Institut für Visualisierung und Interaktive Systeme

Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Bachelorarbeit

Filterung bibliographischer Daten mittels eines erweiterten “Parallel Sets”-Ansatzes

Marcel Warbeck

Studiengang:	Softwaretechnik
Prüfer/in:	Prof. Dr. Thomas Ertl
Betreuer/in:	Dipl.-Math. Martin Baumann, M.A. Moritz Knabben, M.Sc.
Beginn am:	20. August 2018
Beendet am:	20. Februar 2019

Kurzfassung

Die Visualisierung historischer Metadaten ist keine triviale Aufgabe. Sollen diese Daten allerdings untersucht werden, wird eine passende Visualisierung mit guten Interaktionsmöglichkeiten benötigt. Im Rahmen dieser Arbeit wird für multidimensionale, multivariate Datensätze basierend auf gesammelten Metadaten historischer Zeitungsartikel des Oceanic Exchanges-Projekts ein Prototyp entwickelt, der Parallel Sets als Grundlage einsetzt, um diese Daten darzustellen. Außerdem wird der Prototyp um die Funktionalität erweitert, diese Daten zur Laufzeit gruppieren zu können, damit große Datenmengen gut analysiert werden können. Eine abschließende Untersuchung des Prototypen mithilfe Expertenbefragungen zeigt, dass der Prototyp bereits für Auszüge solcher Daten geeignet ist.

Abstract

Visualization of historical metadata is no trivial task. In order to analyze such data, a suitable visualization with good interaction methods is required. A prototype for analyzing multidimensional, multivariate data based on metadata collected within the scope of the Oceanic Exchanges project will be developed in this thesis. It uses Parallel Sets as a groundwork and extends these to be able to group data together at runtime using appropriate visualizations to analyze big chunks of data. At the end of the thesis, feedback from experts will be evaluated, showing that the tool is suitable for such datasets.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Listings

1 Einführung

Diese Arbeit beschäftigt sich mit der Visualisierung multidimensionaler und multivariater Datensätze mithilfe erweiterter Parallel Sets.

1.1 Motivation

Es gibt viele verschiedene Ansätze, Daten zu visualisieren. Oft bestehen Datensätze, die in realen Umgebungen erstellt oder gesammelt werden, allerdings aus mehreren verschiedenen Typen und sind daher nicht einheitlich beschreib- oder visualisierbar. Zusätzlich bestehen diese aus mehr Informationen als Testdaten, weshalb sie oft nicht ohne Probleme oder andere manuelle Vorarbeit übersichtlich dargestellt werden können.

Parallel Sets bietet eine Grundlage, um Daten unterschiedlicher Art durch wenig Aufwand verstehen und analysieren zu können. Sollte man allerdings einen Datensatz mit einer großen Anzahl an Werten untersuchen wollen, wird man bei den Parallel Sets an deren Grenzen stoßen, denn die Visualisierung wird für den Nutzer so zu komplex oder verzweigt, als dass man sich noch zurechtfinden kann. Es entsteht Visual Clutter.

1.2 Zielsetzung

Das Ziel der Arbeit ist, den oben erwähnten Parallel Sets-Ansatz so zu erweitern, sodass Partitionierung der Daten innerhalb der Visualisierung möglich wird und es nicht notwendig ist, diese extern zu bearbeiten. Hierbei wird für diese Idee ein Prototyp entwickelt, der Möglichkeiten bietet, die Daten, die durch andere bekannte und im Kontext sinnvolle Visualisierungen je nach Datentyp repräsentiert werden, sowohl manuell, als auch automatisch zu gruppieren. Die Eingabedaten werden dabei in einem für Parallel Sets üblichen Format, erwartet, indem diese bereits in ihre Dimensionen und Kategorien aufgeteilt werden.

Der spezielle Anwendungsfall dieser Arbeit und des Prototypen wird auf Datensätzen, welche aus historischen Zeitungsartikeln extrahiert wurden, basieren, die verschiedene Einträge enthalten, unter anderem Veröffentlichungsdatum und -ort. Für diese speziellen Einträge soll mit dem Prototypen eine benutzerfreundliche Möglichkeit gegeben sein, diese Daten auf eine Ansicht zu reduzieren, mit der der Nutzer verschiedene Aspekte des Datensatzes einfach untersuchen kann. Ihm stehen hierbei Möglichkeiten zur Verfügung, mit den Dimensionen zu interagieren und die Ansicht auf den unterliegenden Datensatz anzupassen.

2 Grundlagen

Um auf die Problemlösung verständlich hinleiten zu können, werden zunächst die Grundlagen, die dafür notwendig sind, erläutert. Hierbei wird auf die Grundlagen der Visualisierung eingegangen und anschließend die Parallel Sets anhand eines Beispiels erklärt.

2.1 Grundlagen der Visualisierung

Die Analyse von Daten jeglicher Art ist aus der heutigen Zeit nicht mehr wegzudenken. Egal, ob es sich dabei um Nutzerstatistiken einer Internetseite, anfallende Daten einer medizinischen Untersuchung oder Ergebnisse einer Analyse mehrerer historischer Texte handelt, fallen meist Datensätze mit mehreren hunderttausenden Einträgen an. Diese Daten lassen sich ohne Hilfsmittel bei dieser Menge nicht mehr von Hand untersuchen. Die Visualisierung kümmert sich um genau dieses Problem.

Um Missverständnisse zu vermeiden, werden die folgenden Begriffe einmal definiert:

- Ein **Wert** bezeichnet eine rohe Information
- Ein **Datensatz** ist eine Sammlung aus Werten, die aus Dimensionen und Kategorien aufgebaut sind; diese lassen sich als Tabelle anordnen
- Eine **Dimension** entspricht den Attributen eines Datensatzes; vergleichbar mit dem Titel einer Spalte einer Tabelle
- Eine **Kategorie** bezeichnet einen Wert innerhalb eines Datensatzes unter einer Dimension; vergleichbar mit einem Eintrag innerhalb einer Tabelle
- Als **Zeile** wird eine Kombination der Werte aller Dimensionen eines Datensatzes bezeichnet

Visualisierung ist im grundlegenden Sinne die graphische Repräsentation von Informationen. Hier handelt es sich im speziellen um Informationsvisualisierung. Diese ist nach Card et. al. definiert als “das Nutzen Computer-unterstützter, interaktiver, visueller Repräsentationen abstrakter Daten, um Erkenntnisse zu vereinfachen” [CMS09]. Daher ist das Ziel, eine anschauliche Möglichkeit zu bieten, mit rohen Informationen umzugehen. Um dieses Ziel zu ermöglichen, kann man sich der Visualisierungspipeline, welche in Abb. 2.1 nach Card et. al. [CMS09] zu sehen ist, bedienen.

Gesammelte Rohdaten sind unübersichtlich. Daher müssen diese Daten zuerst strukturiert werden, damit die Daten auch dem Format entsprechen, mit welchem man arbeiten kann. Dieser Schritt wird generell automatisiert erledigt und hierbei hat der Nutzer nur beschränkt Interaktionsmöglichkeiten. Daraus entsteht dann der vorbereitete Datensatz,

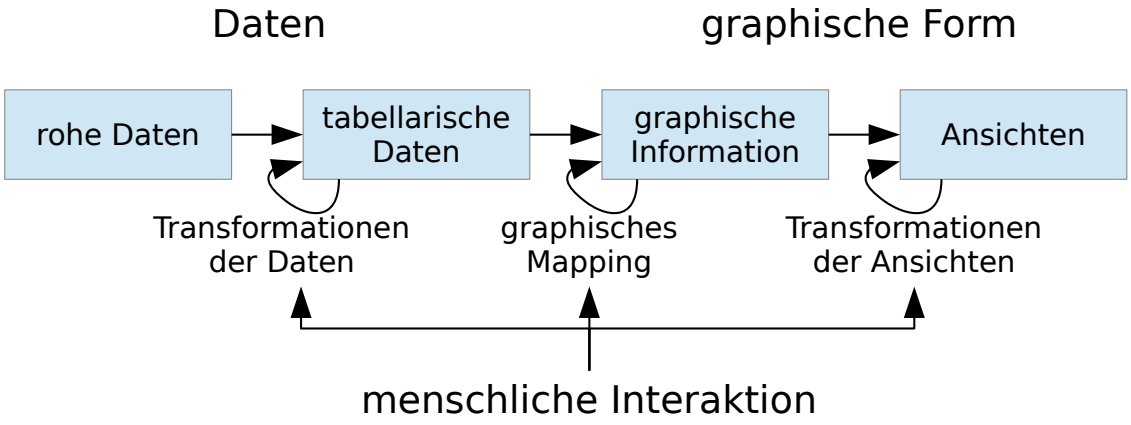

Abbildung 2.1: Abbildung der Visualisierungspipeline, welche den Prozess von rohen Daten zur Visualisierung darstellt, nach Card et. al.[Car99]

meist in tabellarischer Form, auf dem dann weitere Transformationen angesetzt werden können. Hierbei hat der Nutzer vollständige Kontrolle darüber, wie die Daten aufgebaut werden sollen, aus denen die Visualisierung besteht. Nachdem die Daten vom Nutzer transformiert wurden, ist das Mapping der nächste Schritt. Die vorbereiteten Daten werden hier mit den geometrischen Informationen abgebildet. Dementsprechend wird entschieden, welcher Aspekt der Daten die möglichen graphischen Variablen repräsentiert. Hierbei lassen sich für Position, Größe, Farbe, Textur oder Form die Abbildung der Daten bestimmen. Sobald dies abgeschlossen ist, entsteht aus den abgebildeten Informationen eine Ansicht auf die Daten. Diese kann dann vom Nutzer mit Transformationen beeinflusst und angepasst werden, zum Beispiel über Kameratransformationen.

Bei der Abbildung der Daten auf die geometrischen Informationen ist es relevant zu wissen, auf welche Art und Weise diese angeordnet werden können, um diese korrekte Eigenschaften zuweisen zu können. Für dieses Problem lassen sich Skalen verwenden, die eine bestimmte Ordnung für verschiedene Datentypen definieren. Nach Stevens [Ste+46] gibt es vier verschiedene Arten von Skalen, welche in folgender genannter Reihenfolge aufeinander aufbauen.

Die Nominalskala ist eine Skala, die sich auf die Gleichheit zweier Elemente beruft. Hierbei ist keine Beschränkung auf den Typen der Werte gegeben, solange man diese als gleich oder ungleich auswerten kann. Als Beispiel kann hier der Name der Herausgeber eines Zeitungsartikels betrachtet und daraus eine Nominalskala erstellt werden. Haben mehrere Zeitungsartikel denselben Herausgeber, werden diese derselben Gruppe innerhalb der Nominalskala zugewiesen.

Im Vergleich zur Nominalskala berücksichtigt die Ordinalskala eine natürliche Ordnung der Elemente. Es muss also möglich sein, die Elemente mithilfe von Vergleichsoperatoren ordnen zu können. Als Beispiel wäre hier die Skala der Schulnoten zu nennen. Diese besteht aus sechs Werten, die sich durch eine Ordnung beschreiben lassen, zum Beispiel gut $>$ ausreichend. Hier lässt sich auch weiterhin die Gleichheit feststellen (gut $=$ gut).

Intervallskalen werden bei Differenzen oder Intervallen eingesetzt, die sich innerhalb zweier Skalen durch eine lineare Transformation übertragen lassen. Zeitpunkte sind hierbei ein Beispiel für eine Intervallskala, denn es lassen sich die Unterschiede zwischen zwei Zeitpunkten feststellen oder ein Intervall zwischen einem Start- und Endzeitpunkt bilden [Ste+46].

Die Verhältnisskalen lassen sich bei Werten einsetzen, die alle vorherigen Eigenschaften aufweisen und zusätzlich auch bei Verhältnissen gleichgesetzt werden können. Dies ist dann der Fall, wenn man einen Wert einer Skala mit einer Konstante multiplizieren kann, um den entsprechenden Wert auf einer anderen Skala zu erhalten. Außerdem muss ein absoluter Nullpunkt auf dieser Skala definiert sein. Um das Beispiel der Intervallskalen zu ergänzen, kann man hier die Zeitdauer betrachten. Diese lassen sich im Verhältnis zueinander angeben (eine Dauer von zwei Tagen ist die Hälfte der Dauer von vier Tagen) und es ist ein absoluter Nullpunkt vorhanden (Dauer von null Tagen) [Ste+46].

Bei dem für diese Arbeit entwickelten Prototypen werden auch andere Visualisierungen verwendet, die als Unterstützung zum Filtern der Daten dienen. Als Hauptvisualisierung für jegliche Art von Daten wird das Säulendiagramm verwendet. Hierbei wird der Wertebereich eines Säulendiagramm horizontal abgebildet. Jedem dieser Werte wird dann ein Balken zugewiesen, dessen Höhe der Häufigkeit des Werts innerhalb des angezeigten Datensatzes entspricht. Deshalb bietet sich das Säulendiagramm als Standardvisualisierung der unterliegenden Datensätze an, mit der sich dieser Prototyp beschäftigt.

