SVEN TERESNIAK · GERHARD HEYER · GERIK SCHEUERMANN · FLORIAN HOLZ

Visualisierung von Bedeutungsverschiebungen in großen diachronen Dokumentkollektionen

Große diachrone Dokumentkollektionen sind eine der wichtigsten Informationsquellen in Politik, Wirtschaft, Wissenschaft und so ziemlich allen Bereichen des öffentlichen Lebens. Eine wichtige wissenschaftliche Problemstellung bei der Nutzung dieser Ressourcen stellt dabei die geeignete Extraktion der in den Dokumenten behandelten Themen dar. Eine geeignete Visualisierung kann dabei helfen, größere Datenmengen – und damit eine große Anzahl behandelter Themen – für einen Benutzer handhabbar zu halten, und kann darüber hinaus die Grundlage für eine tiefere, interaktive und visuell unterstützte Analyse der Daten bilden. Nachfolgend soll ein neues Maß vorgestellt werden, welches die Bedeutungsveränderung von Termen über die Zeit misst. In diesem Artikel werden dabei der aktuelle Arbeitsstand und die ersten Ergebnisse des Teilprojektes »Topology-based Visual Analysis of Information Spaces« dokumentiert und ein Ausblick auf weitere Arbeiten im Sinne des Visual Analytics gegeben.

1 Einleitung

Große Kollektionen von Textdokumenten, wie sie beispielsweise mit dem *New York Times Annotated Corpus*[1] (NYTC) und anderen archivierten Zeitungskorpora vorliegen, enthalten vielfältige temporale Informationen, die sich auf *Ereignisse*, *Geschichten* oder *Themen* (*events*, *stories* bzw. *topics*) beziehen. Wir verwenden diese Begriffe analog zu [Allan 2002] wie folgt: *Ereignis*: Ein berichtetes Auftreten in Raum und Zeit inklusive aller unvermeidbaren Konsequenzen, bspw. bestimmte Wahlen, Unfälle, Naturkatastrophen, etc.; *Geschichte*: Ein thematisch kohäsiver Textausschnitt, welcher Daten mit mindestens zwei unabhängige, inhaltlich nicht identische Aussagen über bestimmtes Ereignis enthält; *Thema*: Ein Ursprungsereignis oder -aktivität inklusive aller abgeleiteten und direkt darauf bezogenen Fakten, Ereignisse und Aktivitäten.

Als Vorstufe einer visuellen Inhaltsanalyse ist zuerst einmal eine geeignete visuelle Repräsentation der in den Daten enthaltenen Themen zu finden. Diese Themen aufzuspüren, ihren Verlauf und ihre Entwicklung zu verfolgen, sind Ziele des *topic detection and tracking* [Allan et al. 1998, Allan 2002]. Dabei sollen innerhalb einer Dokumentkollektion relevante (wichtige, aussagekräftige) Terme identifiziert und zu bestimmten Zeiträumen oder Events in Beziehung gesetzt werden. In der Gegenrichtung sollen inhaltlich zusammenhängende Zeiträume oder Ereignisse möglichst passend verschlagwortet werden.

[1] `http://www.ldc.upenn.edu/`

Um o. g. relevante und/oder neue Terme in Textdatenströmen zu identifizieren, gibt es vielfältige Ansätze, von denen nachfolgend exemplarisch drei vorgestellt werden sollen, um das Spektrum der Verfahren in aller Kürze zu beleuchten. [Swan & Allan 1999, Swan & Allan 2000, Kumaran & Allan 2004] bewerten die Relevanz von Termen anhand sog. *multiple document models* und Schwellwerten auf tf-idf-Basis[2] über Segmenten des Textdatenstroms.

Kleinberg entwickelt in [Kleinberg 2002] den Begriff der *burstiness* von Termen, indem er einen kantengewichteten endlichen Automaten nutzt, um die Sensitivität für Häufungen relevanter Terme besser steuern zu können, welche erfahrungsgemäß in ihrer Auftretensfrequenz auch außerhalb interessanter Zeiträume Schwankungen unterliegen.

In [Wang & McCallum 2006] werden Kookkurrenzmuster von Termen und deren lokale Verteilungscharakteristik über die Zeit verwendet, um Themen über die Zeit zu identifizieren. Dieser Ansatz nutzt die Menge von Kookkurrenten eines Terms als Repräsentation des Themas bezüglich eines Zeitfensters.

Allgemeiner betrachtet müssen Themen nicht zwangsläufig nur bestimmte Events behandeln, sie können stattdessen auch mehr die Sicht des Autors oder der Gesellschaft auf eben diese Events wiederspiegeln. Diese Sichtweise oder Einstellung bezüglich bestimmter Ereignisse sich über die Zeit verändern. In der natürlichen Sprache besteht ein nicht unwesentlicher Anteil aus derartigen (subjektiven wie objektiven) Bewertungen und Einschätzungen, welche über die Zeit Veränderungen unterworfen sein können. Eine Beobachtung dieser Verschiebungen in der öffentlichen Wahrnehmung von Dingen kann sehr hilfreich sein, wenn man historische Entwicklungen bestimmen möchte. In unserem Ansatz betrachten wir neben der Termfrequenz den globalen Kontext der Terme als zusätzliche Dimension, um die Relevanz und die Bedeutung von Termen zu bestimmen. Nähere Details werden in Abschnitt 2 gegeben. Wir nutzen die beiden angesprochenen Dimensionen, um einen Überblick über die Themen in der Kollektion zu ermitteln.

