*Erschienen in: Datenbank-Spektrum ; 11 (2011), 3. - S. 195-206*

# Interaktive explorative Suche in großen Dokumentbeständen

**Gerhard Heyer · Daniel Keim · Sven Teresniak · Daniela Oelke**

**Zusammenfassung** Im klassischen Paradigma des Information Retrievals steht das Finden von Dokumenten im Vordergrund, die Informationen bzw. Fakten enthalten, die dem vermuteten Informationsbedürfnis des Nutzers entsprechen. Dabei stellt der Nutzer solche Anfragen an das Informationssystem, von denen er annimmt, dass dazu eindeutige Antworten im Informationssystem vorhanden sind, die lediglich zurückgeliefert oder gefunden werden müssen. In vielen Fällen ist der Benutzer aber weniger an den Fakten selber interessiert, als vielmehr daran, wie über Fakten berichtet wird: Über welche Fakten wird berichtet? Nach welchen Kriterien werden Fakten ausgewählt? Wie werden Fakten bewertet? Welche Konzeptualisierungen der Anwendungsdomäne werden vorausgesetzt? Und wie ändern sich Bewertungen und Konzeptualisierungen über die Zeit? Der vorgestellte Ansatz skizziert eine mögliche Lösung für die explorative Suche in großen Datenmengen.

## 1 Einleitung

Im klassischen Paradigma des Information Retrievals steht das Finden von Dokumenten im Vordergrund, die Informationen bzw. Fakten enthalten, die dem vermuteten Informationsbedürfnis des Nutzers entsprechen. Dabei stellt der Nutzer solche Anfragen an das Informationssystem, von denen er annimmt, dass dazu eindeutige Antworten im Informationssystem vorhanden sind, die lediglich zurückgeliefert oder gefunden werden müssen. In vielen Fällen ist der Benutzer aber weniger an den Fakten selber interessiert, als vielmehr daran, wie über Fakten berichtet wird: Über welche Fakten wird berichtet? Nach welchen Kriterien werden Fakten ausgewählt? Wie werden Fakten bewertet? Welche Konzeptualisierungen der Anwendungsdomäne werden vorausgesetzt? Und wie ändern sich Bewertungen und Konzeptualisierungen über die Zeit? Rechercheaufgaben im Bereich des Journalismus, des Technology Mining oder in den eHumanities, um nur einige zu nennen, sind Beispiele für diese Art von Suchaufgaben, die allgemein als *explorative Suche* bezeichnet werden. Im Gegensatz zum klassischen Muster des *query-and-retrieve*, bei dem ein Benutzer auf der Grundlage eines ihm bekannten Domänenmodells möglichst schnell und erschöpfend die für sein Informationsbedürfnis relevanten Dokumente erhalten möchte, benötigt er im Fall der explorativen Suche Unterstützung,

1. um sich mit den Inhalten der Dokumentkollektion überhaupt erst vertraut zu machen,
2. die für sein Rechercheinteresse möglicherweise relevanten Terme zu identifizieren sowie
3. verschiedene Pfade zur Erkundung der Dokumentkollektion zu erkennen und zu verfolgen.

Mit dem nachfolgenden Bericht möchten wir einerseits den Stand der Forschung zur explorativen Suche (eng. *exploratory seach*) zusammenfassen sowie andererseits einen eigenen Ansatz zur explorativen Suche unter Verwendung eines neuen Maßes zur Berechnung auffälliger oder „interessanter“ Terme sowie geeigneter Visualisierungen vorstellen, die u. a. im Rahmen des DFG Schwerpunktprogramms Scalable Visual Analytics entstanden sind. Hierbei wird für die explorative Suche in zwei Schritten verfahren:

1. Auswahl von Zeiträumen und Texten, die betrachtet werden sollen und Berechnung von „interessanten“ Termen nach verschiedenen Berechnungsverfahren wie z. B. tf·idf oder dem Maß der Kontextvolatilität. Im Ergebnis

G. Heyer · D. Keim · S. Teresniak (✉) · D. Oelke
University of Leipzig, Leipzig, Deutschland
e-mail: teresniak@informatik.uni-leipzig.de

erhält der Benutzer eine Liste von „interessanten“ Termen, aus denen er interaktiv die für ihn interessanten Terme zur weiteren Analyse auswählen kann.
2. Nach Auswahl eines (oder mehrerer) Term(s/e) aus der Liste möglicherweise „interessanter“ Terme wird dem Benutzer für den ausgewählten Zeitraum der Aktivitätswert der Terme farblich dargestellt, so dass er erkennen kann, zu welchen Zeiträumen die ausgewählten Terme besonders „interessant“ sind. Sofern das Maß der Kontextvolatilität verwendet worden ist, ist damit ersichtlich, zu welchen Zeiten der Term besonders stark diskutiert worden ist. Anhand der Kookkurrenzgraphen der ausgewählten Terme kann sich der Nutzer dann ein erstes Bild davon machen, welche Themenaspekte sich in den fraglichen Zeiträumen besonders stark verändert haben und sich bei Bedarf zu einzelnen Themen die entsprechenden Dokumente heraussuchen.

Wir behandeln zunächst den Begriff der explorativen Suche, stellen seine wesentlichen Aspekte heraus und geben einen kurzen Überblick über verfügbare Literatur und Systeme. Sodann stellen wir verschiedene Konzepte für die Berechnung statistisch auffälliger Terme vor, insbesondere Neuigkeit, Burstiness, Interessantheit und Kontextvolatilität. Unter Verwendung des Maßes der Kontextvolatilität stellen wir abschließend auf der Grundlage des New-York-Times-Korpus ein praktisches Anwendungsbeispiel vor.

## 2 Explorative Suche

Das klassische Paradigma des Information Retrievals basiert auf der Annahme, dass der Nutzer ein klar definiertes Informationsbedürfnis hat, weswegen er überhaupt das Information Retrieval System benutzt [17]. Für die Suche im Web zählt Broder dazu neben dem Aufruf von Webseiten, die bestimmte Informationen enthalten (informational need), auch das Bedürfnis, eine bestimme Webseite zu erreichen (navigational need) oder eine bestimmte Web-basierte Aktivität auszuführen (transactional need) [1]. Für die Beantwortung einer Nutzeranfrage wird bei der Umsetzung dieses Paradigmas in den gängigen Suchmaschinen aus den zu durchsuchenden Dokumenten ein Schlagwortindex und eine invertierte Datei aufgebaut; die Nutzeranfrage wird nach verschiedenen Beantwortungsstrategien mit den verfügbaren Dokumenten (oder Textpassagen) abgeglichen und die am besten passenden werden dem Nutzer meist in Form einer Ergebnisliste als Antwort auf seine Suchanfrage zurückgeliefert. Für das Nachschlagen von Informationen, sog. lookup searches, ist das klassische Paradigma des Information Retrieval besonders geeignet. Komplexe Suchanfragen können durch analytische Suchstrategien, wie sie noch heute z. B. in der Patent- oder Bibliotheksrecherche verwendet werden, bearbeitet werden. Dabei werden sorgfältig ausgewählte Suchterme in einer präzisen Syntax miteinander verbunden, um möglichst vollständige und eindeutige Suchergebnisse zu generieren, die nur mit geringem Aufwand weiter bearbeitet werden müssen. Zwar kann das Nachschlagen von Informationen im Rahmen des klassischen Paradigmas des Information Retrievals nicht zuletzt dank der großen Suchmaschinen im Netz als sehr erfolgreich bezeichnet werden, aber mit der zunehmenden Verfügbarkeit großer digitaler Dokumentkollektionen im Web entwickeln sich neue Informationsbedürfnisse, die über das bloße Nachschlagen von Informationen hinausgehen und unter Verwendung von Versuch-und-Irrtum-Taktiken vielmehr darauf abzielen, Suchterme interaktiv einzugrenzen und zu bestimmten Informationsquellen zu navigieren. Als gutes Beispiel für eine interaktive Analyse mit hoher Relevanz zum vorliegenden Artikel sei der *Time Explorer* von Matthews et al. erwähnt, welcher als Teil des Projekts *LivingKnowledge*[1] entstanden ist [14].

Mit dem Begriff der explorativen Suche werden in der Literatur alle Formen von Suchanwendungen bezeichnet, bei denen der Nutzer iterativ verschiedene Arten von Informationen zusammentragen und bewerten muss und seine Suchanfragen entsprechend seinem Informationsbedürfnis und den bisher erhaltenen Informationen ggf. neu formulieren kann. Die explorative Suche verbindet also analytische Suchstrategien mit interaktiven browsing Strategien [13]. Sie ermöglicht es damit in besonderer Weise, eine Suche in der Tiefe eines Informationsraumes mit der Suche in der Breite zu verbinden [22]. Typische Anwendungen der explorativen Suche finden sich beim Lernen oder Erforschen von Zusammenhängen. Empirische Studien belegen, dass die explorative Suche tatsächlich ein Informationsbedürfnis und Suchverhalten beschreibt, das von den gängigen Suchmaschinen nicht direkt unterstützt wird [2, 15].