Als spezielle Visualisierung für Orte eignet sich eine Karte. Hierbei werden die Koordinaten, die in Längen- und Breitengrad vorliegen und so die Oberfläche einer dreidimensionalen Kugel beschreiben, mithilfe einer Kartenprojektion auf die Oberfläche einer zweidimensionalen Ebene transformiert. So kann man die abstrakten Koordinaten in einem anschaulichen Umfeld untersuchen. Für Daten dieser Art wird eine Rektangularprojektion verwendet, um die Koordinaten innerhalb eines Rechtecks darzustellen.

2.2 Parallel Sets

Nachdem im vorherigen Abschnitt einige Visualisierungen für Datensätze genannt wurden, wird nun Parallel Sets erklärt. Denn sollte man einen Datensatz analysieren wollen, bei dem man die Zusammenhänge verschiedener Dimensionen untersuchen will, reicht eine einfache Visualisierung nicht mehr aus, um diesen ohne Verlust von Informationen untersuchen zu können. In Tabelle 2.1 ist in Auszug eines Beispieldatensatzes in Form einer Tabelle zu sehen.

Die Dimensionen dieses Datensatzes bestehen aus verschiedenen Typen:

- **Geschlecht** hat zwei mögliche Werte, “männlich” und “weiblich”
- **Studiengang** hat eine Menge an Werten zur Verfügung, zum Beispiel “Informatik”
- **Geburtsdatum** ist eine Datumsangabe im Format `Jahr-Monat-Tag`
- **Geburtsort** besteht aus einer Kombination von Längen- und Breitengrad des Orts, getrennt durch ein Semikolon

Tabelle 2.1: Ein Beispieldatensatz mit verschiedenen Datentypen

Geschlecht	Studiengang	Geburtsdatum	Geburtsort
männlich	Informatik	1993-02-18	48.782;9.177
weiblich	Mathematik	1993-06-02	48.782;9.177
männlich	Physik	1996-03-23	48.741;9.266
weiblich	Informatik	1994-04-04	52.508;13.285
weiblich	Mathematik	1996-03-23	40.731;-73.935

Diese Daten lassen sich unabhängig voneinander oder in Kombination zweier Spalten visualisieren, zum Beispiel, um einen möglichen Zusammenhang zwischen Studiengang und Geburtsdatum zu untersuchen.

Mit Parallel Sets existiert eine Möglichkeit, mehrdimensionale Daten integriert darzustellen. Hierbei werden die Rohdaten pro Spalte betrachtet und danach die Verbindungen der Spalten zueinander hergestellt. Die Parallel Sets-Visualisierung selbst zeigt sowohl Relationen zwischen den Dimensionen eines Datensatzes, als auch Häufigkeitsinformationen an. Sollten Kategorien mehrmals vorkommen, werden diese aggregiert und entsprechend ihrer relativen Häufigkeit visuell skaliert. Den kompletten Datensatz des in Tabelle 2.1 gezeigten Ausschnitts kann man in Abb. 2.2 als Parallel Sets visualisiert sehen.

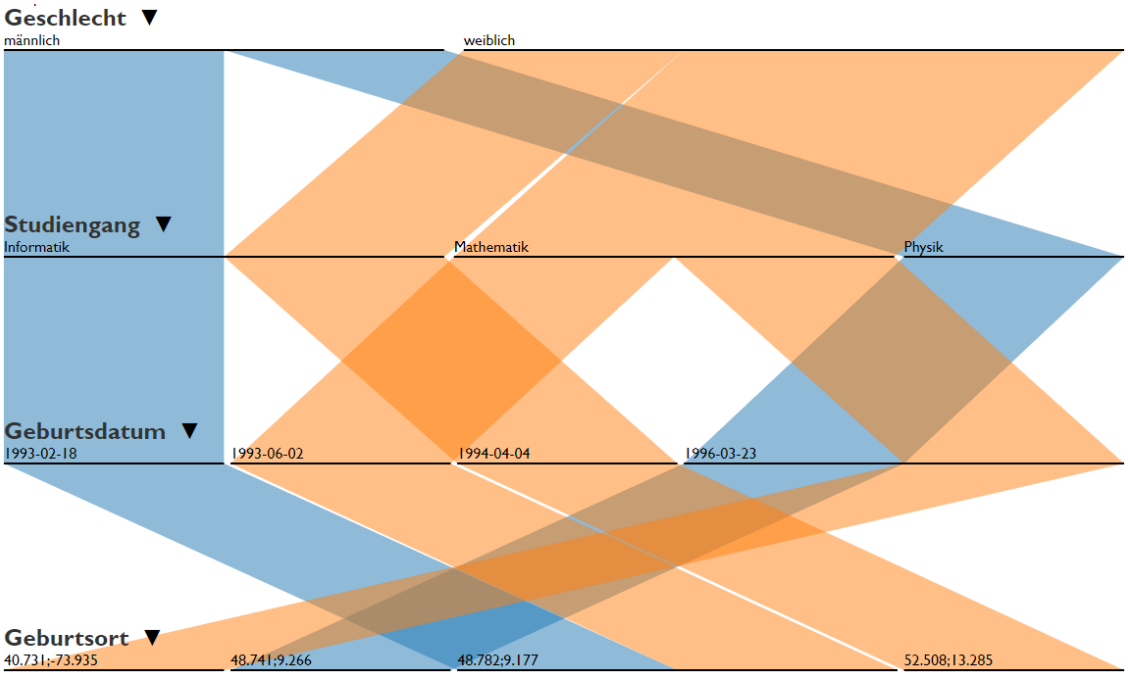

Abbildung 2.2: Parallel Sets-Visualisierung des Beispieldatensatzes aus Tabelle 2.1

Die Dimensionen werden hierbei parallel zueinander und vertikal absteigend entsprechend der Reihenfolge innerhalb der Daten dargestellt, können aber mittels Interaktion umgeordnet werden. Die Kategorien einer Dimension werden horizontal innerhalb des Abschnitts der Dimension angeordnet. Die oberste Dimension weist jeder ihrer Kategorien eine Farbe zu, die sich dann über die Verbindungen zwischen den Dimensionen bis zur untersten

Dimension hin verteilen.

Die Breite einer Kategorie wird durch das Verhältnis der in ihr vorhandenen Kategorien zu der Gesamtanzahl der Werte der Dimension festgelegt. So erkennt man die relative Häufigkeit einer Kategorie durch ihre Breite. Innerhalb der Kategorien selbst wird dann die nächste Trennung vorgenommen. Es werden alle Reihen des Datensatzes untersucht, die aus der aktuellen Kategorie stammen, und überprüft, welche Kategorie der darauffolgenden Dimension in dieser Reihe vorkommt. Zum Beispiel entsteht für die Zeile "männlich → Informatik" des Datensatzes aus Tabelle 2.1 ein Balken, der eine Verbindung der Kategorie "männlich" zu der Kategorie "Informatik" aufbaut. Hierbei gibt die Breite des Balkens, wie schon innerhalb der Dimension, die relative Häufigkeit der Trennung in die darauffolgende Kategorie an. Dies wird pro Dimension für jede Kategorie wiederholt, sodass sich am Ende das in Abb. 2.2 zu sehende Balkenmuster ergibt.

Parallel Sets ist aber nicht die einzige Visualisierungsart für diese Zusammensetzung an multidimensionalen Daten. Die Inspiration für die Parallel Sets, die Parallel Coordinates [ID90], sind ebenfalls eine Möglichkeit, Daten wie solche in Tabelle 2.1 zu visualisieren. Hierbei ähnelt sich die Visualisierungsart. Bei den Parallel Coordinates werden Dimensionen ebenfalls als parallel zueinander liegende Achsen dargestellt. Vergleichbar mit dem Parallel Sets-Ansatz werden Punkte einer Dimension auf der Achse angesetzt und als Verbindungen zwischen den Achsen ausgeführt. Im Vergleich zu den Parallel Sets die Datenpunkte einzeln angezeigt und innerhalb der Dimensionen verbunden anstelle der zusammengefassten Häufigkeitsinformationen. In Abb. 2.3 ist der Beispieldatensatz aus Tabelle 2.1 einmal als Parallel Coordinates dargestellt, um den Vergleich zu der Visualisierung als Parallel Sets in Abb. 2.2 zu sehen.

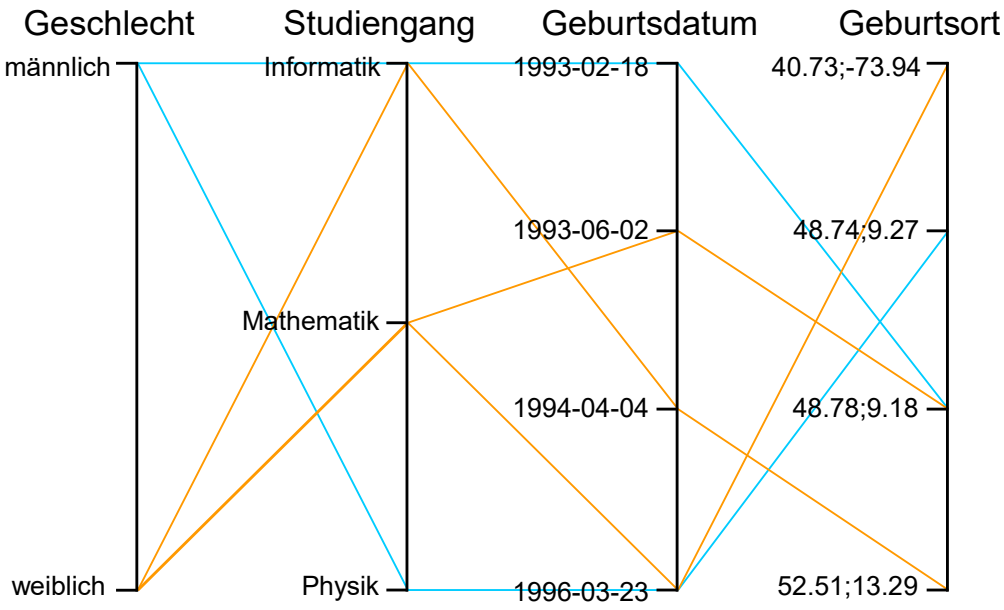

Abbildung 2.3: Parallel Coordinates-Visualisierung des Beispieldatensatzes aus Tabelle 2.1

3 Verwandte Arbeiten

Da die Grundlagen etabliert sind, werden nun einige verwandte Arbeiten genannt, die dem Ansatz oder der Problemlösung dieser Arbeit ähneln, und erklärt, inwiefern diese die Arbeit beeinflusst haben oder in welcher Verbindung sie zu der Arbeit stehen.

3.1 Parallel Sets

Parallel Sets wurden von Kosara et al. [KBH06] vorgeschlagen.

Die Parallel Sets wurden hierbei als Ansatz zur Visualisierung kategorialer Daten entwickelt. Sie basieren auf der Visualisierungstechnik der Venn-Diagramme, da diese bereits Datenhäufigkeiten zum Visualisieren einsetzen. Genau diesem Problem sollen sich die Parallel Sets annehmen. Nicht jede Visualisierungsmethode ist darauf ausgelegt, mit kategorialen Datensätzen zu arbeiten, da diese keine Häufigkeitsinformationen in die Visualisierung integrieren. Daher wurden die Parallel Sets entwickelt. In Abschnitt 2.2 wurden bereits die grundlegenden Eigenschaften, sowie der Aufbau und Hintergrund der Parallel Sets erklärt, daher werden hier noch die wichtigen Punkte für den Prototypen dieser Arbeit aufgegriffen.

Es stehen verschiedene Interaktionsmöglichkeiten mit der von Kosara et. al. vorgeschlagenen Parallel Sets-Visualisierung zur Verfügung. Zum einen die Auswahl und das Hervorheben von Kategorien. Hierbei wird die vom Nutzer gewählte Kategorie hervorgehoben, um sich beim Interagieren mit der Visualisierung auf eine bestimmte Kategorie fokussieren zu können. Außerdem erlaubt der Ansatz interaktives Filtern, indem man Kategorien, die uninteressant für die aktuelle Untersuchung sind, aus dem Bereich der Parallel Sets entfernen kann, um den Platz für die im Fokus stehenden Kategorien freizumachen. Oder man fasst mehrere Kategorien zu einer zusammen, wenn diese Eigenschaften besitzen, die für die Analyse interessant sind. Außerdem kann man sowohl Dimensionen, als auch Kategorien einer Dimension umsortieren. Somit kann man unter anderem Kategorien oder Dimensionen, die Ähnlichkeiten aufweisen, nebeneinander bzw. untereinander anordnen, um einen besseren Überblick zu erhalten. Allerdings lassen sich Kategorien auch automatisch sortieren.

Eine besondere Interaktion ist das Zusammenfassen von Kategorien aus unterschiedlichen Dimensionen, im Artikel “dimension composition” [KBH06] genannt. Hierbei werden alle Kategorien einer Dimension, die in eine Kategorie der darunterliegenden Dimension laufen, als einzelne neue Kategorien angelegt. Die restlichen werden unter einer übrigen Kategorie zusammengefasst. So erhält man kompaktere Informationen über die Kategorien der ersten Dimension, wenn diese die Kategorie der zweiten Dimension ohnehin gemeinsam haben.

