Institut für Visualisierung und Interaktive Systeme

Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Bachelorarbeit Nr. 164

Entwicklung interaktiver Techniken für die Zusammenfassung visueller Dokumenträume

Florian Prager

Studiengang:	Informatik
Prüfer/in:	Prof. Dr. Thomas Ertl
Betreuer/in:	Dr. Steffen Koch, Dipl. Phys. Qi Han
Beginn am:	4. August 2014
Beendet am:	3. Februar 2015
CR-Nummer:	H.3.3, H.5.2, I.3.6

Kurzfassung

Durch die zunehmende Vergrößerung der Datensätze, mit denen in unserer heutigen Zeit gearbeitet wird, gestaltet sich die Suche nach Informationen als eine schwierige und zugleich wichtige Aufgabe. Eine Suche mittels Schlüsselwörtern ist für spezifische Fragestellungen ein zielführendes Verfahren, doch ist es für andere weniger spezifische Aufgabenstellungen keine geeignete Methode. Zum Beispiel kann eine Dokumentensammlung nicht durch die Suche mit Schlüsselwörtern auf thematische Ausrichtung untersucht werden. Die interaktive Exploration von Dokumenten, die mittels einer Visualisierung abstrahiert wurden, ist eine mögliche Lösung für diese Art von Aufgabenstellungen. Hierzu wurde im Rahmen dieser Arbeit eine interaktive Lupe entwickelt, um die auf einer Ebene dargestellten Dokumente zu explorieren. Die Lupe fasst die Dokumente zusammen und zeigt nach Termgewichtungen ausgewählte Stichworte an. Diese Auswahl kann vom Nutzer gesteuert und die Exploration somit optimiert werden. Dazu wurden bestehende Techniken der Textzusammenfassung und Informationsvisualisierung analysiert und die Funktionalität des entwickelten Prototyps anhand eines Expertenfeedbacks evaluiert.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Verzeichnis der Listings

1 Einleitung

Die Suche ist eine wichtige Funktionalität, wenn es um die Untersuchung von Datensätzen geht. Sie wird von vielen Anwendungen und technischen Geräten angeboten und von Nutzern speziell bei großen Datensätzen genutzt, um gezielt Inhalte in den Datensätzen zu finden. Um diese Aufgabenstellung zu realisieren, müssen Suchanfragen mit spezifischen Schlüsselwörtern formuliert werden. Diese Methode ist bestens geeignet für konkrete Anfragen, doch werden Nutzern schnell die Grenzen aufgezeigt, wenn weniger spezifische Suchaufgaben gefordert sind. Möglicherweise sind die Inhalte der Datensätze nicht bekannt und eine gezielte Schlüsselwortanfrage ist somit keine Option. Auch die thematische Ausrichtung einer Dokumentensammlung herauszufinden ist eine dieser Suchanfragen.

Für diese Problemstellung werden häufig explorative Ansätze verwendet, bei denen die Dokumente in einer gemeinsamen Ebene dargestellt werden. Die Dokumente werden durch Symbole beziehungsweise Glyphen, zum Beispiel kleine Kreise, repräsentiert um somit von Nutzern interaktiv exploriert werden zu können. Damit bei der Erforschung der Dokumente die thematische Ausrichtung auch in der visuellen Darstellung zu finden ist, werden Symbole von ähnlichen Dokumente räumlich näher beieinander platziert. Dazu wird die paarweise Ähnlichkeit zwischen den Dokumenten bestimmt und eine entsprechende Platzierung berechnet. Um resultierende Cluster oder einfach nur ausgewählte Dokumente zu untersuchen, werden den Nutzern Interaktionstechniken angeboten. Zum Grundgerüst gehören das Verschieben, sowie das Vergrößern und Verkleinern der Ansicht. Auch die Anzeige von detaillierten Informationen zu entsprechenden Dokumenten ist eine häufig implementierte Funktion, ebenso wie die Platzierung von repräsentativen Textlabels.

Diese interaktiven Hilfestellungen reichen jedoch nicht aus um speziell große Dokumentensammlungen bestmöglich zu untersuchen. Eine Lösung um auf Abruf interessante Details zu den Dokumenten zu erhalten und um auch auf die statische Platzierung von Textlabeln zu verzichten, ist die in dieser Arbeit entwickelte Lupe. Diese Lupe bietet die Möglichkeit bestimmte Dokumente zu markieren und effektiv zusammenzufassen. Dazu werden bis zu zehn ausgewählte Terme im räumlichen Kontext zur Lupe angezeigt. Die Auswahl erfolgt anhand der zuvor bestimmten Termgewichtungen und des vom Nutzer eingestellten Filters. Es werden alle Terme der Dokumente, die mittels der Lupe markiert wurden, anhand der vom Nutzer gewählten Termgewichtung sortiert. Um diese entstandene Liste weiter anzupassen, kann mit der als Histogramm dargestellten Filterfunktion ein Auswahlkriterium festgelegt werden. Dazu wird aus allen Termen, der zu untersuchenden Dokumentensammlung, ein nach dieser Termgewichtung sortiertes Wörterbuch gebildet. Der Nutzer kann nun einen Bereich im Wörterbuch wählen, in dem alle Terme liegen müssen, um neben der Lupe angezeigt werden zu können. Diese Funktion ist interessant, da nicht grundsätzlich die inhaltlich relevantesten Terme an der Spitze des Wörterbuches liegen. Da es aber genau diese zu finden gilt, ist die Filterfunktion ein essentieller Teil der Interaktionstechnik.

Im folgenden Kapitel „Grundlagen“ werden grundsätzliche Algorithmen und Begriffe erklärt, die wichtig für das Verständnis der Arbeit sind. In Kapitel 3 „Verwandte Arbeiten“ werden ähnliche Konzepte vorgestellt und ein Einblick in den bisherigen Forschungsstand im Themenbereich gegeben. Anschließend wird in Kapitel 4 das Konzept vorgestellt und die Implementierung dieser in Kapitel 5. Kapitel 6 enthält die Evaluation der entwickelten Software. Ein Ausblick und eine Zusammenfassung über die Arbeit werden in Kapitel 7 vorgestellt.

2 Grundlagen

Das folgenden Kapitel beinhaltet alle Formeln und Algorithmen, die für diese Arbeit benötigt werden und auf denen diese basiert. Zusätzlich werden wichtige Begriffe für das grundsätzliche Verständnis erklärt.

2.1 Vektorraum-Modell

Das Vektorraum-Modell („Vector Space Model“) ist die Darstellung von Dokumenten als Vektoren und deren Platzierung in einem gemeinsamen hochdimensionalen Vektorraum. Unterstützt durch Termgewichtungen ist das Vektorraum-Modell ein grundlegendes Verfahren im Fachgebiet Information-Retrieval [MRS08].

2.1.1 Dokumentvektor

Die Darstellung von Dokumenten als Vektoren ist ein essentieller Grundgedanke dieser Arbeit um spätere Berechnungen auf den Dokumenten auszuführen. In [MRS08] wird ein Dokumentvektor wie folgt erstellt: Jedes Wort, im folgenden Term genannt, eines Dokuments wird zu einem Element im Vektor. Zusätzlich erhält jeder Term, jetzt Element eines Vektors, noch eine Termgewichtung. In dieser Arbeit werden für die Gewichtung die Termfrequenz und der TF-IDF Wert genutzt, die beide später in diesem Abschnitt erklärt werden. Für spätere Berechnungen werden diese Dokumentvektoren nicht nur aus den Termen des entsprechenden Dokuments erstellt, sondern jeder Term im Wörterbuch wird zu einem Element im Vektor. Dieses Wörterbuch wird aus der kompletten Dokumentensammlung erstellt. Bei Termen, die nicht im Dokument enthalten sind, ist die Termfrequenz und der TF-IDF Wert gleich null.

Hierzu ist noch zu sagen, dass im Normalfall bei der Vektorerstellung, sowohl aus einzelnen Dokumenten als auch aus Wörterbüchern, sogenannte Stoppworte herausgefiltert werden. Stoppworte („Stop Words“) werden entfernt, da diese hauptsächlich für die Syntax beziehungsweise Grammatik von Sätzen benötigt werden und deshalb nicht mittels dieser Worte auf den Inhalt geschlossen werden kann [MRS08]. Stoppworte sind zum Beispiel Artikel, Präpositionen und Ähnliche. Auch Zahlen sind ohne den Kontext als eigenständiger Term inhaltlich nicht relevant und können herausgefiltert werden. Falls nur mit der TF-IDF Gewichtung gearbeitet werden soll, ist diese Filterung nicht zwingend erforderlich, da Stoppworte bei diesem Verfahren grundsätzlich niedrige Werte bekommen, aufgrund des hohen Vorkommens in jedem Dokument.

2.1.2 Kosinus-Ähnlichkeit

Die Kosinus-Ähnlichkeit („Cosine Similarity“) ist ein Vergleichsmaß, um die Ähnlichkeit zwischen zwei Dokumentvektoren zu bestimmen. Mittels der Ähnlichkeit kann analysiert werden, wie nah Werte beieinander liegen. Für die spätere Platzierung der Punkte wird diese Ähnlichkeit benötigt. Nach Manning et al. [MRS08] ist die Kosinus-Ähnlichkeit Sim für zwei Dokumente, d_1 und d_2, wie folgt definiert, wenn „$\cdot$“ das Skalarprodukt ist und $|\vec{V}(d)|$ die Euklidische Norm:

$$Sim(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|} \tag{2.1}$$

2.1.3 Euklidischer Abstand

Der Euklidische Abstand („Euclidean Distance“) kann als Vergleichsmaß genutzt werden, um die Distanz, beziehungsweise den Abstand, zwischen zwei Dokumentvektoren zu bestimmen. Daraus kann auf die Ähnlichkeit, oder besser gesagt die Unähnlichkeit, zweier Dokumentvektoren geschlossen werden, da große Abstandswerte zu kleinen Ähnlichkeitswerten führen und umgekehrt. Der Euklidische Abstand wird nach Manning et al. [MRS08] mit folgender Gleichung berechnet, bei der d_1 und d_2 zwei Dokumente sind:

$$|\vec{d_1} - \vec{d_2}| = \sqrt{\sum_{i=1}^{M}(d_{1i} - d_{2i})^2} \tag{2.2}$$

2.2 Termgewichtung

Termgewichtungen sind, wie vorher erwähnt, ein wichtiger Teil für Berechnungen und Vergleiche mit Dokumentvektoren in dieser Arbeit. Dabei wird jedem Term ein Wert zugewiesen, um diese vergleichbar zu machen. Es gibt diverse Methoden um diese Werte zu generieren, die in [MRS08] erklärt werden. Im Folgenden werden nur Gewichtungen vorgestellt, die für diese Arbeit von Interesse sind: Die Termfrequenz, die Dokumentfrequenz und die TF-IDF Gewichtung.

2.2.1 Termfrequenz

Die Termfrequenz („Term Frequency“, TF) beschreibt nach Manning et al. [MRS08] das Vorkommen eines Terms innerhalb eines Dokuments. Bei der Termfrequenz wird davon ausgegangen, dass alle Terme in einem Dokument gleich wichtig sind und somit diejenigen, die häufiger in einem Dokument erscheinen, als wichtiger für den Inhalt angesehen werden. Bei dieser Gewichtungsmethode spielt die Reihenfolge, in der die Terme in einem Dokument erscheinen, keine Rolle, sondern nur die Häufigkeit mit der diese auftreten ist entscheidend. Für einzelne Dokumente kann die Termfrequenz problemlos genutzt werden, doch kann es speziell für eine Sammlung an Dokumenten zu Problemen kommen, wenn ausschließlich die Termfrequenz genutzt wird. Denn für den Fall, dass Dokumente der Sammlung aus einem ähnlichen Themenbereich stammen, überschneiden sich diese inhaltlich und

dies spiegelt sich in der Termfrequenz der einzelnen Terme wieder. Für einen solchen Fall sind Terme mit einer sehr hohen Termfrequenz, die gleichzeitig in vielen oder gar allen Dokumenten vorkommen (siehe Dokumentfrequenz), weniger zur Charakterisierung geeignet. Dieses Problem kann mittels der Dokumentfrequenz behoben werden.

2.2.2 Dokumentfrequenz

Um die Dokumentfrequenz („Document Frequency“, DF) zu bilden, werden die Dokumente gezählt, in denen das Erscheinen eines Terms größer oder gleich eins ist [MRS08]. Anders als bei der Termfrequenz ist es hier nicht entscheidend, wie oft ein Term in einem Dokument vorkommt, sondern nur ob ein Term in einem Dokument auftaucht. Die Verwendung der Dokumentfrequenz ist somit nur für Sammlungen von Dokumenten sinnvoll und nicht für einzelne Dokumente. Mit dieser Methode kann ein Eindruck über die Verteilung eines Terms in einer komplette Sammlung gewonnen werden. Terme, die nicht in jedem Dokument auftauchen, charakterisieren in diesem Zusammenhang ein einzelnes Dokument besser als diejenigen, die eine hohe Dokumentfrequenz haben. Zum Beispiel ist in einem Datensatz mit Ausarbeitungen aus dem Bereich der Visualisierung die Dokumentfrequenz für das Wort „Visualisierung“ sehr hoch, aber inhaltlich nicht relevant, da schon im Vorhinein bekannt ist, dass die Dokumente aus dem Fachgebiet der Visualisierung stammen.

Mithilfe dieser Dokumentfrequenz kann die inverse Dokumentfrequenz („Inverse Document Frequency“, IDF) gebildet werden. Die inverse Dokumentfrequenz kann mittels folgender Gleichung berechnet werden. Hierbei ist N die Anzahl der Dokumente in einer Sammlung und DF die Dokumentfrequenz für einen Term t:

$$IDF_t = \log\left(\frac{N}{DF_t}\right) \tag{2.3}$$

Diese inverse Dokumentfrequenz ist niedrig für Terme, die häufig erscheinen und hoch für die Terme, die eine kleine Dokumentfrequenz besitzen.

2.2.3 TF-IDF

Mit einer Kombination der Termfrequenz und der inversen Dokumentfrequenz kann nun eine Gewichtung für einen Term in jedem Dokument erstellt werden, die sogenannte Termfrequenz - inverse Dokumentfrequenz („Term Frequency - Inverse Document Frequency“) oder kurz TF-IDF. Die TF-IDF Gewichtung wird für einen Term t in einem Dokument d wie folgt berechnet [MRS08]:

$$TF - IDF_{t,d} = TF_{t,d} * IDF_t \tag{2.4}$$

Der TF-IDF Wert ist am höchsten für Terme, die häufig in einer kleinen Anzahl an Dokumenten vorkommen. Woraus geschlussfolgert werden kann, dass die Relevanz und inhaltliche Aussagekraft dieser Terme hoch ist. Wenn Terme nicht so häufig in einzelnen Dokumenten vorkommen oder auf viele Dokumente verteilt sind, dann ist die TF-IDF Gewichtung dieser Terme niedriger und am niedrigsten für Terme, die in nahezu allen Dokumenten erwähnt werden und in diesen auch nur selten.

