Visuelle Textanalyse

Interaktive Exploration von semantischen Inhalten

Christian Rohrdantz · Steffen Koch
Charles Jochim · Gerhard Heyer
Gerik Scheuermann · Thomas Ertl
Hinrich Schütze · Daniel A. Keim

Einleitung

Sowohl die fortschreitende Digitalisierung und Automatisierung im Bereich redaktioneller Veröffentlichungen und in der Wissenschaft und Forschung als auch die stetig wachsende Zahl nutzergenerierter Inhalte im Web 2.0 führen zu einer enormen textuellen Informationsflut. Dabei übersteigt die Größe einzelner Dokumentsammlungen schnell das, was ein Mensch in annehmbarer Zeit erfassen kann. Während je nach Dokumentsammlung und Informationsbedürfnis ganz spezielle Herausforderungen entstehen, bleibt in allen Fällen die gemeinsame Datengrundlage: Information und Wissen liegen als natürlichsprachlicher Text vor. Anders als bei anderen Einsatzgebieten visueller Datenanalyse liegt hier ein Hauptproblem in der Verarbeitung der, zumindest aus maschineller Sicht, weitgehend unstrukturierten und ambiguen Textdaten. Die Herausforderungen liegen zum einen in der Modellierung und Erfassung semantischer Aspekte und Inhalte und zum anderen in der Vermittlung der automatisch erfassten Information an den Nutzer.

Einige elementare Textanalyseschritte lassen sich bereits heute effizient mit automatischen Algorithmen lösen, die oft auf statistischen Verfahren beruhen. Man denke beispielsweise an die Suche nach Dokumenten aus einer großen Sammlung, welche im Bezug auf eine einfache Suchanfrage möglicherweise relevant sind. Automatische Verfahren stoßen jedoch dann an ihre Grenzen, wenn für die Lösung einer Analyseaufgabe semantisches Textverständnis erforderlich ist oder eine Analyse explorativer Natur ist, das heißt das Ziel der Analyse im Voraus nicht genau feststeht.

In solchen Fällen gilt es zunächst einmal automatische Verfahren zu entwerfen, welche interessante und relevante Aspekte der Daten auswerten. Darauf aufbauend muss es dem Nutzer ermöglicht werden, mit den Ergebnissen dieser Verfahren zu interagieren, um zur Lösung der Analyseaufgabe zu kommen. Interaktionsmöglichkeiten zwischen menschlichen Benutzern und automatischen Verfahren zur Verarbeitung von Massendaten sind ein fester Bestandteil von Ansätzen aus dem Bereich der visuellen Datenanalyse.

Im weiteren Verlauf des Artikels wird zuerst das Forschungsfeld der visuellen Textanalyse genauer beschrieben und dann werden beispielhaft

DOI 10.1007/s00287-010-0483-x

Christian Rohrdantz · Daniel A. Keim
Arbeitsgruppe Datenbanken, Datenanalyse und Visualisierung, Fachbereich Informatik und Informationswissenschaft, Universität Konstanz,
Universitätsstrasse 10, 78464 Konstanz
E-Mail: {christian.rohrdantz, daniel.keim}@uni-konstanz.de

Steffen Koch · Thomas Ertl
Institut für Visualisierung und Interaktive Systeme,
Universität Stuttgart,
Universitätsstrasse 38, 70569 Stuttgart
E-Mail: {steffen.koch, thomas.ertl}@vis.uni-stuttgart.de

Charles Jochim · Hinrich Schütze
Institut für Maschinelle Sprachverarbeitung,
Universität Stuttgart,
Azenbergstr. 12, 70174 Stuttgart
E-Mail: jochimcs@ims.uni-stuttgart.de, hs999@ifnlp.org

Gerhard Heyer
Abteilung Automatische Sprachverarbeitung,
Institut für Informatik, Universität Leipzig,
Johannisgasse 26, 04103 Leipzig
E-Mail: heyer@informatik.uni-leipzig.de

Gerik Scheuermann
Abteilung Bild und Signalverarbeitung,
Institut für Informatik, Universität Leipzig,
Johannisgasse 26, 04103 Leipzig
E-Mail: scheuermann@informatik.uni-leipzig.de

Zusammenfassung
Methoden und Techniken zur automatischen Verarbeitung und inhaltlichen Erfassung großer Mengen an Textdokumenten haben in den vergangenen Jahren enorm an Bedeutung gewonnen. Während einerseits die Verfügbarkeit und der Zugang zu digitalisierten Textdokumenten bis dato in ungeahntem Maße gestiegen sind, erweist sich die Erfassung des semantischen Inhalts solcher Dokumentsammlungen als problematisch. Dem expandierenden Forschungsfeld der visuellen Textanalyse und Textvisualisierung kommt dabei eine Schlüsselrolle bei der Lösung von Problemstellungen aus der Praxis zu. Anhand aktueller Anwendungsbeispiele und einem Überblick über den Stand der Forschung erläutert dieser Artikel die vielfältigen Möglichkeiten, die sich durch visuelle Textanalyse ergeben.

drei Anwendungsbereiche der visuellen Textanalyse vorgestellt, die eine hohe praktische Relevanz haben und denen vonseiten der Wissenschaft und Wirtschaft großes Interesse zuteilwird. Dabei handelt es sich um die visuelle Analyse a) von Themenentwicklungen in Nachrichten über die Zeit, b) von Meinungen, die im Internet geäußert werden und c) von Patentdatenbanken.

Visuelle Textanalyse

Unter visueller Textanalyse versteht man im Allgemeinen die Kombination automatischer Textanalyseverfahren und interaktiv manipulierbarer Visualisierungen. Vereinfacht gesagt geht es in erster Linie darum, den Benutzer von der Notwendigkeit zu befreien, große oder sogar alle Teile des verfügbaren Texts zu lesen, um die (für ihn) relevanten Inhalte einer Textsammlung zu erfassen und zu überblicken. In vielen Fällen wäre ein sequenzielles Lesen des gesamten Textkorpus sogar unmöglich, da dies viel Zeit erfordern würde. Computer können dahingegen annähernd beliebig große Mengen von Text verarbeiten, statistisch analysieren und riesige Resultatmengen speichern. Allerdings fehlt dem Computer die menschliche Fähigkeit, den präsentierten Text inhaltlich zu verstehen.