Der Prototyp dieser Arbeit wird auf diesem Vorschlag basieren. Die “dimension composition” wird nicht implementiert, da sie für die Analyse der Daten nicht notwendig ist und die Kategorien durch manuelle Auswahl bereits so reduziert werden, dass es keinen Nutzen mehr davon trägt. Das manuelle Umsortieren von Kategorien wird aufgrund spezieller automatischer Sortierungen nach den Interaktionen nicht möglich sein.

3.2 Analyse multidimensionaler Daten

Die Analyse multidimensionalen Daten wurde ebenfalls von Lex et. al. [LSP+10] mithilfe eines eigenen Visualisierungsframeworks untersucht, welches sich Caleydo nennt.

Caleydo ist ein Visualisierungsframework, welches mit einem Fokus auf Untersuchung biomolekularer Daten entwickelt wurde. Diese Daten müssen oft in Gruppen eingeteilt werden, da die Daten zusammenhängen und so eine Analyse Sinn ergibt und übersichtlich durchgeführt werden kann. Allerdings kommt es zu Problemen, wenn Daten, die sich zu Beginn an ähneln, am Ende der Daten divergieren. Diese werden dann in unterschiedliche Gruppen eingeteilt, obwohl gerade der Punkt, an dem diese auseinandergehen, interessant ist. Die von Lex et. al. entwickelte Matchmaker-Visualisierungstechnik soll dieses Problem lösen [LSP+10].

Der Fokus liegt hierbei auf einer “Focus+Context”-Technik [LSP+10], die mehrere, voneinander unabhängig gruppierte Dimensionen nach Ähnlichkeiten aufteilt und diese dann zum Vergleich visualisiert. Hierzu werden die gruppierten Dimensionen wie bei Parallel Coordinates parallel zueinander angeordnet und gleiche Wertpaare zwischen den Gruppen miteinander verbunden. Allerdings werden die parallel angeordneten Gruppen nicht mit einfachen Linien repräsentiert, sondern mit Heatmaps, die durch Farbkodierung bereits selbst Informationen über den Inhalt der Gruppe darstellen. Außerdem werden die Verbindungen zwischen den Gruppen nicht über Geraden wie bei Parallel Coordinates, sondern über speziell geformte Kurven, die Hilfspunkte in der Nähe der Gruppen nutzen, um den Ausgangs- bzw. Eintrittspunkt deutlich zu machen.

Die Matchmaker-Technik von Lex et. al. [LSP+10] hat ähnliche Ansätze zu der Erweiterung der Parallel Sets, die für diesen Prototypen erarbeitet werden. Die Parallel Sets wurden bei der Entwicklung von Matchmaker in Betracht gezogen, da sie dem endgültigen Aufbau sehr ähneln. Da aber die Möglichkeit zur Anzeige einzelner Elemente fehlt, wurde eine eigene Lösung entwickelt. Der Prototyp dieser Arbeit wird die Parallel Sets um eine Übersicht der darunterliegenden Daten erweitern, aber nicht konkret um eine Ansicht einzelner Elemente. Daher ist Matchmaker für die Arbeit nicht in Betracht gezogen worden.

3.3 Interaktive Analyse Mengen-typisierter Daten

Die Visualisierung Mengen-typisierter Daten ist für die Daten, die mit dem Ansatz dieser Arbeit dargestellt werden sollen, nicht irrelevant. Von Freiler et. al. [FMH08] wurde die Analyse Mengen-typisierter Daten bereits untersucht.

Mengen-typisierte Daten können im Vergleich zu nominalen oder kategorialen Daten nicht ohne Umwege visualisiert werden. Daher stellt Freiler et. al. das “set’o’gram” vor. Dieser Visualisierungsansatz bietet eine intuitive Methode, mit Mengen-typisierten Daten umzugehen. Die zugrunde liegende Visualisierung basiert auf einem Säulendiagramm. Jede Säule repräsentiert eine mögliche Kategorie innerhalb der Mengen-typisierten Dimension. Die Höhe jeder Säule entspricht dann der Gesamtanzahl, wie oft dieses Element in allen Mengen vorkommt. Innerhalb der Säulen selbst wird dann eine weitere Einteilung vorgenommen.

Jede Säule enthält Blöcke, die von unten nach oben die Kombination mehrerer Elemente der Menge darstellt und dessen Höhe die Anzahl an den Kombinationen darstellt. Der unterste Block innerhalb einer Säule gibt die Kategorie der Säule an und wie häufig diese allein vorkommt, der Block darüber dann diese Kategorie kombiniert mit einer beliebigen anderen Kategorie, der nächste Block dann die Kombination der Kategorie mit zwei beliebigen anderen. Dies wird entsprechend bis zum höchsten Block wiederholt. Die Breite der Blöcke variiert, damit man die einzelnen Abschnitte einfach auseinander halten kann.

Bei dem Prototypen dieser Arbeit werden ebenfalls Daten visualisiert, die vom Typ Mengen-typisierte Daten sind. Hierbei werden komma-separierte Einträge innerhalb eines Feldes als Menge an Einträgen behandelt. Diese innerhalb des Parallel Sets als einzelne Kategorien zu visualisieren, ist nicht korrekt, da die Parallel Sets sonst Werte, die mehrmals vorkommen, nicht korrekt als Häufigkeiten darstellen. Hierbei hilft das set’o’gram in Verbindung mit dem zu entwickelnden Ansatz des Browsers, beschrieben in Abschnitt 4.1.3, um solche Daten korrekt in ihre Bereiche einteilen zu können. Das set’o’gram wird allerdings nicht im Rahmen dieses Prototypen implementiert, sondern für einen späteren Zeitpunkt angesetzt.

4 Konzept

Mit dem bisherigen Wissen kann der Prototyp konzeptioniert werden. Hierbei werden zuerst die Anforderungen genannt und dann der Entwurf erstellt.

4.1 Anforderungen

Der zu entwickelnde Prototyp muss einigen Anforderungen gerecht werden, um die Problemstellung, welche in Abschnitt 1.2 beschrieben wurde, zu lösen. Diese werden in folgenden Unterpunkten anhand der Grundlagen erklärt und darauf basierend der Entwurf angefertigt.

4.1.1 Basis

Die Basis des Prototypen ist eine Erweiterung der vorgeschlagenen Parallel Sets von Kosara et. al. [KBH06]. Hierbei werden sowohl die Visualisierung der Parallel Sets, als auch die Visualisierungen zum Filtern der Werte innerhalb einer Dimension verwendet.

Wie in Abschnitt 2.2 beschrieben, eignet sich die Parallel Sets-Visualisierung, um multidimensionale, kategoriale Datensätze darzustellen. Dies funktioniert und ist für den Nutzer anschaulich, solange die Datensätze wenige Dimensionen und Kategorien besitzen, da so in der Visualisierung wenige Überschneidungen vorhanden sind und man alles auf einen Blick sieht. Sollte der Datensatz allerdings viele Dimensionen enthalten, kann die Visualisierung, wie in Abb. 4.1 mit einem realen Datensatz zu sehen, durch exponentielles Wachstum der Überschneidungen unübersichtlich werden.

Bei Datensätzen mit vielen Zeilen soll nun der Prototyp eingesetzt werden, um diese Daten zur Laufzeit zu gruppieren und damit auftretenden Visual Clutter zu reduzieren. Dies wird über Binning durch den Nutzer innerhalb der Dimensionen gelöst. Der Aufbau der Visualisierung des Parallel Sets wird wie in Abschnitt 2.2 beschrieben durchgeführt, wie in Abb. 2.2 zu sehen. Als Erweiterung des Aufbaus wird jede Dimension um die Funktion erweitert, den sogenannten Browser zu öffnen, welcher das Binning ermöglicht und ab Abschnitt 4.1.3 näher beschrieben wird.

4.1.2 Binning

Das Binning soll automatisch beim Laden der Datensätze durchgeführt werden und manuell beim Interagieren mit einer Dimension möglich sein. Beim Laden der Daten wird ein automatisches Binning durchgeführt. Hierbei werden ab mehr als acht Kategorien Trennungen

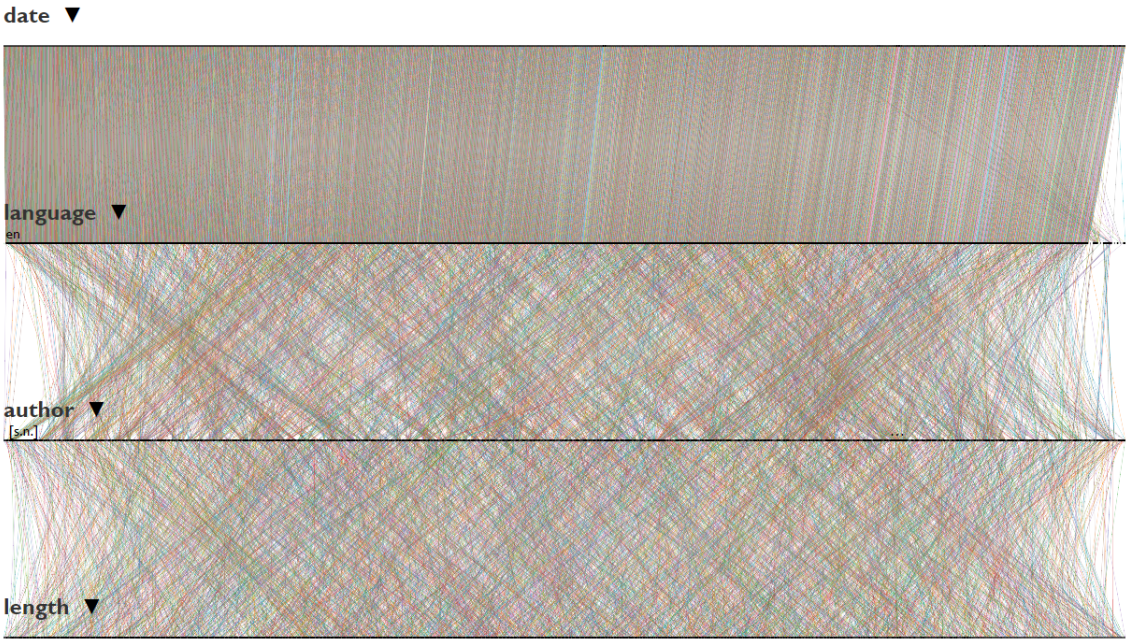

Abbildung 4.1: Beispiel von Parallel Sets für einen unübersichtlichen, realen Datensatz mit 10000 Einträgen, der publizierten Text in deren Veröffentlichungsdatum, Sprache, Herausgeber und Länge gliedert.

eingeführt, welche die Kategorien äquidistant aufteilt, um die Gesamtanzahl an Kategorien der Parallel Sets übersichtlich zu halten. Diese werden nach der Erstellung wie vom Nutzer erstellte Bins behandelt und können daher nachträglich angepasst werden. Für den Nutzer werden hierbei für unterschiedliche Datentypen angepasste Möglichkeiten zur Erstellung der Bins bereitgestellt. Diese Bins sind an eine Dimension gebunden und werden wie eine Kategorie behandelt.

4.1.3 Interaktion

Mit dem Prototyp werden verschiedene Möglichkeiten der Interaktion bereitgestellt. Diese bestehen zum einen aus Interaktionen der Parallel Sets selbst und zum anderen aus der Gruppierung der Daten.

Die angezeigten Dimensionen und deren Reihenfolge können interaktiv verändert werden. Einstellungen für eine Dimension, wie etwa die aktuellen Bins, werden gespeichert, auch wenn die Dimension aktuell nicht sichtbar ist. Standardmäßig werden alle Dimensionen des Datensatzes angezeigt. Allerdings kann mithilfe einfacher Auswahl die aktuelle Menge an angezeigten Dimensionen angepasst werden. Das Verändern der Reihenfolge erfolgt über eine Drag&Drop-Geste. Die Änderung der Dimensionsreihenfolge hat Auswirkungen auf die Verbindungen zwischen den Dimensionen, aber ändert weder die Daten, noch die Filter, die bereits vom Nutzer angewandt wurden.

Die Filterung der Daten erfolgt über den Browser. Hierbei bezieht sich Browser auf den ausklappbaren Teil einer Dimension, der die Interaktion mit dieser über verschiedene

Ansichten und Interaktoren bietet. Jedes in der Arbeit folgende “Browser” bezieht sich hierbei auf den Browser der Visualisierung und nicht auf den Webbrowser. Falls doch, wird dieser explizit erwähnt.

Der Browser selbst dient als Schnittstelle zwischen dem Parallel Set und dem Nutzer, indem der Nutzer Veränderungen an der Ansicht der Daten vornehmen kann. Für die zu bearbeitende Ansicht einer Dimension stellt der Browser eine Anzahl an Ansichten und davon abhängigen Interaktoren zur Verfügung, die speziell zu dem Datentypen der Dimension passen. Zwischen den Ansichten auf die Daten der Dimension lässt sich frei hin- und herschalten, während die Interaktoren je nachdem, ob aktuell ein Binning vorhanden ist, bereits passend ausgewählt werden. Diese Interaktoren stellen die Mittel zur Verfügung, die Daten zum Beispiel via Brushing in mehrere Bins einzuteilen.