Die Haupttrends in den aktuellen Arbeiten zur explorativen Suche betreffen zum einen die domänenspezifische Strukturierung von Informationen und deren interaktive Eingrenzung in Form einer facettierten Suche auf der Basis von domänenspezifischen Thesauren und Semantic Web Technologien wie SPARQL [3, 6] oder Topic Maps [21]. Zum anderen ist die explorative Suche eine paradigmatische Anwendung der Visual Analytics für die interaktive Textvisualisierung, indem passende Visualisierungen es ermöglichen, einen ersten Überblick über sehr große Textmengen zu generieren, der sukzessive weiter vertieft werden kann (drill down), oder Textmerkmale zu erkunden, die für eine weiterführende Dokumentanalyse verwendet werden sollen [10, 16].

Im Folgenden wollen wir uns auf die explorative Suche in Textkollektionen konzentrieren, die mit einem Zeit-

[1] http://livingknowledge.europarchive.org/.

stempel versehen sind, wie z. B. die Ausgaben einer Tageszeitung. Hier besteht die konkrete Aufgabe zunächst darin, auf der Grundlage einer vorgegebenen Dokumentkollektion dem Nutzer ähnlich einer Term-Dokument-Matrix interaktiv eine Term-Zeit-Matrix zu generieren, welche diejenigen Terme enthält, die in den Textdaten für einen bestimmten Zeitpunkt oder Zeitraum charakteristisch sind. Der Nutzer kann sich dann für ausgewählte Terme die für die Verwendung dieser Terme charakteristischen Zeitpunkte oder -räume anzeigen lassen (z. B. „Wendehals“), oder umgekehrt, für ausgewählte Zeitpunkte oder -räume die für diese Zeiten besonders charakteristischen Terme (z. B. „11. September 2001“) [5].

## 3 Konzepte und Maße für die Berechnung auffälliger Terme

Um die für einen Zeitpunkt oder -raum charakteristischen Terme zu identifizieren, ist die reine Termfrequenz genau so wenig hinreichend, wie für die Indexierung einzelner Dokumente. Es bietet sich daher zunächst ein Indexierungsverfahren, wie z. B. tf·idf, an, mit dem die Terme aus den Dokumenten des ausgewählten Zeitraums im Vergleich zu allen Dokumenten extrahiert werden können (vgl. dazu [12, 19]). Die zeitspezifische Relevanz von Termen wird hier genauso berechnet wie ihre dokumentenspezifische Relevanz, auch wenn das Verfahren es erlaubt, in Bezug auf einen bekannten Sachverhalt neue Dokumente herauszufinden, die bisher nicht bekannte Aspekte des Sachverhalts thematisieren [18]. Insofern für die Dokumente aus bestimmten Zeiten nicht nur einzelne Terme, sondern auch das gemeinsame Auftreten von Termen charakteristisch ist, beziehen Wang und McCallum auch Kookkurrenzmuster in ihre Analysen mit ein und betrachten deren zeitliche Häufung [23]. Um das gehäufte Auftreten von Termen zu einem bestimmten Zeitpunkt bzw. über einen bestimmten Zeitraum zu modellieren (Burstiness), hat Jon Kleinberg ein Modell auf der Grundlage gewichteter endlicher Automaten entwickelt, das es ermöglicht, aufgrund der zeitlichen Ausdehnung von Termen über einen bestimmten Zeitraum hinweg Zeiten und Terme miteinander zu korrelieren [11]. Weitere Arbeiten zur Identifikation auffälliger Terme in diachronen Dokumentkollektionen behandeln Aspekte, wie überraschend oder interessant ein Term ist. Ein Term kann als überraschend gelten, wenn sein Auftreten zu Widersprüchen mit dem Vorwissen oder bestimmten Erwartungen des Nutzers führt [4]. Die Interessantheit von Termen kann daran bemessen werden, wir stark ein Term in einer Dokumentkollektion während eines bestimmten Zeitraums diskutiert wird. Dabei gilt die Grundregel: Je kontroverser eine Diskussion, desto interessanter [7].

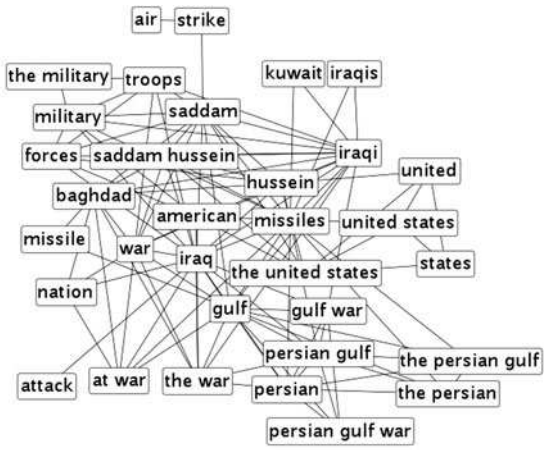
Im Idealfall stehen für die explorative Suche verschiedene Maße oder Werkzeuge zur Verfügung, die der Nutzer für seine Zwecke auswählen kann. Das Explorationsziel des Nutzers ist normalerweise vorab nicht bekannt. Deshalb kommen interaktiv verwendbare Analysemethoden zum Einsatz, welche im Sinne des Visual-Analytics-Ansatzes kombiniert werden können. Der stetige Wechsel zwischen visuellen und automatischen Vorgängen stellt dabei besondere Anforderungen an die verwendeten Verfahren, beispielsweise an Berechnungsgeschwindigkeit von Analysen und Aussagekraft von Visualisierungen.

Um die Wirkung eines Maßes zu erläutern und seinen Nutzen für die explorative Suche zu verdeutlichen, wollen wir uns aus Gründen der Verständlichkeit im Folgenden auf das Maß der Kontextvolatilität konzentrieren. Grundannahme dieses Maßes ist die Intuition, dass unterschiedliche Kontexte eines Terms einen Hinweis darauf geben, in welcher Hinsicht ein Term verwendet wird. So verdeutlichen die nachfolgenden Abb. 1(a) bis 1(c) die unterschiedlichen Verwendungskontexte des Terms *iraq* anhand seiner Kookkurrenzen (gerechnet mit dem loglikelihood Maß) aus dem New-York-Times-Korpus im März und Dezember 2003 sowie im Mai 2004.

Je kontroverser nun während eines bestimmten Zeitraums der Sachverhalt diskutiert wird, für den ein Term steht, desto häufiger dürften sich auch seine Kontexte, d. h. seine Kookkurrenzen, verändern. Diese Veränderung der Kontexte eines Terms lässt sich mit dem Maß der Kontextvolatilität messen (Algorithmus in Abb. 2, weitere Details in [20]).[2]

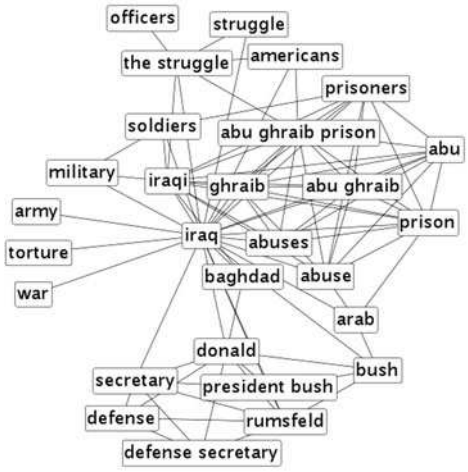
Für die gesamte Dokumentkollektion eines zeitindizierten Textkorpus werden dabei zunächst zeitscheibengenau die Ränge aller Kookkurrenten eines Terms ermittelt und der Rang jedes Kookkurrenten bezüglich des o. g. Signifikanzmaßes bestimmt. Die Kontextvolatilität eines Terms errechnet sich dann als der durchschnittliche Variationskoeffizent der Ränge seiner Kookkurrenten über einen vorgegebenen Zeitraum. Für das New-York-Times-Korpus werden aus praktischen Gründen die Veränderungen in den Rangpositionen auf der Grundlage eines 30-Tage-Fensters gerechnet. Tage, an denen ein Term nicht auftritt, werden ignoriert. Die für den Zeitraum 1.1.2004 bis 31.12.2004 im New-York-Times-Korpus mit diesem Maß errechneten hochvolatilen Terme sind in Tab. 1 angegeben. Offenbar finden sich in dieser Liste eine Vielzahl von Themen, welche die

[2] Die Volatilität als ein Gradmesser der Kontroverse zielt dabei nicht auf die Polarität der Aussagen ab, d. h. die innere Einstellung des Sprechers zur Sache, wie sie die Sentiment-Analyse untersucht, sondern auf die Unterschiede in den Standpunkten der am Diskurs beteiligten Parteien. Je stärker sich diese Standpunkte unterscheiden, umso stärker weichen auch die verwendeten Vokabulare voneinander ab, mittels derer diese Standpunkte vertreten werden. Was für den einen *unvermeidbar* ist, kann für den anderen *unvertretbar* sein.