Ausgehend davon ist ein geeignetes Zusammenspiel maschineller Textverarbeitung und menschlicher Kognition eine gute Lösung, um große Dokumentmengen zu explorieren. Grundlegende Visualisierungen können helfen, einen Überblick über die Datenlage zu erhalten. Meist lässt sich der konkrete Informationsbedarf von Nutzern anhand ihrer Anforderungen und den verfügbaren Daten identifizieren. Dann gilt es zuerst einmal automatische Verfahren zu entwickeln, die in der Lage sind, die zur Befriedigung des konkreten Informationsbedarfs relevanten Informationen zu extrahieren. Die reine Extraktion ist jedoch ohne eine geeignete Vermittlung der Informationen wertlos. Hier kommen wieder Visualisierungen ins Spiel, die sich hervorragend dazu eignen, Information zu vermitteln. Schließlich ist das menschliche Auge der Informationseingangskanal mit der größten Bandbreite. Statische Bilder können aber nur bedingt helfen, denn oft gilt es, visuellen Auffälligkeiten auf den Grund zu gehen, indem andere Sichten auf die Daten gewählt werden oder auch auf entsprechende Stellen im Volltext zugegriffen wird. Diese Interaktionen sind notwendige Schritte, um Verständnis zu fördern und Hypothesen generieren zu können. Denn immer ist auch die Möglichkeit gegeben, dass es bei automatischen Verfahren zu Fehlern kommt, die sich in der Ambiguität und Komplexität natürlicher Sprache begründen.

Automatische Methoden zur Textanalyse haben ihren Ursprung häufig in den Feldern der Computerlinguistik und des Natural Language Processing [8], dem Information Retrieval [13] sowie der jüngeren Disziplin des Text Minings [4]. Im methodischen Teil weisen diese Felder große Überlappungen auf. Ein Beispiel sind Parser, die Wortarten, Satzstrukturen und grammatikalische Abhängigkeiten ermitteln, wie beispielsweise der häufig verwendete Stanford Parser (http://nlp.stanford.edu/software/lex-parser.shtml). Solche Verfahren beruhen zumeist auf statistischen Ansätzen und Modellen, die mit manuell annotierten Texten trainiert wurden. Außerdem gibt es eine Vielzahl von Methoden, die versuchen, Textsemantik aus der Syntax abzuleiten und die sich dabei auf manuell erstellte Regelsätze und Wissensbasen stützen. Ein populäres Beispiel für eine Wissensbasis ist die von Psychologen erstellte lexikalische Datenbank Wordnet (http://wordnet.princeton.edu/). Ein prominentes und verbreitetes Verfahren aus dem Information Retrieval ist die Latent Semantische Analyse [12], die unter anderem bei der Identifikation von Themen und Konzepten helfen kann.

Abstract
Methods and techniques for the automatic processing and content exploration of large amounts of text documents have increasingly gained importance over the last years. While the availability and accessibility of digitalized text documents is greater than ever, one major problem is to capture the semantic content of document collections. The growing research field of visual text analysis and text visualization plays a key role in solving such real-world problems. This article is dedicated to outlining the opportunities that arise with visual text analysis by introducing innovative application examples and providing a research overview.

Eine generelle Tendenz ist, dass sich mit steigender Textmenge die Effektivität statistikbasierter Methoden im Vergleich zu komplexen linguistischen Ansätzen klar verbessert. Außerdem sind statistische Verfahren in der Regel deutlich effizienter und spielen deshalb bei den großen Textmengen, die heutzutage zur Verfügung stehen, die Hauptrolle. Nichtsdestotrotz werden auch einfache linguistisch motivierte Methoden eingesetzt, wie etwa eine morphologische Reduktion von Wörtern auf Wortstämme oder Grundformen. Abgesehen von spezialisierten Methoden der Textanalyse kommen auch allgemeine Methodiken des Data Mining [5], wie Klassifikationsverfahren, Clusteralgorithmen und Assoziationsregeln, zum Einsatz.

In der visuellen Textanalyse gilt es, bestehende Methoden weiterzuentwickeln, um sie eng mit geeigneten interaktiven Visualisierungen zu koppeln. Dabei sollte eine Skalierbarkeit im Bezug auf die zu verarbeitende Datenmenge gewährleistet sein, vor allem was die Laufzeitkomplexität der automatischen Verfahren angeht. Häufig müssen Verfahren zudem auf spezielle Aufgaben zugeschnitten werden. Einige konkrete Beispiele für den Einsatz visueller Textanalyse werden in den folgenden Abschnitten näher beschrieben.

Visuelle Analyse von Themenveränderungen über die Zeit

Durch die immer weitergehende Benutzung des Internets, besonders des WWW, als Kommunikationsmedium in nahezu allen Bereichen des täglichen Lebens, wuchs und wächst die Menge an Text, auf die potenziell in kurzer Zeit zugegriffen werden kann, schnell und stark an. Google und Co. helfen dabei, bestimmte Informationen in Teilen dieser Datenflut zu finden, aber wie findet man Neuigkeiten, ein bestimmtes Thema betreffend? Wie behält man den Überblick über neue Entwicklungen und Trends, beispielsweise in seiner Branche? Eine überblicksartige visuelle und interaktiv verfeinerbare Aufbereitung speziell nach Neuigkeitswert gefilterter Daten kann helfen.

Ein geflügeltes Wort über das Internet lautet "the net never forgets". Neue wie alte Informationen vermischen sich und meist ist das Alter oder Erstellungsdatum von Dokumenten nicht klar ersichtlich oder nicht indexiert. Eng verwandt ist die Beobachtung von Communities und Äußerungen von Nutzern in Blogs, Webseiten, Twitter und über beliebige andere Kanäle. Nachrichtenmeldungen, z. B. von Onlinezeitungen, sind ebenfalls Informationsströme von beträchtlichem Volumen. In diesem Zusammenhang hat sich die Medienresonanzanalyse oder Media Monitoring als Wirtschaftsfaktor am Markt etabliert. Es geht darum, Auswirkungen von Aktionen einzelner Akteure (Firmen, NGOs oder beliebige andere Vereinigungen) im Netz zu beobachten oder eine Art „Frühwarnsystem" zu bieten, um aufkeimende Ressentiments (z. B. gegen das eigene Unternehmen) oder auch neue Entwicklungen beim Konkurrenten zeitnah zu erfassen.

Nachfolgend soll ein Ansatz zur Erkennung von Neuigkeiten als „heiß diskutierten Themen" vorgestellt werden. Die Idee ist dabei vom etablierten Forschungsschwerpunkt Topic Detection and Tracking [1] insofern verschieden, als dass nicht nur neue Themen aus großen Datenströmen gefiltert werden sollen, sondern auch neue Nuancen und Informationen innerhalb von etablierten Themen entdeckt werden können. Hierbei wird vorausgesetzt, dass ein Neuigkeitsgewinn vor allem dann feststellbar ist, wenn sich der Verwendungskontext eines Wortes stark verändert.

Vom Novelty-Detection-Track der Text Retrival Conference (TREC) unterscheidet sich der hier vorgestellte Ansatz in dem Punkt, dass das System ohne Aufsicht (unsupervised) und ohne vordefiniertes Wissen (knowledge free) arbeitet, um zum einen eine bessere Skalierbarkeit auf großen Datenmengen zu erreichen und zum anderen die Arbeit für

das Training beim Wechsel auf neue Datenquellen zu minimieren. In der Novelty Detection ist das Ziel, bezüglich eines Themas möglichst relevante Dokumente aus einer Datenkollektion zu extrahieren und in diesen genau die Passagen (meist Sätze) zu markieren, welche einen hohen Neuigkeitswert besitzen, also bisher unbekannte Informationen bieten [17]. Der vorliegende Ansatz legt wenig Gewicht auf das Retrievalproblem, identifiziert neben neuen Informationen jedoch auch neue Kombinationen von alten Informationen.