Eine Veränderung des Kontextes eines Terms über die Zeit reflektiert damit eine sich verändernde Benutzung des Terms und deutet auf eine Bedeutungsveränderung dieses Terms hin. Die Stärke der Veränderung (oder die Stabilität) dieser Bedeutungsverschiebung eines Terms beschreibt demnach direkt, wie stark (oder schwach) die typischen Nutzer dieses Terms dessen Bedeutung zustimmen. Als Analogie stelle der Leser sich den börsenbasierten Wertpapierhandel vor, bei dem die Schwankungen des

[2] Tf-idf ist eine Standardmethode der Termgewichtung im Information Retrieval und wird aus dem Produkt von Termfrequenz *tf* und Inverser Dokumentfrequenz *idf* berechnet. Die Termfrequenz wertet die Wichtigkeit eines Terms im Dokument, während die Inverse Dokumentfrequenz den Term bezüglich der kompletten Dokumentkollektion wertet. Damit bevorzugt tf-idf innerhalb eines Dokuments häufige Terme, die in nur wenigen anderen Dokumenten häufig auftreten.

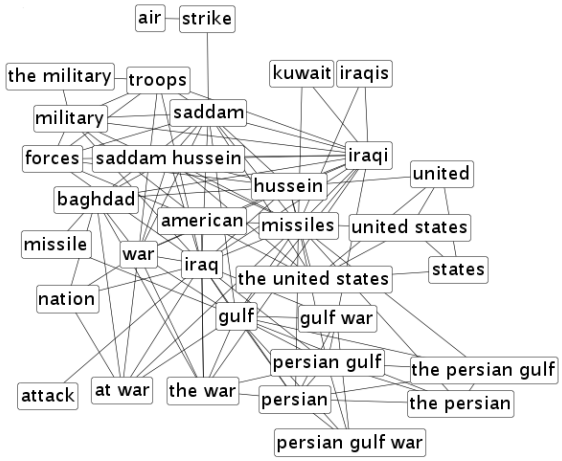
Tab. 1: Die 30 signifikantesten Kookkurrenzen mit Signifikanzwerten (multipliziert mit 10^6 und auf 3 Stellen abgeschnitten) im globalen Kontext von »abu ghraib« am 10. Mai 2004.

prisoners 0,346, abuse 0,346, secretary 0,259, abu ghraib prison 0,259, iraqi 0,247, rumsfeld 0,221, military 0,218, prison 0,218, bush 0,210, prisoner 0,200, photographs 0,183, donald 0,183, secretary of defense 0,174, prisons 0,174, photos 0,174, the scandal 0,174, interrogation 0,163, naked 0,163, mistreatment 0,163, under 0,162, soldier 0,154, saddam 0,154, armed 0,154, defense 0,143, the bush 0,140, senate 0,140, videos 0,130, torture 0,130, arab 0,130, captured 0,130

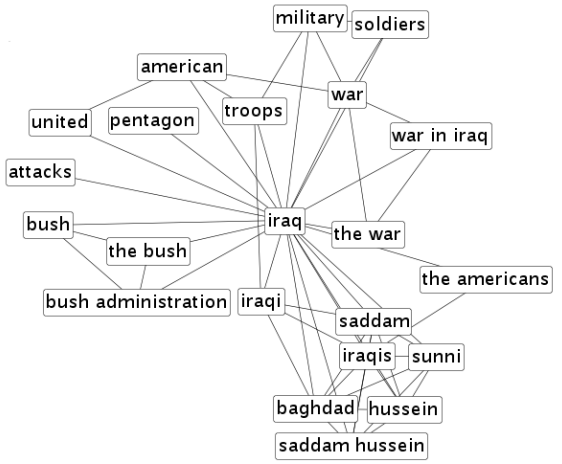
Kurses ein Indikator dafür sind, wie der Markt den korrekten Preis für ein Asset sucht. Je stärker die Schwankungen, desto mehr Unsicherheit herrscht über den korrekten Wert. In der Ökonometrie ist für diese Zwecke (als ein Risikomaß) die Volatilität definiert. Übertragen auf die globalen Kontexte von Termen beschreiben wir nachfolgend ein Volatilitätsmaß, welches Themen und deren Veränderungen über die Zeit anhand globaler Kontexte zu identifizieren hilft. Dafür werden im folgenden Abschnitt zuerst die nötigen Begriffe und Annahmen erläutert, bevor in Abschnitt 3 der Algorithmus im Detail vorgestellt wird.