(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

Meinungsmacher in den USA im Jahr 2004 bewegt haben, so u. a. die Präsidentschaftswahl und die Geschehnisse im Irak. Interessiert man sich gar für eine Übersicht über einen Zeitraum von zehn Jahren, treten noch deutlicher populäre wie auch umstrittene Terme in den Vordergrund, wie Tab. 2 beispielhaft für den Zeitraum 1990 bis 2000 zeigt. In Abschn. 5.2 wird die Analyse des Präsidentschaftswahlkampfs erneut aufgegriffen und detaillierter betrachtet.

Werden aus dem gleichen Datenbestand für den gleichen Zeitraum die auffälligsten Terme mit Hilfe des tf·idf-Maßes gerechnet, so kann leicht verglichen werden, welche Terme von beiden Verfahren und welche Terme nur von dem einen oder dem anderen Verfahren als auffällig bzw. interessant zurückgeliefert werden. Um tf·idf auf einem Zeitscheibenkorpus berechnen zu können, wird jeweils eine Zeitscheibe als ein Dokument behandelt, also alle Artikel eines Tages zusammengefasst. Wie die folgenden Tab. 3, 4 und 5 zeigen, gibt es unter den Top-100 mit beiden Maßen gerechneten auffälligsten Termen nur eine geringe Übereinstimmung – lediglich fünf Wörter kommen in den Rankings beider Maße unter den Top-100 vor (22 Wörter unter den Top-500). Beide Maße bevorzugen offenbar deutlich unterschiedliche Terme: Während sich mit dem Maß tf·idf vor allem linguistisch auffällige Terme ermitteln lassen, etwa in Form von Abkürzungen oder Eigennamen von Personen und Firmen, finden sich in der Liste der mit dem Maß der Kontextvolatilität gerechneten Terme eher Bezeichner für Ereignisse oder Themen wie z. B. „olympics“, „homeland security“ oder „gay marriage“. Die Anzahl der Terme, die von beidem Maßen hoch gerankt werden, wächst auch bei Betrachtung der Rankings über die ersten 100 Stellen hinaus nicht schneller an, wie Abb. 3 verdeutlicht.

Insgesamt finden sich unter den mit dem Maß der Kontextvolatilität berechneten Termen auffällig viele Terme, die im oben skizzierten Sinne kontroverse Themen benennen. Diese Terme sind unabhängig von der Termfrequenz. Auch für niederfrequente Terme kann das Maß der Kontextvolatilität daher gut verwendet werden, um frühzeitig Veränderungen im Verwendungskontext eines Terms zu identifizieren. Neben den hochvolatilen, kontroversen Themen können im Umkehrschluss aber auch solche Terme identifiziert werden, die offenbar unstrittig sind und gewissermaßen einen politischen oder gesellschaftlichen Konsens darstellen. So wurde in der New York Times in den 1990er Jahren (neben vielen weiteren) der Begriff „climate change“ in relativ statischen Kontexten verwendet; obwohl der Klimawandel diskutiert wurde, finden sich offenbar in den Kontexten, in denen der Begriff zu Anfang der 1990er Jahre in den USA diskutiert worden ist, kaum neue oder kontroverse Aspekte. Dies änderte sich schlagartig, als Al Gore an der Seite Clintons das Thema in der amerikanischen Politik verwurzelte und nach 2000 seine Stellung und Popularität zur Förderung des Umweltthemas in den USA nutzte.

1. Berechne alle signifikanten Kookkurrenzen $C_{\mathrm{o}}(t)$ des Terms $t$ im Gesamtkorpus.
2. Berechne alle signifikanten Kookkurrenzen $C_{\mathrm{t}_i}(t)$ für jedes Zeitscheibenkorpus $\mathrm{t}_i$ für Term $t$.
3. Für jeden kookkurrenten Term $c_{\mathrm{o},t,j} \in C_{\mathrm{o}}(t)$ aus dem Gesamtkorpus, berechne die Rangserie $\mathrm{rank}_{c_{\mathrm{o},t,j}}(i)$, um die jeweiligen Ränge der Kookkurrenten $c_{o,t,j}$ in den jeweiligen globalen Kontexten von $t$ innerhalb der Zeitscheiben $\mathrm{t}_i$ zu erhalten.
4. Berechne den Variationskoeffizient (CV) über die Rangserie $\mathrm{CV}_i(\mathrm{rank}_{c_{\mathrm{o},t,j}}(i))$ für jeden kookkurrenten Term in $c_{\mathrm{o},t,j} \in C_{\mathrm{o}}(t)$.
5. Berechne das arithmetische Mittel über die so erhaltenen Varianzwerte, um die Volatilität des Terms $t$ zu erhalten:

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

**Abb. 2** Der Algorithmus zur Volatilitätsberechnung eines Termes $t$. Durch Variation der Zeitscheiben $t_i$ (Schritt 3) kann die Volatilität für verschiedene Zeitfenster berechnet werden

**Tab. 1** Die 30 volatilsten Terme für 2004

| |
|---|
| the convention, the reach, john edwards, touchdown, republican convention, caucuses, howard dean, national convention, republican national convention, election day, presidential debate, abu ghraib, ghraib, inning, turnout, quarterback, democratic convention, hurricane, delegates, innings, abu, holiday, the holidays, thanksgiving, running mate, falluja, the holiday, contests, interrogation, democratic national convention |

**Tab. 2** Die 30 interessantesten Terme im Zeitraum 1990–2000. Neben Staatsoberhäuptern und anderen bekannten Politikern kann das Aufkommen der Internettechnologien ebenso beobachtet werden, wie die beiden Kriegsschauplätze Bosnien und Kuwait

| |
|---|
| bill clinton, bob dole, bosnia, bush administration, clinton, clinton adminstration, e-mail, gingrich, giuliani, gorbachev, inning, internet, kuwait, lewinski, newt, newt gingrich, on-line, pataki, president bush, president clinton, rudolph, the bush, the clinton administration, the internet, the recession, the web, touchdown, touchdowns, web, web site |

**Tab. 3** Die 100 volatilsten Terme für 2004, welche *nicht* unter den Top-100 im tf·idf-Ranking zu finden sind

| |
|---|
| presidential debate, election day, the convention, touchdown, caucuses, turnout, the reach, hurricane, howard dean, abu, new year, fenway, league championship series, john edwards, inning, debates, united states senate, halloween, republican convention, innings, quarterback, fenway park, battleground, running mate, receiver, national convention, iowa caucuses, democratic convention, the new year, division series, incumbent, incumbents, moderator, gephardt, same-sex marriage, contested, the terrorists, assists, american people, undecided, tax cuts, pitches, gay marriage, electoral votes, abortion, counterterrorism, wesley, voter, caucus, the american people, democratic national convention, end zone, the capture, re-elected, abu ghraib prison, touchdowns, kerry campaign, college football, pedro, the democrats, home run, holiday, republican national convention, mentioning, the red sox, commander in chief, mad cow disease, hampshire, yankee stadium, football conference, interrogation, the struggle, the cardinals, ballots, nader, johnny damon, insurance companies, linebacker, the world series, new mexico, electoral, this administration, appointments, challenger, storms, schilling, curt schilling, inspectors, jeter, damon, ballot, adequate, the holidays, new hampshire, dugout |

**Tab. 4** Das Top-100-Ranking für 2004, ermittelt mit tf·idf, ohne Terme, die auch unter den 100 volatilsten Termen sind

| |
|---|
| csx, ferreira, khalilzad, kennan, izetbegovic, rhubarb, gallego, clarett, parmalat, messner, brando, de beers, chief justice rehnquist, tivo, dr. dean, hollinger, mientkiewicz, justice rehnquist, smarty, lord black, smarty jones, reit, peoplesoft, yea, backman, newmont, eldredge, bower, quattrone, fastow, paxson, eckersley, sadr, vick, kerry, custer, federer, falun gong, denotes, mr. franklin, anchovies, kaczynski, dayne, yukos, chrebet, falun, celebrex, nortel, ephedra, kibbutz, google, global crossing, amgen, aspirin, redstone, chalabi, title ix, kerik, ebbers, skakel, fugard, newell, adelphia, aventis, mondesi, vioxx, grasso, weprin, rna, rigas, gazprom, guidant, leffler, genentech, perle, mr. marcus, manton, cloning, haft, molson, venturi, mayfield, bacanovic, black holes, beckham, yushchenko, haider, barbera, tbs, autism, ganz, analog, kuchma, lebed, schaffer |