Von Wörtern zu Bedeutung zu Bedeutungsverschiebung

Bevor das Vorgehen erläutert wird, sollen ein paar Grundannahmen und Definitionen dargelegt werden, auf denen das Modell fußt. Die Bedeutung eines Wortes sei durch die Verwendung dieses Wortes innerhalb seines globalen Kontexts gegeben. Die Fundamente des vorliegenden Ansatzes sind hierbei im Strukturalismus zu finden, welcher durch Ferdinand de Saussure begründet wurde [2]. Genauer sei der globale Kontext eines Wortes die Menge aller statistisch signifikanten Kookkurrenzen, also genau der Wörter, welche statistisch gesehen auffallend häufig gemeinsam mit dem betreffenden Wort innerhalb eines Textfensters auftreten. Übliche Fenstergrößen sind direkte Nachbarn, Sätze, Absätze, Dokumente. Statistische Auffälligkeit bedeutet, dass zwei Wörter häufiger gemeinsam auftreten, als dies durch Zufall zu erwarten wäre.

Jedes Wort hat somit zu jedem Zeitpunkt, an welchem es verwendet wird, Kookkurrenzen, die den Kontext bilden. Wird ein Wort nun über die Zeit betrachtet mit neuen oder unerwarteten Wörtern kombiniert, beispielsweise weil Themen kontrovers diskutiert werden oder neue Fakten in die Diskussion eingebracht werden, dann ändert sich der Kontext dieses Wortes im obigen Sinne. Die Änderung der Kontexte ist maschinell und auch für große Textmengen messbar. Die Beobachtungen der Verwendung von Wörtern in der Vergangenheit prägen die erwartete Verwendung in der Zukunft, was eben gerade das Messen der Abweichungen von dieser Erwartung ermöglicht. Populäre deutsche Beispiele, die eine Wandlung durchlebt haben, sind in jüngster Vergangenheit „Terrorismus“, „Manager“ und „Globalisierung“.

Das Maß, welches für jeden Zeitpunkt angibt, wie stark ein Wort von dessen erwarteter thematischer Verwendung abweicht, haben wir Volatilität genannt. Dieser Begriff ist der Ökonometrie entliehen, in welcher die Volatilität als Risikomaß ein Gradmesser für die Schwankung von Werten von Assets darstellt. Die Bedeutungsvolatilität misst die Schwankung in den Bedeutungen von Wörtern. Allerdings sind die Berechnungsgrundlagen deutlich andere. Für detailliertere Informationen und den Algorithmus sei auf [7] oder [18] verwiesen. Das Verfahren basiert darauf, dass die nach Signifikanz sortierten Kookkurrenzen von Wörtern über die Zeit verglichen werden.

Experimente

Das Vorgehen soll am Beispiel von Nachrichtentexten der New York Times verdeutlicht werden. Für Analysen wurden alle Artikel dieser Zeitung verwendet, welche zwischen 1987 und 2007 erschienen. Diese Artikel wurden als Datenstrom betrachtet, die auf täglicher Basis (inkl. Sonntagsausgabe) bereit standen. Wie oben beschrieben können nun für jeden Tag die Kookkurrenzen eines jeden Wortes berechnet werden (wobei Stoppwörter und sehr seltene Wörter nicht berücksichtigt werden). Ändert sich die Verwendung eines Wortes – beispielsweise „Iraq“ – über die Zeit, indem es in anderen Zusammenhängen genannt wird, soll dies dem Benutzer mitgeteilt werden, um eine tiefere Analyse innerhalb des interessanten Zeitpunkts vornehmen zu können. Für das Wort „Iraq“ kann dies bedeuten, dass in den 1990er-Jahren „Iran“, „Kuwait“, „Bill Clinton“ und anderen Terme relevant sind, wohingegen zu Zeiten des zweiten Golfkriegs eher „Terrorism“, „George W. Bush“ und „Abu Ghraib“ starke Kookkurrenzen sind. Dieses Beispiel ist stark vereinfacht, verdeutlicht aber das Prinzip, dessen Wirksamkeit auf den realen Daten nachgewiesen werden kann. So ließ sich u. a. zeigen, dass die Vorfälle im Gefängnis von Abu Ghraib von Mahmud Ahmadinedschad als Propagandawerkzeug missbraucht und die Fotos aus Abu Ghraib in New York in Museen ausgestellt wurden. Der letztgenannte Fakt war Teil einer sehr kurzen Meldung und wäre, würde man lediglich auf Wortfrequenzen schauen, um populäre Entwicklungen zu errechnen, nicht aufgefallen. Dass die Berechnung der Volatilität von Wörtern unabhängig von der reinen Termfrequenz ist, ist ein wichtiges Ergebnis der Modellierung.

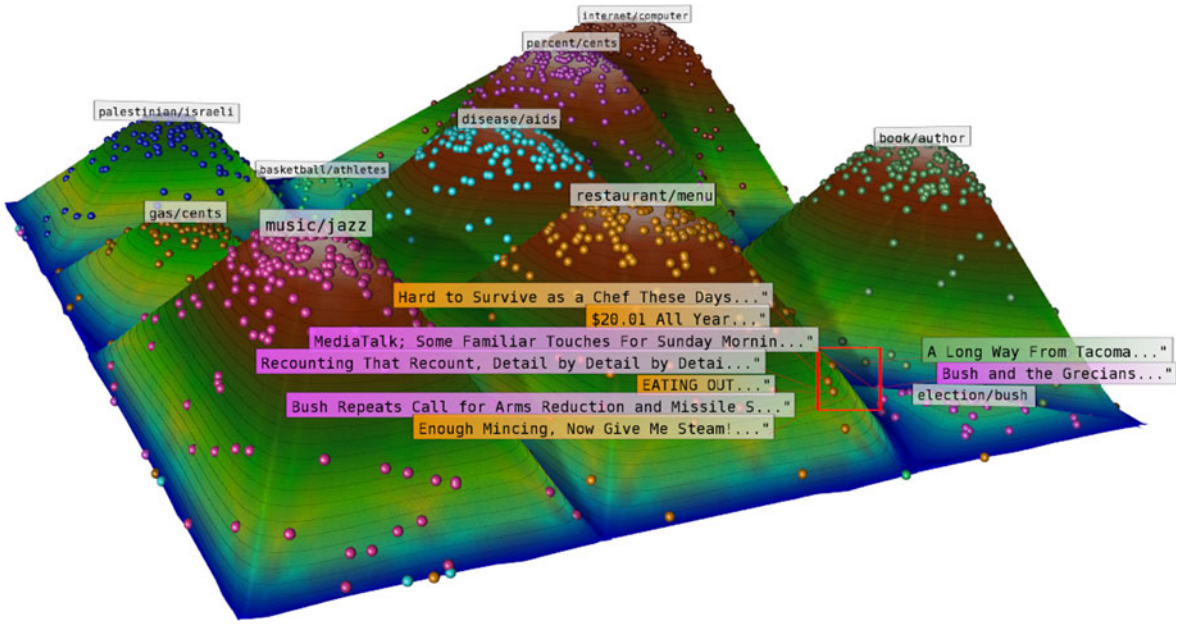

Abb. 1 Themenlandschaft zu 1896 Zeitungsartikeln aus der New York Times für eine einzelne Zeitscheibe im Jahr 2001. Kugeln und Berge repräsentieren Dokumente und Themen. Die zwei wichtigsten Worte aller Dokumente eines Berges dienen als Bergbeschriftung (weisen also auf das Thema hin). Dokumente in der beweglichen Linse werden mit ihren Titeln beschriftet