2.3 t-SNE

Der t-SNE Algorithmus, „t-distributed Stochastic Neighbor Embedding“, wurde von Laurens van der Maaten und Geoffrey Hinton zur Reduktion von hochdimensionalen Daten entwickelt [MH08]. t-SNE versucht eine nicht lineare Platzierung zwischen hochdimensinalen Datenpunkten und einem niedrigdimensionalen Raum zu finden, bei der die paarweisen Abstände so gut wie möglich erhalten bleiben. Als Ergebnis liegen Datenpunkte in 2D oder 3D vor, die zur Visualisierung genutzt werden können. Hierbei wird trotz der Dimensionsreduktion die Eigenschaft, dass Punkte die nahe beieinanderliegen eine größere Ähnlichkeit als entfernte haben, beibehalten. t-SNE ist eine Abwandlung des „Stochastic Neighbor Embedding“ (SNE), welches von Hinton und Roweis [HR02] entwickelt wurde. In dieser Arbeit wird der in [MH08] beschriebene Ansatz von t-SNE verwendet. Angepasste Versionen von t-SNE können in [Maa09] und [Maa13] gefunden werden.

Die ersten Schritte von t-SNE bauen auf dem SNE Verfahren auf. Zuerst erfolgt die Umwandlung der hochdimensionalen Euklidischen Abstände oder paarweisen Ähnlichkeiten zwischen zwei Datenpunkten in bedingte Wahrscheinlichkeiten. Diese bedingten Wahrscheinlichkeiten entsprechen der Wahrscheinlichkeit, dass ein Datenpunkt einen anderen als Nachbar wählt. Diese Wahrscheinlichkeit ist hoch für ähnliche Punkte und niedrig für unähnliche. Zwischen dieser konstruierten Wahrscheinlichkeitsverteilung und der auf gleiche Weise erstellten Wahrscheinlichkeitsverteilung der Punkte im niedrigdimensionalen Raum wird die Kullback-Leibler-Divergenz gebildet. Diese Diskrepanz wird von t-SNE minimiert, um die Ähnlichkeiten zwischen den Datenpunkten bestmöglich zu erhalten und die Platzierung der Punkte zu erstellen.

Erwähnenswert ist, dass die von t-SNE generierten Ergebnisse sich in jedem Durchlauf ändern und sich somit die Punkteverteilungen unterscheiden. Speziell bei der Visualisierung der Punkte als Streudiagramm, kann dies zu unterschiedlich gut geeigneten Ergebnissen führen.

3 Verwandte Arbeiten

Für diese Arbeit sind im Wesentlichen folgende drei Themengebiete von Bedeutung: Textzusammenfassung, Informationsvisualisierung und Interaktive Lupen. In diesem Kapitel werden grundlegende Erkenntnisse sowie bestehende Ansätze und Techniken, die für diese Arbeit relevant sind, vorgestellt und analysiert.

3.1 Textzusammenfassung

Um die Textzusammenfassungstechnik in dieser Arbeit verstehen zu können, müssen zunächst einige grundlegende Begriffe geklärt werden. Nach Radev et al. [RHM02] wird Zusammenfassung folgendermaßen definiert. Eine Zusammenfassung wird aus einem oder mehreren Texten erstellt. Das Ergebnis ist ein Text, der am Umfang des Ausgangstextes gemessen, normalerweise nicht mehr als halb so lang ist, jedoch alle wichtigen Informationen enthält. Die Herausforderung hierbei ist, den Umfang zu reduzieren ohne dabei den Informationsgehalt zu verringern. Nach [RHM02] kann die Textzusammenfassung mit vier unterschiedlichen Verfahren realisiert werden. Die erste Methode stellt die Extraktion („Extraction“) dar. Hierbei steht das Finden der inhaltlich wichtigen Textstellen im Vordergrund, welche in der Zusammenfassung wörtlich übernommen werden. Bei der Abstraktion („Abstraction“) werden die wichtigen Informationen in einem neuen Text wiedergegeben. Die Fusion („Fusion“) ist ein Verfahren, bei dem extrahierte Textstellen in einer schlüssigen Weise zusammengefügt werden. Die letzte Methode zur Zusammenfassung von Texten ist die Kompression („Compression“). Dabei wird der Umfang des Textes durch Herausfiltern von unwichtigen Textstellen reduziert. Alle vier Verfahren sind sowohl für das Zusammenfassen von einzelnen Dokumenten, als auch für mehrere Dokumente geeignet. Die in dieser Arbeit verwendete Methode zur Zusammenfassung ist als eine Art der Extraktion anzusehen, auch wenn am Ende kein zusammenhängender Text entsteht.

In frühen Arbeiten zur Zusammenfassung von wissenschaftlichen Texten wurden verschiedene Ansätze verwendet, um nennenswerte Sätze zu extrahieren. Luhn [Luh58] analysierte Texte unter dem Gesichtspunkt der Frequenz von Worten und Sätzen. Baxendale [Bax58] konzentrierte sich auf die Position der Sätze im Text, während Edmundsen [Edm69] Schlüsselwörter zur Textzusammenfassung nutzte. Im Laufe der Zeit wurden auch verschiedene Arbeiten veröffentlicht, die sich nicht nur auf das Zusammenfassen von wissenschaftlichen Texten, sondern auch auf das Zusammenfassen von anderen Texten, zum Beispiel Nachrichten, konzentrieren. In dieser Arbeit werden ebenso Nachrichtentexte zusammengefasst, da diese nicht themenspezifisch sind und somit keine besonderen fachspezifischen Vorkenntnisse von den Nutzern gefordert werden. Eine Studie kann deshalb mit unterschiedlichen Nutzergruppen durchgeführt werden. Die Texte werden, wie nach Luhn [Luh58], auf Basis der Frequenz von Worten analysiert, doch bleiben während der Verarbeitung der Texte nur die Schlüsselwörter erhalten.

Das und Martins geben in ihrem Artikel „A Survey on Automatic Text Summarization“ [DM07] einen Überblick über Techniken und Ansätze der automatischen Textzusammenfassung. Diese beschreiben zunächst die Verfahren zur Zusammenfassung von Einzeltexten. Darauf wird hier nicht weiter eingegangen, da in dieser Arbeit mehrere Dokumente zusammengefasst und dargestellt werden, auch wenn zuerst die einzelnen Dokumente einer Art Zusammenfassung unterliegen, um daraus Dokumentvektoren bilden zu können.

Im heutigen Zeitalter gewinnt das Zusammenfassen von mehreren Dokumenten zu einem informativen prägnanten Text immer mehr an Bedeutung. Viele Entwicklungen auf diesem Gebiet sind für Nachrichten zu finden. Nach Das und Martins [DM07] ist das von McKeown und Radev [MR95] entwickelte Programm „SUMMONS“ das erste System zur Zusammenfassung von mehreren Texten. Es erstellt aus mehreren Artikeln zu einem Themengebiet eine knappe Zusammenfassung der relevanten Informationen. Probleme ergeben sich jedoch, wenn Dokumente aus thematisch weiter entfernten Themengebieten stammen. McKeown et al. [MKH+99] verbesserte daher das System insofern, dass zunächst das Thema der Texte über Ähnlichkeiten von Textstellen identifiziert wird. Ebenso werden in dieser Arbeit die einzelnen Dokumente nach ihrem Thema sortiert, genauer gesagt werden die Ähnlichkeiten bestimmt, um bei der Platzierung in der Ebene Clusterbildungen und Muster zu erzeugen. Im Unterschied zu den Arbeiten von McKeown, bei welchen die Zusammenfassung als neuer Text gegeben ist, werden in dieser Arbeit repräsentative Schlüsselworte als Zusammenfassung ausgegeben.

Eine gängige Methode um inhaltlich relevante Schlüsselworte zu finden, ist Texte als einzelne Terme darzustellen, dabei die Stoppworte herauszufiltern und die erhaltenen Terme zu gewichten. Erkenntnisse dazu und allgemein zum Finden von inhaltlich relevanten Kernsätzen wurde von Chuang et al. [CMH12] veröffentlicht. Der Schwerpunkt der Arbeit liegt darauf, geeignete Terme und Kernsätze zu finden, um diese zur Visualisierung zu nutzen. Verglichen wurden die Ergebnisse mit den Termen die vom Menschen gewählten wurden und den Inhalt somit bestmöglich wiedergeben.

Die Gewichtungsmethoden die unter anderem in dieser Arbeit zum Einsatz kommt beruhen auf der Frequenz der Terme. Von besonderer Bedeutung ist daher die Erkenntnis von Luhn [Luh58], dass die repräsentativsten Terme nicht sehr oft die mit den höchsten oder die mit den niedrigsten Frequenzen sind, sondern eher die Terme mit mittleren Frequenzwerten.

3.2 Informationsvisualisierung

Um die Textdaten den Nutzern zugänglich zu machen, müssen die Daten visuell dargestellt werden. Informationsvisualisierung stellt die Verbindung zwischen den abstrakten Daten und der menschlichen Wahrnehmung dar. Visualisierung bietet Nutzern die Möglichkeit die Daten zu interpretieren und Informationen zu gewinnen. Nach Card et al. [CMS99] ist Informationsvisualisierung der Gebrauch von computergestützten, interaktiven, visuellen Repräsentationen von abstrakten Daten um die Wahrnehmung zu erweitern. Die abstrakten Daten in dieser Arbeit sind die Textdaten. Diese einzelnen zusammengefassten Texte müssen dargestellt werden, um mit der interaktiven Lupe nach bestimmten Kriterien analysiert werden zu können. Daraus können inhaltliche Informationen von Nutzern gewonnen werden.

Es gibt verschiedene Möglichkeiten die aus der Textzusammenfassung erhaltenen Schlüsselworte darzustellen. Eine Möglichkeit ist die sogenannte Schlagwortwolke („Tag Cloud“), zu sehen in Abbildung 3.1. Hierbei werden die gewichteten Terme meist zweidimensional angeordnet. Die Gewichtung spiegelt sich beispielsweise in der Schriftgröße wider. Somit sind die für den Text relevantesten Terme für den Betrachter der Schlagwortwolke leicht zu erkennen. Es wurden inzwischen auch Methoden entwickelt, um mit Schlagwortwolken nicht nur einzelne Dokumente, sondern auch Dokumentsammlungen darzustellen.

Abbildung 3.1: Beispiel einer Schlagwortwolke [BGN08].

Es gibt verschiedenste Arten Informationen zu visualisieren. Diese hängen mit den Daten zusammen, die visualisiert, und den Aufgaben, die gelöst werden sollen. Shneiderman [Shn96] spricht von insgesamt sieben Aufgaben und von folgenden sieben Datentypen: 1-, 2-, 3-dimensionale Daten, zeitabhängige und n-dimensionale Daten, Baum und Netzwerk Daten. Für jeden dieser Datentypen gibt es Visualisierungsmethoden, die geeigneter sind als andere.

Mögliche Darstellungen für eindimensionale Daten sind beispielsweise Kreisdiagramme („Pie Charts“) oder Säulendiagramme („Bar Charts“), zu sehen in Abbildung 3.2.

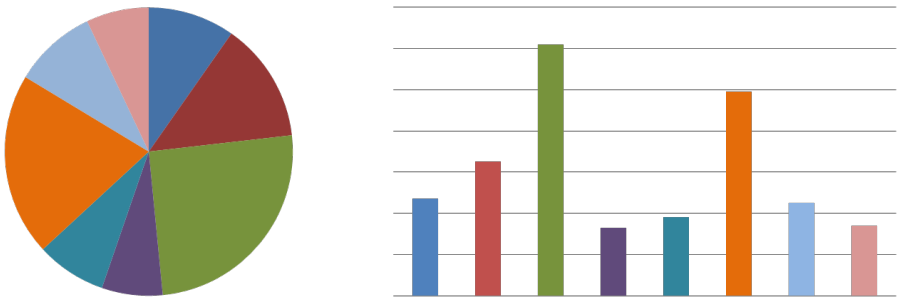

Abbildung 3.2: Mögliche Darstellungsformen für eindimensionale Daten. Links ist ein Kreisdiagramm zu sehen und rechts das entsprechenden Säulendiagramm.

Bei zweidimensionalen Daten werden Streudiagramme („Scatter Plots“), zu sehen in Abbildung 3.3, Matrizen oder Ähnliches genutzt, da die zweite Komponente nicht ohne Einschränkungen in den vorherigen Diagrammtypen dargestellt werden kann.

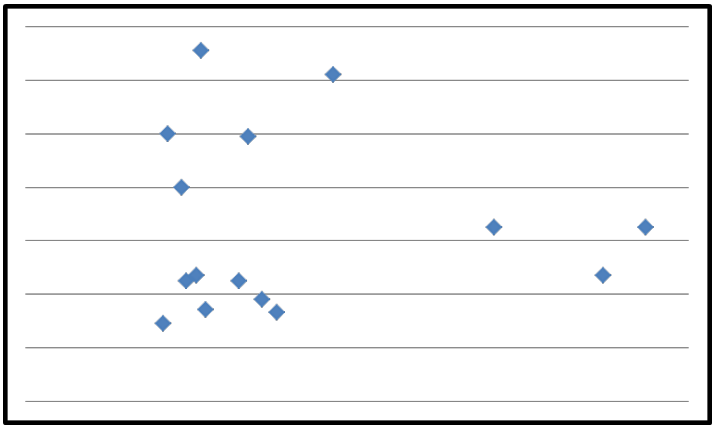

Abbildung 3.3: Beispiel eines Streudiagramms.

Werden nun n-dimensionale Daten betrachtet, gestaltet sich die Darstellung schwieriger. Es gibt unterschiedliche Ansätze, wie Parallele Koordinaten („Parallel Coordinates“), um dieses Problem zu lösen, doch soll hier nicht weiter auf Darstellungen von Daten mit n Komponenten eingegangen werden.

In dieser Arbeit wird jeder einzelne Text als kleiner Kreis in einer Ebene dargestellt, woraus ein Streudiagramm resultiert. Einer der ersten Ansätze, bei dem Textdaten auf diese Art und Weiße dargestellt werden, ist von Wise et al. [WTP+95]. Dabei werden große Dokumentensammlungen spatialisiert, um exploriert werden zu können. Die Dokumente werden als hochdimensionale Vektoren dargestellt und mittels einer Dimensionreduktionsmethode und der Bestimmung der Ähnlichkeit auf eine Ebene projiziert. Dort können die Daten mit Interaktion von Nutzern analysiert werden.

Um die Schlüsselwörter der Texte zu erhalten, werden die kleinen Kreise, wie bereits erwähnt, in dieser Arbeit mittels einer interaktiven Lupe weiter analysiert. Das Kombinieren von Visualisierungsmethoden und Interaktionstechniken wie beispielsweise einer Lupe ist nicht unüblich. Durch Interaktion können Schwächen von Darstellungsformen behoben oder die Visualisierung unterstützt werden.

3.3 Interaktive Lupen

In der Visualisierung muss häufig mit überfüllten Darstellungen gearbeitet werden, aufgrund der großen und stetig wachsenden Datensätze. Um trotz dieser großen Datenmengen die Möglichkeit zum Auslesen und Erkennen wichtiger Daten zu erhalten, wird ein interaktiver Ansatz zur Exploration der Visualisierung benötigt. In dieser Arbeit zum Verarbeiten der visualisierten Textdaten. Dazu wird eine in der Visualisierung gängige Methode angewandt, der Gebrauch einer interaktiven Lupe. Diese Lupen

gewährleisten, während der Ausführung nach Bedarf eine alternative visuelle Repräsentation für einen bestimmten Bereich auf dem Bildschirm, dem Bereich unter der Lupe. Lupen gibt es in verschiedenen Formen mit den unterschiedlichsten Funktionen. Ein einfaches Beispiel für die Funktion einer Lupe ist die Lupe mit Vergrößerungsfunktion. Sie stellt die Visualisierung unter der Lupe vergrößert dar, nach dem Prinzip der Vergrößerungslupen in der realen Welt. Eine ausführliche Erklärung des erwähnten Problems und eine Zusammenfassung über den aktuellen Stand der Forschung im Bereich interaktive Lupen kann in „A Survey on Interactive Lenses in Visualization“ [TGK+14] von Tominski et al. gefunden werden.