2 Begriffsbildung

Wie in [Heyer et al. 2008] sei eine *Wortform* die flektierte Form eines Wortes, während ein *Wort* die Äquivalenzklasse aller zugehörigen Wortformen darstellt. *Term* ist ein variabler Begriff, der die gerade betrachtete Art von Einheiten bezeichnet. Er kann sich also sowohl auf Wortformen wie auch auf Wörter oder auch Wortmengen wie z. B. Synsets (Mengen synonymer Wörter) beziehen, je nachdem, auf welcher Abstraktionsebene und nach welcher Vorverarbeitung ein Text betrachtet und weiterverarbeitet wird. Ein *Token* bezeichnet ein Wort an einer bestimmten Stelle im Text, wohingegen ein *Type* ein Wort unabhängig von dessen konkreter Verwendung, Form, Ort etc. bezeichnet. Analog dazu verwenden wir den Begriff des *Themas* als Menge von Wörtern, die ein Ereignis, eine Situation, usw. beschreiben (welche sich wiederum aus dem globalen Kontext des Namens des Themas ergeben). Weiterhin sei ein *Konzept* die Äquivalenzklasse aller semantisch verwandten Wörter. Der *globale Kontext* einer Themenbezeichnung ist die Menge aller statistisch signifikanten Kookkurrenzen innerhalb des betrachteten Korpus. Jede betrachtete Zeitscheibe für sich wird als abgeschlossenes Korpus betrachtet. Die Menge *statistisch signifikanter Kookkurrenzen* einer Wortform errechnet sich aus dem gemeinsamen Auftreten dieser Wortform mit andere Wörtern innerhalb eines vordefinierten Textfensters. Übliche Fenstergrößen sind sind direkte Nachbarn, Sätze, Absätze, Dokumente. Der Signifikanzwert wird dabei auf Grundlage des Log-Likelihood-Maßes nach [Dunning 1993] wie in [Heyer et al. 2008] berechnet und anschließend bezüglich der Korpusgröße normiert. Diese Signifikanzwerte dienen im hier vorgestellten Ansatz nur zum Ranken der Kookkurrenzen. Die absoluten Zahlenwerte spielen keine Rolle. Tabelle 1 zeigt zur Veranschaulichung den globalen Kontext des Wortes »abu ghraib« aus dem New-York-Times-Korpus für den 10. Mai 2005, wobei die Zahlen in Klammern die Signifikanzwerte (multipliziert mit 10^6) darstellen, die für das Ranking verwendet werden (vgl. Abb. 2).



(a) 21. März 2003



(b) 14. Dezember 2003

(c) 6. Mai 2004

Abb. 1: Die Kookkurrenzgraphen für »iraq« zeigen die Veränderungen im globalen Kontext zu ausgewählten Zeitpunkten in den Jahren (a) 2003 und (c) 2004.

Der globale Kontext kann alternativ als ein Graph betrachtet werden, in welchem ein Wort und dessen Kontextwörter die Knoten bilden und die Kanten zwischen diesen Wörtern die Signifikanzwerte als Gewichte tragen. Dabei werden nur statistisch signifikante Werte berücksichtigt. Die Abbildungen 1(a)–(c) zeigen die zeitliche Veränderung der Kookkurrenzen – und damit der globalen Kontexte – des Wortes »iraq«, basierend auf dem Korpus der New York Times (siehe Abschnitt 4). Die Graphen sollen verdeutlichen, wie eine sich verändernde Berichterstattung über die Zeit in den Kookkurrenzdaten niederschlägt.

3 Berechnung der Volatilität

Die Verarbeitung von großen und sehr großen Dokumentkollektionen führt zu mehreren Herausforderungen, welche es dem Benutzer prinzipiell erschweren, auf einzelne Dokumente zuzugreifen, speziell wenn die Existenz oder Position des Dokuments diesem User nicht bekannt sind. State-of-the-art-Benutzerschnittstellen auf große Dokumentkollektionen stellen derzeit zum einen Indexe dar, wie beispielsweise Google und andere Suchmaschinen sie bieten und welche auf dem Indexieren aller oder der statistisch relevanten Terme beruhen. Zum anderen kommen strukturierte Kataloge zum Einsatz, welche die Kollektion über annotierte und reichhaltige Metadaten, die für jedes Element gepflegt werden, durchsuch- oder browsbar machen.

Das größte Problem ist jedoch die üblicherweise enorme Datenmenge selbst und die Komplexität der Analyse selbiger. So liegt die Zeit- und Platzkomplexität bei der Berechnung der globalen Kontexte (beispielsweise Satzkookkurrenzen) in $O(n^2)$, wobei n der Anzahl der Types entspricht, welche üblicherweise zwischen einer und zehn Millionen liegt. Mit *Medusa* steht ein Framework zur Verfügung, um auch auf größeren Datenmengen signifikante Kookkurrenzen in akzeptabler Zeit zu berechnen (vgl. [Büchler 2006]).

Aus diesem Grund sind die Berechnungen selbst, wie schon das Finden geeigneter Maße für Term- und Kookkurrenzgewichtung, nicht trivial. Deshalb basieren heutzutage viele Analysen auf Termfrequenzen – wie beispielsweise die Termgewichtung tf-idf – die sich effizient auch auf großen Kollektionen berechnen lassen.

Wir gehen bei der Interaktion mit großen Zeitscheibenkorpora einen neuen Weg. Offensichtlich ist es unmöglich, die Informationen aller Dokumente der Kollektion auf einmal zu visualisieren. Denn nimmt man zum Beispiel einen Bildschirm mit der Auflösung $1\,280 \times 1\,024$ ($= 1\,310\,720$ Pixel) an, stünden für jedes der 1,6 Millionen Artikel des New-York-Times-Korpus gerade mal $0,82$ Pixel zur Verfügung. Es ist also notwendig, einem Nutzer aggregierte Sichten auf die Daten anzubieten, welche – *details on demand* – eine interaktive explorative Durchdringung der Kollektion ermöglichen. Dabei soll der Nutzer möglichst wenig eingeschränkt werden, was die Ziele und Möglichkeiten seiner Datenexploration betrifft.