4.1.4 Ansichten des Browsers

Es werden mehrere Ansichten für den Browser des Prototyps implementiert, um ein Mittel zur Lösung des Problems zu bieten und dem Nutzer eine Auswahl zu geben, wie die Daten analysiert und gefiltert werden sollen. Die Ansichten, die für den Prototypen gewählt wurden, sind ein Balkendiagramm zur Veranschaulichung der Datenhäufigkeiten unabhängig vom Datentyp, eine Karte zur Darstellung von Koordinaten in Längen- und Breitengrad und einen Kalender für Datumsangaben. Jede dieser Ansichten bietet angepasste Interaktionmöglichkeiten, um dem Nutzer die Manipulation der Bins möglichst zu vereinfachen.

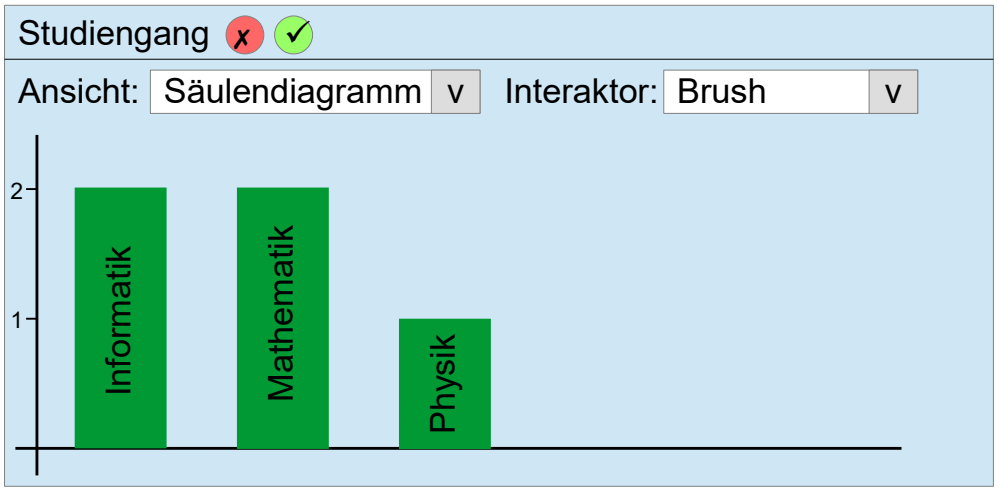

Abbildung 4.2: Browserkonzept des Balkendiagramms anhand der “Studiengang”-Dimension des Beispieldatensatzes aus Tabelle 2.1

Das Balkendiagramm, in Abb. 4.2 anhand der Beispieldaten aus Tabelle 2.1 als Konzept dargestellt, ist für jegliche Art von Daten, die in das Parallel Set geladen werden, geeignet, da es sich hierbei auf die Datenhäufigkeiten einer einzelnen Kategorie bezieht.

Sollte die Dimension aus nicht kategorialen Daten bestehen, werden diese zuerst in künstliche Kategorien umgewandelt und dann dargestellt. Hierbei repräsentiert jeder Balken eine einzelne Kategorie und dessen Höhe die Häufigkeit der Kategorie innerhalb des geladenen Datensatzes. Die Breite des Balkens hat hier keine Bedeutung. Innerhalb des Balkens wird der Wert der Kategorie angezeigt, damit auch nicht-kontinuierliche Daten für den Nutzer gut lesbar angezeigt werden können. Das Balkendiagramm unterstützt zwei Interaktionsmöglichkeiten.

Zum einen gibt es die Möglichkeit, mithilfe eines Auswahlbereichs bestimmte Kategorien auszuwählen, welche dann als neue, eigene Kategorie des Parallel Sets eingeführt werden. Alle Kategorien, die links bzw. rechts des mit dem Brushes ausgewählten Bereichs liegen, werden jeweils in einen eigenen Bin eingeteilt, der diese dann sammelt. Im Parallel Set werden daher die Bins immer in dieser Reihenfolge angezeigt: Daten kleiner als der gewählte Bereich, der gewählte Bereich selbst, Daten größer als der gewählte Bereich. So steht der fokussierte Bereich immer im Zentrum.

Zum anderen wird das Partitionieren der Datensätze auf Grundlage des Balkendiagramms unterstützt. Hierbei werden beim Hinzufügen einer Partition von der ersten bis zur letzten Kategorie all diejenigen bis zum gesetzten Trennpunkt in einen Bin zusammengefasst. Dies wird für jede gesetzte Partition wiederholt, bis man am Ende der Kategorien angekommen ist. Hierbei werden die Partitionen innerhalb des Parallel Sets so angeordnet, wie sie erstellt werden. Diese Trennungen kann der Nutzer an beliebigen Stellen platzieren.

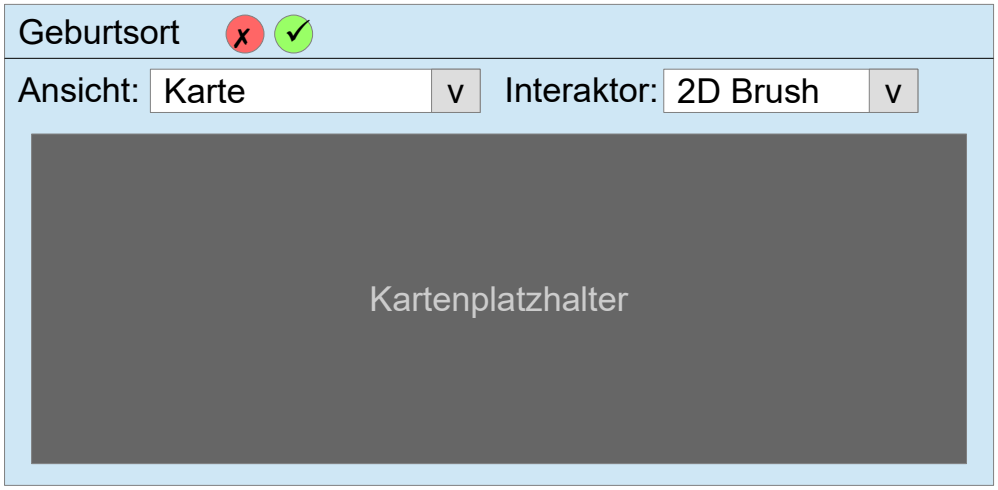

Abbildung 4.3: Browserkonzept der Kartenansicht mit Platzhalter für die später verwendete Kartengrafik

Die Karte wird bei Daten, die einer Ortsangabe entsprechen, ausgewählt und dargestellt. Das Konzept des Browsers dafür wird in Abb. 4.3 gezeigt. Hierbei wird der Umriss der Erde mit einer Rektangularprojektion dargestellt. Die Kategorien des Datensatzes werden dann als Kreise auf dieser angezeigt. Die x- und y-Koordinate entspricht dem projizierten Breiten- und Längengrad der Kategorie, der Radius dann der Häufigkeit. Sollte also derselbe Ort also mehrfach vorkommen, wird dieser angepasst größer auf der Karte dargestellt. Die

Karte bietet als Interaktionsmöglichkeit einen zweidimensionalen Auswahlbereich auf den Koordinaten der Karte an. Dieser lässt den Nutzer einen rechteckigen Bereich an Breiten- und Längengraden auswählen, nach denen der Datensatz gefiltert und zusammengefasst werden soll. Alle Datenpunkte, die sich nicht innerhalb dieses Bereichs befinden, werden in einen zusätzlichen Bin gefiltert, damit Fokussierung auf den gewählten Bereich ermöglicht wird.

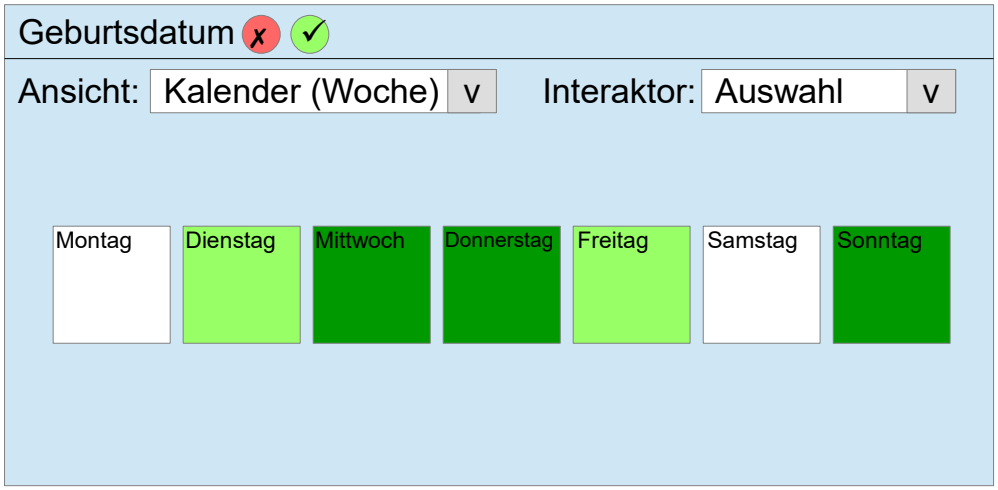

Abbildung 4.4: Browserkonzept des Kalender, als Beispiel an dem Kalender für eine Woche; je intensiver der Grünwert des Tags ist, desto mehr Datensätze beinhalten diesen Tag

Der Kalender wird bei Dimensionen, die Datumsangaben repräsentieren, verwendet. Die Darstellungsmöglichkeiten des Kalenders bestehen aus der Übersicht einer Woche, eines Monats oder eines kompletten Jahres. Das Konzept der Darstellung wird für eine Woche in Abb. 4.4 dargestellt. Hierbei werden die Daten auf die gewünschten Teile des Datums heruntergebrochen. Zum Beispiel werden bei der Ansicht einer Woche alle Tage aufgelistet, aus den unterliegenden Daten die entsprechenden Wochentage ausgelesen und diese dann via Farbkodierung der Häufigkeit dargestellt. Diese Farbkodierung bildet die Häufigkeit von weiß auf grün ab. Die Interaktion mit dieser Art an Visualisierung wird über Auswahl an interessanten Bereichen oder einzelnen Punkten gelöst. Hierbei hat der Nutzer die Möglichkeit, einen speziellen Tag zu fokussieren oder mehrere einzelne Tage in einen einzigen Bin zusammenzufassen.

4.2 Entwurf

Da nun die Anforderungen genannt und ausgeführt wurden, werden Details zum Entwurf erklärt.

4.2.1 Datenformat

Damit die Parallel Sets die gewünschten Daten verarbeiten können, müssen diese lediglich in einer *Comma-Separated Values*-Datei vorliegen. Hierbei gilt die Struktur, die in RFC 4180[1] spezifiziert ist, allerdings sollte die Kopfzeile mit den Namen der einzelnen Spalten immer vorhanden sein. Also gilt für die Daten die Struktur, welche in Listing 4.1 dargestellt ist.

```
Geschlecht,Studiengang,Geburtsdatum,Geburtsort
männlich,Informatik,1993-02-18,48.782;9.177
weiblich,Mathematik,1993-06-02,48.782;9.177
männlich,Physik,1996-03-23,48.741;9.266
weiblich,Informatik,1994-04-04,52.508;13.285
weiblich,Mathematik,1996-03-23,40.731;-73.935
```

Listing 4.1: Erwartete Struktur der Daten anhand des Beispiels in Tabelle 2.1.

Anhand dieses Datensatzes sieht man bereits die Struktur, die beim Laden entsteht. Die Titelzeile der Daten enthält die Dimensionsnamen. Die Werte jeder Dimension bilden die Kategorien. Sollten Werte einer Dimension kontinuierlich sein, werden diese in Kategorien abgebildet. Kategoriewerte können hierbei mehrfach vorkommen, sollten allerdings vom selben Datentypen sein, um optimale Ergebnisse mit dem Prototypen zu erzielen. Zum Beispiel sollte die Dimension "Geburtsdatum" im Beispiel aus Listing 4.1 nur aus Geburtsdaten in dem gegebenen Format bestehen. Sollten in Datensätzen Lücken vorhanden sein, macht dies keinen Unterschied. Der Prototyp markiert fehlende Werte innerhalb einer Dimension mit einer speziellen Konstante.

Der Prototyp unterstützt die Analyse und entsprechende Vorverarbeitung der Daten. Hierbei wird auf die Struktur, die Art der Daten und bestimmte Schlüsselwörter in den Dimensionsnamen geachtet. Beim Laden werden die ersten, nicht leeren Daten einer Dimension auf den Datentypen dieser untersucht. Hierbei wird zwischen Standarddaten (Zeichenketten), Zahlen, Datums- und Zeitangaben und Ortsangaben via Koordinaten unterschieden. Dies ist allerdings leicht erweiterbar. Sollten innerhalb der Daten URLs[2] vorkommen, wird die Zeichenkette des Links auf den Host heruntergekürzt, da der Host meist den Unterschied bei solchen Daten darstellt und der Nutzer damit in der Visualisierung selbst den Überblick behält. Wenn in den Dimensionsnamen Spalten gefunden werden, die sowohl Latitude, als auch Longitude beschreiben, werden diese zusammengefasst und als Ortsangabe behandelt. Als Ortsangaben werden sonst Dimensionen gekennzeichnet, deren Kategorien aus `latitude;longitude` bestehen.