Visualisierung

Die Errechnung der Bedeutungsverschiebung ist nur ein erster Schritt im Analyseprozess, welcher für sich als den Benutzer unterstützend entwickelt wird. Ganz im Sinne des Visual Analytics sollen die großen und teilweise widersprüchlichen Daten zuerst vorgefiltert und anschließend in interaktiver Weise präsentiert werden. Die Visualisierung soll als Werkzeug und „Fenster in die Daten“ dienen, somit nicht nur das Ergebnis darstellen, sondern einen Indikator liefern, der weitere Recherchen anstößt, vereinfacht oder beschleunigt.

Nachdem nun die Volatilität für beliebige Wörter genau die Zeiträume identifiziert hat, in denen Interessantes (im Sinne von neuen Entwicklungen) zu erwarten ist, musste ein Weg gefunden werden, um diese Informationen einfach erfassbar darzustellen. Das Informationsbedürfnis und das Rechercheziel des Benutzers können dem System a priori nicht bekannt sein, was einen flexiblen Umgang mit den Daten erfordert, die wiederum in großem Umfang vorliegen (und mit der Zeit weiter anwachsen können). Dem Nutzer wird deshalb überblicksartig die thematische Zusammensetzung der Dokumentkollektion unter Hervorhebung der bedeutungsvolatilen Bereiche präsentiert, um ihm daraufhin die Möglichkeit zu eröffnen, Teilbereiche der Daten für tiefere Analysen zu vergrößern. Die Zoomstufe bestimmt dabei den Detailgrad der dargestellten Informationen bis hinunter auf die Dokumente selbst.

Hierfür ist aktuell eine Visualisierung in Arbeit, welche thematisch verwandte Dokumente zu Themenbergen zusammenfasst und Dokumente mit neuen Entwicklungen auszeichnet. Der Durchmesser der Themenberge spiegelt dabei die Wichtigkeit des Themas wider. Erste Ergebnisse für einen einzelnen Zeitschritt sind in [15] nachzulesen und Abb. 1 zu entnehmen.

Es ist geplant, die Visualisierung als sedimentartige 3D-Landschaft zu erweitern, in welcher einzelne Zeitscheiben den jeweiligen Sedimenten entsprechen. Die Themen der Dokumente werden in jeder Schicht zweidimensional ausgelegt und die dritte Dimension (die Höhe der übereinander angeordneten Schichten) repräsentiert dann die Zeit, um die Verläufe der Themen geografisch darstellen zu können. Vereinfacht ausgedrückt ist eine dreidimensionale Version des recht populären Themerivers angedacht [6], bei der die Geologie von Sedimenten als visuelle Metapher für die Analyse der Daten durch den Nutzer dient.

Diese Landschaft, in der man sich sowohl thematisch als auch zeitlich bewegen kann, befindet

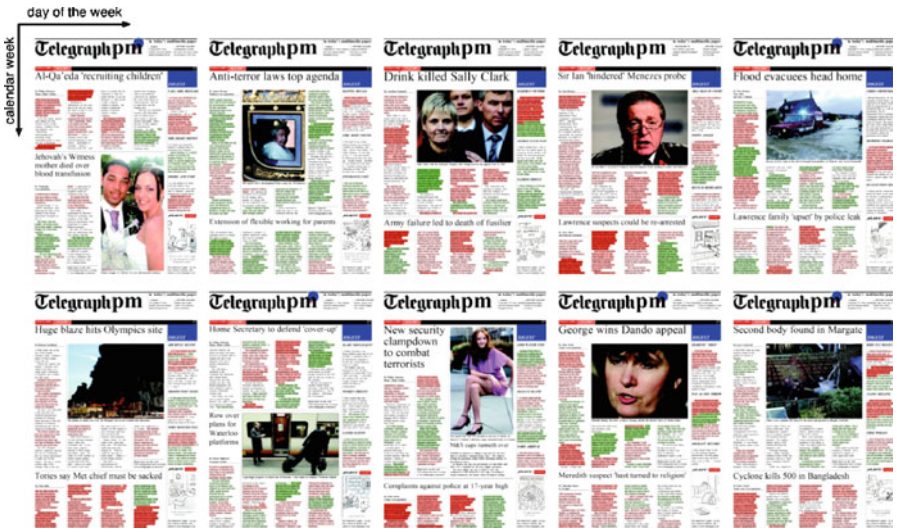

Abb. 2 Titelblätter einer Tageszeitung. Sätze nach positivem (grün) oder negativem (rot) Inhalt eingefärbt, nachgedruckt aus [9]

sich gerade in aktiver Entwicklung. Nach Fertigstellung kann dem Nutzer ein Werkzeug an die Hand gegeben werden, womit große Dokumentkollektionen interaktiv erkundet werden können. Dabei werden die Grundsätze der Informationsvisualisierung, „overview first", „zoom and filter" und „details on demand", berücksichtigt [16]: Zuerst einen groben Überblick über den Inhalt der Kollektion geben und dann tiefer gehende, präzise Analysen ermöglichen, ohne dass der Nutzer alle Artikel zu seinem Wunschthema sichten muss.

Visuelle Analyse von Meinungen

Während unzufriedene Kunden ihren Unmut über gewisse Produkte und Dienstleistungen früher vorwiegend direkt in einem eingeschränkten Bekanntenkreis kommuniziert haben, stellen heutzutage immer mehr Leute ihre Produktbewertungen für jeden zugänglich ins Internet. Die schiere Menge dieser Produktbewertungen, wie sie beispielsweise auf amazon.de zu finden sind, lässt Firmen schnell den Überblick über die Meinungen verlieren, die im weltweiten Netz über ihre einzelnen Dienstleistungen oder Produkte vorherrschen. Positive Kundenbewertungen wirken schnell einmal umsatzsteigernd, wohingegen negative Resonanz im Extremfall katastrophale Auswirkungen auf die Geschäftsbilanz einer Unternehmung haben kann, wenn sie nicht rechtzeitig detektiert und Gegenmaßnahmen eingeleitet werden.