Für diesen Zweck identifizieren wir die relevantesten Terme, derart, dass diese möglichst prägnant jeweilige Entwicklungen in Teilbereichen des Zeitscheibenkorpus wiederspiegeln. Wir definieren dafür das Maß der Volatilität eines Wortes, welches die Veränderung des globalen Kontexts dieses Wortes quantifiziert und demzufolge auch die Veränderung in der Verwendung dieses Wortes anzuzeigen imstande ist. Damit können wir einen Überblick

1. Erstelle ein Korpus, in dem die Daten aller Zeitscheiben vereinigt sind.
2. Berechne für dieses Korpus die signifikanten Kookkurrenzen $C(t)$ für jedes Term t.
3. Berechne alle signifikanten Kookkurrenzen $C_i(t)$ für jedes Zeitscheibenkorpus zu einer Zeitscheibe T_i für jeden Term t.
4. Für jeden Kookkurrenten $c_{t,j} \in C(t)$ berechne die Serie der Ränge $\mathrm{rank}_{c_{t,j}}(i)$ von $c_{t,j}$ in der Kookkurrenzliste von t über alle Zeitscheiben T_i.
5. Berechne die Varianz der Liste von Rängen $\mathrm{Var}_i(\mathrm{rank}_{c_{t,j}}(i))$ für jeden Kookkurrenten $c_{t,j} \in C(t)$.
6. Die Volatilität eines Termes $\mathrm{Vol}(t)$ errechnet sich als arithmetisches Mittel über diese Varianzwerte

$$\begin{aligned}\mathrm{Vol}(t) &= \operatorname*{avg}_j \left(\operatorname*{Var}_i \left(\mathrm{rank}_{c_{t,j}}(i) \right) \right) \\ &= \frac{1}{|C(t)|} \sum_j \operatorname*{Var}_i \left(\mathrm{rank}_{c_{t,j}}(i) \right) \ .\end{aligned}$$

Abb. 2: Der Algorithmus zur Berechnung der Volatilität.

über Themen geben, die stark diskutiert wurden oder anderweitig starken Veränderungen in ihrer Benutzung über die Zeit unterworfen waren. Dieser Überblick dient dann als Einstieg in die Kollektion.

Die Datenbasis unserer Analyse stellt eine Menge von Zeitscheibenkorpora dar. Jedes Korpus entspricht einem Tag und enthält sämtliche Artikel der Druckausgabe der New York Times. Die Bedeutungsveränderung von Wörtern wird durch einen Vergleich der globalen Kontexte dieser Wörter über die einzelnen Zeitscheiben ermittelt.

Das Maß, welches für diese Aufgabe genutzt wird, nennen wir *Volatilität*. Es ist vom gleichnamigen Maß der Ökonometrie abgeleitet, wird jedoch anders berechnet. Eine Übersicht über verschiedene Ansätze zur Volatilitätsberechnung gibt [Taylor 2007]. Prinzipiell werden die Wörter in den globalen Kontexten (signifikante Kookkurrenzen) nach den ihnen zugeordneten Signifikanzwerten sortiert und anschließend diese Sortierungen der einzelnen Zeitscheiben verglichen (siehe Abschnitt 2). Die Grundidee dabei ist, dass die Bedeutungsveränderung eines bestimmten Wortes dazu führt, dass es über die Zeit betrachtet in (leicht) anderen Kontexten Verwendung findet, sich also dessen Kookkurrenzen und Signifikanzwerte ebenfalls verändern. Der genaue Algorithmus zur Volatilitätsberechnung ist in Abbildung 2 angegeben.

Um die Zeitkomplexität unseres Algorithmus zu reduzieren, werden nur die über das gesamte Korpus wichtigsten Kookkurrenzen $C(t)$ eines Wortes t betrachtet. Dazu werden die globalen Kontexte des Wortes ermittelt, nachdem alle Zeitscheiben in ein großes Korpus vereinigt wurden – im Fall NYTC entspricht der Umfang rund 7 500 Tage, also etwa 20 Jahre. Anhand dieses Gesamtkorpus werden die signifikantesten Kookkurrenzen $C(t)$ über einem bestimmten Schwellwert für die zu analysierenden Wörter t ausgewählt und dann bei der Zeitscheibenanalyse als einzige berücksichtigt. Über die Kookkurrentenfilterung hinaus spielt das Gesamtkorpus im Analyseprozess keine Rolle. Kookkurrenzen gelten als statistisch signifikant, wenn sie a) mindestens zwei mal im Korpus auftreten und b) ihr über Log-Likelihood berech-

Tab. 2: Die Beziehung zwischen Volatilität und globalem Kontext.

Anzahl Kookkurrenzen	Volatilität: niedrig	Volatilität: hoch
gleich	hochfrequente Terme: - Stopwörter - niedrige/mittlere »statische Konzepte«	zyklische oder arbiträre Konzepte, z. B. »Montag«, »Stadtzentrum« usw.
ungleich	aufsteigende und absteigende Konzepte, z. B. »Globalisierung« usw.	stark dynamische Begriffe: - niederfrequent (»schwache Signale«) - hochfrequent

neter Signifikanzwert über einem Schwellwert liegt. Aktuell wird der Schwellwert so gewählt, dass rund die Hälfte der Kookkurrenzen mit Anzahl größer eins entfernt werden. Aus sprachstatistischer Sicht ist dies eine überaus vorsichtige Filterung.