4.2.2 Benutzeroberfläche

Die Nutzeroberfläche ähnelt den Parallel Sets von Kosara et al. [KBH06]. In Abb. 4.5 wird die Übersicht der Oberfläche gezeigt, sobald man einen beliebigen Datensatz geladen hat.

[1] https://tools.ietf.org/html/rfc4180
[2] spezifiziert in RFC 3986, https://tools.ietf.org/html/rfc3986

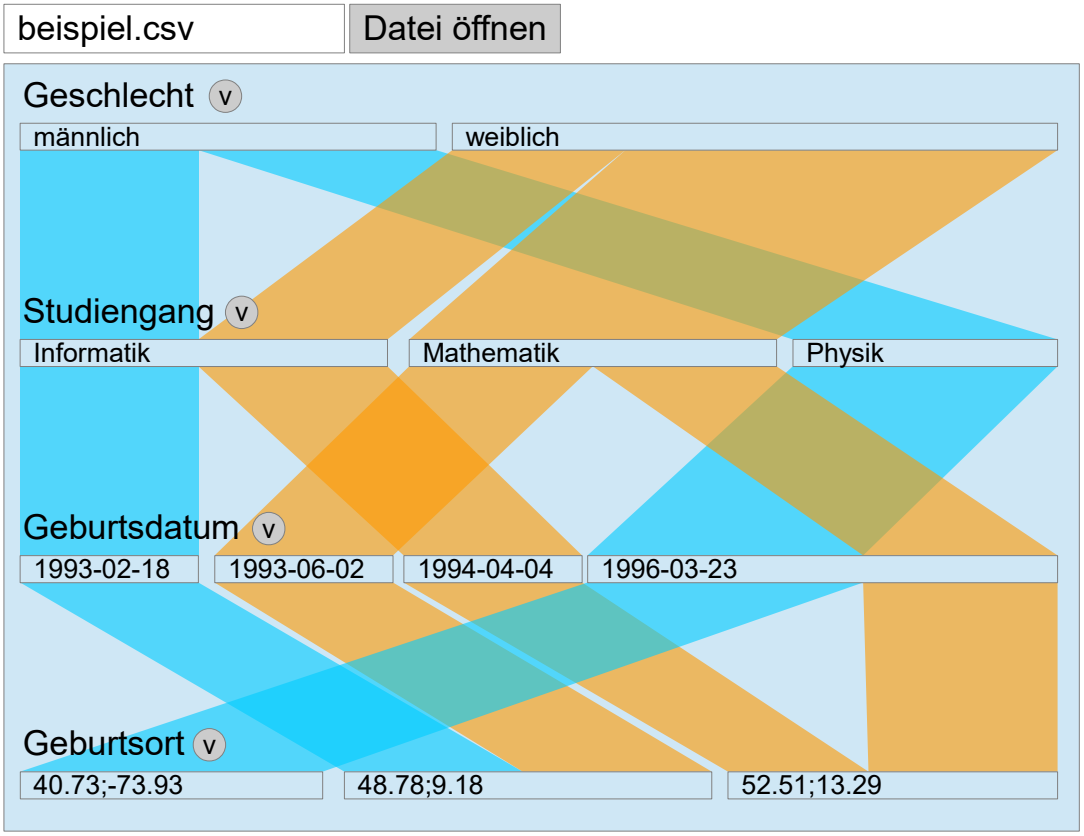

Abbildung 4.5: Konzept der Nutzeroberfläche nach Laden eines Datensatzes anhand des Beispieldatensatzes aus Tabelle 2.1, leicht abgeändert

Über der Visualisierung befinden sich die Kontrollelemente, welche ein- und ausgeklappt werden können, um der Visualisierung den Hauptfokus zu geben. Diese Kontrollelemente bestehen aus der Auswahl der aktuell angezeigten Dimensionen und einem Expertenmodus zum Hinzufügen von Bins.

Darunter befindet sich das Parallel Set. Hierbei sieht man, wie zuvor beschrieben, die Dimensionen als vertikal angeordnete Ebenen, während die Kategorien den horizontalen Raum innerhalb einer Dimension füllen. Unter den Kategorien beginnen die Balken zwischen den Dimensionen, die durch die Häufigkeiten der Werte entstehen. Jede Dimension besitzt ihren eigenen Namen, der direkt durch die Daten gegeben ist. Sollte es sich bei der Kategorie allerdings um einen Bin handeln, wird dies durch die Namen der Kategorien ersichtlich (zum Beispiel: “1993-02-18 - 1994-04-04”).

Neben dem Namen jeder Dimension befindet sich ein Pfeil, der mithilfe eines Klicks den Browser für diese Dimension öffnet. Dieser bietet die Möglichkeit, in der Visualisierung mit den Daten zu interagieren und diese mithilfe von Bins zu aggregieren. Dieser wird in Abb. 4.6 gezeigt. Die verschiedenen Ansichten des Browsers, welche in Abschnitt 4.1.4 aufgezählt und gezeigt wurden, werden dynamisch für die vom Nutzer fokussierte Dimension ausgewählt. Für manche Dimensionen stehen mehrere Ansichten und Interaktionsmöglichkeiten zur

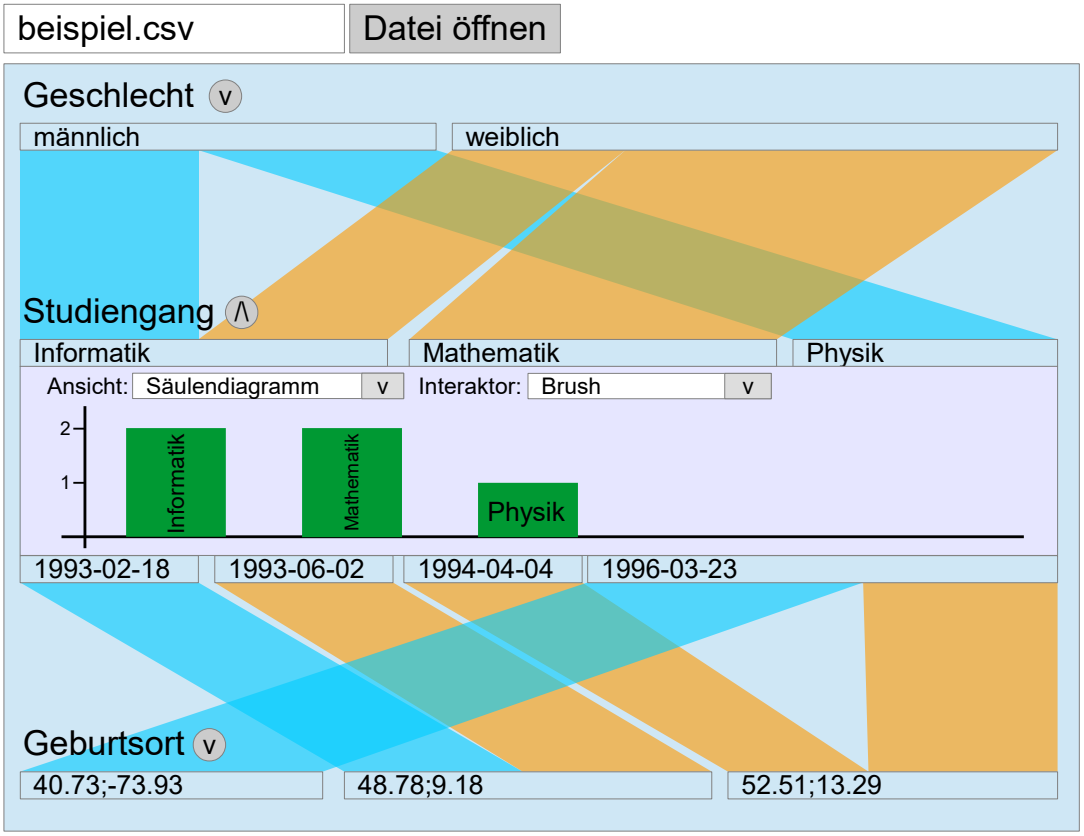

Abbildung 4.6: Konzept des Browsers für eine Dimension, in diesem Fall für die “Studiengang”-Dimension mit der Ansicht des Säulendiagramms des Beispieldatensatzes beschrieben in Tabelle 2.1

Verfügung.

5 Implementierung

Da das beschriebene Konzept die Einzelheiten des Prototyps erklärt und dargelegt hat, wird hier nun die Implementierung vorgestellt.

5.1 Technologie

Der Prototyp dieser Arbeit wird mithilfe von aktuellen Webtechnologien, HTML, CSS und JavaScript, entwickelt. Die einzige Bibliothek, die hierbei zum Einsatz kommt, ist D3.js[1], welche für das Erzeugen der Visualisierung zuständig ist und einige Hilfsmethoden bietet. Der Rest des Prototypen wurde in JavaScript auf Standard des ECMAScript 2018[2] und auf Grundlage der objektorientierten Programmierung, sofern in JavaScript möglich, entwickelt.

5.2 Oberfläche

Die Oberfläche des Prototypen besteht beim initialen Öffnen aus einer Aufforderung, den gewünschten Datensatz zu laden. Nachdem dies vom Nutzer getan wurde, baut sich in der Mitte des Bildschirms die Visualisierung der Parallel Sets auf. In Abb. 5.1 sind die relevanten Abschnitte der Benutzeroberfläche, wie in Abschnitt 4.2.2 gezeigt, aufgebaut. Hierbei dient der obere Abschnitt mit dem Titel “Options” dafür, die aktuell angezeigten Dimensionen auszuwählen und bietet einen Expertenmodus für das manuelle Erstellen der Bins, während sich darunter die Visualisierung der Parallel Sets befindet.

5.3 Browser

Die Implementierung des Browsers erfolgt über eine von den Parallel Sets selbst unabhängige Schnittstelle. Durch einen Druck auf den Pfeil neben dem Namen einer Dimension wird der Browser mit dieser Dimension geöffnet. Dieser wählt dann die passende Ansicht und den passenden Interaktoren für die vom Nutzer gewählte Dimension aus.

[1] Unterstützung für Version 5.6.0, https://github.com/d3/d3/releases/tag/v5.6.0
[2] http://www.ecma-international.org/ecma-262/9.0/index.html

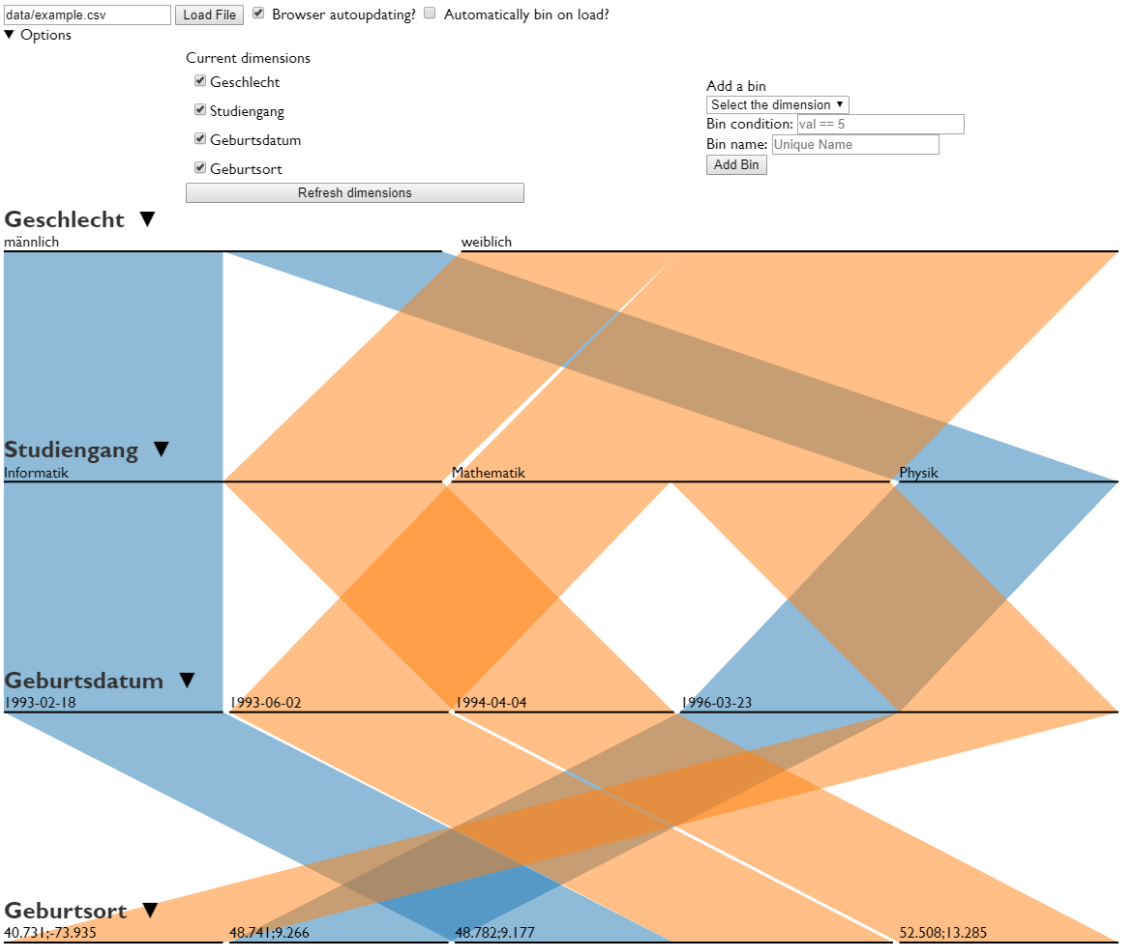

Abbildung 5.1: Übersicht des Prototypen nach Öffnen des Beispieldatensatzes beschrieben in Tabelle 2.1

5.3.1 Ansichten

Wie in Abschnitt 4.1.4 beschrieben, wurden verschiedene Ansichten auf die Daten implementiert, aus welchen, je nach Datentyp der Dimension, die Passenden für die aktuelle Situation ausgewählt und dem Nutzer zur Verfügung gestellt werden.