Ähnliche Probleme tun sich auf, wenn es etwa für Unternehmen, Parteien, Politiker oder andere Personen des öffentlichen Lebens darum geht, die Reaktionen der Presse auf ihre Handlungen und Aussagen im Einzelnen zu analysieren. Oft kann es hilfreich sein, aktiv auf Meldungen zu reagieren, die einzelne Medien aufgegriffen haben, um so einem größeren negativen Presseecho vorzubeugen. Auch im Rückblick sind Analysen der Pressedynamik interessant, um aus Kommunikationspannen oder anderen Fehlern seine Lehren für die Zukunft ziehen zu können.

In beiden skizzierten Fällen – Produktbewertungen und Meldungen über Institutionen bzw. Personen – kommt es darauf an, Meinungen und andere positiv bzw. negativ konnotierte Aussagen, die Nutzer im Internet äußern, zu erfassen und zusammenzufassen. Neben klassischen Internetmedien, Webseiten und Blogs stehen dabei auch neuartige Kommunikationskanäle wie RSS und Twitter Feeds im Fokus. Im Folgenden soll auf zwei praxisrelevante Anwendungsszenarien eingegangen werden: erstens auf die detaillierte Zusammenfassung von Onlinekundenresonanzen zu gewissen Produkten, und zweitens auf die zeitliche Analyse von RSS News Feeds zum US-Präsidentschaftswahlkampf 2008. Dabei gilt es, Meinungen und ähnliche Äußerungen in Onlinemedien automatisch zu detektieren und zu visualisieren, um so einen Ausgangspunkt für weitere interaktive Analysen zu bieten. Ein erster einfacher Ansatz ist es, Sätze nach dem Vorkommen positiv oder negativ konnotierter Wörter, wie in Abb. 2 einzufärben.

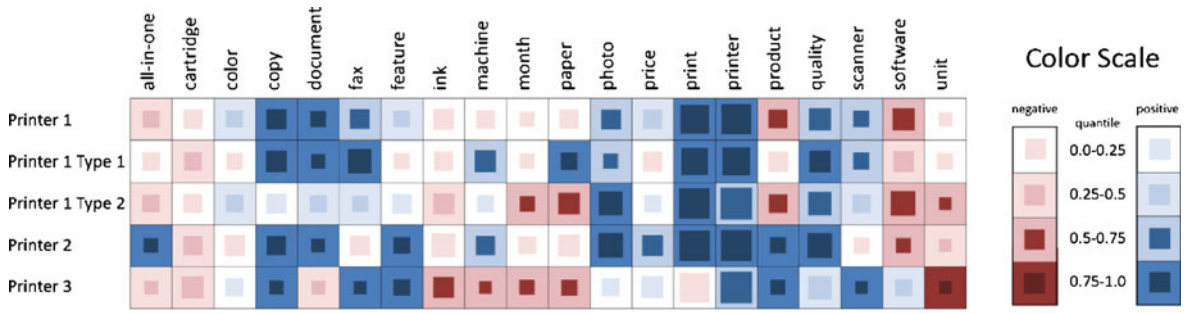

Abb. 3 Summary Report für verschiedene Drucker, nachgedruckt aus [14] (© 2009 IEEE)

Typischerweise basieren solche Ansätze zur Meinungsanalyse auf vordefinierten Listen, die Wörter enthalten, welche positive bzw. negative Assoziationen hervorrufen. In manchen Fällen hängt die Polarität einzelner Wörter auch von der Textdomäne bzw. von dem Objekt, auf das sie sich beziehen, ab. So ist ein Adjektiv wie „klein“ wohl eher als positiv zu erachten, wenn es sich auf ein mobiles Gerät bezieht und als eher negativ, wenn es um die Größe eines Bildschirms geht. Es wurden zahlreiche automatische Verfahren entwickelt, die die Domänen- und Objektabhängigkeit mancher Meinungswörter aus thematisch einschlägigen Textkollektionen lernen sollen. Um zu ermitteln, welche Meinungswörter sich auf welche Objekte beziehen, werden häufig syntaktische Dependenzen oder Muster verwendet, oder es wird einfach der Wortabstand als Bezugswahrscheinlichkeit herangezogen. Die Anwendung von Heuristiken und die Komplexität und Mehrdeutigkeit natürlicher Sprache führen dazu, dass es in Einzelfällen immer wieder zu Fehleinordnungen kommen kann. Nicht zuletzt deswegen ist eine Auswertung durch den Nutzer wichtig.

Onlineproduktbewertungen

Firmen sind nicht nur daran interessiert, ob Kunden ihre Produkte oder Dienstleistungen mögen, sondern im Speziellen auch daran, was genau daran gut ankommt oder abgelehnt wird. Es ist durchaus möglich, dass ein Kunde ein gewisses Produkt als positiv beurteilt, aber mit einem Detail oder Aspekt davon trotzdem unzufrieden ist. Bei der Analyse von Produktbewertungen sollten also die Beurteilungen einzelner Produktmerkmale oder -attribute einzeln betrachtet werden. In [14] wird ein ganzheitliches Verfahren vorgestellt, das von der automatischen Identifikation von Produktattributen und der Meinungszuordnung bis hin zur Gruppierung von gleichartigen Kundenprofilen das gesamte Analysespektrum abdeckt. Zur Identifikation von Produktattributen wurde ein Verfahren zur Extraktion diskriminierender Terme entwickelt [10]. Die Meinungsidentifikation basiert auf dem Abgleich mit Listen von Wörtern, die positive bzw. negative Assoziationen hervorrufen. In einem folgenden Schritt werden anschließend, basierend auf einer heuristischen Distanzfunktion, Meinungsäußerungen zu Produktattributen zugeordnet. Ein vergleichender Überblick über die Analyseresultate von Onlinemeinungen zu Produktattributen für verschiedene Drucker ist Abb. 3 zu entnehmen. Dieser sogenannte Summary Report ist eine kompakte und differenzierte Repräsentation der wesentlichen Meinungen zu bestimmten Produkten, die vollautomatisch aus tausenden von Kundenkommentaren erstellt wird. Jede Zeile der Matrix entspricht dabei einem bestimmten Produkt. Jede Spalte repräsentiert ein gewisses Produktattribut. So lässt sich dann die Zusammenfassung der Kommentare zu den Produkten, aufgeteilt nach Produktattributen, erfassen und zwischen verschiedenen Produkten vergleichen. Die Farbe einer Matrizenzelle zeigt dabei, ob ein Produktattribut zu den eher positiv (blau) oder negativ (rot) beurteilten gehört. Die Farbintensität weist auf den Prozentsatz der Kommentare hin, die ein Produktattribut als eher positiv bzw. negativ beurteilt haben. Die Größe des inneren Quadrats der Zelle zeigt den Anteil der Kommentare, die das entsprechende Attribut erwähnt hatten – je größer dieser Anteil, desto größer das Quadrat. Auf die konkreten Kommentarstellen, die einer Matrizenzelle zugrunde liegen, lässt sich dann interaktiv zugreifen, um so etwa mögliche Probleme im Detail verstehen zu können.