Nach Tominski et al. können interaktive Lupen nach verschiedenen Kriterien kategorisiert und analysiert werden. Eine Möglichkeit ist die Betrachtung der Lupen im Bezug auf die Form, Größe und Position, also die Eigenschaften der Lupen. Die unterschiedlichen Datensätze, auf denen operiert wird, stellen andere Ansprüche an Lupen und sind somit ein weiterer Analysepunkte für Lupen. Das dritte Kriterium ist die Funktionalität der Lupen, die stark mit dem Ziel verbunden ist, welches durch den Einsatz einer interaktiven Lupe erreicht werden soll.

Zuerst ist eine Einordnung der speziell entwickelten Lupentechnik sinnvoll. Dazu eignet sich die Kategorisierung von Cockburn et al. [CKB08] in Überblick+Detail, Zoom und Fokus+Kontext Schnittstellen. Grundsätzlich fallen Lupen in die Kategorie Überblick+Detail. Der Überblick wird durch die Ansicht gegeben, auf der die Daten dargestellt sind und die Lupe bewegt wird. Der Bereich innerhalb der Lupe kann dementsprechend als die detaillierte Ansicht gesehen werden. Diese beiden Ansichten sind normalerweise räumlich getrennt dargestellt. Bei einer Lupe gibt es jedoch diese klassische räumliche Trennung auf der X-Y-Ebene der dargestellten Daten nicht. Vielmehr trennen sich die Übersicht und die detaillierte Ansicht auf der Z-Achse, da die Lupe über den Daten bewegt wird und diese manipuliert. Die Lupe in dieser Arbeit dient mehr als eine Art Auswahltechnik und nur zur indirekten Anzeige von Details zu den Daten. Es gibt zwar eine visuelle Rückmeldung an den Nutzer über die markierten Daten, aber keine Vergrößerung oder Ähnliches. Jedoch wird die komplette Dokumentensammlung auf einer Ebene dargestellt und eine detaillierte Ansicht der Dokumente wird in Form von Stichworten neben der Lupe auf einer gemeinsamen Ebene angezeigt. Dieser Aspekt der entwickelten Lupe kann aber auch als eine Fokus+Kontext Technik angesehen werden. Die Stichworte befinden sich zusammen mit den Dokumenten in einer Ansicht und sind somit gleichzeitig für den Nutzer wahrnehmbar. Sie bleiben im Kontext und werden nicht aus dem Fokus der Ansicht genommen. Die Kriterien für die Zoom Einordnung treffen auf die entwickelte Interaktionstechnik nicht zu. Es gibt zwar die Funktionalität die Ansicht der spatialisierten Dokumente zu vergrößern und zu verkleinern, doch diese Funktion ist nicht zwingend erforderlich für die entwickelte Lupe. Zusammenfassend kann gesagt werden, dass die Lupe nicht in eine einzelne Kategorie eingeteilt werden kann, sondern Aspekte unterschiedlicher Kategorien vereint.

Erste Ansätze für interaktive Lupen wurden von Bier et al. [BSP+93] unter dem Namen „magic lenses“ und „toolglass widgets“ veröffentlicht. Diese Lupen können die Darstellung der Objekte, die unter der Lupe liegen, verändern. Es können zusätzliche hilfreiche Daten hinzugefügt oder störende entfernt werden. Auch die Möglichkeit nicht sichtbare Informationen, zum Beispiel durch Vergrößerung, aufzudecken, ist gegeben. Die Lupen sind als dynamische bewegliche Filter implementiert und der Grundstein für heutige interaktive Lupen.

Robertson und Mackinlay entwickelten und veröffentlichten 1993 eine Lupe zur Arbeit mit Dokumenten [RM93], zu sehen in Abbildung 3.4. Diese operiert ebenfalls auf Dokumentdaten, wie die in dieser

Arbeit umgesetzte Lupe, doch mit dem Fokus auf der Darstellung der kompletten Dokumentseiten und nicht einer alternativen Repräsentation der Dokumente.

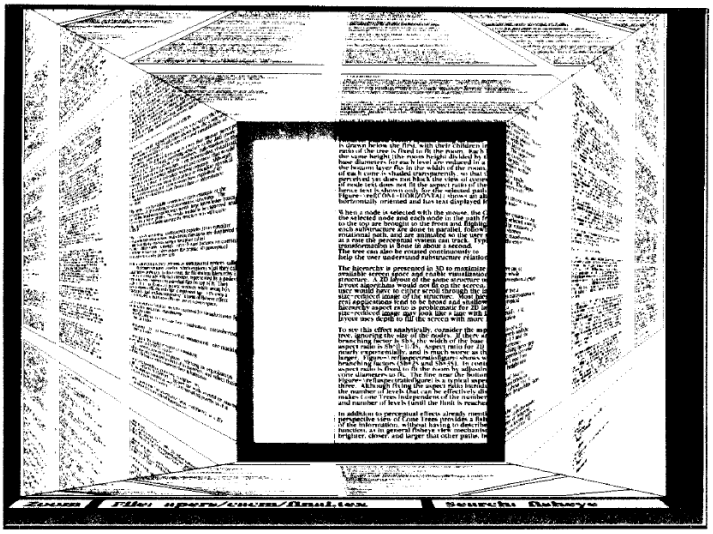

Abbildung 3.4: Die von Robertson und Mackinlay entwickelte Lupe namens „Document Lens“ [RM93].

Die Seiten werden nebeneinander als Übersicht dargestellt und mittels einer rechteckigen Lupe, der „Documet Lens“, können Teilbereiche dieser Übersicht vergrößert werden, ohne den globalen Kontext zu verlieren. Dieser interessante Aspekt bleibt auch in dieser Arbeit nicht unbeachtet. Dies ist jedoch nicht der Fall bei gewöhnlichen Vergrößerungslupen oder Fisheye-Lupen. Diese Fokus+Kontext Ansicht vergrößert den Bereich unter der Lupe und zieht die umliegenden Dokumentenseiten in die Länge, um sie an die Ränder der Lupe zu binden. Der dadurch entstehende Pyramidenstumpf ermöglicht zwar den Erhalt des globalen Kontextes, doch kann der anliegende Text durch diese Art der Darstellung schnell für Nutzer unleserlich werden.

Eine weitere Fokus+Kontext Technik, entwickelt von Chang und Collins [CC13], ist ein Ansatz zur Textvisualisierung. Die Besonderheit ist, dass abstrakte Textdaten und spatialisierte Daten mittels einer Lupe miteinander verbunden werden und in einer gemeinsamen Ansicht dargestellt werden. Die Lupe kann in diesem speziellen Fall über einem 3D Modell eines Autos bewegt werden, um bestimmte Teile des Modells auszuwählen, wie in Abbildung 3.5 zu sehen ist. Das Modell ist eine graphische Repräsentation der verarbeiteten Textdaten, die zu den korrespondierenden Teilen zugeordnet wurden. Dadurch bleibt für die Analysten der Bezug zu Daten der realen Welt erhalten.

Da in dieser Arbeit Nachrichtendaten exploriert werden, aber auch andere Daten den Grundstein bilden können, wurde für die Darstellung eine abstraktere Repräsentation gewählt. Ein Streudiagramm hat somit keinen konkreten Bezug zur realen Welt mehr, eignet sich jedoch für die unterschiedlichsten Daten. Um aus der von Cang und Collins entwickelten Darstellung die zugrundeliegenden Textdaten zu erhalten, können mit der Lupe Teilbereiche ausgewählt werden. Die Informationen zu diesen ausgewählten Daten werden in einem eigenen Bereich direkt neben der Lupe angezeigt. Wie in dieser Arbeit werden die Daten der ausgewählten kreisförmigen Glyphen, oder Autoteile, direkt neben der

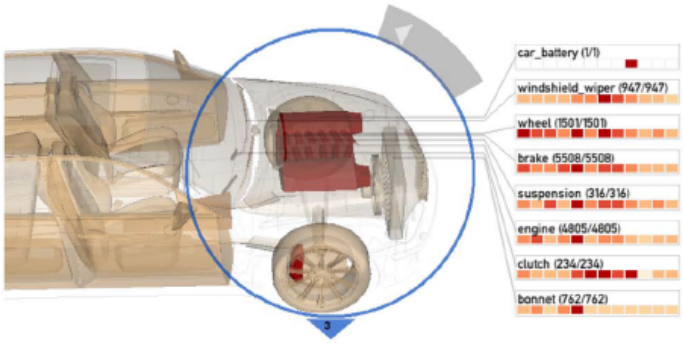

Abbildung 3.5: Die von Cang und Collins entwickelte Lupe zur Exploration von Textdaten, die auf das Modell eines Autos transferiert wurden [CC13].

Lupe angezeigt, um den Kontext nicht zu verlieren und Nutzer müssen ihr Aufmerksamkeit nicht auf zwei unterschiedliche Ansichten aufteilen.

Die interaktive Lupe, auf der diese Arbeit aufgebaut wurde, ist die von Heimerl et al. [HKBE12] entwickelte „Term Lens“, zu sehen in Abbildung 3.6. Die „Term Lens“ dient für diese Arbeit als eine Art Grundgerüst. Erweitert wurde sie in dieser Arbeit um die visuell dargestellte Filterfunktion. Zusätzlich sind auch weitere kleinere Veränderungen getätigt worden. Die Lupe von Heimerl et al. wird zur Dokumentenexploration genutzt. Die Dokumente sind als Streudiagramm dargestellt. Sobald mit der Lupe über Dokumente gefahren wird, werden bis zu zehn Terme mit den höchsten Dokumentfrequenzen neben der Lupe angezeigt. Als zusätzliche Information erhält der Nutzer die absolute Dokumentfrequenz in Bezug auf die Anzahl der Dokumente, die unter der Lupe liegen, neben jedem der dargestellten Terme.

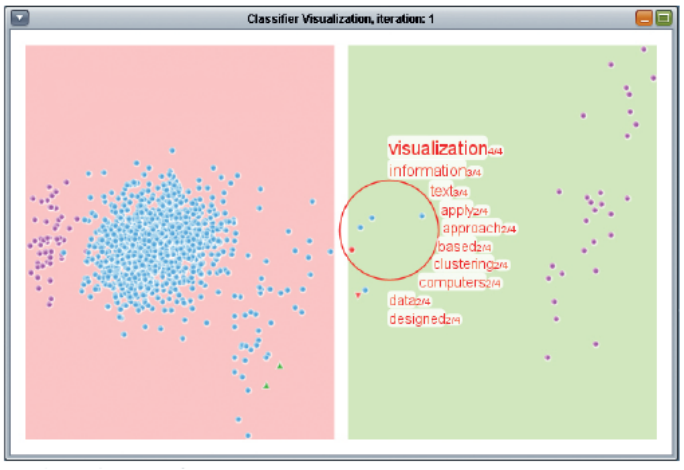

Abbildung 3.6: Die Hauptansicht des von Heimerl et al. entwickelten Programms mit der interaktiven Lupe, der „Term Lens“ [HKBE12].

Diese Anzeige der Dokumentfrequenz wurde in dieser Arbeit nicht auf die Art und Weiße umgesetzt, da die Frequenz der Terme in der als Histogramm enthaltenen Filterfunktion dargestellt wird. Eine weitere neu implementierte Funktionalität ist die Auswahl der Frequenz. Es kann zwischen der Dokumentfrequenz und der Termfrequenz gewechselt werden, um die Exploration möglichst facettenreich zu gestalten. Auch die entwickelte Filterfunktion hilft bei einer besseren Analyse der thematischen Ausrichtung der Dokumente. Die alleinige Anzeige der zehn Terme mit der höchsten Dokumentfrequenz reicht dazu nicht aus. Nach Luhn [Luh58] kann, wie schon zuvor erwähnt, drauf geschlossen werden, dass die für den Inhalt relevantesten Terme weder die mit den höchsten Frequenzen, noch die mit den niedrigsten sind. Tendenziell liegen die relevanten Terme in der Mitte. Aufgegriffen wurde diese These auch von Chuang et al. [CMH12] um unwichtige Worte aus Texten herauszufiltern.

4 Konzepte

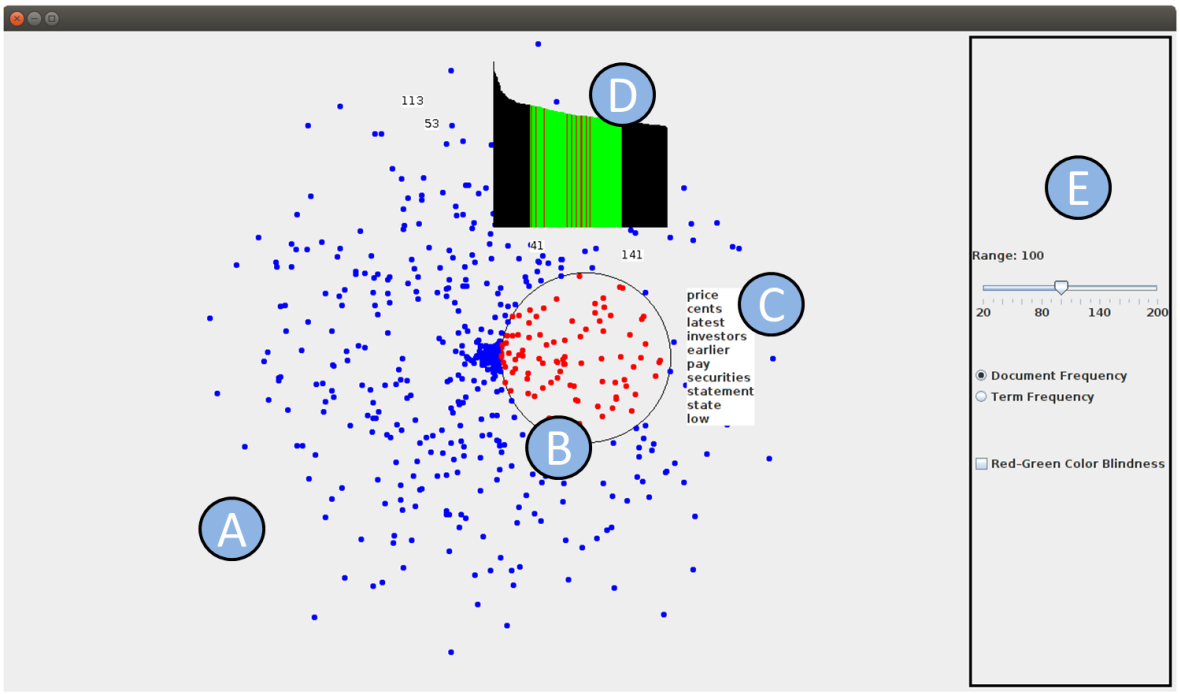

Abbildung 4.1: Das in dieser Arbeit entwickelte Programm besteht aus dem Dokumenten-Panel (A), der interaktive Lupe (B), dem Stichwort-Panel (C), dem Histogramm-Panel (D) und dem Auswahl-Panel (E).

Die grundsätzliche Idee für das Konzept ist die Entwicklung einer interaktiven Technik um Dokumente möglichst gut zusammenzufassen. Speziell sollen bestimmte Dokumente von Nutzern markiert werden können, damit diese automatisch zusammengefasst werden. Zu diesen sollen Informationen im direkten Kontext der Interaktionstechnik angezeigt werden.