**Tab. 5** Schnittmenge aller Terme unter den Top-100-Rankings 2004 für tf·idf und der Kontextvolatilität. Die Anzahl von Termen, die von beiden Maßen berechnet werden, ist sehr gering. Die Rangpositionen der Terme in den jeweiligen Rankings sind sehr unterschiedlich

| |
|---|
| falluja, martínez, allawi, abu ghraib, ghraib |

## 4 Aspekte der visuellen Analyse für die explorative Suche

Für die Analyse ist nicht nur von Interesse, ob für einen Term Veränderungen im Verwendungskontext festgestellt werden können, sondern auch zu welchen Zeitpunkten solche Veränderungen stattgefunden haben. Darüber hinaus kann der Vergleich der Volatilitätskurven verschiedener Terme Aufschluss darüber geben, welche Zusammenhänge zwischen den Begriffen bestehen. Auf diese Weise könnten Entwicklungen ermittelt werden, die zeitgleich oder zeitlich nur knapp versetzt stattfinden. Wenn Wörter als Repräsentanten für Themen innerhalb eines beliebigen Zeitraumes eine ähnliche Veränderung in der Kontextvolatilität aufweisen, könnte dies u. a. folgende Ursachen haben:

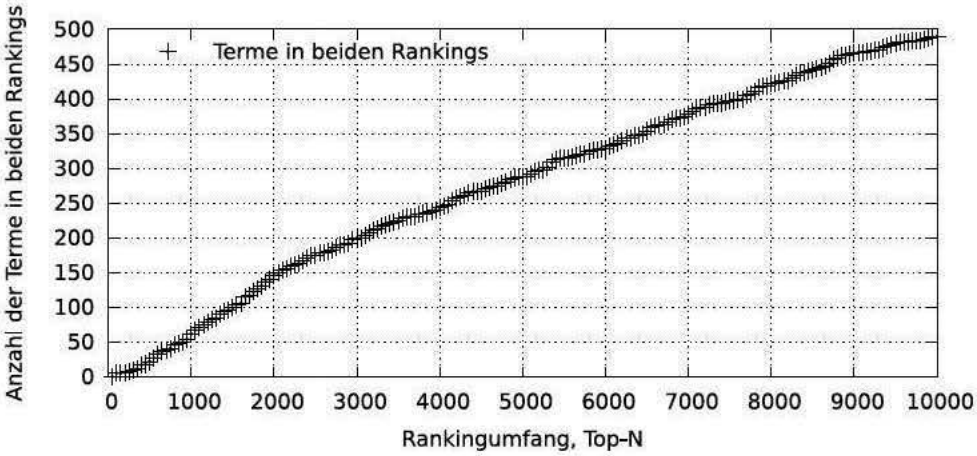



**Abb. 3** Anteil der Terme, die in beiden Rankings – tf·idf und Kontextvolatilität – vorkommen, in Abhängigkeit vom Rankingumfang. Im Mittel lässt sich knapp jedes 20. Wort in beiden Rankings finden

- *Kausalität*: Bestimmte Entwicklungen ziehen andere Entwicklungen nach sich (die quasi zeitgleich ablaufen können), über die ebenfalls berichtet wird, beispielsweise das letzte schwere Erdbeben in Japan und die Kernschmelze in Fukuchima II.
- *Kalkül*: Mechanismen der Berichterstattung werden mitunter missbraucht oder für eigene Zwecke ausgenutzt, wie z. B. die Bekanntgabe von negativen Nachrichten jeweils zum 11. September oder zum Superbowl in den USA (sog. *Burying*). Thematisch haben die Begriffe wenig oder keine gemeinsame Basis, jedoch ist der Zeitpunkt der Lancierung der negativen Meldung nicht zufällig.
- *Absicht*: Medien werden aus verschiedensten Gründen gelenkt und neben der objektiven Berichterstattung auch zur Meinungsmache verwendet. Dies gilt nicht nur für totalitäre Staaten mit Medienmonopolen und Zensur. Manche Kontroversen können deshalb als gesteuert und zielgerichtet angesehen werden (z. B. Schmutzkampagnen, die Personen mit negativen Themen in Verbindung bringen).
- *Zufall*: Verschiedene Entwicklungen laufen parallel, stehen jedoch in keinem direkten Verhältnis zueinander. Beispielsweise ein Erdbeben in Asien und Wahlen in Europa.
- *Andere Effekte*: Denkbar ist auch, dass andere als die oben genannten Ursachen für ähnliche Volatilitätsverläufe unterschiedlicher Terme verantwortlich sein können, wie beispielsweise Synonymie („Papst" vs. „Heiliger Vater"). Eine nähere Untersuchung solcher Effekte steht aktuell noch aus.

Etwas vorgreifend sei hier auf Abb. 7 verwiesen, um ein Beispiel für parallele Trends zu zeigen. John Kerry (Term nicht in der Abbildung) wurde während des US-Präsidentschaftswahlkampfs 2004 mit dem Fakt konfrontiert, dass er 1991 gegen den (aus amerikanischer Sicht) ersten Golfkrieg (*gulf war*) stimmte. Später änderte Kerry seine Einstellung zum ersten Golfkrieg. Dieser Meinungswechsel wurde von Bush in den beiden Debatten (*presidential debate*) ausgenutzt. Der gezeigte Peak in der Kontextvolatilität ist der einzige für *gulf war* in 2004.

Welche Muster im zeitlichen Verlauf für einen Nutzer von Interesse sind, hängt von der jeweiligen Analyseaufgabe ab und kann oftmals nicht vorab spezifiziert werden. Eine vollautomatische Bearbeitung ist daher nicht vorteilhaft bzw. nicht möglich. Abhilfe kann hier geschaffen werden, indem die entsprechenden Verläufe visuell dargestellt werden und dem Benutzer anschließend zur explorativen Auswertung präsentiert werden, während ihm interaktiv nutzbare visuelle Analysewerkzeuge zur Hand gegeben werden. Dem im Umgang mit Medien geübten Benutzer wird damit der Einstieg in eine Recherche erleichtert. Neu gewonnene Einsichten und Hypothesen können anschließend in einer weiterführenden Medienrecherche gezielt untersucht werden.

Hierfür sind verschiedene Arten der Darstellung denkbar. Eine der gängigsten Repräsentationen von zeitlichen Verläufen sind Liniendiagramme. Sie sind intuitiv interpretierbar, haben aber den Nachteil, dass sie nicht gut skalieren hinsichtlich der Menge der gleichzeitig dargestellten Verläufe. Schon wenn vergleichsweise wenige Verläufe ins gleiche Diagramm gezeichnet werden, nimmt die Lesbarkeit deutlich ab. Auch die einzelnen Verläufe untereinander zu zeichnen schafft keine wirkliche Abhilfe, da ein einzelnes Diagramm eine zu große vertikale Ausdehnung hat, als dass viele Verläufe untereinander gezeichnet werden könnten. Die Ausnutzung des zur Verfügung stehenden Platzes, beispielsweise auf einem Bildschirm, ist vergleichsweise gering.

Als alternative Darstellungform bietet es sich an, die Ausschläge der Kurve nicht über die Auslenkung der gezeichneten Punkte in $y$-Richtung (d. h. auf der vertikalen Achse) anzuzeigen, sondern stattdessen über Farbe zu codieren. Jeder Datenpunkt wird dazu als kleines Rechteck (nachfolgend Pixel) dargestellt, dessen Farbe den Datenwert repräsentiert. Da die Rechtecke maximal bis zur Größe eines einzelnen Bildschirm-Pixels geschrumpft werden können, spricht man auch von pixel-basierten Visualisierungen [9].