Die Beziehung zwischen der Volatilität eines Terms und dessen globalem Kontext kann wie in Tabelle 2 dargestellt werden. Es ist zu erwarten, dass z. B. die Volatilität der Wortform »Montag« recht hoch ist, da die Wochentage (und andere zyklisch wiederkehrende Zeitreferenzen) hochgradig ambig sind, da durch die fehlende weitere Spezifizierung nicht klar ist, auf welchen konkreten Tag Bezug genommen wird. Gleiches gilt für ambige Ortsangaben ohne genauere Spezifikation, wie »Stadtzentrum«, »Rathaus«, etc., welche ebenfalls nicht disambiguiert werden (d. h. bei denen nicht die Extension, d. h. die konkret bezeichnete Entität identifiziert wird). Aus diesem Grund erscheint uns derzeit eine Konzentration auf die Volatilitätswerte von Wörtern mit schwankender Anzahl von Kookkurrenzen über die Zeit (die untere Zeile in Tabelle 2) als vorrangig. Stopwörter oder besonders seltene Wörter besitzen i. d. R. recht konstante Anzahlen von Kookkurrenzen und sind primär in die obere Tabellenzeile einordenbar.

4 Experimente

Nachfolgend präsentieren wir Ergebnisse der Experimente bis dato, die auf Basis des New-York-Times-Korpuses (NYTC) durchgeführt wurden. Die Eckdaten des Korpus werden in Tabelle 3 angegeben. Erste Tests wurden für Deutsch auf den Daten des Projekts Deutscher Wortschatz[3] durchgeführt und erbrachten vergleichbare Resultate (vgl. [Heyer et al. 2009]). In einem ersten High-Level-Versuch war das Ziel, in bestimmten variablen Zeiträumen »stark diskutierte« Themen zu identifizieren. Dazu wurden aus dem 20 Jahre abdeckenden NYTC die rund 50 000 häufigsten Terme extrahiert und die Wortliste folgendermaßen modifiziert:

- Stoppwörter wurden entfernt und das Korpus wurde in Kleinschreibung transformiert.
- Es wurden alle Terme mit Rang größer 50 000 entfernt (das entspricht etwa 850 oder weniger Belegstellen in 20 Jahren NYTC)
- Es wurden alle Artikelnamen aus der Wikipedia, die maximal drei Leerzeichen enthalten und mindestens Rang 50 000 besitzen, zur Wortliste zugefügt (dadurch konnten Kookkurrenzen zu »new york city« gefunden werden, statt nur zu »new«, »york« und »city«)
- Zahlen, Nummern und ähnliches wurden entfernt (z. B. »101st«, »20,000«, etc.)

Die entstandene Wortliste enthält als häufigstes Wort »people« und als seltenstes Wort »benes« (ein Nachname). Wie in Abschnitt 3 beschrieben wurde für jedes der rund 50 000 Wortformen die im Gesamtkorpus (20 Jahre) signifikantesten Kookkurrenzen ermittelt. In einem zweiten Schritt wurden anschließend für jedes Wort dieser Kookkurrenzliste dessen Volatilität errechnet.

In Abbildung 3 ist der Verlauf der Volatilität für »abu ghraib« von Januar 2003 bis Dezember 2006 dargestellt. Die Volatilität wurde dabei tagesweise mit einem 30-Tage-Fenster berechnet, d. h. für die Volatilität an einem Tag wurden die Ränge der Kookkurrenten über den Zeitraum der letzten 30 Tage davor betrachtet (vgl. Abbildung 2). Daneben ist in Abbildung 3 zusätzlich die Frequenz des Wortes »abu ghraib« dargestellt. Diese Frequenz ist ebenfalls ein 30-Tage-Mittel, berechnet aus den absoluten Frequenzen (Auftretenshäufigkeiten) von »abu ghraib« in den einzelnen Ausgaben der New York Times der jeweils letzten 30 Tage.

Die klar erkennbaren Spitzen der Volatilitätskurve sind leicht mit bestimmten Ereignissen und der jeweilig damit verbundenen Diskussion in Zusammenhang zu bringen. Die erste Spitze im Mai 2004 geht auf die erstmalige Berichterstattung über die Folterfotos und -videos aus dem Gefängnis in Abu Ghraib zurück. Das schlägt sich auch direkt in den für diese Tage berechneten Kookkurrenzen von »abu ghraib« nieder, wie man in Tabelle 1 beispielhaft für den 10. Mai 2004 sehen kann. Da »abu ghraib« vorher überhaupt nicht in der Dikussion war, nimmt neben der Frequenz auch die Volatilität stark zu, da sich der globale Kontext stark ändert.

Im weiteren Verlauf sieht man, daß die Frequenz insgesamt abnimmt, aber schwankt. Die Schwankungen stehen allerdings in keinem Zusammenhang mit thematischen Entwicklungen, wie man in der Dastellung daran sieht, daß die weiteren Spitzen der Volatilität nicht mit denen der Frequenz korrellieren. Die Volatilitätsspitze im April 2005 wird durch einen Selbstmordanschlag auf Abu Ghraib am 5. April verursacht. Diese thematische Verschiebung in der Erwähnung von »abu ghraib« führt nicht zu einer erweiterten Berichterstattung in der New York Times und ist trotzdem als Veränderung im Kontext meßbar. Die dritte Volatilitätsspitze im November und Dezember 2005 wird durch eine Wanderausstellung, die unter anderem in New York gastierte, verursacht. Dort wurden Bilder aus Abu Ghraib zusammen mit anderen wie zur Weimarer Republik und dem 2. Weltkrieg gezeigt. Diese Thematik schlägt sich auch deutlich im globalen Kontext von »abu ghraib« am 20. November, als die Berichterstattung über diese Ausstellung einsetzte, nieder (vgl. Tabelle 4). Auch dieses Er-

Tab. 3: Daten des untersuchten Korpus NYTC.