Alle für den Prototypen implementierten Ansichten sind in Abb. 5.2 zu sehen. Von oben nach unten sind diese das Säulendiagramm, der Kalender für das Aggregieren einer Woche und die Karte. Die Implementierung der Ansichten unterscheidet sich nur in der Anordnung der Elemente von den in Abschnitt 4.1.4 beschriebenen und gezeigten Ansichten. Aus Zeitgründen wurde für den Kalender nur die Ansicht auf eine Woche implementiert. Die Ansichten für einen Monat bzw. ein Jahr sind nicht verfügbar.

5.3.2 Interaktoren

Gebunden an die Ansichten des Browsers sind deren Interaktoren, welche die Möglichkeit bieten, das Binning für die unterliegenden Daten durchzuführen. Für jede Ansicht existiert mindestens ein Interaktor. Dieser bietet die Schnittstelle für den Nutzer, die Gruppierungen in einer für die Ansicht zurechtgeschnittenen Art zu erstellen. Die Interaktoren unterscheiden sich in der Implementierung nicht von den in Abschnitt 4.1.4 beschriebenen Funktionen.

Das Balkendiagramm bietet zwei Möglichkeiten, Gruppierungen zu erstellen. Diese sind in Abb. 5.3 gezeigt. Zum einen kann der Nutzer mithilfe einer Auswahl einen Bereich des Balkendiagramms auswählen, der dann als neue Kategorie in die Parallel Sets eingefügt wird und all die Werte sammelt, die der Bereich erfasst. Dieser Bereich bietet die Möglichkeit, die Randwerte direkt innerhalb der Auswahl zu sehen, während die transparente Fläche das Balkendiagramm darunter sichtbar macht. Zum anderen hat der Nutzer auch die Möglichkeit, mithilfe der Partitionierung eigene Grenzen für Bereiche zu setzen, die dann die Kategorien von dem kleinsten Wert bis zur eingefügten Grenze aggregiert, diese in einen Bereich einfügt und dies wiederholt, bis die größte Kategorie erreicht ist.

Die Karte bietet einen zweidimensionalen, rechteckigen Auswahlbereich als Interaktor an. Dieser sammelt alle Längen- und Breitengrade, die innerhalb des Bereichs liegen, in einer neuen Kategorie der Parallel Sets. Alle Kategorien, die nicht innerhalb des ausgewählten Bereichs liegen, werden in einer “Andere”-Kategorie gesammelt, sodass der Nutzer den Fokus auf dem ausgewählten Bereich hat.

Der Kalender bietet die Möglichkeit, verschiedene Tage auszuwählen, als seinen Interaktor an. Hierbei kann der Nutzer beliebig viele Tage des Kalenders auswählen, welche dann zusammen in eine neue Kategorie der Parallel Sets eingefügt werden. Die ausgewählten Tage sind dann hervorgehoben, sodass der Nutzer erkennt, welche Tage gerade ausgewählt sind. Übrige Kategorien werden in einer “Andere”-Kategorie gesammelt.

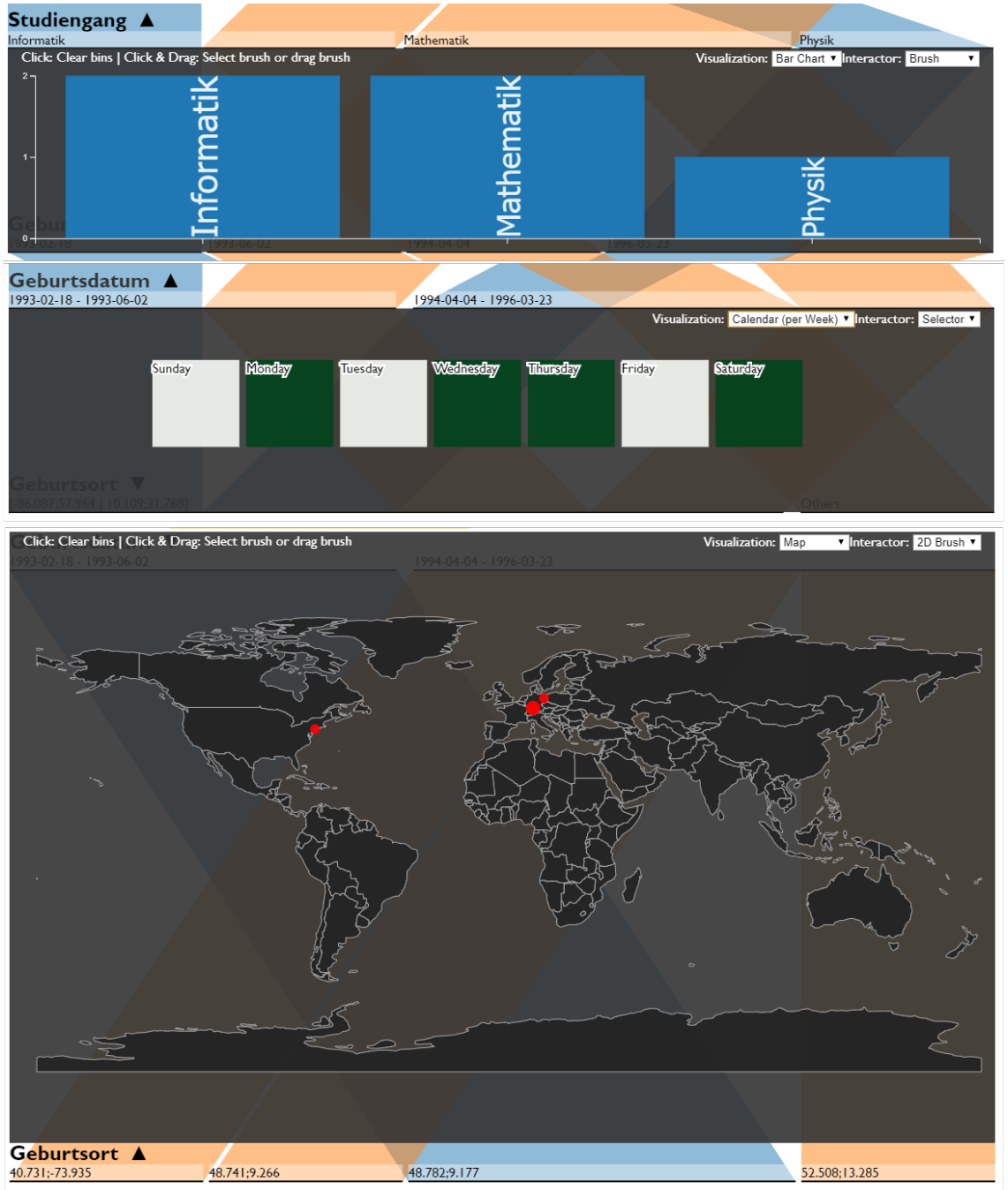

Abbildung 5.2: Alle implementierten Ansichten des Browsers anhand des Beispieldatensatzes beschrieben in Tabelle 2.1. Von oben nach unten: Säulendiagramm, Kalender, Karte

Abbildung 5.3: Die Interaktoren für das Balkendiagramm des Browsers, oben die Gruppierung mithilfe einer Bereichsauswahl, unten das Partitionieren

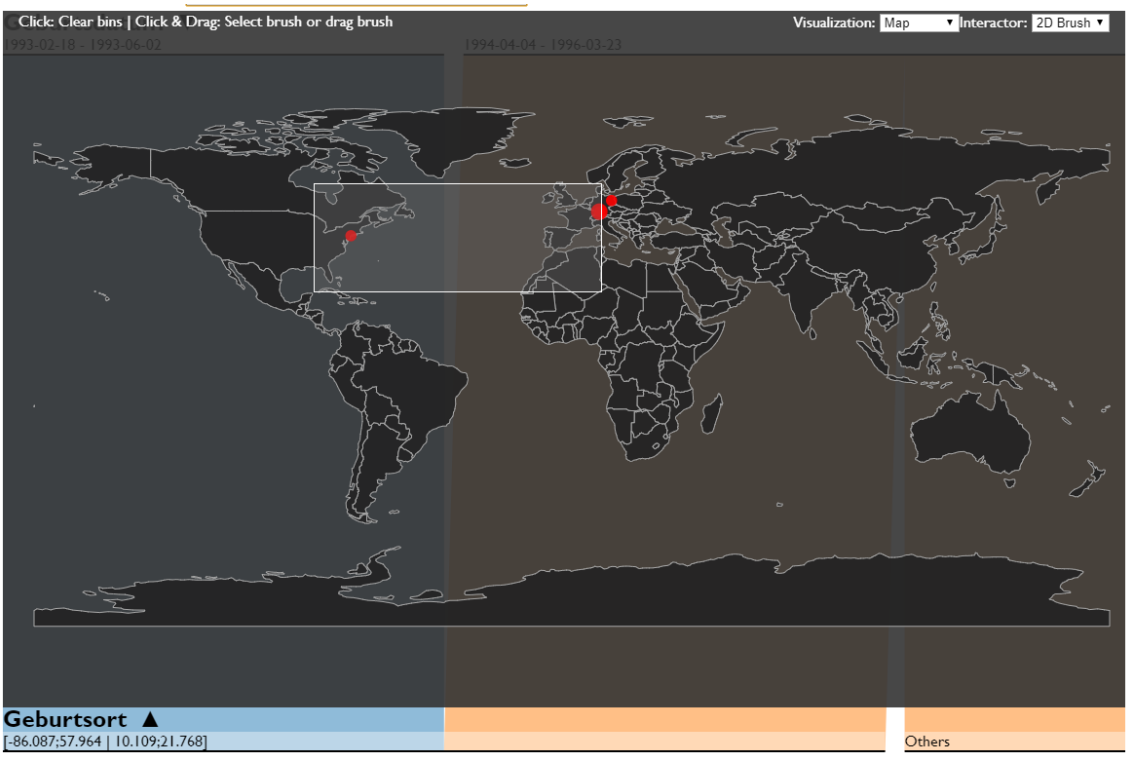

Abbildung 5.4: Der Interaktor für die Karte des Browsers mit einem ausgewählten Bereich

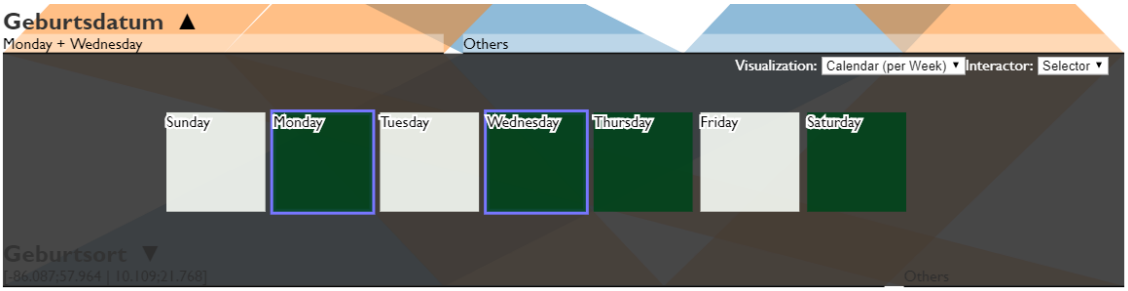

Abbildung 5.5: Der Interaktor für den Kalender des Browsers, hier für die Kalenderansicht einer Woche

6 Ergebnisse

Die Ergebnisse der Arbeit werden in diesem Kapitel durch drei Anwendungsfällen mithilfe eines realen Datensatzes demonstriert. Außerdem wird eine Befragung mehrerer Experten analysiert und zum Abschluss des Kapitels werden interessante Aspekte hervorgehoben und Probleme, die aufgefallen sind, diskutiert.

6.1 Beschreibung des Datensatzes

Der Datensatz, der für die Anwendungsfälle und Expertenbefragungen verwendet wird, entspricht einem Auszug des Datensatzes des Problems. Dieser wird von dem Oceanic Exchanges-Projekt [1] zur Verfügung gestellt und bietet daher die Basis, als Referenz des Lösungsansatzes zu dienen. Er besteht aus Metainformationen über historische Zeitungsartikel.