Für die Darstellung und detaillierte Untersuchung der Dokumentensammlung soll ein explorativer Ansatz verwendet werden. Dazu sollen die Dokumente als kleine Kreise auf einer zweidimensionalen Ebene dargestellt werden. Die kreisförmigen Glyphen sollten nach inhaltlicher Ähnlichkeit der Dokumente verteilt werden. Dadurch können sich Cluster bilden und Muster gefunden werden. Mit einer nutzergesteuerten interaktiven Lupe sollen diese platzierten Glyphen exploriert werden können. Dabei sollen ausgewählte Glyphen markiert werden können, deren Inhalt zusammengefasst und in einem Panel neben der Lupe angezeigt wird. Dadurch wird es Nutzen ermöglicht, auch den Inhalt

großer Datensätze zu erfassen, ohne die Texte lesen zu müssen. Zusammengefasst werden sollen die Dokumente anhand der Term- und Dokumentfrequenz. Der anzuzeigende Inhalt soll dabei durch Stichworte repräsentiert werden. Ein Filter, der als Histogramm in einem weiteren Panel neben der Lupe dargestellt werden soll, ermöglicht die Festlegung von Auswahlkriterien, nach denen die Stichworte gewählt werden, anzupassen. Nur Terme aus einem Bereich zwischen zwei bestimmten Frequenzen werden im Stichwort-Panel angezeigt.

Eine detaillierte Beschreibung der Konzepte wird in den folgenden Abschnitten gegeben.

4.1 Dokumenten-Panel

In Abbildung 4.1(A) ist das Dokumenten-Panel zu sehen. Das Dokumenten-Panel ist die Fläche auf der mit der interaktiven Lupe operiert wird und die visuell repräsentierten Daten exploriert werden können. Die Dokumente werden als zweidimensionales Streudiagramm auf dem Dokumenten-Panel angezeigt. Dabei wird jedes Dokument aus dem Datensatz durch einen Punkt platziert und als kleiner Kreis dargestellt. Informationen können von Nutzern leichter aus visuell dargestellten Daten entnommen werden, als aus unverarbeiteten. Große Dokumentensammlungen können vom Menschen nicht manuell in kurzer Zeit analysiert werden. Durch diese zweidimensionale Art der Darstellung ist es möglich große Datensammlungen kompakt darzustellen und mit verschiedensten Interaktionen zu verbinden.

Die Verteilung der kreisförmigen Glyphen steht für die Ähnlichkeit zwischen den Dokumenten zueinander. Somit können die Abstände zwischen den Glyphen als Maß für die Ähnlichkeit beziehungsweise Unähnlichkeit gesehen werden. Glyphen die nah beieinander liegen, haben eine höhere Ähnlichkeit als weit voneinander entfernte. Das heißt, wenn zwei gleiche Dokumente im Datensatz vorhanden wären, würden deren Kreise auf einen gemeinsamen Punkt auf dem Dokumenten-Panel fallen. Diese Art der Verteilung soll zu Clusterbildungen führen und die thematische Ausrichtung der Dokumente widerspiegeln. Speziell für große Dokumentensammlungen ist dies eine Möglichkeit die Daten übersichtlich darzustellen und Nutzern Informationen zu geben, die zuvor nicht erkennbar sind.

Die Kreise werden um einen gedachten Punkt, der im Zentrum des Dokumenten-Panels liegt gezeichnet, um eine gleichmäßige Verteilung zu erhalten. Zusätzlich, um den gegebenen Platz möglichst gut auszunutzen, werden die Werte der zugrundeliegenden Punkte in Richtung des Randes skaliert. Bei der Limitierten Anzahl an Pixeln auf einem Bildschirm ist es speziell bei Darstellungen mit vielen repräsentativen Kreisen sinnvoll, diese über den gesamten vorhandenen Platz zu verteilen, um kein visuelles Durcheinander zu produzieren. Nichtsdestotrotz muss darauf geachtet werden, dass zu Beginn kein Kreis außerhalb des sichtbaren Bereiches des Dokumenten-Panels platziert wird. Es soll damit sichergestellt werden, dass der komplette Datensatz für Nutzer beim Programmstart zu erkennen ist und diese einen guten Überblick bekommen. Die Nutzer können später das Streudiagramm selbständig mittels Interaktion auf relevante und interessante Teilbereiche analysieren und die Ansicht anpassen.

Begründet durch die Tatsache, dass die Verteilung der repräsentativen Kreise, in dieser Arbeit mit t-SNE generiert, sich in jedem Durchlauf verändert und es auch zu ungünstigen Ergebnissen kommen kann,

bietet es sich an die Verteilungen zu speichern. Eine Verteilung mit einem oder mehreren Ausreißern, das heißt es existieren Kreise, deren zugrundeliegende Punkte Werte besitzen, die stark von denen der anderen abweichen, ist tendenziell ungeeignet für die Repräsentation als Streudiagramm, speziell, wenn nicht ausreichend viele Interaktionsmöglichkeiten gegeben sind. Denn für einen solchen Fall würde die Darstellung aller repräsentativen Kreise im sichtbaren Bereich des Dokumenten-Panels zu starken Überdeckungen im Zentrum des Streudiagramms führen, exemplarisch zu sehen in Abbildung 4.2. Doch können gerade diese Ausreißer interessant sein, denn es gilt herauszufinden, inwiefern sich die Werte deutlich von den anderen unterscheiden und wie es dazu kommt. Um diese Ausreißer nicht zu verlieren und trotzdem keine starke Überdeckung in der Visualisierung zu erhalten, könnten die Ausreißer vor der Anzeige bestimmt werden. Die kreisförmigen Glyphen dieser Ausreißer würde außerhalb des sichtbaren Bereiches platziert werden und im sichtbaren Bereich die sich ähnlicheren Glyphen. Damit die Ausreißer trotzdem gefunden werden können, könnten kleine Pfeile am Rand des Dokumenten-Panels als Indikatoren angebracht werden. Mit zusätzlichen Interaktionen kann zu diesen Ausreißern navigiert werden. Diese Funktion zum Ausreißer finden und verarbeiten, wurde in dieser Arbeit nicht implementiert.

Zusätzlich muss durch die Sicherung in eine externe Datei, die Verteilung für einen gleichbleibenden Datensatz nicht bei jedem Programmstart neu berechnet werden und somit kann die Laufzeit optimiert werden. Auch falls mehrere Nutzer mit der gleichen erstellten Platzierung arbeiten sollen, kann diese externe Datei genutzt und ausgetauscht werden.

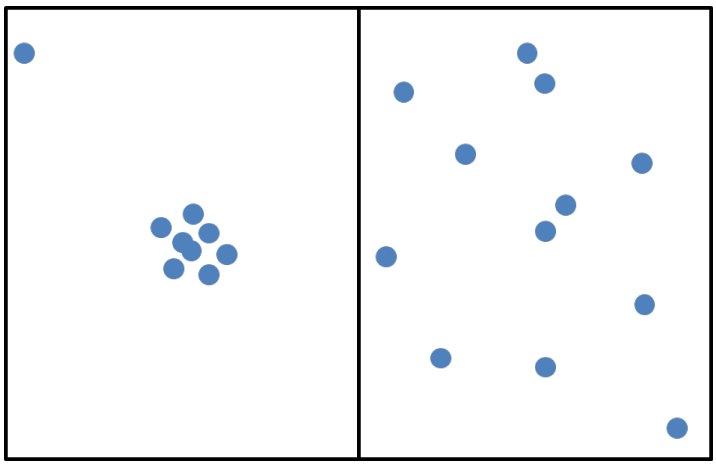

Abbildung 4.2: Schematische Darstellungen einer Punkteverteilung. Jeder Punkt wird durch einen kleinen Kreis repräsentiert. Auf der linken Seite enthält die Verteilung einen Ausreißer, die Verteilung auf der rechten Seite nicht.

Allen kreisförmigen Glyphen wird eine einheitliche dunkelblaue Farbe zugewiesen, damit diese sich vom Hintergrund und auch von der Lupe abheben. Bei Kontakt mit der interaktiven Lupe wird die Farbe der Glyphen geändert. Details und Aspekte im Bezug auf die Lupe werden im nächsten Abschnitt erklärt. Die Größe der Glyphen ist so festgelegt, dass es problemlos möglich ist die einzelnen Kreise zu erkennen. Dennoch sollte darauf geachtet werden, dass der Durchmesser der Kreise nicht zu groß gewählt wird, um Überschneidungen und Überlagerungen zwischen den Kreisen auf ein

Minimum zu reduzieren. Nichtsdestotrotz können Streudiagramme mit großen zugrundeliegenden Datensätzen schnell überladen sein und eine Betrachtung von kleineren Teilbereichen ist nur bedingt möglich. Um dieses Problem zu beheben kann eine Zoom-Funktion verwendet werden, die im Rahmen dieser Arbeit nicht umgesetzt wurde. Diese Zoom-Funktion ermöglicht eine Vergrößerung, beziehungsweise Verkleinerung, des Streudiagramms, genauer gesagt der Abstände zwischen den Kreisen. Der Mittelpunkt der Zoom-Funktion ist der Mauszeiger. Wenn der Mauszeiger sich nicht über dem Panel befindet, hat das aktivieren der Zoom-Funktion keinen Effekt. Um diese Zoom-Funktion zu unterstützen kann das vergrößerte Streudiagramm mittels Pfeiltasten in alle Richtungen verschoben werden. Dadurch wird speziell bei großen Datensätzen die Suche nach Mustern oder Bereichen, die für genauere Betrachtungen von Interesse sein können, vereinfacht. Um interessante Bereiche gezielt auszuwählen und das Streudiagramm zu explorieren, kann die interaktive Lupe genutzt werden.

4.2 Interaktive Lupe

Die interaktive Lupe ist das Herzstück dieser Arbeit und der Grundstein für alle folgenden Konzepte, zu sehen in Abbildung 4.1(B). Die Lupe wird gezeichnet, sobald sich der Mauszeiger über dem Dokumenten-Panel befindet. Die Lupe ist kreisförmig nach dem Beispiel der Vergrößerungslupen aus der realen Welt. Die Interaktionstechnik dieser Lupe beruht darauf, dass sie an den Mauszeiger und dessen Bewegungen gebunden ist. Der Mauszeiger bildet dabei den Mittelpunkt des Kreises, das heißt der Bereich, der durch die Lupe beeinflusst wird, liegt immer um den Mauszeiger herum.

Die essenzielle Aufgabe der Lupe ist die Auswahl und Markierung von Dokumenten, beziehungsweise den repräsentativen Kreisen. Die Fläche unter der Lupe, das heißt der Bereich innerhalb des gezeichneten Kreises, bestimmt welche Glyphen weiterverarbeitet werden, um den Nutzern zusätzliche Informationen über die korrespondierenden Dokumente zu liefern. Die Auswahl wird durch Interaktion von den Nutzern bestimmt. Glyphen, die nicht unter der Lupe liegen werden in diesem Schritt nicht weiter beachtet. Erst wenn die Lupe verschoben wird, ändern sich die für weitere Verarbeitungen relevanten Glyphen.

Die Lupe dient als visuelles Bindeglied zwischen den Nutzern und der Auswahl von Dokumenten. Grundlage sind die auf der Ebene platzierten Dokumente. Um diese zusammenfassen zu können, müssen mehrere repräsentative Kreise markiert werden. Ohne den Einsatz der Lupe müssten die Kreise zum Beispiel mit dem Mauszeiger angeklickt werden. Diese Aufgabe wäre zeitaufwendig und bei großen Datensätzen manuell nicht durchführbar. Die Lupe ermöglicht es Nutzern, nur durch die Bewegung mit dem Mauszeiger, Dokumente zu markieren. Diese Alternative ist wesentlich schneller und flexibler. Eine Einschränkung für die Lupe ist jedoch, dass zu markierende Kreise räumlich nah beieinander liegen müssen. Dieses Problem spielt aber in dieser Arbeit keine größere Rolle, da die Dokumente nach ihrer Ähnlichkeit zueinander platziert werden. Die Problemstellung inhaltlich ähnliche Dokumente zu finden und zu analysieren ist somit ohne weiteres lösbar. Der Inhalt eines Dokuments wird durch alle Terme, die im entsprechenden Dokumentvektor stehen, repräsentiert. Beispielsweise kann aus den Glyphen, beziehungsweise Dokumentvektoren, die unter der Lupe liegen, die Dokumentfrequenzen gebildet werden und eine bestimmte Auswahl der sortierten Stichworte im Stichwort-Panel angezeigt werden. Nähere Details zu diesem Konzept im folgenden Abschnitt.

Eine weitere optische Anforderung an die Lupe ist die Liniendicke der Kreisform. Diese ist dünn gewählt, damit möglichst wenig vom kreisförmigen Rand der Lupe überdeckt wird. Zusätzlich wird die Lupe über den kleinen Kreisen des Streudiagramms auf das Dokumenten-Panel gezeichnet, um eine Auswahl und Markierung der repräsentativen Kreise zu vereinfachen, da die Lupe nicht unter dem Streudiagramm verschwindet. Die interaktive Lupe wird in einem festgelegten Radius um den Mauszeiger gezeichnet. Dieser ist so gewählt, dass es möglich ist kleinere Teilmengen von Glyphen unter der Lupe einzuschließen, um diese für genauere Untersuchungen auszuwählen. Theoretisch ist es sinnvoll solche kleinen Teilbereiche oder Cluster des Streudiagramms mit der Lupe zu untersuchen. Dementsprechend ist der Lupenradius zu Beginn des Programms gewählt. Nichtsdestotrotz wird den Nutzern die Funktionalität geboten, mittels Mausrad, den Radius der Lupe beliebig zu vergrößern und zu verkleinern. Mit dieser Funktionalität und der Zoom-Funktion des Dokumenten-Panels ist es Nutzern möglich, auch nur einzelne Glyphen oder den kompletten Datensatz auf einmal mit der Lupe einzuschließen.

Als visuelle Rückmeldung für die Nutzer werden die ausgewählten Kreise in einer anderen Farbe dargestellt. Die dunkelblaue Farbe wird zu einem Rot, wenn ein Kreis von der Lupe eingeschlossen ist. Dadurch werden eindeutig die markierten Kreisen von den anderen abgegrenzt. Auf den ersten Blick ist dieser Funktion kein größeres Gewicht zuzuordnen, doch ist es vor allem bei Interaktionen notwendig den Nutzern Rückmeldung zu geben. Dadurch wissen Nutzer, dass etwas passiert und sie Einfluss auf das Programm nehmen können.

Um Nutzern zu helfen mit dem später vorgestellten Auswahl-Panel zu interagieren, besteht die Möglichkeit die Lupe an einem bestimmten Punkt festzusetzen. Ohne diese Funktionalität müsste die Lupe von einer interessanten Region gezwungenermaßen entfernt werden, nur um die Auswahlkriterien näher einzustellen. Durch diese Methode kann die Lupe vom Mauszeiger losgelöst werden und wird erst wieder an dessen Bewegungen angepasst, wenn die Nutzer dies für nötig erachtet. Eine mögliche Alternative zu dieser Funktion ist die Vergabe von Tastenkombinationen für alle Auswahlmöglichkeiten. Wenn die Maus nicht zwangsweise zum Bedienen der Auswahlkriterien benötigt wird, muss sie nicht festgesetzt werden. In dieser Arbeit wurden diese Tastenkombinationen weggelassen, um die Einstiegshürde niedrig zu halten. Doch sind Tastenkombinationen speziell für erfahrene Nutzer von großer Bedeutung, um schnell mit dem Programm interagieren zu können.

Die weiteren Verarbeitungsschritte, die von den ausgewählten Dokumente durchlaufen werden und die dazugehörigen Funktionalitäten der interaktiven Lupe, werden in den folgenden Abschnitten bei dem entsprechenden Panel, auf dem die Anzeige der Informationen erfolgt, erklärt.