Sprache	englisch
Zeitraum	Januar 87 – Juni 07
Anz. Zeitscheiben	7 475
Anz. Dokumente	1,65 Mio.
Anz. Tokens	1,2 Mrd.
Anz. Types	3,6 Mio.
Anz. sig. Kookkurenzen	29,5 Mrd.
Größe (reiner Text)	5,7 GB

[3] http://wortschatz.uni-leipzig.de/

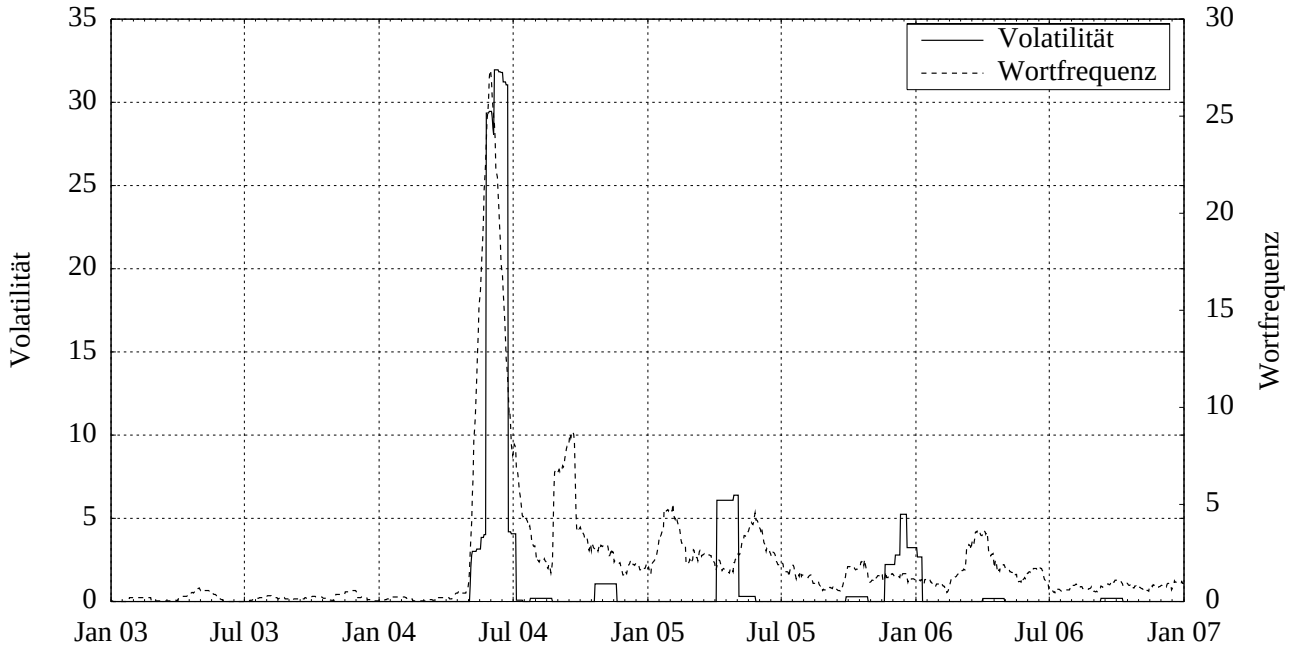


Abb. 3: 30-Tages-Volatilität und über 30 Tage gemittelte Frequenz von »abu ghraib« von 2003 bis 2006 auf Basis des NYT Korpus.