Im Ausschnitt des Datensatzes, der als Tabelle in Tabelle 6.1 dargestellt ist, befinden sich folgende Spalten:

- **date** beschreibt das Veröffentlichungsdatum des Zeitungsartikels
- **language** enthält eine durch Kommata getrennte Liste an Sprachen, in denen der Zeitungsartikel verfasst wurde
- **author** enthält alle Herausgeber des jeweiligen Artikels
- **length** beschreibt die Länge des Artikels in Wörtern
- **source** gibt einen Link zur Originalquelle des Artikels an, meist in Bildform
- **place** enthält den Namen des Orts der Publikation
- **latitude** und **longitude** fassen den Ort der Publikation in Längen- und Breitengrad zusammen; diese werden automatisch zu einer Dimension verbunden, die sich **lat-long** nennt und beide Informationen durch ein Semikolon trennt

Die Beispiele der Anwendungsfälle werden mit einem Auszug des Datensatzes mit 2500 Einträgen dargestellt. Die Expertenbefragung erfolgt zum Feststellen der Skalierbarkeit mit einem Auszug mit 100 und 2500 Einträgen des Datensatzes.

[1] https://oceanicexchanges.org/

Tabelle 6.1: Ein Ausschnitt des realen Datensatzes des Oceanic Exchanges-Projekts, “source”, “latitude” und “longitude” leicht gekürzt

date	**language**	**author**	**length**	**source**	**place**	**latitude**	**longitude**
1876-05-19	en	Warren & Martin	473	http://chroniclingamerica.loc.gov/...	Hickman County, Kentucky, USA	36.690	-88.959
1882-03-23	en	J.W. Swindells & Co.	393	http://chroniclingamerica.loc.gov/...	Dallas, Dallas County, Texas, USA	32.776	-96.797
1883-06-19	en	[Whitmore Bros.]	1189	http://chroniclingamerica.loc.gov/...	Memphis, Shelby County, Tennessee, USA	35.149	-90.052
1886-02-13	en	Herald Steam Print. House	1517	http://chroniclingamerica.loc.gov/...	LA, Los Angeles County, California, USA	34.054	-118.243
1872-12-03	en	[s.n.]	7253		Milwaukee, Milwaukee County, Wisconsin, USA	43.035	-87.923
...	...	...	...	...	...	...	...

6.2 Anwendungsfälle

Es wurden drei Anwendungsfälle ansteigender Komplexität erstellt, um die Funktionalität des Prototypen zu untersuchen. Im Folgenden sind die Schritte dargestellt, die mit dem Prototypen zum Beantworten der Fragen genutzt werden können.

Die erste Frage beschäftigt sich mit dem Zusammenhang zweier Dimensionen. In diesem Fall wird nach einer Artikellänge innerhalb eines bestimmten Veröffentlichungszeitraums gesucht: “Wie viele Artikel sind zwischen 1840 und 1870 veröffentlicht worden und maximal 400 Wörter lang?”. Um diese Frage beantworten zu können, muss zuerst der Datensatz geladen werden. Die Benutzeroberfläche nach dem Laden ist in Abb. 6.1 zu sehen.

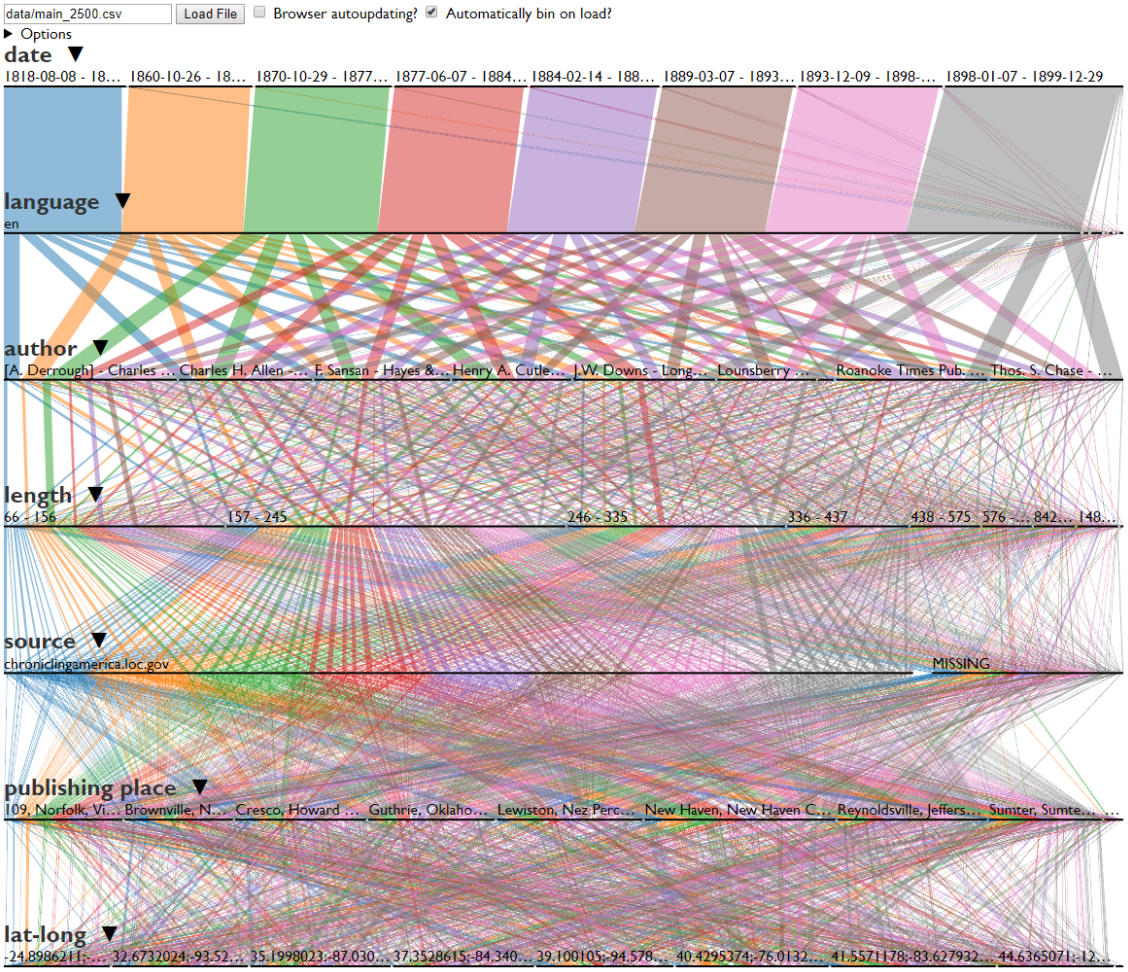

Abbildung 6.1: Die Benutzeroberfläche nach dem Öffnen des Datensatzes zur Vorbereitung der Beantwortung von Frage 1

Danach können die sichtbaren Dimensionen auf die beiden in der Frage relevanten Dimensionen, **date** und **length**, reduziert werden. Dies erfolgt über die Liste der Dimensionen innerhalb der Optionen, zu sehen in Abb. 6.2. Daraufhin kann der Browser der beiden Dimensionen dazu verwendet werden, die Kategorien zu den geforderten Bereichen zu

gruppieren. Die Einstellungen für die Dimensionen sind in Abb. 6.3 zu sehen, links der Browser der **date**-Dimension, rechts der Browser der **length**-Dimension.

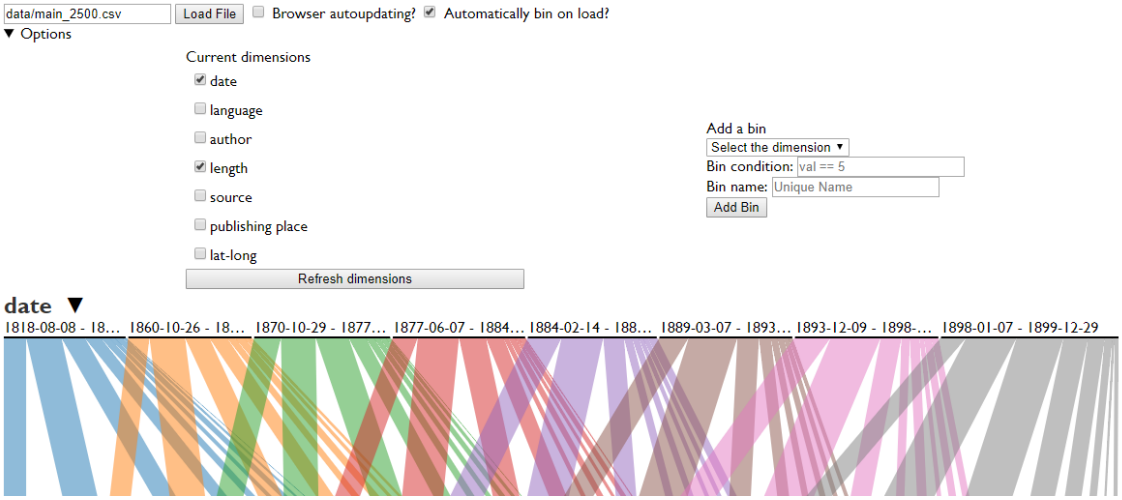

Abbildung 6.2: Die Einschränkung der sichtbaren Dimensionen zur Vorbereitung der Beantwortung von Frage 1, für diesen Fall **date** und **length**

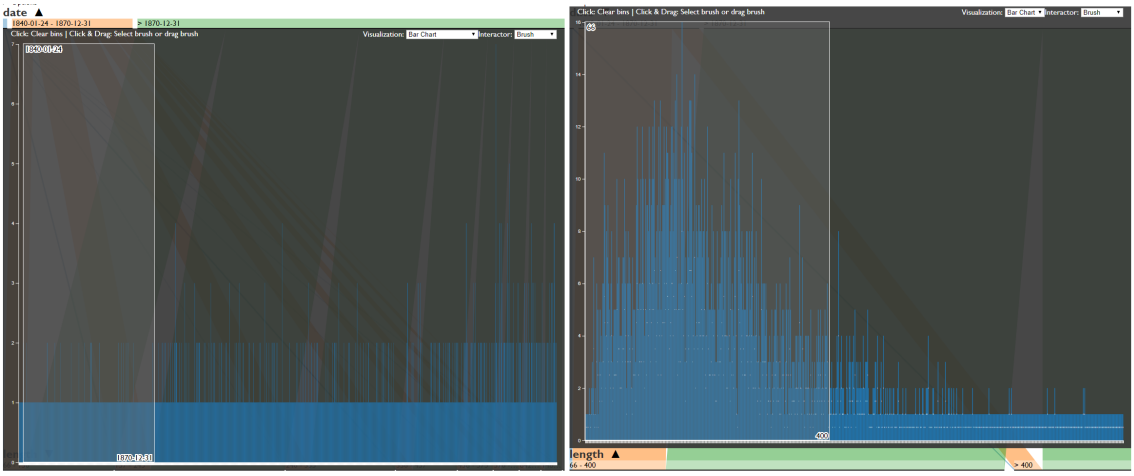

Abbildung 6.3: Die Einstellungen des Browsers für beide Dimensionen zur Beantwortung von Frage 1

Um abschließend die Frage beantworten zu können, verfolgt man die Verbindung zwischen der Kategorie “1840-01-24 - 1870-12-31” der **date**-Dimension zu der Kategorie “66 - 400” der **length**-Dimension und liest das Ergebnis am Hinweis, der beim Highlighten des Balkens auftaucht, in Abb. 6.4 zu sehen, ab.