4.3 Stichwort-Panel

Das Stichwort-Panel steht in direktem inhaltlichem und lokalem Zusammenhang mit der Lupe. Inhaltlich gesehen werden zusätzliche Informationen über die Dokumente im Stichwort-Panel angezeigt, deren repräsentative Kreise unter der Lupe liegen. Der lokale Zusammenhang zur interaktiven Lupe entsteht dadurch, dass das Stichwort-Panel direkt neben der Lupe gezeichnet wird.

Das Stichwort-Panel wird auf der rechten Seite der Lupe auf das Dokumenten-Panel gezeichnet, zu sehen in Abbildung 4.1(C). Dadurch müssen Nutzer ihre Aufmerksamkeit nicht aufteilen und können

sich auf das Streudiagramm und die angezeigten Informationen im Stichwort-Panel gleichzeitig konzentrieren. Bei getrennten Ansichten können die Informationen zwar großflächiger dargestellt werden, doch geht dabei der Kontext der markierten repräsentativen Kreise verloren. Die Informationen in dieser Arbeit werden als Stichworte angezeigt und verbrauchen somit keinen großen Platz und können problemlos neben der Lupe angezeigt werden.

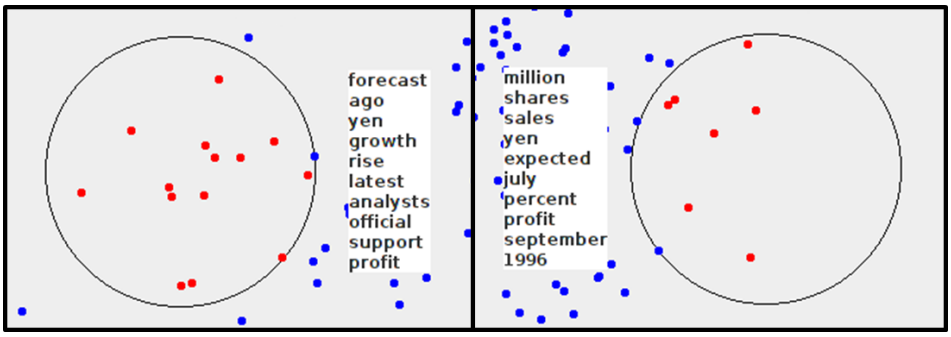

Abbildung 4.3: Auf der linken Seite ist das Stichwort-Panel neben der Lupe im Standardfall dargestellt. Im rechten Bild ist das Stichwort-Panel auf die linke Seite verschoben worden.

Um auch Dokumente untersuchen zu können, die am rechten Rand des Dokumenten-Panels platziert wurden, wird das Stichwort-Panel auf die linke Seite der Lupe verschoben, zu sehen in Abbildung 4.3. Wenn dies nicht der Fall wäre, würden, sobald das Stichwort-Panel, komplett oder auch nur teilweise, nicht mehr im sichtbaren Bereich des Dokumenten-Panels liegt, die Informationen verloren gehen, die im Stichwort-Panel angezeigt werden. Wenn der Radius der Lupe sehr groß gewählt wird, kann es dazu kommen, dass das Stichwort-Panel auf beiden Seiten neben der Lupe nicht dargestellt werden kann ohne abgeschnitten zu werden. Da in diesem Spezialfall die Fläche innerhalb der Lupe sehr groß wird, ist es kein Problem das Stichwort-Panel innerhalb der Lupe darzustellen. Dieser Spezialfall wurde im Rahmen dieser Arbeit nicht umgesetzt, da er nicht beim intendierten Gebrauch der Lupe auftritt.

Der Hintergrund des Stichwort-Panels ist durchsichtig um die Überdeckung mit dem Streudiagramm auf ein Minimum zu reduzieren und somit den Kontext der markierten Glyphen nicht zu verlieren. In diesem Stichwort-Panel werden bis zu zehn Terme angezeigt, abhängig davon, wie viele oder ob Glyphen unter der Lupe liegen. Die dargestellten Terme werden weiß hinterlegt, damit sie für Nutzer immer noch lesbar sind, auch wenn sie direkt über Teilen des Streudiagramms liegen.

Es gibt zwei Konzepte wie, beziehungsweise welche, Terme im Stichwort-Panel angezeigt werden. Beide Konzepte beruhen darauf, dass der Inhalt von Dokumenten mit repräsentativen Schlüsselwörtern wiedergegeben werden kann. Da schon bei der Erstellung der Dokumentvektoren Stoppworte und Ähnliches entfernt werden, liegen in diesem Schritt nur inhaltlich relevante Terme vor. Um nun mehrere Dokumente zusammenzufassen, werden die Terme vereint und sortiert angezeigt. Dadurch werden Stichworte, die repräsentativ für alle zusammengefassten Dokumente stehen, durch Termgewichtung hervorgehoben. Somit können Nutzer auf den gemeinsamen Inhalt von Dokumenten schließen.

Das erste konkrete Konzept beruht auf der Termfrequenz. Die Idee ist den Inhalt der Dokumente anhand der gesammelten Termfrequenzen zu analysieren. Dafür werden alle repräsentativen Kreise, die unter der Lupe liegen, separat von den restlichen bearbeitet. Nur diese vom Nutzer gewählten Kreise werden in diesem Schritt verarbeitet. Alle Terme, der hinter den Kreisen liegenden Dokumente, werden gesammelt und in einer Liste gespeichert. Als zusätzliche Information wird neben jedem Term die entsprechende Termfrequenzen vermerkt. Es werden keine doppelten Einträge erstellt und bei Termen, die in mehr als einem Dokument vorkommen, werden die Termfrequenzen addiert und in einem gemeinsamen Eintrag gesammelt. Die Einträge in der Liste werden absteigend nach der Termfrequenz sortiert. Einträge mit gleicher Termfrequenz werden in dieser Arbeit lexikographisch sortiert.

Das zweite Konzept wird mittels der Dokumentfrequenz umgesetzt. Die Idee der inhaltlichen Analyse ist in diesem Konzept die gleiche, außer dass die Dokumentfrequenzen für die einzelnen Terme über den ausgewählten Dokumenten bestimmt wird. Um dieses Konzept umzusetzen werden ähnlich zu dem vorherigen Ansatz die Terme in einer Liste gespeichert. Jedoch werden bei dieser Methode nicht die Termfrequenzen gespeichert, sondern es werden die Dokumentfrequenzen der einzelnen Terme bestimmt. Nach diesen werden die Einträge in der Liste absteigenden sortiert. Einträge mit gleicher Dokumentfrequenz werden lexikographisch sortiert.

Die Prinzipien der Termgewichtung wurden schon im Grundlagen Kapitel erklärt. Auf diesen basieren beide vorgestellten Konzepte. Die Terme werden nach der jeweiligen Gewichtung sortiert, um die relevantesten Terme zu finden. Zum Beispiel werden bei der Termfrequenz Terme, die häufig in einem Dokument auftauchen, als wichtig angesehen. Deswegen werden die Terme absteigende sortiert und angezeigt.

Genauer gesagt werden im Stichwort-Panel die ersten zehn Terme aus der Liste, von oben nach unten, angezeigt. Diese Terme entsprechen den Termen mit der höchsten Termfrequenz, beziehungsweise Dokumentfrequenz. Der Ansatz, dass die wichtigsten Terme, diejenigen mit den höchsten Frequenzen sind, wird im nächsten Abschnitt genauer beleuchtet. Deswegen wird auch die Anzeige der ersten zehn Terme mit der Einführung der Filterfunktion, im nächsten Abschnitt, angepasst.

Um auch die Spezialfälle abzudecken, werden diese hier näher beleuchtet. Für den Fall, dass die Liste weniger als zehn Terme enthält, werden alle Terme aus der Liste angezeigt, weiterhin in sortierter Reihenfolge. Dieser Fall tritt nur bei sehr kurzen Dokumenten ein. Zusätzlich gibt es noch die Spezialfälle, dass nur ein oder kein Dokument unter der Lupe liegt. Wenn kein Dokument markiert ist, werden naheliegenderweise bei beiden Konzepten keine Terme im Stichwort-Panel angezeigt. Für den Fall, dass ein Dokument unter der Lupe liegt, werden bei dem auf der Termfrequenz basierenden Konzept, wie auch für den Standardfall, bis zu zehn Terme mit den höchsten Termfrequenzen angezeigt. Beim Konzept, basierend auf der Dokumentfrequenz, werden keine Terme angezeigt, da die Dokumentfrequenz aller Terme eins ist und somit die inhaltliche Aussagekraft im Vergleich fehlt.

Damit Nutzer die Frequenz (Term- und Dokumentfrequenz) der angezeigten Terme besser unterscheiden können, wäre es möglich die Schriftgröße der angezeigten Terme anzupassen. Grundsätzlich ist sonst davon auszugehen, dass Terme, die im Stichwort-Panel weiter oben angezeigt werden, eine höhere Frequenz besitzen. Durch die Anpassung der Schriftgröße könnten Nutzer beispielsweise erkennen, ob alle Terme die gleiche Frequenz haben, oder sich unterscheiden. Beide Fälle können durchaus auftreten.

Wie zuvor schon erwähnt ist es nicht zielführend immer nur die zehn Terme mit den höchsten Frequenz anzuzeigen. Deswegen können Nutzer mit dem Histogramm-Panel interagieren und einen Filter erstellen, um die Anzeige des Stichwort-Panels zu beeinflussen.

4.4 Histogramm-Panel

Das Histogramm-Panel ist in Abbildung 4.1(D) zu sehen und ist die visuelle interaktive Darstellung einer Filterfunktion.

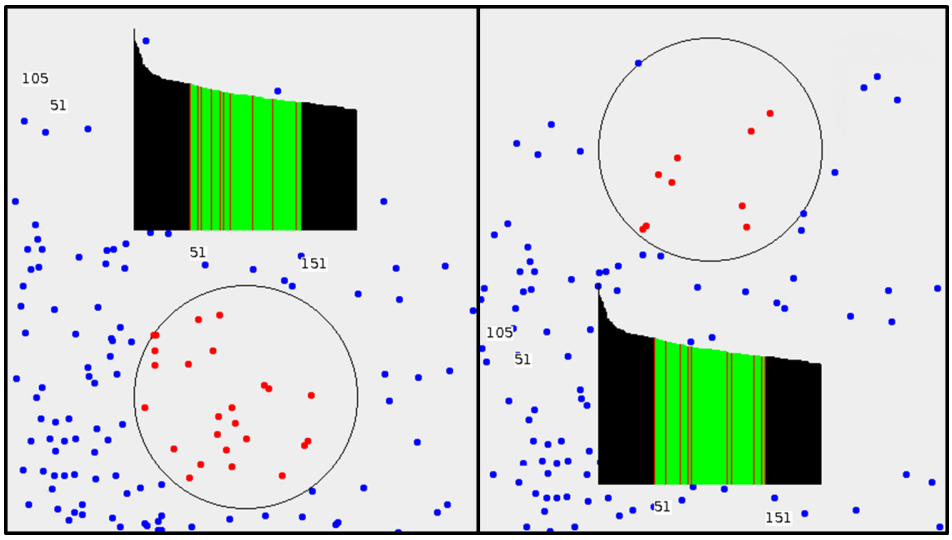

Abbildung 4.4: Links ist das Histogramm-Panel im Normalzustand über der Lupe zu sehen. Rechts ist das Histogramm-Panel auf der Unterseite der Lupe, da die Lupe zu nah an den oberen Rand gekommen ist.

Das Histogramm-Panel befindet sich über der Lupe und wird genauso wie das Stichwort-Panel verschoben, wenn es in Kontakt mit dem Rand des Dokumenten-Panels kommt, zu sehen in Abbildung 4.4. Nur erfolgt die Verschiebung in diesem Fall von oben nach unten und nicht von rechts nach links.

Das Histogramm-Panel zeigt, wie der Name schon suggeriert, ein Histogramm an. Ein Histogramm ist die visualisierte Darstellung einer Häufigkeitsverteilung und repräsentiert in dieser Arbeit einen Filter. Das zu filternde Element ist die Liste mit Termen, die für die Anzeige im Stichwort-Panel genutzt wird. Dazu wird ein Wörterbuch mit allen Termen des spatialisierten Datensatzes erstellt. Je nachdem, welches Konzept von den Nutzern gewählt wurde, wird zusätzlich die aufaddierte Termfrequenz oder die bestimmte Dokumentfrequenz für jeden Term gespeichert. Das heißt wenn die Termfrequenz für die Sortierung des Stichwort-Panels gewählt wird, ist die Termfrequenz auch für die Erstellung des Wörterbuchs zu verwenden. Das Konzept für die Sortierung im Stichwort-Panel und der Ansatz für das Wörterbuch des Histogramm-Panels lässt sich nicht getrennt voneinander wählen. Das gewählte Konzept wird immer für beide Panel genutzt.

Die grundsätzlich Idee ist es den Nutzern einen Bereich wählen zu lassen, aus dem die Terme, die im Stichwort-Panel angezeigt werden, stammen müssen. Dieser Auswahlbereich kann so gewählt werden, dass nicht immer die zehn Terme mit der höchsten Frequenz angezeigt werden, sondern zum Beispiel auch aus dem unteren Bereich des Wörterbuchs stammen und somit Vertreter kleiner Frequenzen sind. Es wird davon ausgegangen, dass speziell die Terme mit mittleren Frequenzwerten am aussagekräftigsten für den Inhalt eines Dokuments sind, wie auch schon in Kapitel 3 belegt. Deswegen ist die Funktionalität des Histogramm-Panels essentiell, um die thematische Ausrichtung der Dokumente bestmöglich zu untersuchen.

Konkreter kann zu dieser Arbeit gesagt werden, dass den Nutzern die Möglichkeit gegeben wird einen Bereich, der zwischen 20 und 200 Terme umfasst, zu wählen. Die 20 Terme repräsentieren eine feingranulare Filterung, während die 200 Terme einen breiten Bereich abdecken. Das heißt, wenn der größte Bereich an der Spitze des Wörterbuchs platziert wird, führt dies mit hoher Wahrscheinlichkeit zu den gleichen zehn angezeigten Termen, wie bei einem Gebrauch der Lupe ohne Filterfunktion. Als Standardeinstellungen wird zu Beginn ein Bereich von 100 Termen gewählt, welcher die obersten 100 Terme des Wörterbuchs umfasst. Zur Laufzeit können Nutzer den Umfang und die Position des Auswahlbereichs regulieren.

Um die Funktion zusammenzufassen: Die Liste für das Stichwort-Panel wird, wie im vorherigen Abschnitt erläutert, erstellt. Dann wird diese Liste von oben, das heißt bei der höchsten Frequenz startend, durchlaufen und für jedes Wort wird überprüft, ob es im gewählten Bereich des Wörterbuchs liegt. Falls zehn Terme gefunden wurden oder die Liste komplett durchlaufen ist, werden die gefilterten Terme im Stichwort-Panel angezeigt.