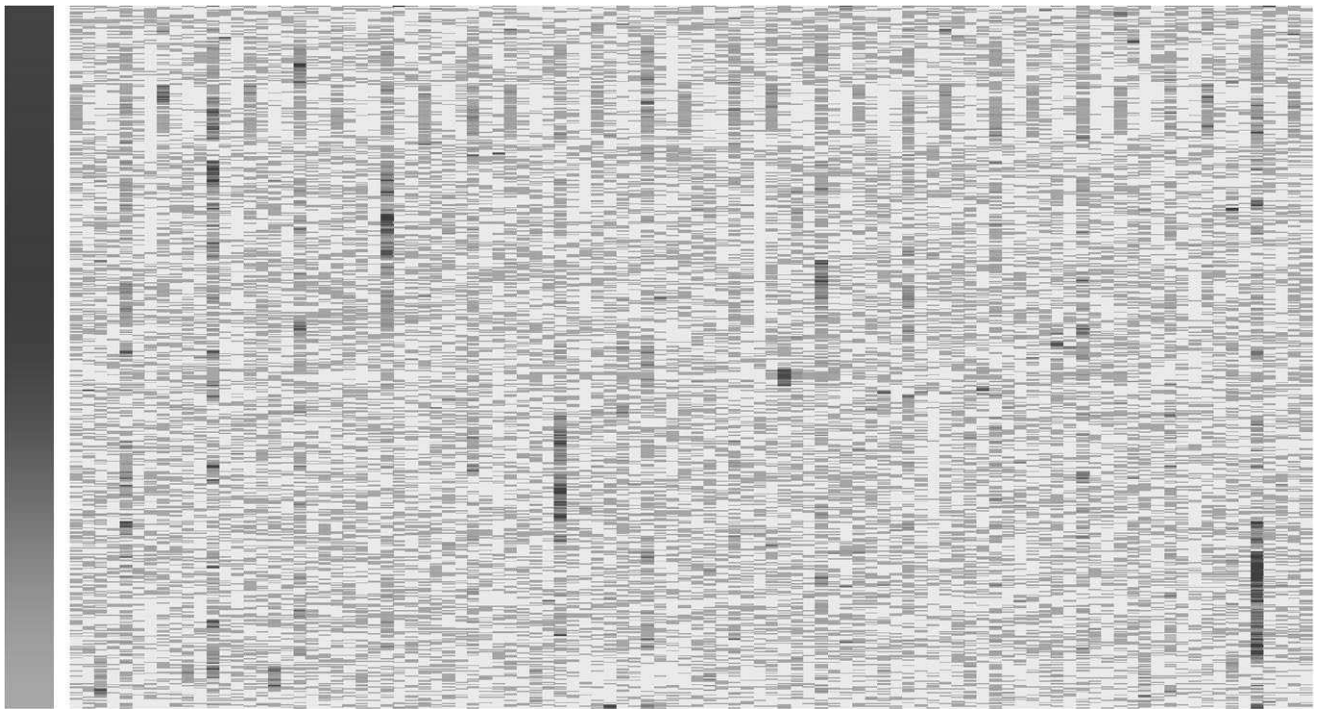


**Abb. 4** Erster Entwurf (Mockup) einer effizienten Visualisierung einer größeren Datenmenge. Jede Zeile stellt den Verlauf eines Wortes dar, jede Spalte eine Zeitscheibe. In der Visualisierung sind Gruppen von Wörtern mit ähnlichen Volatilitätsverläufen zu erkennen

Für die Anordnung der einzelnen Datenpunkte gibt es verschiedene Möglichkeiten. Eine naheliegende Variante ist, die Pixel sequentiell in einer langen Reihe hintereinander zu setzen. Ein solches Layout hat den Vorteil, dass die einzelnen Reihen sehr gut miteinander vergleichbar sind. Abbildung 4 zeigt ein Beispiel für eine solche Darstellung. Eine einzelne Zeile stellt den zeitlichen Verlauf der Volatilitätswerte für einen bestimmten Term dar. Die kleinen Rechtecke, aus denen sich die Zeilen zusammensetzen, repräsentieren den Volatilitätswert für einen bestimmten Tag, wobei im Beispiel ein Gesamtzeitraum von einem halben Jahr gezeigt wird. Gut zu erkennen ist das periodisch wiederkehrende Muster in den ersten Zeilen der Visualisierung. Optisch sichtbar werden hier jeweils die Sonntagsausgaben der analysierten New York Times, welche die Geschehnisse der vergangenen Woche umfangreich zusammenfassen, aufarbeiten und mit anderen Geschehnissen in Zusammenhang bringen.

Darüber hinaus zeigen sich Cluster von Termen, die zum gleichen Zeitpunkt eine Spitze in der Volatilitätskurve aufweisen. Eine Inspektion dieser Term-Cluster kann Aufschluss geben über parallel verlaufende Entwicklungen, die Kontextfluktuation betreffend. Damit solche Muster leicht zu erkennen sind, ist es wichtig, die Terme bezüglich der zu analysierenden Zeiträume in eine sinnvolle Reihenfolge zu bringen. Ähnliche Werteverläufe – die zudem noch invariant gegenüber Skalierung und Verschiebung sind – bezüglich beliebiger Zeiträume für viele Zehntausende Terme zu berechnen, ist nicht trivial und aktuell noch nicht gelöst. Bis dato existieren einige vielversprechende Ansätze, jedoch noch keine zufriedenstellende Lösung für dieses Problem.

Auch diese Darstellung skaliert nur begrenzt, wenngleich sie im Hinblick auf die Anzahl unterschiedlicher Verläufe, die parallel gezeigt werden können, eine deutliche Verbesserung im Vergleich zu Liniendiagrammen darstellt. Die Skalierungsgrenze ist bei pixel-basierten Visualisierungen spätestens dann erreicht, wenn die Anzahl abgebildeter Datenwerte der Anzahl Pixel auf dem Bildschirm entspricht. Sollen darüber hinaus noch weitere Daten dargestellt werden, muss entweder aggregiert werden oder die Darstellung muss in mehrere Abschnitte geteilt werden.

Alternative Anordnungen bieten sich an, wenn es wichtiger ist, möglichst lange Datenreihen abbilden zu können als eine möglichst hohe Anzahl an Verläufen gleichzeitig darzustellen. Würde man am Ende einer Zeile einfach umbrechen und die Datenreihen in der nächsten Zeile fortsetzen, so würde das die Vergleichbarkeit zwischen den Zeitreihen für unterschiedliche Terme erschweren. Außerdem ergeben sich unter Umständen visuelle Artefakte, da Rechtecke, die in der zweidimensionalen Darstellung direkt aneinander angrenzen in der eigentlichen Sequenz weit voneinander entfernt sein können.

Gelöst werden kann dieses Problem, indem kleine Gruppen von Rechtecken gebildet werden. In Abb. 5 wird dies im unteren Bereich sichtbar. Jeweils die sieben aufeinanderfolgenden Datenwerte einer Woche sind hier als Gruppe an-

geordnet (d. h. der Umbruch erfolgt bereits nach jeweils drei Rechtecken). Ein Vergleich zwischen unterschiedlichen Termen ist somit noch problemlos möglich. Abschnitt 5 geht näher auf die Grafik und ihre mögliche Interpretation ein.

## 5 Systembericht Prototyp Explorative Suche

Im Rahmen des Visual-Analytics-SPP[3] der DFG wurde ein Prototyp entwickelt, um über einer größeren Textsammlung eine explorative Suche zu ermöglichen. Dabei wurden die Überlegungen, welche im vorigen Kapitel skizziert wurden, konsequent umgesetzt, um die bestmögliche Ausnutzung des Bildschirms bei gleichzeitig intuitiver Benutzbarkeit zu erreichen.

Wichtige Anforderungen an das zu erstellende System waren (a) eine gute Skalierbarkeit, (b) eine intuitive grafische Ausgabe der Informationen und (c) ein interaktiver Umgang mit den Daten, der Abkehr von der bloßen Visualisierung berechneter Ergebnisse, hin zum effizienten Interagieren mit den Daten über die bereitgestellten Werkzeuge.

Nachfolgend werden die für den Prototyp verwendete Datenquelle und der Prototyp aus Nutzersicht beschrieben, anschließend wird dessen Verwendung für die explorative Suche anhand einiger Beispiele skizziert.

### 5.1 Verwendete Datenquelle

Wie bereits genannt, findet derzeit als Datenquelle das *New York Times Annotated Corpus*[4] Verwendung, welches aus allen Artikeln der gleichnamigen Tageszeitung im Zeitraum 1.1.1987 bis 19.8.2007 besteht. Die im Korpus enthaltenen Metadaten werden nicht verwendet, um eine spätere Anwendbarkeit des vorgestellten Ansatzes auf beliebige (textuelle) Datenquellen gewährleisten zu können. Lediglich der Zeitstempel eines jeden Artikels wird benutzt, um ein Zeitscheibenkorpus mit 7.475 Tageszeitscheiben aufzubauen. Jede Zeitscheibe umfasst dabei etwa 300 (unter der Woche) bis 800 Dokumente (Sonntagsausgabe, mit Wochenrückblick). Um die für die Volatilitätsberechnung notwendigen globalen Kontexte zu ermitteln, wurden alle statistisch signifikanten Kookkurrenzen für alle knapp 7.500 Zeitscheiben berechnet, was zu annähernd 30 Milliarden gewichteten Wortpaaren führte. Details der Berechnung stehen nicht im Fokus dieses Artikels und können in [8] nachgelesen werden. Für Stoppwörter und seltene Wörter wurden keine Kookkurrenzdaten ermittelt. Die Kookkurrenzdaten, welche statistische Informationen über die Signifikanz von Wortpaaren innerhalb einer jeden Zeitscheibe darstellen, dienen als Grundlage für die Berechnung der Kontextvolatilität. Optimierte Datenstrukturen ermöglichen einen schnellen Zugriff auf die umfangreiche Datenbasis, um die Kontextvolatilität interaktiv verwenden zu können.

Mit Hilfe der Kontextvolatilität ist es nun möglich, für jeden Term zu jedem Zeitpunkt den Grad der Kontextfluktuation zu berechnen. Als interessant gelten dabei vor allem die Zeiträume, in welchen die Kontextvolatilität stark ansteigt, da dies auf eine neuartige oder überraschende Verwendung des jeweiligen Terms hinweist. Dem gegenüber ist ein konstant hoher Volatilitätswert an sich nicht von großem Interesse, da es viele Wörter gibt, die schon von der Erwartung her eine hohe Kontextfluktuation aufweisen und eine diesbezügliche Berechnung kaum neue Einblicke in die Daten gewähren würde. Hierzu zählen Stoppwörter ebenso wie hochfrequente Nomen wie „people“, „percent“, „city“ etc., welche aufgrund ihrer Funktion, Ambiguität oder einfach wegen ihres sehr allgemeinen semantischen Begriffs in vielen verschiedenen Kontexten und Domänen eingesetzt werden.