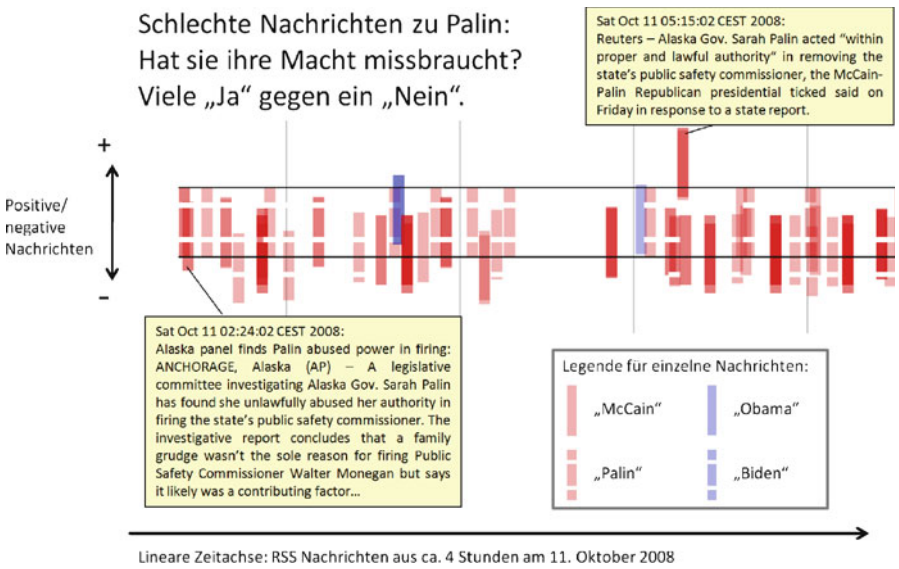

Abb. 4 Visualisierung von einzelnen RSS-Nachrichten im Zeitverlauf zum US-Präsidentschaftswahlkampf 2008, siehe [19]

Onlinenachrichten im Wahlkampf

Nachrichten zu Parteien oder Politikern können entscheidenden Einfluss auf die Stimmungslage und Entscheidungen von Wählern haben. Ein gekonnter Umgang mit den Medien ist für Politiker in der heutigen Zeit sehr wichtig. Gerade in den USA spielt die mediale Inszenierung von Politik im Wahlkampf eine große Rolle. Um die Stimmungslage der Medienlandschaft visuell zu analysieren, wurden während des letzten Monats des Präsidentschaftswahlkampfes in den USA im Jahr 2008 50 einflussreiche einschlägige RSS-Feeds zu diesem Thema abonniert und automatisch verarbeitet. Insgesamt kamen so über 23.000 einzelne Kurznachrichten zusammen. Anhand einer intuitiven visuellen Abbildung von Datenmerkmalen, insbesondere positiv und negativ assoziierten Meldungen, konnten verschiedene entscheidende Themen identifiziert werden [19]. Abbildung 4 greift beispielhaft eines dieser Themen auf. Nachrichten zum Lager der Demokraten, repräsentiert durch Rechtecke, sind blau gefärbt, Nachrichten zum Lager der Republikaner entsprechend rot. Die vertikale Verschiebung der Rechtecke zeigt die Polarität der Meldungen an: Positive Nachrichten sind nach oben, negative nach unten verschoben. Abbildung 4 zeigt einen etwa vierstündigen Zeitabschnitt, in dem negative Nachrichten zu den Republikanern dominieren, genauer gesagt zur republikanischen Kandidatin für die Vizepräsidentschaft.

Durch interaktives Hervorheben inhaltlich ähnlicher Meldungen kann auch der Verlauf bzw. die Verbreitung spezieller Themen sichtbar gemacht werden. Durch die Selektion einzelner Nachrichtenquellen wird deutlich, ob bei einigen Medien eine politische Tendenz auszumachen ist.

Visuelle Analyse von großen Dokumentkollektionen

Für einige Domänen stehen Dokumentkollektionen zur Verfügung, die besonderen Vorgaben bezüglich Struktur, Inhalt und vorhandenen Metainformationen genügen müssen. Ein Beispiel hierfür sind Patentkollektionen, deren Bedeutung insbesondere durch die Globalisierung der vergangnen Jahre stark zugenommen hat. Patentanalyse ist daher mittlerweile ein nicht mehr wegzudenkender Arbeitsschritt für viele Experten aus unterschiedlichen Bereichen wie Patentämtern, Forschungs- und Entwicklungsabteilungen von Firmen, Rechtsabteilungen und Kanzleien und für die Wissenschaft. Gleichzeitig wächst die Menge der Patentanmeldungen. Laut Statistiken der World Intellectual Property Organization [21] wurden im Jahr 2007 weltweit 1,85 Mio. Patente angemeldet, was einem Zuwachs von 3,7 % gegenüber dem Vorjahr entspricht. Über öffentliche Patentdatenbanken [3] waren 2007 mehr als 60 Mio. Patentdokumente abrufbar, von denen weltweit 6,3 Mio. gültig waren. Obwohl Patentdaten mittlerweile in elektronischer Form vorliegen, gibt es eine Vielzahl an Problemen, welche die