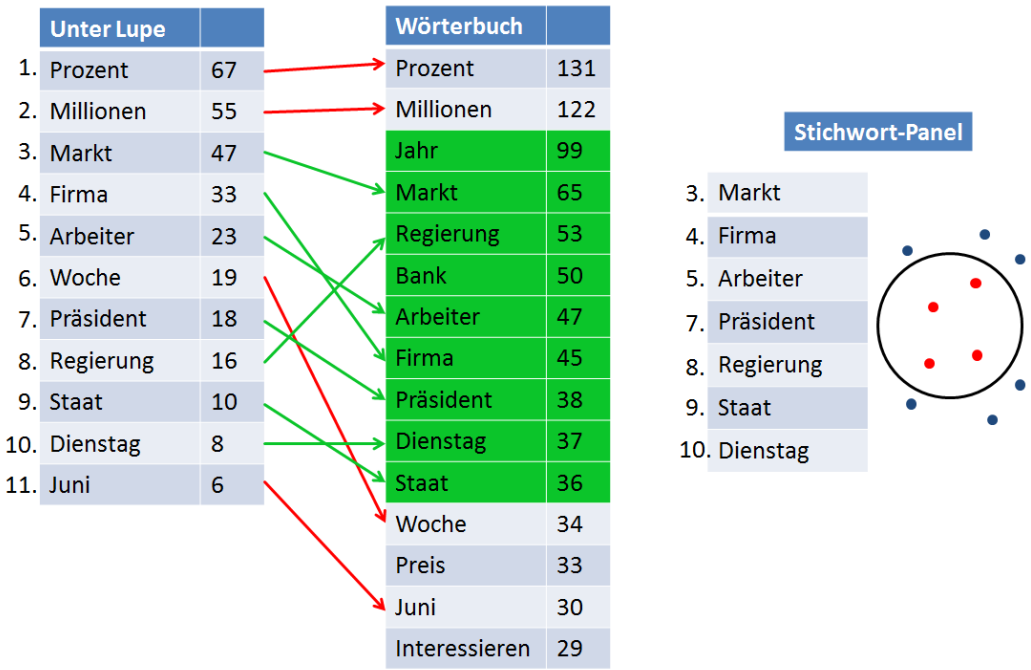

Abbildung 4.5: Schematische Darstellung der Filterfunktion anhand eines Beispiels.

Ein schematisches Beispiel für die Filterfunktion kann in Abbildung 4.5 gesehen werden. Auf der linken Seite befindet sich die Liste mit den Termen unter der Lupe. Zusätzlich ist die Frequenz angezeigt. In der Mitte ist das aus dem kompletten Datensatz generierte Wörterbuch zu sehen. Die grünen Zellen markieren den Auswahlbereich. Auf der rechten Seite ist die Lupe mit dem Stichwort-Panel abgebildet. Die Pfeile zwischen der Liste auf der linken Seite und dem Wörterbuch stellen die einzelnen Verbindungen zwischen zwei gleichen Termen dar. Wenn der Pfeil rot ist, liegt der Term nicht im Auswahlbereich des Wörterbuchs. Ein grüner Pfeil soll symbolisieren, dass der entsprechende Term im Auswahlbereich des Wörterbuchs liegt. Diese Terme werden daher im Stichwort-Panel neben der Lupe angezeigt. Die Reihenfolge entspricht der Sortierung in der linken Liste. Hierbei kann auch erkannt werden, dass die Reihenfolge, in der die Terme im Auswahlbereich stehen nicht ausschlaggebend ist.

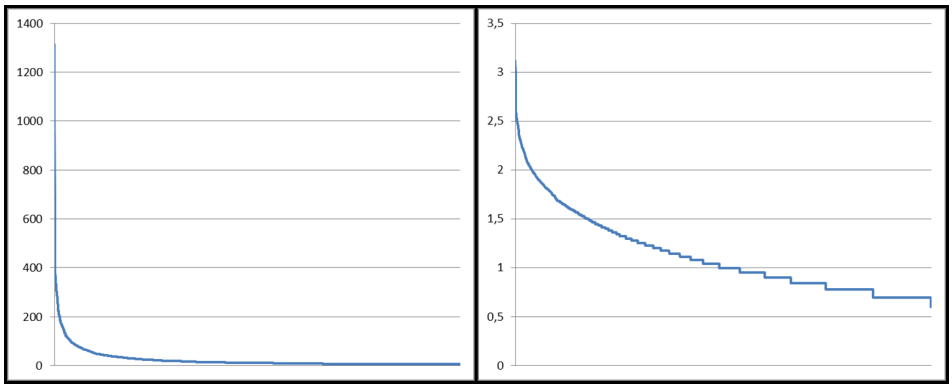

Abbildung 4.6: Links zu sehen ist eine Zipf Verteilung für einen exemplarischen Datensatz mit 5000 Termfrequenzen. Rechts ist der gleiche Datensatz zu sehen, doch sind die Werte mit dem Logarithmus zur Basis 10 verrechnet.

Im Histogramm wird jeder Term durch eine Säule visuell dargestellt. Die Höhe der Säulen steht für die Frequenz des entsprechenden Terms. Die Höhe wird zusätzlich an die Größe des Histogramm-Panels angepasst. Da es sich bei dieser Anordnung der Terme um eine Zipf Verteilung handelt, wird eine logarithmische Skalierung gewählt. Das Zipfsche Gesetz („Zipf's Law“) besagt, dass bei einer nach der Frequenz sortierten Liste von Termen in einem Dokument, der Rang mit der Frequenz des Terms speziell verteilt zusammenhängt. Es ergibt sich für einen Term auf dem Rang i mit der Frequenz F_i folgende Gleichung, bei der $\propto$ für den proportionalen Zusammenhang steht [MRS08]:

$$F_i \propto \frac{1}{i} \tag{4.1}$$

Das heißt, der zweithäufigste Term hat eine halb so große Frequenz wie der häufigste, der dritt häufigste ein drittel und so weiter. Da diese schnell abfallende Verteilung, zu sehen in Abbildung 4.6 links, bei einer linearen Skalierung zu schlechten Ergebnissen für das Histogramm führt, es gibt wenige sehr hohe Werte und viele sehr kleine, ist die Anwendung eines Logarithmus geeignet. In Abbildung 4.6 rechts ist dabei klar zu erkennen, dass die Funktion auf die der Logarithmus angewendet wurde, nicht so stark abfällt und somit die Werte näher beieinander liegen. Damit Nutzer die Frequenz, trotz der Skalierungen, aus dem Histogramm auslesen können, werden links neben der Zeichnung zwei

Werte dargestellt. Die Werte entsprechen dem Term mit der höchsten Frequenz im Auswahlbereich und demjenigen mit der niedrigsten. Die Werte sind auf der gleichen Höhe wie das Ende der Säule des entsprechenden Terms, um schnell zugeordnet werden zu können.

Der Auswahlbereich wird in grün dargestellt und die anderen Säulen in schwarz. Es werden maximal 50 weitere Terme links und rechts vom Auswahlbereich angezeigt. Dadurch bleibt der lokale Kontext erhalten, auch wenn der globale nicht angezeigt wird. Mögliche Ideen zu Änderungen in dieser Hinsicht sind dem Ausblick zu entnehmen. Um trotzdem die globale Orientierung nicht zu verlieren wird unter dem Histogramm der Rang im sortierten Wörterbuch für die höchste und niedrigste Frequenz im Auswahlbereich angezeigt. Die Terme, die als Stichworte im Stichwort-Panel angezeigt werden, werden als rote Säulen dargestellt. Diese roten Säulen helfen auch dabei den Auswahlbereich optimal festzulegen, da Nutzer wissen, in welchem Teil des Auswahlbereiches die angezeigten Terme liegen. Verschiedene Ansichten des Histogramm-Panels in unterschiedlichen Zuständen sind in Abbildung 4.7 zu sehen.

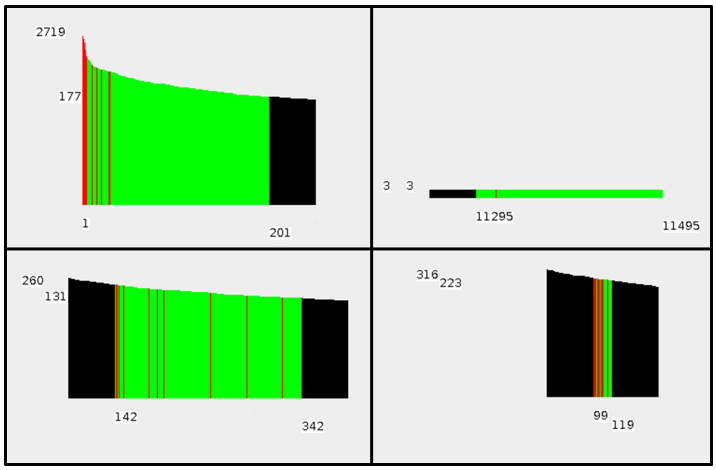

Abbildung 4.7: Links oben is das Histogramm am oberen Ende des Wörterbuchs zu sehen mit einem Auswahlbereich von 200. Rechts oben befindet sich das Histogramm am unteren Ende mit einem Auswahlbereich von 200. Links unten ist eine mittlere Ansicht mit einem Auswahlbereich von 100 zu sehen. Rechts unten befindet sich ein 20 Terme große Auswahlbereich in der Mitte des Wörterbuches.

4.5 Auswahl-Panel

In Abbildung 4.1(E) ist das Auswahl-Panel zu sehen. Dieses Panel bietet den Nutzern die in den vorherigen Kapitel erwähnten Auswahlmöglichkeiten an. Zum einen wird die Größe des Auswahlbereiches textuell angezeigt, zum anderen kann mittels eines Schiebereglers diese Größe manuell festgelegt werden. Auch die Auswahl, ob die Termfrequenz oder die Dokumentfrequenz zur Verarbeitung der Dokumente genutzt werden soll, kann hier von den Nutzern festgelegt werden. Standardmäßig wird die Dokumentfrequenz genutzt.

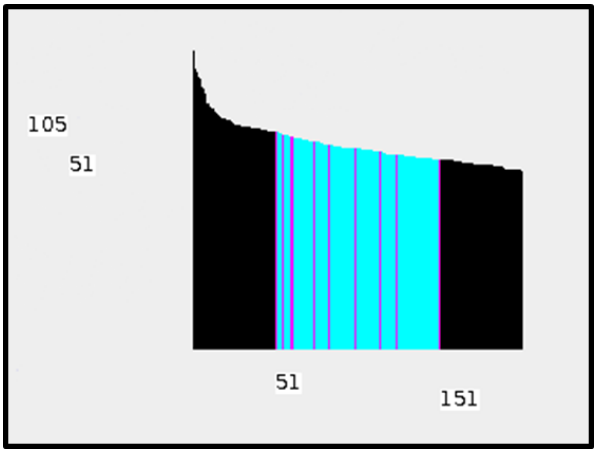

Abbildung 4.8: Das Histogramm mit der veränderten Farbdarstellung für Nutzer mit Rot-Grün-Schwäche.

Damit auch Nutzer mit einer Rot-Grün-Schwäche im Auswahlbereich die roten Striche auf grünem Hintergrund sehen können, gibt es die Möglichkeit diese Farbwahl zu ändern. In Abbildung 4.8 ist das angepasste Histogramm zu sehen.

5 Implementierung

In diesem Kapitel wird die Umsetzung der Konzepte anhand des entwickelten Prototypen näher beschrieben. Die Implementierung erfolgte in Java Version `JavaSE-1.7`. Alle Zeichnungen wurde mit der Klasse `Graphics` realisiert.

5.1 Datensatz

Die Dokumente müssen für die weitere Verarbeitung als Dokumentvektoren dargestellt werden. Dabei wird jeder Term zu einem Element im Vektor, wobei Stoppworte herausgefiltert werden. Für jeden Term werden die Termfrequenz und der TF-IDF Wert errechnet und gespeichert. Es gibt auch die Möglichkeit die Dokumentvektoren aus dem kompletten Wörterbuch der Sammlung zu erstellen. Der Vorteil dabei ist, dass bei Vergleichen von zwei Dokumentvektoren jeder Term durch die gleiche Dimension des Vektors repräsentiert wird. Bei einem großen Wörterbuch ist dies jedoch nicht empfehlenswert, da selbst ein Dokument mit nur ein paar Worten zu einem großen Dokumentvektor umgeformt wird und nahezu jedes Element eine null enthält.

Ein Beispiel für den Aufbau eines Dokumentvektors, der in eine externe Datei gespeichert wurde, kann Listing 5.1 entnommen werden. In der ersten Spalte werden die Termfrequenzen gespeichert, in der zweiten der TF-IDF Wert und in der dritten der entsprechende Term. Die Spalten sind jeweils durch einen Tabulator getrennt und Zeilen werden durch Enter beendet.

Listing 5.1 Beispielhafter Dokumentvektor generiert aus dem RCV1 Datensatz.

```
1     0.07233808021899067 accounts
1     0.07496300228474244 acquire
2     0.0953815399668045 acquisition
1     0.08319485739079886 activists
1     0.046837147499974446 amp
1     0.055462829484117 annual
1     0.06243688604797957 approved
1     0.0767112546634457 attorney
1     0.07194591440062494 benefit
2     0.054435684656537066 billion
14    0.24592217556392856 blue
2     0.10960431191470538 brown
1     0.1273535006720092 burry
1     0.06980421147767844 businesses
...
```

Der in dieser Arbeit verwendete Datensatz ist der Reuters Corpus Volume I (RCV1). Dieser enthält über 800.000 Nachrichtendaten, von denen Untermengen verwendet wurden. Nähere Informationen über diesen Datensatz können der Ausarbeitung von Lewis et al. entnommen werden [LYRL04].

5.2 Dokumenten-Panel

Die Dokumente, die als hochdimensionale Dokumentvektoren dargestellt werden, werden mittels dem in Abschnitt 2.3 vorgestellten t-SNE Algorithmus in eine zweidimensionale Form gebracht. Dafür werden die paarweisen Kosinus-Ähnlichkeit zwischen den Dokumentvektoren bestimmt und in einer Matrix gespeichert. Für die Erstellung der Matrix werden die TF-IDF Werte zweier Vektoren mit der Gleichung 2.1 verrechnet. Bei der Variante mit den kompletten, aus dem Wörterbuch erstellten, Dokumentvektoren können die Vektoren einfach Dimension für Dimension miteinander verrechnet werden. Für den Fall, dass die Vektoren nur aus dem entsprechenden Dokument erstellt worden sind, müssen die Dimensionen der Vektoren auf gleiche Terme überprüft werden. Terme, die nur in einem Vektor vorhanden sind, werden mit null verrechnet. Zur Erstellung der Ähnlichkeitsmatrix kann auch der Euklidischer Abstand zwischen den Dokumentvektoren berechnet werden. dabei muss aber darauf geachtet werden, dass der Abstand, also die Unähnlichkeit, berechnet wird.

Durch die von t-SNE durchgeführte Dimensionsreduktion können die Dokumentvektoren auf der 2D-Ebene, dem Dokumenten-Panel (realisiert durch ein `JPanel`), dargestellt werden. Dabei dient eine Koordinate als X- und die zweite als Y-Koordinate. Die Ergebniskoordinaten liegen um den Punkt $(0, 0)$, das heißt der Ursprung für das Streudiagramm sollte in der Mitte des Dokumenten-Panel liegen. Damit werden die Glyphen am besten über das Panels verteilt. Dazu wird auf alle X-Koordinaten das Ergebnis von `Dokumenten-Panel.getWidth()/2;` addiert. Für die Y-Koordinate wird `Dokumenten-Panel.getHeight()/2;` verwendet.