Analog dazu sind Wörter mit konstant niedriger Kontextvolatilität oft ebenso vernachlässigbar. Diese Wörter deuten entweder darauf hin, dass sie nur in unstrittigen Kontexten vorkommen (z. B. eng definierte Fachbegriffe) oder die Datenbasis bezüglich dieser Wörter zu klein ist und deshalb statistische Analysen unzuverlässig sein können.

Daraus folgt, dass für die explorative Suche ebenjene Begriffe und Zeiträume vom System ermittelt werden müssen, welche eine starke Veränderung der Kontextvolatilität aufweisen. Im Prototypen kann ein Nutzer deshalb aktuell einen Zeitraum angeben, innerhalb dessen der Nutzer eine Recherche durchführen möchte. Das System berechnet für alle Wörter und alle Zeitscheiben im Zeitraum die Kontextvolatilität und gibt die unstetesten Terme in Form einer Tagcloud zurück. Farbe und Größe eines Terms in der Tagcloud repräsentieren gleichsam den Rang des Terms innerhalb der Top-N-Liste: das am größten geschriebene Wort weist die höchste Kontextvolatilität auf.

Da gerade Anstiege (oder generell unsteter Charakter) im Volatilitätsverlauf auf interessante Veränderungen hinweisen, wird die Wichtigkeit des Terms als Varianz über die Volatilitätswerte im betrachteten Zeitraum berechnet. Neben der Varianz sind noch einige andere Verfahren implementiert, liefern jedoch subjektiv schlechtere Ergebnisse.

Abbildung 5 zeigt die Ausgabe des Prototypen, nachdem der Nutzer einen zu analysierenden Zeitraum (hier: 2004) angegeben hat. Die Tagcloud mit interessanten Termen befindet sich im oberen Drittel der Visualisierung und kann u. a. im Umfang über Parameter gesteuert werden. Die Tagcloud entspricht in der Berechnung der Liste in Tab. 3 für den jeweils ausgewählten Zeitraum. Ein Klick auf die Begriffe in der Tagcloud visualisiert die jeweiligen Volatilitäts-

[3] http://www.visual-analytics.de/.

[4] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T19.

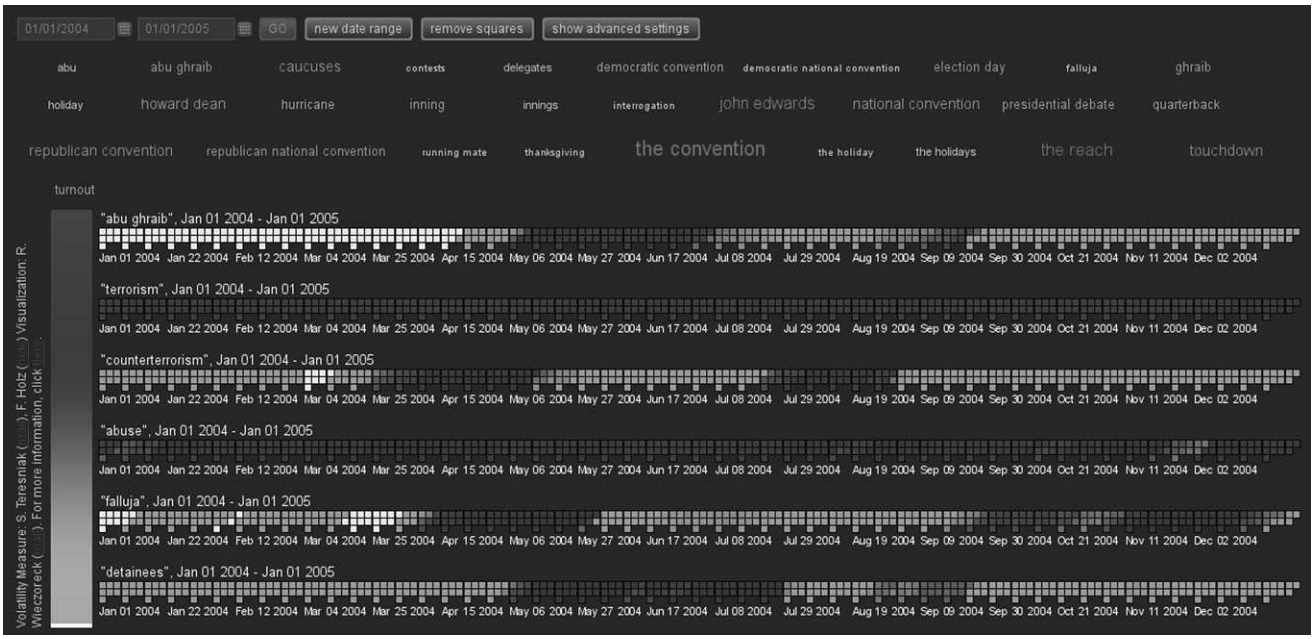



**Abb. 5** Screenshot des Prototyps. Einige Terme wurden manuell zugefügt, andere aus der Top-20-Liste übernommen. Auffällig sind die sich abwechselnden Peaks für *abu ghraib* und *counterterrorism*

verläufe, der Übersichtlichkeit halber gruppiert. (Die Gruppierung wird weiter unten beschrieben.) Beliebige andere Begriffe können über ein Freitextfeld eingegeben werden. Niedrige Volatilitätswerte werden von hell=niedrig bis dunkel=hoch angegeben.[5] Interessant sind nun Zeiträume, in denen sich die Volatilitätswerte – und damit die Farbe – stark verändern. Eine graue Färbung bedeutet, dass im Zeitraum keine Daten vorliegen. Für die Zukunft ist angedacht, dass überhaupt nur Zeiträume angezeigt werden, in denen starke Farb- und damit Wertveränderungen vorliegen. Dieses Feature ist noch nicht komplett implementiert und deshalb im Screenshot nicht zu sehen. Nach Fertigstellung werden dann potentiell mehrere verschiedene „interessante“ Zeiträume für jeden Term angezeigt und mit Datum markiert.

Die Werte – durch Farbpixel repräsentiert – sind im Screenshot in Gruppen zu je sieben Werten (eine Woche) gruppiert und innerhalb einer jeden Gruppe standardmäßig als Z-Pattern angeordnet. Eine Alternative ist die lineare Anordnung der Werte mit einem fest definierten Zeilenumbruch. Beide Möglichkiten der Auslegung können ohne Neuberechnung der Werte jederzeit im Client geändert werden, beispielsweise auf 30-Tage-Gruppierung (6×5 Pixel bei linearer Auslegung der Pixel. Eine 1-Tage-Gruppierung ergibt demnach eine einzelne Linie von Pixeln für jedes dargestellte Wort.

Die angezeigten Pixel offenbaren durch Mouseover den genauen Volatilitätswert und das Datum, welches dieses Pixel repräsentiert. Ein Klick auf ein Pixel zeigt Dokumente an, in welchem der jeweilige Term vorkommt. Hierbei gibt es jedoch urhebberrechtliche Fragestellungen zu beachten, die noch zu klären sind.

## 5.2 Beispiele der Explorativen Suche

Nachdem in den vorangegangenen Abschnitten der Ansatz zur Berechnung interessanter Zeiträume und die grafische Steuerung der Berechnungen inkl. der Visualisierung der Ergebnisse besprochen wurde, soll nun expemplarisch auf einige Beispiele eingegangen werden, die die Verwendung des Prototyps für die explorative Suche verdeutlichen. Die Verwendung der New-York-Times-Daten erleichtert unserer Ansicht nach das Verständnis und die Bewertung der Ergebnisse, da viele ermittelte „interessante“ Terme zu Themen mit nationaler und internationaler Tragweite zuzuordnen sind. Es bleibt anzumerken, dass das beschriebene System prinzipiell für jegliche Textkollektionen funktioniert, sofern die einzelnen Dokumente mit Zeitstempeln versehen sind.

Als Beispiel soll eine Top-down-Analyse der Daten für das Jahr 2004 gelten. Ausgangspunkt ist der Flash-basierte Client (siehe Abb. 5), in dem der Nutzer initial ein Start- und Enddatum für den zu analysierenden Zeitraum angibt, in unserem Beispiel vom 1.1.2004 bis zum 1.1.2005. Ein Klick

[5] Aufgrund des Schwarz-weiß-Drucks des Magazins sind möglicherweise in diesem Artikel beschriebene visuelle Effekte weniger deutlich sichtbar, speziell die recht wichtige Unterscheidung zwischen dem Farbverlauf und Weiß für fehlende Daten. In der Anwendung kann zwischen einer Darstellung im RGB- und HSI-Farbraum umgeschaltet werden.