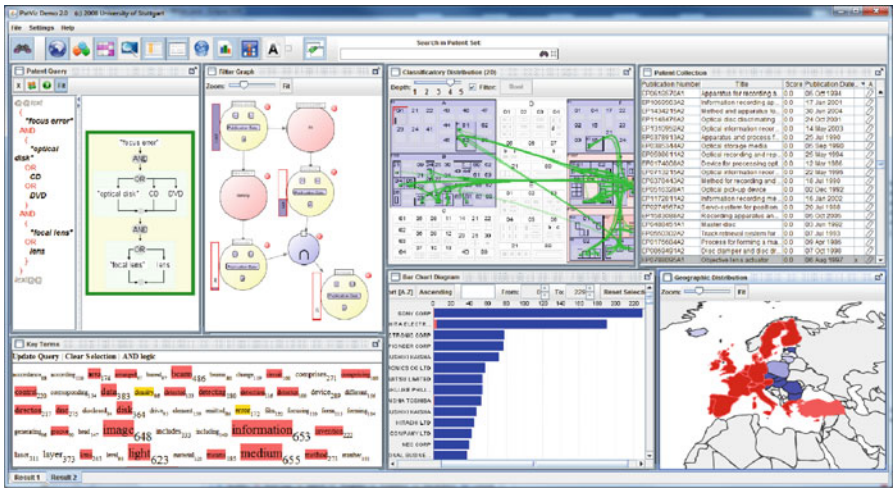

Abb. 5 Unterschiedliche koordinierte Darstellungen in PatViz

Suche und Analyse von Patentdaten erschweren. Zum einen liegt der größte Teil eines Patents als natürlichsprachlicher Text vor, was die aus der Computerlinguistik und dem Information Retrieval bekannten Probleme mit sich bringt. Zudem finden häufig domänen- und fachspezifische Begriffe Verwendung und insbesondere die rechtlich relevanten Patentansprüche sind in komplexer Rechtssprache verfasst. Beides verschlechtert die Lesbarkeit von Patentdokumenten für Nichtspezialisten und stellt weitere Herausforderungen im Hinblick auf die Patentsuche. Auch die Anforderungen an die Suche selbst unterscheiden sich grundsätzlich von der Suche im Internet. Während bei letzterer meist einige wenige „gute" Ergebnisse den Informationsbedarf eines Nutzers erfüllen, ist es bei der Patentsuche oft notwendig, *alle* relevanten Dokumente zu finden – anderenfalls drohen schwerwiegende finanzielle Konsequenzen. Neben Patentinhalten wie Text und Abbildungen stehen außerdem eine ganze Reihe von bibliografischen Daten und weiteren Metadaten zur Verfügung. Diese beinhalten unter anderem Informationen zu Erfinder, Anmelder, Einreichungs- und Veröffentlichungsdatum, rechtlichem Status, Gültigkeitsbereich und die Einordnung in ein Patentklassifikationssystem wie beispielsweise der International Patent Classification (IPC) [22]. Patentdaten sind daher hochdimensional, multimodal, mehrdeutig und heterogen. In der Praxis führt dies zu iterativen und oftmals langwierigen Vorgehensweisen bei der Patentsuche. Eine weitgehende Automatisierung ist bei der Patentanalyse aus den genannten Gründen nicht möglich. Dennoch ist die Unterstützung von Anwendern durch Methoden der maschinellen Sprachverarbeitung, des Information Retrieval, des maschinellen Lernens und des Data Minings notwendig – einerseits, um damit die Ausführungszeit für einzelne Analyseaufgaben zu verbessern, und andererseits, um Patentdokumente nach nutzerspezifizierten Gesichtspunkten zu kategorisieren. Visual Analytics bietet Techniken und Methoden, um diese Unterstützung zu realisieren.

Bislang werden in kommerziellen Produkten für die Patentrecherche Visualisierungsmöglichkeiten hauptsächlich im Berichtswesen eingesetzt. In den wenigsten Fällen existieren jedoch Lösungen, die den gesamten Analysezyklus, angefangen bei der Suche, einer ersten Analyse der Ergebnisse, der Verfeinerung der Suche und weiteren Analysen bis hin zu einem für den Rechercheur zufriedenstellenden Ergebnis abbilden. Die Verfeinerung von Suchanfragen ist notwendig, um die gewünschten relevanten Patentdokumente aufzufinden und führt in der Praxis zu sich wiederholenden, zeitaufwendigen Analyse- und Suchzyklen. Mit „PatViz" wurde ein Werkzeug entwickelt, das diesen iterativen Analysevorgang visuell unterstützt und dessen interaktive Steuerung durch Benutzer ermöglicht. Das umfasst die grafische Erzeugung von Suchanfragen genauso wie die Visualisierung von Suchergebnismengen. Mithilfe unterschiedlicher Sichten, die in eine Multiple-Coordinated-View-Umgebung eingebettet sind, kann der Problemraum hinsichtlich verschiedener Aspekte untersucht werden. Die unterschiedlichen Ansichten (Abb. 5) sind hierbei über das zugrunde liegende Datenmodell so verknüpft, dass die Selektion eines

Aspektes und damit die Einschränkung der aktuell betrachteten Patentmenge in allen anderen Ansichten reflektiert wird (Brushing und Linking). Dies vereinfacht die Analyse hinsichtlich spezifischer Patenteigenschaften bereits stark. Weiterhin lassen sich Selektionen aus den unterschiedlichen Perspektiven mittels eines graphbasierten Werkzeuges über Boole-Operationen kombinieren, wodurch der Einfluss komplexer Einschränkungen auf die aktuell angezeigte Patentdokumentmenge untersucht werden kann. Um den Analysezyklus zu schließen, gibt es die Möglichkeit, Erkenntnisse, die aus den verschiedenen Sichten gewonnen wurden, per Drag-and-Drop in eine Verfeinerung der Suchanfrage einfließen zu lassen und damit den Such-/Analysezyklus, mit einer erweiterten oder stärker eingeschränkten Anfrage, von Neuem zu beginnen. Die direkte, interaktive Modifikation von Suchanfragen wird dadurch erreicht, dass die Selektionsmächtigkeit in unterschiedlichen visuellen Ansichten auf die Ausdrucksfähigkeit der Anfragesprachen aller Backendsysteme angepasst wurde. Auf diese Art und Weise werden unterschiedliche Systeme für die Volltextsuche, bibliografische Datenbanken, ein semantisches Repository sowie eine Bildähnlichkeitssuche mittels visuellen interaktiven Mechanismen zu einem Visual Analytics System kombiniert. Die Notwendigkeit des Einsatzes unterschiedlicher Backendsysteme ergibt sich aus der Komplexität der Patentdaten und den qualitativen Anforderungen wie z. B. nach hohem Recall. Während sich bibliografische Daten in einer relationalen Datenbank speichern lassen, bieten sich für mit NLP-Methoden aufbereitete Textinformationen andere Speichersysteme wie Text-Repositories und semantische Datenbanken an. Die Aufbereitung erfolgt dabei in Form von Vorverarbeitungsschritten wie der Generierung von mehrsprachigen Suchindexen oder von Semantikextraktion. Die Kopplung von grafischen Benutzungsoberflächen mit den spezialisierten Suchsystemen für Textinformationen wird einerseits über die Bereitstellung einer visuellen Suchschnittstelle und andererseits über den Transport von Textmerkmalen in entsprechende Ansichten realisiert. So ist es beispielsweise möglich, die häufigsten oder die relevantesten Begriffe einzusehen und diese genauso interaktiv für Selektionen und Suchverfeinerungen einzusetzen, wie dies bei bibliografischen Daten der Fall ist. Äquivalent stehen diese Möglichkeiten für die semantische Suche und die Bildähnlichkeitssuche zur Verfügung. Eine ausführliche Beschreibung der PatViz-Oberfläche findet sich in [11].