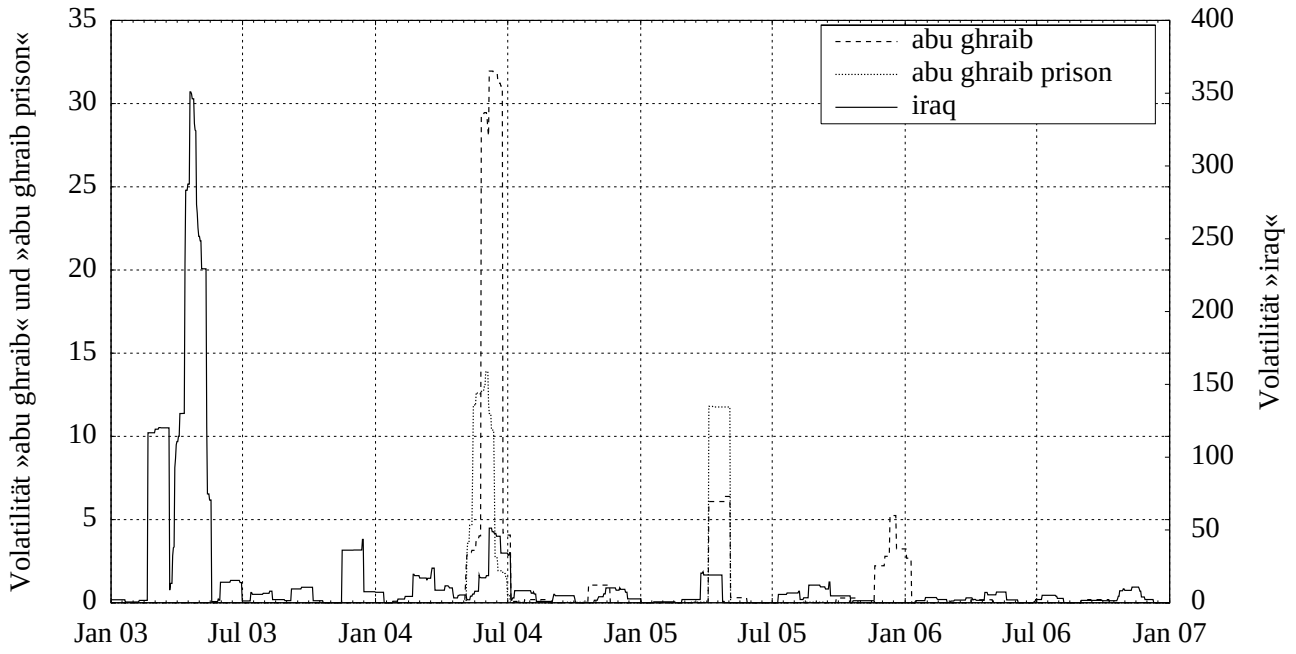


Abb. 4: 30-Tages-Volatilität von »abu ghraib«, »abu ghraib prison« und »iraq« von 2003 bis 2006 auf Basis des NYT Korpus.

eignis führt nicht zu einer häufigeren Nennung von »abu ghraib« in der New York Times, ist aber als Kontextverschiebung deutlich meßbar.

Interessant ist auch die Gegenüberstellung der Volatilitätsgraphen von »abu ghraib«, »abu ghraib prison« und »iraq« wie in Abbildung 4. In der Zeit von Januar 2003 bis Mai 2004 ist Abu Ghraib überhaupt kein Thema, wohingegen die Invasion im Irak sich ab Anfang 2003 deutlich niederschlägt. Dabei sind ganz zu Anfang zwei große Spitzen unterscheidbar. Die erste ist die letzte Phase der Vorkriegsdiskussion, wohingegen sich die Spitze da-

Tab. 4: Die 30 signifikantesten Kookkurrenzen im globalen Kontext von »abu ghraib« am 20. November 2005.

disasters, hook, grosz, international center, finalized, weighty, inkling, complement, partnerships, guggenheim museum, collaborative, the big city, easel, reaped, hudson river museum, blockbuster, enlarging, goya, weimar, art museums, eras, inconvenient, negatives, golub, poughkeepsie, griswold, big city, impressionist, staging, neuberger

nach mit der Invasion im Irak entwickelt. Sie zieht sich bis 20. Mai 2005, woran man aufgrund des 30-Tage-Fensters erkennt, daß bis 20. April, dem Höhepunkt, alle in diesem Zeitraum neuen Aspekte hinzugekommen sind.

Im Zusammenhang mit dem Bekanntwerden der Vorgänge in Abu Ghraib ist zu beobachten, daß in den ersten Tagen eine thematische Verschiebung sich vorrangig bei »abu ghraib prison« zeigt, bis sich dann kurz darauf »abu ghraib« ohne *prison* etabliert. Seitdem ist der Begriff »Abu Ghraib« in der westlichen Welt auf das Gefängnis in der Stadt Abu Ghraib, die rund 1 Mio. Einwohner hat, und die Vorfälle dort verengt.

Hier hat offensichtlich eine Begriffsbildung stattgefunden, denn »abu ghraib prison« ist danach an der Entwicklung der Dikussionen um Abu Ghraib nicht mehr beteiligt. Nur bei der Berichterstattung über den Selbstmordanschlag auf das Gefängnisgebäude wird auch »abu ghraib prison« wieder im neuen des Anschlages Kontext erwähnt, wie man den zusammenfallenden Spitzen im April 2005 sieht. Das es inbezug auf den Irak deutlich mehr thematische Aspekte als bei Abu Ghraib gibt, ist bei »iraq« der Einfluß der thematischen Verschiebung auf die Volatilität sowohl durch das Bekanntwerden der Vorgänge wie auch durch den Anschlag auf das Gefängnisgebäude geringer als bei »abu ghraib« und »abu ghraib prison«. Die Ausstellung zu Abu Ghraib ab November 2005 hat auf den Kontext von »iraq« und auch »abu ghraib prison« keinen meßbaren Einfluß, da in der Berichterstattung fast ausschließlich der inzwischen selbständige Begriff »abu ghraib« erwähnt wird.

5 Weiteres Vorgehen

Dieser Artikel soll eher das Vorgehen und den Stand der Arbeiten im vorliegenden Projekt »Topology-based Visual Analysis of Information Spaces« dokumentieren, als dass abschließende Ergebnisse präsentiert werden. Demzufolge ist noch eine Menge Arbeit zu leisten, um das Vorhaben antragsgemäß abschließen zu können. Die weitere Fokussierung soll nachfolgend kurz skizziert werden. Die nachfolgend beschriebenen Punkte stellen dabei eine Ideensammlung dar.

Clustering Betrachtet man die einzelnen Datenpunkte der Volatilitätskurve als Vektor, in welchem jede Zeitscheibe eine Dimension repräsentiert, ist ein Clustering über das Vektorraummodell möglich. Auf diese Weise können Wörter mit ähnlichen Kurvenverläufen identifiziert werden und evtl. eine Korrelation der Volatilität einzelner Konzepte generell oder in bestimmten Zeiträumen nachgewiesen werden. Welche Clusteralgorithmen oder welche Ähnlichkeitsmaße zum Einsatz kommen, ist bis dato noch nicht abschließend geklärt.