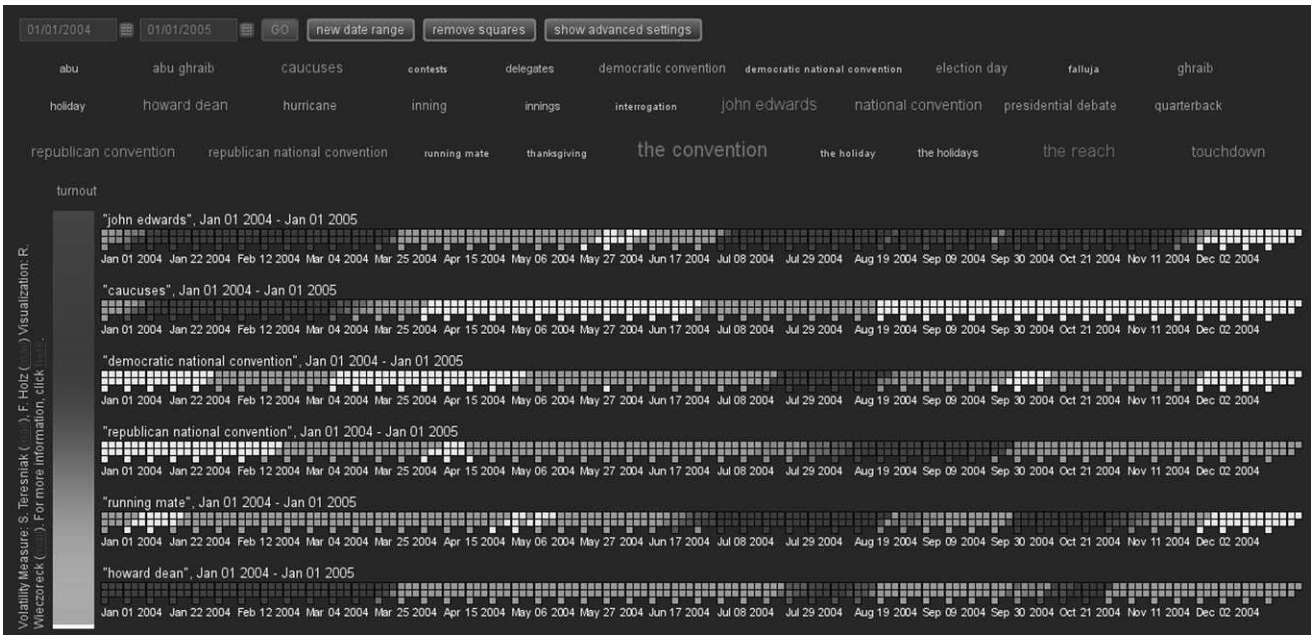



**Abb. 6** Screenshot mit Beispiel einer Recherche zum US-amerikanischen Präsidentschaftswahlkampf 2004

auf den GO-Button am oberen Bildrand liefert nach einigen Augenblicken die (standardmäßig) 20 interessantesten Terme im Zeitraum als Tagcloud (siehe auch Tab. 1).

Abhängig davon, wie gut man sich mit der US-amerikanischen Berichterstattung und dem Wahlsystem auskennt und je nach Recherchezieł können nun einige dieser Terme dem Einstieg in eine tiefere Analyse dienen, oder eigene Terme über ein Freitextfeld ausgewählt werden. Offensichtlich wird jedoch, dass im Zeitraum einige Personennamen hervorgehoben werden: *John Edwards* und *Howard Dean*. Beide spielten als mögliche Kandidaten der Demokraten eine wichtige Rolle in den US-amerikanischen (Vor-)Wahlen in 2004. Sie, wie auch *Dick Gephardt* (unter den Top-50, nicht in der Tagcloud zu sehen), unterlagen im Kandidatenpoker dem späteren Bush-Herausforderer John Kerry (Edwards war *running mate* von Kerry, d. h. Kerrys Wunsch-Vizepräsident). Weitere Begriffe sind direkt mit der damaligen Präsidentschaftswahl verbunden, wie beispielsweise die *caucuses* von Iowa, die verschiedenen *conventions*, wie auch die Debatten (*presidential debates*) der Spitzenkandidaten vor der Wahl selbst. Zweifelsohne war die Präsidentschaftswahl in der amerikanischen Berichterstattung 2004 ein sehr prominentes und kontrovers diskutiertes Thema, für welches alle Leitmedien eingespannt wurden. Der Wahlkampf kann als Einstieg in eine Recherche zur US-amerikanischen Innenpolitik dienen.

Ein anderes Themenfeld in der Tagcloud sind die Geschehnisse im amerikanisch-geführten Gefängnis von Abu Ghraib, in welchem Insassen durch Militär gefoltert und gedemütigt wurden. Der Fall erregte auch in Deutschland großes Aufsehen und bedarf inhaltlich keines Kommentars. Terme wie *abu ghraib prison*, *falluja* oder *interrogation* könnten als Einstieg in eine Recherche zur US-amerikanischen Außenpolitik dienen.

Der Term *hurricane* bezieht sich auf den Hurrikan Katrina, welcher starke Verwüstung im Süden der USA angerichtet hat und zu einer kontroversen Berichterstattung bezüglich des Katastrophenmanagements des damaligen US-Präsidenten Bush geführt hat. Als eine der verheerendsten Naturkatastrophen in der Geschichte der USA ist dieser Hurrikan nachvollziehbar ein wichtiges Ereignis in 2004 und damit ein sinnvoller Einstiegspunkt in weitere Recherchen.

Weitere Ereignisse beziehen sich auf sportliche Events, sowohl im Inland (*innings*, *quarterback*, *touchdown*), die aus europäischer Sicht weniger „sprechend“ sind, wie auch auf die Olympischen Spiele von Sydney.

Manche Terme können nicht klar einer Thematik zugeordnet werden (z. B. *the reach*); diese Fragmente erscheinen im Umfang jedoch nicht störend. Wortbestandteile von Mehrwortbegriffen, wie beispielsweise *abu* als Teil von *abu ghraib*, ist auf die Art und Weise zurückzuführen, wie Mehrwortbegriffe dem System bekannt gemacht wurden: Zum einen wurden Mehrwortbegriffe tendenziell eher übergeneriert, zum anderen wurden für jeden Mehrwortbegriff auch die einzelnen Wortbestandteile selbst als Terme in der Analyse belassen.

Der Nutzer muss sich nunmehr entscheiden, zu welchen Termen (oder welchen Themenkomplexen) er recherchieren möchte.

Die US-Präsidentschaftswahl soll nachfolgend als Beispiel vertieft werden. Werden einzelne Begriffe angeklickt und deren Kontextfluktuation über das Jahr 2004 beobach-

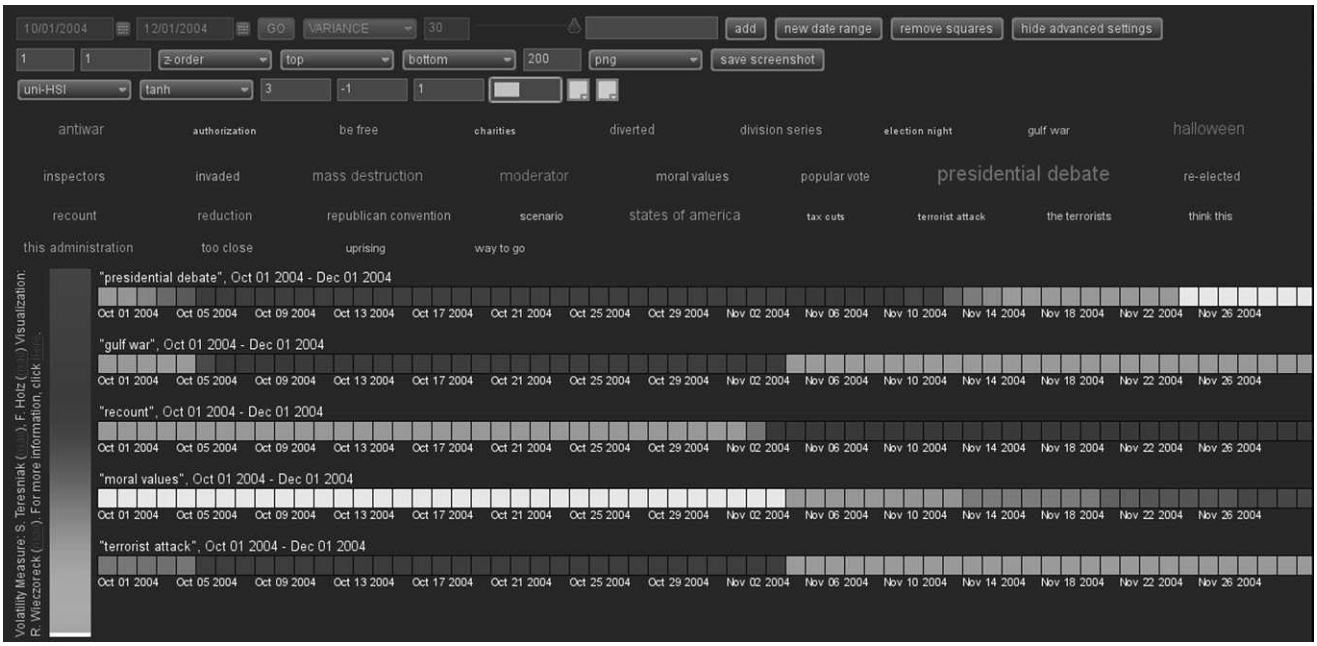



**Abb. 7** Der Zeitraum um die US-Präsidentschaftswahl am 2.11.2004. Terroristische Angriffe (*terrorist attack*) wurden im Wahlkampf mehrfach thematisiert. Der starke Anstieg der Kontextvolatilität dieses Terms korreliert im Zeitraum nicht mit der Termfrequenz

tet, wie in Abb. 6 dargestellt, lassen sich mehrere Beobachtungen machen:

