

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

- *coordinates are optional?*

3.3.3 *Geographic Profiling*

- *M.D. Lieberman and J. Lin. "You are where you edit: Locating Wikipedia users through edit histories." In:* ICWSM'09 *(2009), pp. 106–113*
- *B.J. Hecht and D. Gergle. "On the localness of user-generated content." In:* Proceedings of the 2010 ACM conference on Computer supported cooperative work. *ACM. 2010, pp. 229–232*
- *from other fields such as criminal research:*
  B. Snook et al. "On the complexity and accuracy of geographic profiling strategies." In: Journal of Quantitative Criminology *21.1 (2005), pp. 1–26*
- *feasibility, maybe just as enhancer*

3.3.4 *Consolidation*

- *settle for a resolution*
- *some examples on accuracy for different countries*
- *clustering of origins: areas of influence*

3.4 VISUALIZATION

- *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

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

V5 *Heatmap* des Artikels mit Stellen höchster Aktivität

V6 Landeskürzel für eine gegebene Textstelle

V7 Edit wars on map, linking two or more places

3.4.1 *Goals*

- *A. Kjellin et al. "Evaluating 2D and 3D visualizations of spatiotemporal information." In:* ACM Transactions on Applied Perception (TAP) *7.3 (2010), pp. 1–23*
- *Identify. Characteristics of an object.*
- *Locate. Absolute or relative position.*
- *Distinguish. Recognize as the same or different.*
- *Categorize. Classify according to some property (e.g., color, position, or shape).*
- *Cluster. Group same or related objects together.*
- *Distribution. Describe the overall pattern.*
- *Rank. Order objects of like types.*
- *Compare. Evaluate different objects with each other.*
- *Associate. Join in a relationship.*
- *Correlate. A direct connection.*

3.4.2 *Design*

3.5 DATA MODEL AND SYSTEM OVERVIEW

- *fetch article*
- *get revision history*
- *determine contributions*
- *transform to word attribution*
- *attach georeference*

4 EXPERIMENTS

4.1 DATA SET

- *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 APPLICATION

- *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

Part III

RESULTS

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 PROBLEME

- *Mobile contributions, smartphones*
- *Privacy*

6.2 AUSBLICK

Part IV

APPENDIX

BIBLIOGRAPHY

[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), pp. 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] F.Å. Nielsen. "Wikipedia research and tools: Review and comments." In: (2011).

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

[10] B. Suh et al. "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, pp. 163–170.

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

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

[13] B.T. Adler et al. "Assigning trust to wikipedia content." In: *Proceedings of the 4th International Symposium on Wikis*. ACM. 2008, pp. 1–12.

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

[15] D. Hardy. "Volunteered geographic information in Wikipedia." PhD thesis. UNIVERSITY OF CALIFORNIA, SANTA BARBARA, 2011.

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

[17] J.A. Muir and PC van Oorschot. *Internet geolocation and evasion*. Tech. rep. Citeseer, 2006.

[18] J.A. Muir and P.C.V. Oorschot. "Internet geolocation: Evasion and counterevasion." In: *ACM Computing Surveys (CSUR)* 42.1 (2009), p. 4.

[19] M. Duckham and L. Kulik. "A formal model of obfuscation and negotiation for location privacy." In: *Pervasive Computing* (2005), pp. 152–170.

[20] L. Xiao et al. "Information extraction from the web: System and techniques." In: *Applied Intelligence* 21.2 (2004), pp. 195–224.

[21] O. Etzioni et al. "Unsupervised named-entity extraction from the web: An experimental study." In: *Artificial Intelligence* 165.1 (2005), pp. 91–134.

[22] B.J. Hecht and D. Gergle. "On the localness of user-generated content." In: *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM. 2010, pp. 229–232.

[23] B. Snook et al. "On the complexity and accuracy of geographic profiling strategies." In: *Journal of Quantitative Criminology* 21.1 (2005), pp. 1–26.

[24] A. Kjellin et al. "Evaluating 2D and 3D visualizations of spatiotemporal information." In: *ACM Transactions on Applied Perception (TAP)* 7.3 (2010), pp. 1–23.