Während Teile des PatViz-Systems bereits innerhalb des EU-Projekts PATExpert (www.patexpert.org) [20] entwickelt wurden, richtet sich die aktuelle Forschung innerhalb des Projekts „Scalable Visual Patent Analysis“ (DFG-Schwerpunktprogramm „Scalable Visual Analytics“) auf eine noch engere Verzahnung von Patentanalyse und Methoden der Computerlinguistik. Eine wichtige Zielsetzung ist es, den Analysezyklus dadurch weiter zu beschleunigen und speziell den Recall (Recall bezeichnet das Maß für die Anzahl der gefundenen relevanten Suchergebnisse im Verhältnis zu den in der gesamten Dokumentmenge vorhandenen relevanten Dokumenten) für die Patentsuche weiter zu erhöhen. Die Hauptherausforderung liegt darin, dass die Methoden der statistischen Dokumentanalyse nicht genau genug sind. Auch mit beliebig hohem Rechenaufwand kann man heute nicht die Qualität erreichen, die für die Patentanalyse notwendig wäre. Daher liegt ein weiterer Schwerpunkt auf der Entwicklung neuer, interaktiver, visueller Werkzeuge, die es Benutzern ermöglichen, Methoden der Sprachanalyse problemspezifisch einzusetzen, ohne sie gleichzeitig durch die Komplexität der Verfahren zu überfordern.

Fazit

Die visuelle Textanalyse ist ein hochaktuelles und praktisch relevantes Forschungsfeld, dessen Methodenrepertoire in einer Vielzahl von Anwendungsbereichen erfolgreich eingesetzt werden kann. Bis dato ist eine vollständige Erfassung von Semantik vollautomatisch nicht möglich und sie scheint auch in näherer Zukunft nur schwer vorstellbar. Eine Fokussierung auf interessante Teilaspekte führt hingegen durchaus zu guten Ergebnissen, wie am Beispiel von Patentdatenbanken, der Meinungsanalyse und der Analyse von zeitlichen Themenveränderungen gezeigt werden konnte. Automatische Analysealgorithmen bergen stets eine gewisse Unsicherheit, deshalb ist gerade bei der Analyse von ambiguen und weitgehend unstrukturierten Daten wie natürlichsprachlichen Texten eine enge Einbindung eines menschlichen Experten in die Analyse von großer Wichtigkeit. Das Feld der visuellen Analyse ist geradezu dafür

prädestiniert, solch einen Brückenschlag zwischen Mensch und Maschine zu bewerkstelligen. Die Visualisierung bietet dabei die Schnittstelle und Grundlage zur Interaktion. Gerade in unserer Wissensgesellschaft gibt es einen stetig steigenden Bedarf an visueller Textanalyse. Mit ihrer Hilfe lassen sich bessere und praxisrelevante Lösungen für die Analyse großer Dokumentkollektionen finden, die nicht zuletzt in Industrie und Forschung gewinnbringend eingesetzt werden können.

Literatur

1. Allan J (2002) Introduction to topic detection and tracking. Kluwer Academic Publishers, Norwell, MA, pp 1–16
2. de Saussure F (2001) Grundfragen der allgemeinen Sprachwissenschaft. Walter de Gruyter
3. European Patent Office (2010) Patent information products and services, "products_services_en.pdf". http://www.epo.org/about-us/publications/patent-information/products-services.html, letzter Zugriff 17.9.2010
4. Feldman R, Sanger J (2007) The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press
5. Han J, Kamber M (2006) Data Mining: Concepts and techniques, 2nd edn. Morgan Kaufmann, San Francisco, CA
6. Havre S, Hetzler B, Nowell L (2000) ThemeRiver: Visualizing Theme Changes over Time. In: Proceedings of the IEEE Symposium on Information Visualization 2000, pp 115f.
7. Holz F, Teresniak S (2010) Towards automatic detection and tracking of topic change. In:. Gelbukh A (ed) Proc. CICLing 2010, Iasi: Conference on Intelligent Text Processing and Computational Linguistics, LNCS 6008. Springer LNCS
8. Jurafsky D, Martin JH (2009) Speech and language processing. An introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2nd edn. Prentice Hall Series in Artificial Intelligence
9. Keim DA, Mansmann F, Oelke D, Ziegler H (2008) Visual analytics. Combining automated discovery with interactive visualizations. In: Boulicaut J-F, Berthold MR, Horváth T (eds) Discovery Science, 11th International Conference, DS 2008, Budapest, Hungary, October 13–16, 2008. Proceedings Lect Notes Artif Intell, vol 5255, pp 2–14. Springer, Heidelberg
10. Keim DA, Oelke D, Rohrdantz C (2010) Analyzing Document Collections via Context-Aware Term Extraction. In: 14th International Conference on Applications of Natural Language to Information Systems (NLDB '09). Lect Notes Comp Sci 5723, pp 154–168. Springer, Heidelberg
11. Koch S, Bosch H, Giereth M, Ertl T (2010) Iterative integration of visual insights during scalable patent search and analysis. Vis Comp Graph, IEEE Transactions, vol 99
12. Landauer TK, McNamara DS, Dennis SJ, Kintsch W (2007) Handbook of latent semantic analysis. Erlbaum, Mahwah, NJ
13. Manning CD, Prabhakar R, Schütze H (2008) Introduction to Information Retrieval. Cambridge University Press
14. Oelke D, Hao M, Rohrdantz C, Keim DA, Dayal U, Haug L, Janetzko H (2009) Visual opinion analysis of customer feedback data. In: Proc IEEE Symp Vis Anal Sci Technol (VAST '09), pp 187–194
15. Oesterling P, Heine C, Jaenicke H, Scheuermann G (2010) Visual analysis of high dimensional point clouds using topological landscapes. In: North S, Shen H-W, van Wijk JJ, (eds) IEEE Pacific Visualization 2010 Symposium Proceedings, pp 113–120
16. Shneiderman B (1996) The eyes have it: a task by data type taxonomy for information visualizations. Technical Report UMCP-CSD CS-TR-3665, College Park, Maryland
17. Soboroff I, Harman D (2005) Novelty detection: the TREC experience. In: HLT/EMNLP, pp 105–112
18. Teresniak S, Heyer G, Scheuermann G, Holz F (2009) Visualisierung von Bedeutungsverschiebungen in großen diachronen Dokumentkollektionen. Datenbank-Spektrum 31:33–39
19. Wanner F, Rohrdantz C, Mansmann F, Oelke D, Keim DA (2009) Visual Sentiment Analysis of RSS News Feeds featuring the US Presidental Election in 2008. In: Proceedings of the IUI'09 Workshop on Visual Interfaces to the Social and the Semantic Web (VISSW 2009). http://ceur-ws.org/Vol-443/paper7.pdf, letzter Zugriff 17.9.2010
20. Wanner L, Baeza-Yates R, Brügmann S, Codina J, Diallo B, Escorsa E, Giereth M, Kompatsiaris Y, Papadopoulos S, Pianta E, Piella G, Puhlmann I, Rao G, Rotard M, Schoester P, Serafini L, Zervaki V (2008) Towards content-oriented patent document processing. World Pat Inf 30(1):21–33
21. World Intellectual Property Organization (2009) World Intellectual Property Indicators. http://www.wipo.int/export/sites/www/ipstats/en/statistics/patents/pdf/wipo_pub_941.pdf, letzter Zugriff 17.9.2010
22. World Intellectual Property Organization (2010) International Patent Classification (IPC). http://www.wipo.int/classifications/ipc/en/, letzter Zugriff 17.9.2010