- Trivialerweise wird die zeitliche Abfolge von Themen erkennbar, speziell wenn es sich dabei um Ereignisse handelt. Man sieht beispielsweise, dass der republikanische Nominierungsparteitag (*republican national convention*, 30.8. bis 2.9.2004) zeitlich nach dem demokratischen Parteitag (*democratic national convention*, 26.7. bis 29.7.2004) abgehalten wurde.
- Howard Dean war recht frühzeitg als Spitzenkandidat aus dem Rennen, war jedoch noch weiterhin Thema im Wahlkampf, wenn auch mit niedrigerer Intensität. Weitere Analysen zeigten, dass sich der Frequenzverlauf von *howard dean* signifikant vom Volatilitätsverlauf unterscheidet.
- Obwohl zu den Nominierungsparteitagen der Präsidentschaftskandidat wie auch der Kandidat für das Amt des Vizepräsidenten festgelegt wurde (sog. *running mate*), so waren die Vize trotzdem noch Thema. Begründet werden kann dies u. a. dadurch, dass die Vizekandidaten (Edwards und Cheney) am 5. Oktober 2004 in einer öffentlich ausgestrahlten Debatte diskutierten.

In einem nächsten, tiefer gehenden Schritt würde auf Dokumentebene gearbeitet. Die hierfür benötigte Funktionalität ist noch nicht komplett implementiert und rechtliche Gründe würden einer Veröffentlichung von Dokumentteilen ebenfalls im Wege stehen. Weitere Schritte der Recherche wären das weitere Eingrenzen der betrachteten Zeiträume, wie Abb. 7 zeigt, und die Analyse der dann feingranularer ermittelten interessanten Terme, speziell im Übergang von der Betrachtung von Tageszeitscheiben (New-York-Times-Ausgaben) auf Dokumentebene (Artikel).

Das oben angeführte Beispiel soll verdeutlichen, wie mithilfe von interaktiven, visuellen Verfahren aus großen (tendenziell) unbekannten Dokumentkollektionen die wichtigsten Terme identifiziert, und über diese dann interessante Zeitabschnitte ermittelt werden können.

Nach Fertigstellung des Prototyps und Klärung der urheberrechtlichen Belange ist angedacht, die hier vorgestellte Funktionalität einem breiteren Publikum per Webbrowser oder Webservice zur Verfügung zu stellen.

## 6 Zusammenfassung

In diesem Artikel wurde ein Ansatz für eine explorative Suche über großen Textmengen vorgestellt, indem Mechanismen der Visual Analytics benutzt wurden. Mit dem Maß der Kontextvolatilität existiert ein Maß, um interessante Terme zu gewichten. Dabei werden andere Ergebnisse erzielt als mit dem populären (frequenzbasierten) Termwichtungsmaß tf·idf. Die Kontextvolatilität dient der Top-Down-Analyse von großen Textkollektionen, was am Beispiel von Newsdaten skizziert wurde. In einem interaktiven, grafischen Prozess kann der Nutzer die Recherche verfeinern und so einzelne behandelte Themen eingrenzen, um diese genauer zu untersuchen. Weiß der Nutzer anfangs nicht, welche Aspekte in der Kollektion behandelt werden, so hat er nun eine

Möglichkeit der effizienten Exploration von ihn interessierenden Themenbereichen. Auf diese Weise liefert das hier vorgestellte System eine wertvolle Hilfestellung bei der Erschließung neuer Datenquellen und dem drill-down in neue Thematiken. Die verwendeten Verfahren sind dabei auf beliebige, mit Zeitstempeln versehene Textkollektionen anwendbar. Die dabei gelieferten Ergebnisse – interessante Zeiträume wie auch interessante Terme – können nun für eine tiefere Analyse der Ausgangsdaten genutzt werden. Die im Artikel verwendeten Nachrichtendaten können beispielsweise Journalisten, Historikern oder Medienwissenschaftlern zur Recherche dienen, ohne dass Millionen von Zeitungsartikeln gesichtet werden müssen. Ein aktuelles Anwendungsszenario ist die Analyse von Patent- und Technologiedokumenten zur Unterstützung von Innovationsprozessen in kleinen und mittelständigen Unternehmen.

## Literatur


1. Broder A (2002) A taxonomy of web search. SIGIR Forum 36(2):3–10
2. Couvering EV (2005) Web behaviour: search engines in context
3. Fluit C, Sabou M, van Harmelen F (2005) Ontology-based information visualisation: towards semantic web applications. In: Geroimenko V (Hrsg) Visualising the semantic web, 2nd Aufl. Springer, Berlin
4. Geng L, Hamilton HJ (2006) Interestingness measures for data mining: a survey. ACM Comput Surv 38(3)
5. Gottwald S, Richter M, Heyer G, Scheuermann G (2008) Tapping huge temporally indexed textual resources with wctanalyze. In: Proceedings of the LREC 2008, Marrakech, Morroco
6. Heim P, Ziegler J (2011) Faceted visual exploration of semantic data. In: Achim E, Allan D, Gershon N, Pohl M (Hrsg) Human aspects of visualization. Lecture notes in computer science, Bd 6431. Springer, Berlin, S 58–75
7. Heyer G, Holz F, Teresniak S (2009) Change of topics over time—tracking topics by their change of meaning. In: Proc of int conf on knowledge discovery and information retrieval (KDIR '09)
8. Holz F, Teresniak S (2010) Towards automatic detection and tracking of topic change. In: Gelbukh A (Hrsg) Proc Iaşi: conference on intelligent text processing and computational linguistics (CICLing 2010). Lecture notes in computer science, Bd 6008. Springer, Berlin
9. Keim DA (2000) Designing pixel-oriented visualization techniques: theory and applications. IEEE Trans Vis Comput Graph 6(1):59–78
10. Keim DA, Kohlhammer J, Ellis G, Mansmann F (2010) Mastering the information age—solving problems with visual analytics. In: Eurographics
11. Kleinberg J (2002) Bursty and hierarchical structure in streams. In: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining (KDD '02), New York, NY, USA. ACM, New York, S 91–101
12. Kumaran G, Allan J (2004) Text classification and named entities for new event detection. In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '04), New York, NY, USA. ACM, New York, S 297–304
13. Marchionini G (2006) Exploratory search: from finding to understanding. Commun ACM 49:41–46
14. Matthews M, Tolchinsky P, Blanco R, Atserias J, Mika P, Zaragoza H (2010) Searching through time in the New York Times. In: HCIR Challange 2010
15. Rasmussen E (2009) Characterizing, supporting and evaluating exploratory search. In: Belkin N, Marchionini G (Hrsg) Proceedings of the NSF workshop, information seeking support systems, Chapel Hill, University of Carolina, S 30–32
16. Rohrdantz C, Koch S, Jochim C, Heyer G, Scheuermann G, Ertl T, Schütze H, Keim DA (2010) Visuelle textanalyse. Inform.-Spektrum 33:601–611. 10.1007/s00287-010-0483-x
17. Shneiderman B, Byrd D, Croft BW (1997) Clarifying search: a user-interface framework for text searches. D-Lib Mag
18. Soboroff I, Harman D (2005) Novelty detection: the TREC experience. In: In HLT/EMNLP, S 105–112
19. Swan R, Allan J (1999) Extracting significant time varying features from text. In: Proceedings of the eighth international conference on information and knowledge management (CIKM '99), New York, NY, USA. ACM, New York, S 38–45
20. Teresniak S, Heyer G, Scheuermann G, Holz F (2009) Visualisierung von bedeutungsverschiebungen in großen diachronen dokumentkollektionen. Datenbank-Spektrum 31:33–39
21. Ueberall M, Drobnik O (2007) Facet-based exploratory search in topic maps. In: Maicher L, Garshol LM (Hrsg) Proc fourth int'l conference on topic maps research and applications (TMRA), 2007. Leipziger Informatik-Verbund, Leipzig
22. Waitelonis J, Knuth M, Wolf L, Hercher J, Sack H (2010) The path is the destination—enabling a new search paradigm with linked data. In: Proc of linked data in the future internet at the future internet assembly, S 700
23. Wang X, McCallum A (2006) Topics over time: a non-Markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '06), New York, NY, USA. ACM, New York, S 424–433