Visualisierung Parallel wird an einem Visualisierungsmodell gearbeitet, um die Volatilitäten verschiedener Wörter in ansprechender Weise darzustellen, ohne den Nutzer mit zu vielen optischen Details zu überfordern. Dabei wird eine Metapher genutzt, um die visuelle Darstellung für den Anwender intuitiv verständlich zu halten: Volatilitätssedimente, die in einer 3d-Darstellung Hügel und Täler bilden. Jede (farbige) Sedimentschicht entspricht einem Zeitraum und der Volatilitätsverlauf eines Wortes über die Zeit kann sich als Kernbohrung in diesem Modell gedacht werden. Eine besondere wissenschaftliche Herausforderung besteht in diesem Fall in der geeigneten Auslegung der Wörter in der Fläche.

Browsing & Zooming Sind die Volatilitätswerte auf einer hohen Abstraktionsstufe visualisiert, soll dem Nutzer die Möglichkeit offen stehen, die Gründe zu ermitteln, die zu den ihm präsentierten Werten geführt haben, und zwar derart, dass die relativ komplexe Volatilitätsberechnung selbst für das Verständnis der Ergebnisse nicht notwendig ist. Hierfür sollen die globalen Kontexte der Wörter, dargestellt durch Wortlisten/Kookkurrenzen, Verwendung finden. Veränderungen in diesen Kontexten müssen geeignet dargestellt und die Verschiebung in den Rängen der Wörter in diesen Kontexten muss verdeutlicht werden.

Höhere Momente Da die Volatilität über die Zeit als Kurve darstellbar ist, liegt die Idee nahe, diese Kurve selbst zu charakterisieren, beispielsweise durch Bestimmung der Varianz, Schiefe oder Kurtosis, und auf diese Weise Aussagen über Grupen von Bedeutungsvolatilitäten treffen zu können. Können aufkommende oder abflauende Themen auf diese Weise identifiziert werden?

6 Zusammenfassung

In diesem Artikel haben wir einen neuen Ansatz in der Analyse großer diachroner Dokumentkollektionen vorgestellt, mithilfe dessen die Bedeutungsveränderung von Termen als Repräsentanten von Themen über die Zeit ermittelt werden kann. Dafür wird der Wandel des globalen Kontexts des Terms als Indikator herangezogen. Die auf diese Weise ermittelten Daten bilden die Grundlage für eine interaktive visualisierungsbasierte Analyse (im Sinne des »Visual Analytics«).

Neben den Termen und Themen, die großen Veränderungen unterworfen sind, kann nunmehr untersucht werden, welche Terme eine sehr »stabile« Bedeutung, d. h. stabile Kontexte besitzen. Wir nehmen an, dass diese Terme in gesellschaftlichem Konsens immer mit bestimmten Themen assoziiert sind. Das Ziel unserer Arbeiten stellt eine Infrastruktur zur Analyse großer diachroner Textkorpora dar, die interaktive und intuitive Unterstützung in der Medienresonanzanalyse, im Trend Mining, im Marketing und in den E-Humanities bieten kann.

Literatur

[Allan 2002] Allan, J.: Introduction to topic detection and tracking., pages 1–16 Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[Allan et al. 1998] Allan, J. et al.: Topic detection and tracking pilot study final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, 1998.

[Büchler 2006] Büchler, M.: , Flexibles Berechnen von Kookkurrenzen auf strukturierten und unstrukturierten Daten . Master's thesis, Universität Leipzig, Augustusplatz 10/11, 04109 Leipzig, Germany, 2006 `http://www.eaqua.net/~mbuechler/references/da-mbuechler-public.pdf`.

[Dunning 1993] Dunning, T. E.: Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[Heyer et al. 2009] Heyer, G.; Holz, F.; Teresniak, S.: Change of topics over time – tracking topics by their change of meaning. In *KDIR '09: Proc. of Int. Conf. on Knowledge Discovery and Information Retrieval*, 2009 accepted, in press.

[Heyer et al. 2008] Heyer, G.; Quasthoff, U.; Wittig, T.: Text Mining: Wissensrohstoff Text – Konzepte, Algorithmen, Ergebnisse. W3L-Verlag, 2nd edition, 2008.

[Kleinberg 2002] Kleinberg, J.: Bursty and hierarchical structure in streams. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 91–101, New York, NY, USA. ACM Press, 2002.

[Kumaran & Allan 2004] Kumaran, G.; Allan, J.: Text classification and named entities for new event detection. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 297–304, New York, NY, USA. ACM, 2004.

[Swan & Allan 1999] Swan, R.; Allan, J.: Extracting significant time varying features from text. In *CIKM '99: Proceedings of the eighth international conference on Information and knowledge management*, pages 38–45, New York, NY, USA. ACM, 1999.

[Swan & Allan 2000] Swan, R.; Allan, J.: Automatic generation of overview timelines. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–56, New York, NY, USA. ACM, 2000.

[Taylor 2007] Taylor, S. J.: Introduction to asset price dynamics, volatility, and prediction. In *Asset Price Dynamics, Volatility, and Prediction*, Introductory Chapters. Princeton University Press, 2007.

[Wang & McCallum 2006] Wang, X.; McCallum, A.: Topics over time: a non-markov continuous-time model of topical trends. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433, New York, NY, USA. ACM, 2006.