Um alle repräsentativen Kreise innerhalb des sichtbaren Bereiches des Dokumenten-Panel darzustellen und zusätzlich die Kreise so weitläufig wie möglich zu verteilen, wird ein Skalierungsfaktor berechnet. Dieser Skalierungsfaktor wird berechnet, indem der kleinste und größte X-Wert sowie der kleinste und größte Y-Wert herausgefunden wird. Dadurch werden alle Punkte und später die dazugehörigen Glyphen abgedeckt, da alle Koordinaten kleiner sind als die gefundenen. Mit diesen vier Werten und dem jeweiligen Abstand zum entsprechenden Rand vom Dokumenten-Panel kann ein Faktor für jede Richtung errechnet werden. Der kleinste Faktor wird auf alle Koordinaten der Punkte multipliziert. Es wird der kleinste Faktor gewählt um sicher zu stellen, dass alle Punkte und die dazugehörigen kreisförmigen Glyphen innerhalb des Panels bleiben. Eine schematische Darstellung davon kann in Abbildung 5.1 betrachtet werden. Die Punkte sind in dieser Darstellung nur zu Veranschaulichung als Kreise repräsentiert. Das Zentrum der Kreise stellt die eigentlichen Punkte dar. Das rote Viereck symbolisiert den Bereich, in dem alle Punkte liegen. Bestimmt wird dieser durch die größten beziehungsweise kleinsten X- und Y-Werte. Der Skalierungsfaktor wird von allen Punkten, bei denen ein Pfeil beginnt in die entsprechende Richtung berechnet. Der Skalierungsfaktor, der bei der Berechnung am oberen Rand berechnet wird, würde in diesem Fall gewählt werden, da er der kleinste ist.

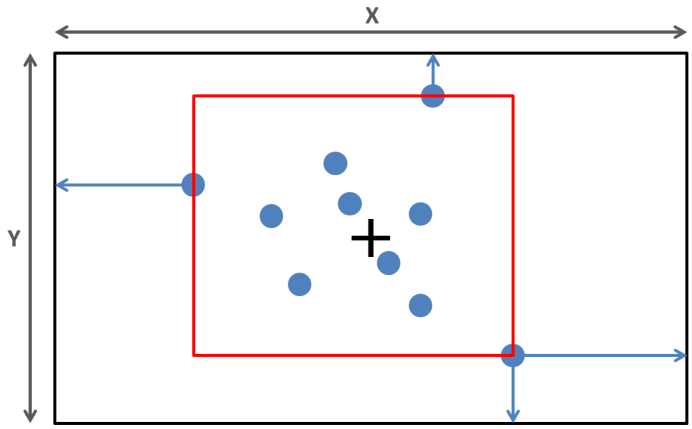

Abbildung 5.1: Beispielhafte Veranschaulichung der Berechnung des Skalierungsfaktors.

5.3 Interaktive Lupe

Die interaktive Lupe ist ein Kreis mit dem Mittelpunkt auf dem gleichen Pixel, auf das auch der Mauszeiger zeigt. Die Lupe erhält die Zeichenkoordinaten vom Mauszeiger mittels eines `MouseMotionListener` und kann mit diesem auch bei jeder Bewegung neu gezeichnet werden. Sie wird auf das Dokumenten-Panel über den kreisförmigen Glyphen gezeichnet. Um den Radius zu verändern kann der Nutzer das Mausrad betätigen. Mit einem `MouseWheelListener` kann abgefragt werden, wann der Radius verändert werden soll. Der wichtigste Teil der Lupe ist herauszufinden welche Glyphen unter der Lupe liegen. Dazu werden einfach die Abstände vom Mauszeigerpixel zu allen Mittelpunkten der Glyphen bestimmt und wenn der Abstand kleiner ist als der Radius, liegt die Glyphe unter der Lupe. Es kann auch eine Schnittpunktberechnung zwischen dem Lupenkreis und den Glyphen durchgeführt werden. Dazu müsste eine Funktion für den Lupenkreis aufgestellt werden. Alle Glyphen, die unter der Lupe liegen, werden in eine Liste geschrieben. Eine visuelle Rückmeldung an den Nutzer erfolgt durch das Zeichnen der Glyphen in rot. Alle anderen Glyphen werden weiterhin in dunkelblau gezeichnet.

Mit der Leertaste kann die Lupe an die derzeitigen Koordinaten gebunden werden. Ab diesem Zeitpunkt wird der Mauszeiger unabhängig von der Lupe bewegt. Die Lupe wird erst wieder bei einem erneuten Bedienen der Leertaste an den Mauszeiger gekoppelt. Der `KeyListener` hierfür ist für das Dokumenten-Panel implementiert, doch wird er an dieser Stelle erwähnt, da die Funktionalität sich auf die interaktive Lupe bezieht.

5.4 Stichwort-Panel

Das Stichwort-Panel wird rechts neben dem Rand der Lupe auf das Dokumenten-Panel als `JPanel` gezeichnet. Der Hintergrund ist durchsichtig, dies kann mit `Stichwort-Panel.setOpaque(false);` realisiert werden. Die Terme werden von oben nach unten im Stichwort-Panel angezeigt. Die Terme

sind durch Textlabel implementiert und haben einen weißen Hintergrund. Wenn die Bounding Box des Stichwort-Panels mit dem Rand des Dokumenten-Panles kollidiert, wird das Stichwort-Panel auf die linke Seite der Lupe verschoben.

Um die maximal zehn anzuzeigenden Terme zu bestimmen, muss die Liste mit den unter der Lupe enthaltenen Glyphen abgerufen werden. Von jeder Glyphe in der Liste werden alle Terme des dazugehörigen Dokuments gesammelt. Dafür werden die Terme der Dokumente nach der Reihe in eine gemeinsame Liste eingefügt. Nun hängt es davon ab, ob die Anzeige basierend auf der Termfrequenz oder der Dokumentfrequenz erfolgen soll. Diese Information kann aus dem Auswahl-Panel ausgelesen werden. Für den Fall der Termfrequenz wird beim Einfügen in die Liste überprüft, ob ein Term schon vorhanden ist und trifft dies zu wird die gespeicherte Termfrequenz mit der des einzufügenden Terms addiert. Falls der Term noch nicht in der Liste enthalten ist, wird der Term als neues Element gespeichert und dazu die entsprechende Termfrequenz aus dem Dokument. Bei der Dokumentfrequenz verhält sich das Verfahren ähnlich, außer dass keine Frequenzen addiert werden, sondern die Dokumentfrequenz des Terms in der Liste um eins erhöht wird, falls der Term schon vorhanden ist. Diese Liste wird absteigend nach der entsprechenden Frequenz sortiert. Zur Anzeige werden die zehn obersten Terme aus der sortierten Liste in Textlabel geschrieben und im Stichwort-Panel vertikal übereinander eingefügt.

5.5 Histogramm-Panel

Das Histogramm-Panel wird oben über dem Rand der Lupe auf das Dokumenten-Panel gezeichnet. Der Hintergrund ist wie auch beim Stichwort-Panel durchsichtig. Bei einer Kollision der Bounding Box mit dem Rand des Dokumenten-Panels, wird das Histogramm-Panel auf die Unterseite der Lupe verschoben.

Die Funktionalität des Histogramm-Panels setzt vor der Anzeige der Terme im Stichwort-Panel ein. Die im vorherigen Abschnitt erstellte Liste mit Termen und deren Frequenzen wird in diesem Schritt gefiltert. Dazu wird die Liste von oben nach unten durchlaufen und für jedes Element wird überprüft, ob es im Auswahlbereich des Wörterbuchs liegt. Alle Terme, die nicht im Auswahlbereich stehen werden aus der Liste gelöscht. Diese ausgedünnte Liste wird nun für die Anzeige im Stichwort-Panel verwendet.

Das Histogramm wird mit pixelbreiten Linien von der Unterseite des Panels nach oben gezeichnet. Jede Linie steht für die Frequenz eines Terms. Um die Zipf Verteilung anzugleichen und alle Linien in das Panel einzupassen, wird der Logarithmus zur Basis 10 auf alle Frequenzen angewandt. Die Höhe des Panels, abzüglich des Randes in dem die Werte geschrieben werden, wird durch das höchste Ergebnis geteilt und darauf alle Werte mit diesem Faktor multipliziert.

Wenn der Auswahlbereich am oberen Ende des Wörterbuchs liegt, können und werden keine Terme mit höheren Frequenzen angezeigt. Falls der Bereich zwischen 1 und 50 Terme vom oberen Ende entfernt ist, wird für alle Terme eine Linie auf die gleiche Art und Weise gezeichnet, nur in schwarz um sie vom Auswahlbereich abzuheben. Für den Fall, dass es mehr als 50 Terme sind, werden die 50 Terme direkt über dem Auswahlbereich gezeichnet, der Rest wird verworfen. Das Gleiche gilt für das untere Ende des Wörterbuchs.

Der Auswahlbereich kann mit der „+“-Taste nach oben, also in Richtung der höheren Frequenzen verschoben werden und mit der „-“-Taste in die andere Richtung. Falls der Auswahlbereich am oberen beziehungsweise unterem Ende ist, wird die Funktion der entsprechenden Taste abgefangen. Der `KeyListener` hierfür ist für das Dokumenten-Panel implementiert, doch wird er an dieser Stelle erwähnt, da die Funktion sich nur auf das Histogramm-Panel bezieht.

5.6 Auswahl-Panel

Das Auswahl-Panel befindet sich rechts neben dem Dokumenten-Panel. Auf diesem Panel ist ein `JSlider` implementiert, von dem der aktuelle Wert abgefragt werden kann, um die Größe des Auswahlbereiches herauszufinden. Zusätzlich befinden sich zwei beschriftete `JRadioButtons` auf dem Panel, mit denen der Nutzer festlegen kann, ob die Termfrequenz oder die Dokumentfrequenz genutzt werden soll. Die Buttons sind in einer gemeinsamen `ButtonGroup`, damit immer nur eine Frequenz ausgewählt werden kann. Eine `JCheckBox` ermöglicht die Anzeige des Histogramms in Farben, die für Nutzern mit Rot-Grün-Schwäche unterscheidbar sind.

6 Evaluation

In diesem Kapitel wird die Evaluation des in dieser Arbeit entwickelten Programms anhand eines Expertenfeedbacks beschrieben. Die Durchführung wurde mit vier Experten aus dem Bereich der Visualisierung umgesetzt.

6.1 Aufgabenstellung und Durchführung

Der Fragebogen, der den Experten vorgesetzt wurde, ist im Anhang zu finden.

Die Durchführung der Evaluation erfolgte nach dem Prinzip eines „Thinking Aloud Tests“, das heißt die Experten wurden dazu aufgefordert während der Durchführung laut zu denken. Die Aussagen der Experten wurden mitgeschrieben, um später analysiert werden zu können.

Zuerst wurden die Experten gebeten ihre Vorkenntnisse im Bezug auf folgenden Themen einzuschätzen: Informationsvisualisierung, Dokumentenrepräsentation und Darstellung hochdimensionaler Daten. Dadurch konnte verifiziert werden, dass die Experten sich auch in den speziellen Themengebieten, auf denen diese Arbeit beruht, gut auskennen. Daraufhin wurden die Experten anhand der ersten Aufgabe an die Funktionalitäten des Programms herangeführt. Dafür wurden die Interaktionsmöglichkeiten des Programms erklärt und das allgemeine Prinzip erörtert. Für die Aufgabe wurde die Lupe über einem kleinen Teilbereich der Visualisierung auf dem Dokumenten-Panel festgesetzt und die Experten sollten den Inhalt der dahinter liegenden Dokumente herausfinden.

In der zweiten Aufgabe sollten die Experten nun selbständig einen Bereich mit Dokumenten wählen, deren Inhalte für sie interessant sind. Dieser Bereich sollte dann in der dritten Aufgabe mit einem weiteren, wiederum eigenständig gewählten, Bereich auf inhaltliche Unterschiede und Ähnlichkeiten analysiert werden.

Daraufhin sollten einzelne Teilfunktionen des Programms bewertet werden. Anschließend wurden Fragen an die Experten gestellt, wie zum Beispiel die Frage, welche Interkationsmöglichkeiten das Programm verbessern könnten, falls diese Informationen nicht schon während der Durchführung genannt wurden.

6.2 Ergebnisse

Alle vier Experten schätzten ihre Vorkenntnisse in allen drei Themenbereichen als gut bis sehr gut ein.

Nützlichkeit folgender Aspekte	Experte 1	Experte 2	Experte 3	Experte 4	∅
Aufgabenlösung	5	3	3	4	3,75
Lupe: Größe	4	5	4	5	4,5
Lupe: Fixierung	5	5	2	5	4,25
Histogramm: Anzeige	5	5	5	4	4,75
Histogramm: Verschiebbung	5	5	5	4	4,75
Histogramm: Größe	3	2	2	3	2,5
Frequenzwechsel	4	4	2	4	3,5

Tabelle 6.1: Tabelle mit den Ergebnissen der Bewertung (Die Nützlichkeit wurde auf einer Skala zwischen 1 und 5 bewertet).

Für die erste Aufgabe wurden die Originaltexte der Dokumente analysiert, um sie mit den Ergebnissen der Experten vergleichen zu können. Den Experten war es möglich die grundsätzliche thematische Ausrichtung der Dokumente und auch manche Details zu bestimmten. Es handelte sich um Finanzberichte aus dem Asiatischen Raum. Spezieller ist in diesen Dokumenten häufig von Verlusten bestimmter Aktien die Rede. Die Experten fanden Terme wie zum Beispiel: „Yen“, „Million“, „Loss“, „Investment“oder auch „Hong“ und „Kong “. Doch ist aufgefallen, dass viele Details auf der Strecke bleiben und der Zusammenhang der gefunden Terme unklar ist. Es war beispielsweise nicht klar, wer oder was Verluste machte.

Die Bewertung der Teilfunktionen und der allgemeinen Funktionalität erfolgte anhand einer Skala der Nützlichkeit, die auf eine Skala von 1-5 reduziert werden kann, bei der 5 Punkte für „höchst nützlich“ stehen und 1 Punkt für „nicht nützlich“. Die detaillierten Ergebnisse sind Tabelle 6.1 zu entnehmen. Die ausformulierte Beschreibung der Tabelle ist im Anhang zu finden.

Die allgemeinen Kommentare zum Programm waren meist positiver Natur. Sehr hilfreich fanden die Experten den verschiebbaren Auswahlbereich und dessen Darstellung als Histogramm. Auch die farbliche Markierung der angezeigten Stichworte im Auswahlbereich des Histogramms war als Anhaltspunkt für die Verschiebung hilfreich. Das Programm wurde als nützlich beschrieben, um einen guten Überblick über die Dokumentensammlung zu erhalten. Für eine detailliertere Exploration wurden von den Experten mögliche Optimierungen und allgemeine Verbesserungsvorschläge gegeben. Im Folgenden werden diese genannt.

Eine konkrete Selektion von einzelnen Dokumenten wurde von mehreren Experten vorgeschlagen. Damit Verbunden wäre eine Anzeige der Originaltexte oder Ähnliches.

Eine zusätzliche Suchfunktion könnte bei einer spezifischeren Exploration nützlich sein. Dabei könnten Nutzer bestimmte Schlüsselwörter in eine Suchleiste eingeben und alle kreisförmigen Glyphen, deren Dokumente die Schlüsselwörter enthalten, würden in einer anderen Farbe dargestellt werden.

Eine bessere Verarbeitung der Terme war auch ein Vorschlag zur Optimierung des Programms. Zusammengehörende Terme könnten als ein Term angezeigt werden. Die Experten stießen zum Beispiel auf die zwei Terme „Hong“ und „Kong“, die zusätzlich auch einen gleichen Frequenzwert besaßen. Hierbei handelte es sich in den Dokumenten offensichtlich um „Hong Kong“. Dies wurde

durch eine Analyse der originalen Dokumente bestätigt. Zusätzlich könnten nicht nur solche Terme als ein gemeinsames Stichwort angezeigt werden, sondern auch Beziehungen zwischen Termen, zum Beispiel von Termen, die häufig in gleichen Sätzen vorkommen, könnten visuell dargestellt werden.

Die Interaktionen mit Maus und Tastatur sind den Experten weder positiv noch negativ aufgefallen. Nichtsdestotrotz versuchten zwei Experten, als die Lupe festgesetzt war, den Auswahlbereich des Histogramms mit der Maus zu verschieben.

Ein Vorschlag war auch weitere Termgewichtungen wie TF-IDF zum Wechsel anzubieten. Dadurch erhalten Nutzer die Möglichkeit zusätzliche Aspekte in die Exploration einfließen zu lassen und die Dokumente noch besser auf den Inhalt zu untersuchen.

Die Experten hatten Schwierigkeiten mit der Größenauswahl des verschiebbaren Auswahlbereiches. Dies kann auch an den Bewertungen der Teilfunktion erkannt werden. Ideen zur Lösung dieses Problems sind dem Ausblick zu entnehmen.

Ein Experte hatte auch den Vorschlag die Terme nicht in einer Liste im Stichwort-Panel anzuzeigen, sondern als Schlagwortwolke. Dadurch könnte die Frequenz und Ähnliches implizit aus der Darstellung entnommen werden. Ein anderer Experte hätte wiederum lieber die konkreten Frequenzwerte neben den Termen zur Verfügung.

Da sich bei der Platzierung der Kreise auf dem Dokumenten-Panel mit t-SNE und einer linearen Skalierung tendenziell eine Anhäufung von Punkten in der Mitte der Visualisierung ergibt, wurde eine andere Art der Skalierung vorgeschlagen, zum Beispiel eine logarithmische.

7 Zusammenfassung und Ausblick

Explorative Ansätze eignen sich für die Untersuchung von Dokumentensammlungen auf thematische Ausrichtung der Dokumente. Dabei werden die Dokumente durch Symbole auf einer gemeinsamen Ebene nach ihrer Ähnlichkeit zueinander platziert. Nutzer werden bei dieser Aufgabe durch Interaktionsmöglichkeiten unterstützt. In dieser Arbeit wird eine interaktive Lupe verwendet, um Teilbereiche der Visualisierung zu analysieren. Um die Konzepte für diese Lupe und weitere Ansätze zu erstellen, wurden bestehende Techniken und Erkenntnisse aus drei Kategorien analysiert: Textzusammenfassung, Informationsvisualisierung und Interaktive Lupen.

Als Ergebnis wurde eine Lupe entwickelt, die auf dem sogenannten Dokumenten-Panel bewegt werden kann. Auf diesem Panel werden die Dokumente als kreisförmige Glyphen dargestellt. Mit der Lupe können von Nutzern Teilmengen an Dokumenten markiert werden, die in lokalem Zusammenhang stehen. Dabei wird die Term- und Dokumentfrequenz aller Terme bestimmt, deren Dokumente unter der Lupe liegen. Diese Terme werden in einer Liste gespeichert und absteigend sortiert. Aus dieser Liste werden die obersten zehn Terme, also diejenigen mit den höchsten Frequenzen, in einem weiteren Panel, dem sogenannten Stichwort-Panel, direkt neben der Lupe angezeigt.

Bei Recherchen erlangte Erkenntnisse zeigen, dass die stetige Anzeige der zehn Terme mit den höchsten Frequenzen, für eine inhaltliche Analyse von Dokumenten, nicht zwingend zielführend ist. Deshalb wird Nutzern eine Filterfunktion in Form eines Histogramms geboten, mit der die angezeigten Terme beeinflusst werden können. Dazu wird aus dem kompletten Datensatz ein Wörterbuch aller Terme erstellt und die dazugehörigen Frequenzen bestimmt. Aus diesem Wörterbuch können Nutzer einen Bereich bestimmt, aus dem alle Terme, die neben der Lupe angezeigt werden, stammen müssen. Dieser Bereich wird im Histogramm-Panel direkt an der Lupe dargestellt. Mit Hilfe dieser Filterfunktion werden zum Beispiel nur Terme mit mittleren Frequenzen im Stichwort-Panel angezeigt. Den Nutzern stehen verschiedene Interaktionen mit dem Histogramm zur Verfügung, wie das Verschieben oder Vergrößern des Auswahlbereiches.

Das entwickelte Programm wurde anhand eines Expertenfeedbacks evaluiert. Die Experten bewerteten das Programm durchschnittlich als gut und nützlich. Verbesserungsvorschläge betrafen vor allem eine mögliche optionale Anzeige der originalen Dokumente als Volltext und eine weiterführende Verarbeitung der Terme.

Ausblick

Der Ausblick ist durch die Ergebnisse des Expertenfeedbacks beeinflusst und befasst sich mit den Optimierungsmöglichkeiten des in dieser Arbeit entwickelten Programms.

Für zukünftige Arbeiten sollte hierbei in Betracht gezogen werden, dass für eine detaillierte Exploration ein Zugriff auf die ursprünglichen Dokumente ermöglicht werden sollte. Dafür würde es sich eignen die Überschriften der Dokumente, die unter der Lupe liegen, neben dem Dokumenten-Panel darzustellen. Bei Interesse können Nutzer zum Beispiel auf eine dieser Überschriften klicken und den ganzen Text angezeigt bekommen.

Um die detaillierte Exploration weiter zu unterstützen, ist die Implementierung einer Art Suchfunktion eine gute Idee. Wie schon bei den Ergebnissen des Expertenfeedbacks erwähnt, können dabei Schlüsselwörter in eine Suchleiste eingegeben werden. Alle kreisförmigen Glyphen, deren Dokumente eines dieser Schlüsselwörter enthalten, werden in einer anderen Farbe dargestellt. Nutzer mit Interesse in eine bestimmte Richtung könnten von dieser Funktionalität profitieren, aber auch Nutzer, die während dem Gebrauch des Programms auf interessante Schlüsselwörter stoßen, könnten diese Suchfunktion als nützlich erachten.

Eine Optimierung für die Verarbeitung der Terme wäre zum Beispiel, dass zusammengehörende Terme, möglicherweise wie in [CMH12], als ein Term angezeigt werden. Dadurch würden auch Namen als ein gemeinsamer Term dargestellt werden.

Zur einfacheren Interaktion könnten mehr Funktionen an die Maus gebunden werden. Zusätzlich könnten die Tastenkombinationen beibehalten werden, um erfahrenen Nutzern eine schnelle Bedienung zu ermöglichen. Zum Beispiel könnte für den Fall, dass die Lupe an einer Stelle fixiert ist, die Größe des Auswahlbereiches mit dem Mausrad verstellt werden und der Auswahlbereich mit dem Mauszeiger per „Drag and Drop“ verschoben werden. Im Normalfall würde das Mausrad weiterhin den Radius der Lupe verändern und Mausbewegungen die Lupe über die Visualisierung der kreisförmigen Glyphen bewegen.

Um die Größenauswahl des Auswahlbereichs zu optimieren, muss das Problem der Experten adressiert werden. Unter anderem ist das Problem darauf zurückzuführen, dass die Bestimmung der Größe über die Anzahl der Terme nicht intuitiv ist. Besser wäre es den Bereich zwischen zwei bestimmten Frequenzwerten wählen zu können und nicht zwischen Rängen von Termen. Damit könnten auch alle Terme mit niedrigen Frequenzen auf einen Blick betrachtet werden, da es tendenziell uninteressant ist sich einzelne Bereiche aus dieser großen Anzahl an niedrigfrequenten Termen anzuschauen.

Eine mögliche Erweiterung für das Programm wäre eine Veränderung am Histogramm-Panel. Interessant wäre es zu prüfen ob es Nutzern helfen würde nicht nur den lokalen Kontext angezeigt zu bekommen, sondern den globalen. Bisher werden nur maximal die fünfzig Terme mit einer höheren und niedrigeren Frequenz neben dem gewählten Auswahlbereich angezeigt. Es könnten aber auch alle Terme angezeigt werden, um den Nutzern noch mehr Informationen zu bieten. Dazu wäre es möglich Terme in einem gewissen Abstand zu wählen und zur Darstellung zwischen den Werten der Terme zu interpolieren. Jedoch gehen dabei die konkreten Werte verloren. Um dieses Problem zu umgehen könnte, wie bisher, ein Ausschnitt des Histogramms an der Lupe angezeigt werden und zusätzlich das komplette Histogramm an einer anderen Stelle platziert werden. Zum Beispiel könnte das komplette Histogramm am unteren Ende des Dokumenten-Panels über die ganze Breite dargestellt werden. Dass dabei keine zu große Überdeckung mit den kreisförmigen Glyphen entsteht, könnte dieses Histogramm transparent gezeichnet werden, bis die Lupe festgesetzt wird. Ab diesem Zeitpunkt könnte das Histogramm vollständig sichtbar dargestellt werden, um Veränderungen und Interaktionen besser zu sehen. Bei dieser Art der Darstellung müsste beachtet werden, ob die Lupe

sich im unteren Bereich des Dokumenten-Panels befindet. Für diesen Fall könnte das Histogramm zum Beispiel an den oberen Rand verschoben werden.

Die Nützlichkeit dieser möglichen Optimierungen und der Erkenntnisse, die aus dem Expertenfeedback gewonnen wurden, sollte anhand einer Studie evaluiert werden.

A. Anhang

Evaluation - Entwicklung interaktiver Techniken für die Zusammenfassung visueller Dokumenträume

1. Wie bewerten Sie Ihre Vorkenntnisse im Bezug auf folgende Themen?

	kein Vorwissen				sehr gute Kenntnisse
Informationsvisualisierung	☐	☐	☐	☐	☐
Dokumentenrepräsentation	☐	☐	☐	☐	☐
Darstellung hochdimensionaler Daten	☐	☐	☐	☐	☐

2. Untersuchen Sie den gegebenen Bereich auf dessen Inhalt.

3. Explorieren Sie die Visualisierung und suchen Sie einen Bereich mit Themen, die Sie interessieren.

4. Wählen Sie einen zweiten Bereich und vergleichen Sie diesen mit dem Bereich aus Aufgabe 3 auf inhaltliche Unterschiede und Gemeinsamkeiten.

5. Wie bewerten Sie die Nützlichkeit des Programms unter folgenden Aspekten?

	nicht nützlich	wenig nützlich	nützlich	sehr nützlich	höchst nützlich
Wie nützlich war das Programm bei der Lösung der Aufgaben?	☐	☐	☐	☐	☐
Wie nützlich fanden Sie die Funktionalität zur Größenänderung der Lupe?	☐	☐	☐	☐	☐
Wie nützlich war es die Lupe fixieren zu können?	☐	☐	☐	☐	☐
Wie nützlich war die Anzeige des Histogramms?	☐	☐	☐	☐	☐
Wie nützlich fanden Sie den verschiebbaren Auswahlbereich?	☐	☐	☐	☐	☐
Wie nützlich fanden Sie die Funktionalität zur Größenänderung des Auswahlbereichs?	☐	☐	☐	☐	☐
Wie nützlich war die Auswahl zwischen Termfrequenz und Dokumentfrequenz?	☐	☐	☐	☐	☐

Literaturverzeichnis

[Bax58] P. B. Baxendale. Machine-made index for technical literature: an experiment. *IBM Journal of Research and Development*, 2(4):354–361, 1958. (Zitiert auf Seite 15)

[BGN08] S. Bateman, C. Gutwin, M. Nacenta. Seeing things in the clouds: the effect of visual features on tag cloud selections. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, S. 193–202. ACM, 2008. (Zitiert auf den Seiten 6 und 17)

[BSP+93] E. A. Bier, M. C. Stone, K. Pier, W. Buxton, T. D. DeRose. Toolglass and magic lenses: the see-through interface. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, S. 73–80. ACM, 1993. (Zitiert auf Seite 19)

[CC13] M.-W. Chang, C. Collins. Exploring entities in text with descriptive non-photorealistic rendering. In *Visualization Symposium (PacificVis), 2013 IEEE Pacific*, S. 9–16. IEEE, 2013. (Zitiert auf den Seiten 6, 20 und 21)

[CKB08] A. Cockburn, A. Karlson, B. B. Bederson. A review of overview+ detail, zooming, and focus+ context interfaces. *ACM Computing Surveys (CSUR)*, 41(1):2, 2008. (Zitiert auf Seite 19)

[CMH12] J. Chuang, C. D. Manning, J. Heer. "Without the clutter of unimportant words": Descriptive keyphrases for text visualization. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 19(3):19, 2012. (Zitiert auf den Seiten 16, 22 und 46)

[CMS99] S. K. Card, J. D. Mackinlay, B. Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999. (Zitiert auf Seite 16)

[DM07] D. Das, A. F. Martins. A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*, 4:192–195, 2007. (Zitiert auf Seite 16)

[Edm69] H. P. Edmundson. New methods in automatic extracting. *Journal of the ACM (JACM)*, 16(2):264–285, 1969. (Zitiert auf Seite 15)

[HKBE12] F. Heimerl, S. Koch, H. Bosch, T. Ertl. Visual classifier training for text document retrieval. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2839–2848, 2012. (Zitiert auf den Seiten 6 und 21)

[HR02] G. E. Hinton, S. T. Roweis. Stochastic neighbor embedding. In *Advances in neural information processing systems*, S. 833–840. 2002. (Zitiert auf Seite 14)

[Luh58] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165, 1958. (Zitiert auf den Seiten 15, 16 und 22)

[LYRL04] D. D. Lewis, Y. Yang, T. G. Rose, F. Li. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004. (Zitiert auf Seite 36)

[Maa09] L. Maaten. Learning a parametric embedding by preserving local structure. In *International Conference on Artificial Intelligence and Statistics*, S. 384–391. 2009. (Zitiert auf Seite 14)

[Maa13] L. van der Maaten. Barnes-Hut-SNE. *arXiv preprint arXiv:1301.3342*, 2013. (Zitiert auf Seite 14)

[MH08] L. Van der Maaten, G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008. (Zitiert auf Seite 14)

[MKH+99] K. McKeown, J. Klavans, V. Hatzivassiloglou, R. Barzilay, E. Eskin. Towards multidocument summarization by reformulation: Progress and prospects. In *AAAI/IAAI*, S. 453–460. 1999. (Zitiert auf Seite 16)

[MR95] K. McKeown, D. R. Radev. Generating summaries of multiple news articles. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, S. 74–82. ACM, 1995. (Zitiert auf Seite 16)

[MRS08] C. D. Manning, P. Raghavan, H. Schütze. *Introduction to information retrieval*, Band 1. Cambridge university press Cambridge, 2008. (Zitiert auf den Seiten 11, 12, 13 und 32)

[RHM02] D. R. Radev, E. Hovy, K. McKeown. Introduction to the special issue on summarization. *Computational linguistics*, 28(4):399–408, 2002. (Zitiert auf Seite 15)

[RM93] G. G. Robertson, J. D. Mackinlay. The document lens. In *Proceedings of the 6th annual ACM symposium on User interface software and technology*, S. 101–108. ACM, 1993. (Zitiert auf den Seiten 6, 19 und 20)

[Shn96] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, S. 336–343. IEEE, 1996. (Zitiert auf Seite 17)

[TGK+14] C. Tominski, S. Gladisch, U. Kister, R. Dachselt, H. Schumann. A survey on interactive lenses in visualization. *EuroVis State-of-the-Art Reports*, S. 43–62, 2014. (Zitiert auf Seite 19)

[WTP+95] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, V. Crow. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Information Visualization, 1995. Proceedings.*, S. 51–58. IEEE, 1995. (Zitiert auf Seite 18)

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Ort, Datum, Unterschrift