Die Interaktion ist nicht auf zwei Dimensionen beschränkt. Die nächste Frage richtet Anforderungen an drei verschiedene Dimensionen: “Wie viele Texte wurden ab 1890 mit einer Länge zwischen 200 und 500 Wörtern im östlichen Teil der USA veröffentlicht?”. Ausgehend davon, dass sich der Prototyp im Zustand des Ergebnisses der letzten Frage befindet, können zuerst die sichtbaren Dimensionen angepasst werden, wie in Abb. 6.5 zu

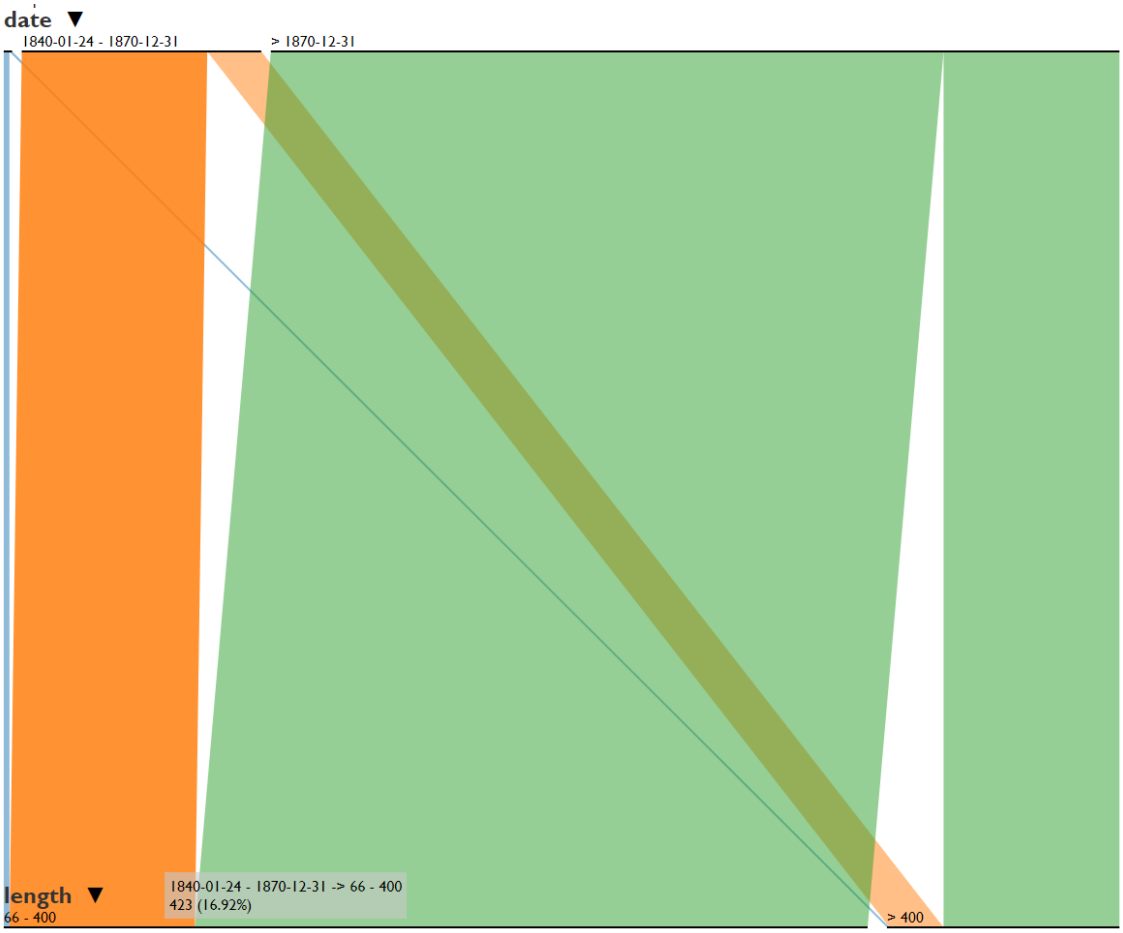

Abbildung 6.4: Das Ergebnis der Gruppierungen beider Dimensionen und das Ergebnis von Frage 1

sehen. Für diese Frage wird die Dimension **lat-long** ebenfalls sichtbar gestellt, damit der Browser für geographische Kategorien genutzt werden kann.

Die Einstellungen für die **date**- und **length**-Dimension müssen angepasst werden, damit sie der Frage entsprechen. Diese neuen Einstellungen sind in Abb. 6.6 zu sehen. Außerdem muss der Bereich für die **lat-long**-Dimension festgelegt werden. Dafür eignet sich die Kartenansicht des Browsers für Längen- und Breitengrade. In Abb. 6.7 sind die Einstellungen für die Karte zu sehen, um dem Bereich der Frage zu entsprechen.

Da nun alle Dimensionen in die gefragten Bereiche eingeteilt wurden, kann durch Verfolgen der entsprechenden Kategorien das Ergebnis abgelesen werden. Hierzu kann man der Verbindung der Kategorie “1890-01-03 - 1899-12-29” der **date**-Dimension zu der Kategorie “200 - 500” der **length**-Dimension und anschließend zur entstandenen Kategorie der **lat-long**-Dimension folgen und das Ergebnis mithilfe des Hinweises ablesen, wie in Abb. 6.8 zu sehen.

Die abschließende Frage stellt ähnliche Anforderungen: “Wie viele Texte wurden am Wochenende (Freitag, Samstag, Sonntag) mit einer Länge von minimal 500 Wörtern ver-

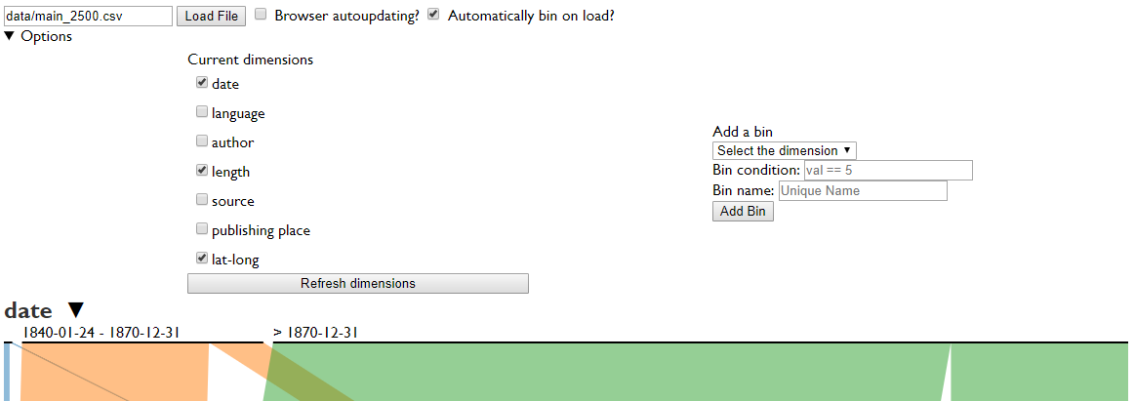

Abbildung 6.5: Die Einschränkung der sichtbaren Dimensionen zur Vorbereitung der Beantwortung von Frage 2

öffentlicht, deren Autorenfeld mit B-J oder N-T beginnt?". Zuerst müssen wieder die Dimensionen angepasst werden. Für diese Frage benötigt man die **date**-, **publisher**- und **length**-Dimension, in Abb. 6.9 dargestellt. Diese kann man außerdem wie in der Fragestellung aufgezählt anordnen, indem man die **author**-Dimension durch eine Drag&Drop-Geste unter die **length**-Dimension schiebt.

Um die Datumsangaben der **date**-Dimension über die Wochentage zu gruppieren, kann die Kalenderansicht des Browsers verwendet werden, wie in Abb. 6.10 zu sehen ist. Die Einstellung für die **length**-Dimension lässt sich wie in den vorherigen Anwendungsfällen mithilfe der Bereichauswahl festlegen. Eine Dimension auf mehrere Bereiche zu gliedern, wird bei dem Browser mithilfe des Partitionierens möglich. Hierbei kann man durch Klicken eine Trennlinie bei B einführen und anschließend bei J einführen, damit die Gruppe an Kategorien für B bis J entsteht. Das Gleiche kann für N bis T wiederholt werden, sodass am Ende eine Ansicht wie in Abb. 6.11 zu sehen ist.

Nachdem dies abgeschlossen wurde, kann man die durch die Gruppierungen entstandenen Kategorien bis nach unten verfolgen. Da in diesem Fall die Ergebnisse zweier Bereiche verlangt ist, muss man hier sowohl die Kategorie der Herausgeber von B bis J, als auch die Kategorie von N bis T beachten. In Abb. 6.12 sind die hervorgehobenen Balken inklusive deren Hinweise zu sehen, die das Ergebnis anzeigen.

6.3 Auswertung der Expertenbefragung

Zur Überprüfung, ob der Prototyp für die Lösung der Problemstellung geeignet ist, wurden zwei Experten gebeten, mit dem Prototypen Anwendungsfälle durchzuführen, die ähnlich zu den in Abschnitt 6.2 beschriebenen Anwendungsfällen aufgebaut waren. Die beiden Experten wurden aus unterschiedlichen Instituten der Universität Stuttgart angefragt - einen Experten des Instituts für Visualisierung und Interaktive Systeme und eine Expertin des Instituts für Literaturwissenschaft. Hierbei dient die Expertin des Instituts für Literaturwissenschaft als Verbindung zum Oceanic Exchanges-Projekt, da diese täglich mit den Daten, für die dieser Prototyp entworfen wurde, arbeitet.

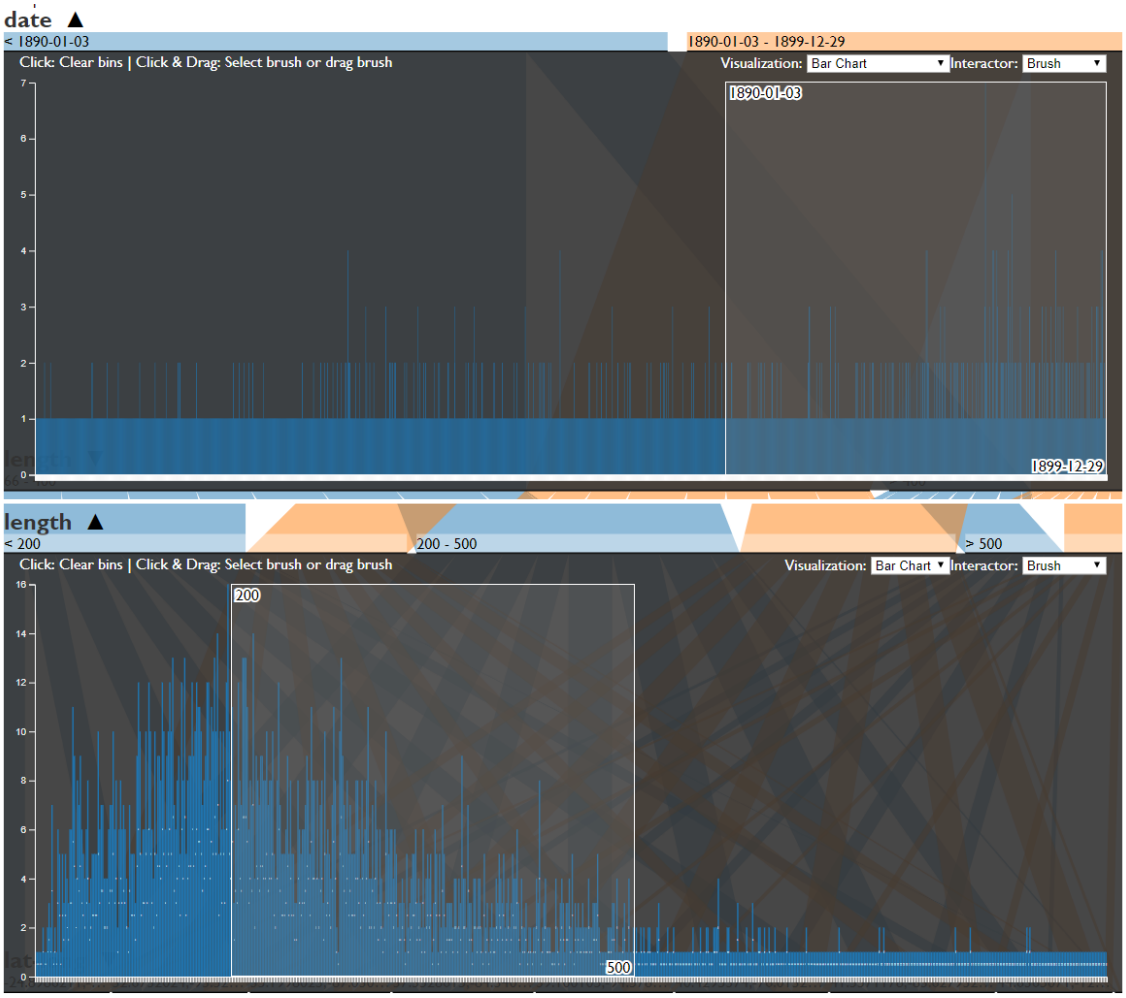

Abbildung 6.6: Die Einstellungen des Browsers für die **date**-Dimension (oben) und **length**-Dimension (unten) zur Vorbereitung der Beantwortung von Frage 2

Das Gespräch begann mit einer Einführung in den Prototypen mithilfe des Beispieldatensatzes aus Tabelle 2.1. Danach wurden verschiedene Fragen gestellt, die den in Abschnitt 6.2 beschriebenen Anwendungsfällen ähneln. Es wurden ähnliche Aufgaben zu den beiden verschieden großen Datensätzen gestellt. Zuerst wurde mit dem Datensatz mit 100 Zeilen, danach mit 2500 Zeilen gearbeitet.

Allgemein wurde der Prototyp von beiden Experten als positiv aufgefasst. Die Arbeit mit dem Prototypen lief für beide Experten nach der Einführung problemlos. Die Expertin des Instituts für Literaturwissenschaft hatte trotz keiner Erfahrung mit Parallel Sets keine Probleme mit dem Verständnis des Prototypen.

Die Arbeit an beiden Datensätzen lief überwiegend flüssig und ohne Probleme. Die Arbeit am kleinen Datensatz war für beide Experten mit dem Browser kein Problem, da die Übersicht über die Daten durch die Anzahl an Einträgen nicht eingeschränkt war. Die Arbeit am großen Datensatz war für den Experten des Visualisierungsinstituts anstrengender und es wurde mehr Zeit pro Aufgabe benötigt. Vor allem, als mehr als zwei Dimensionen durch

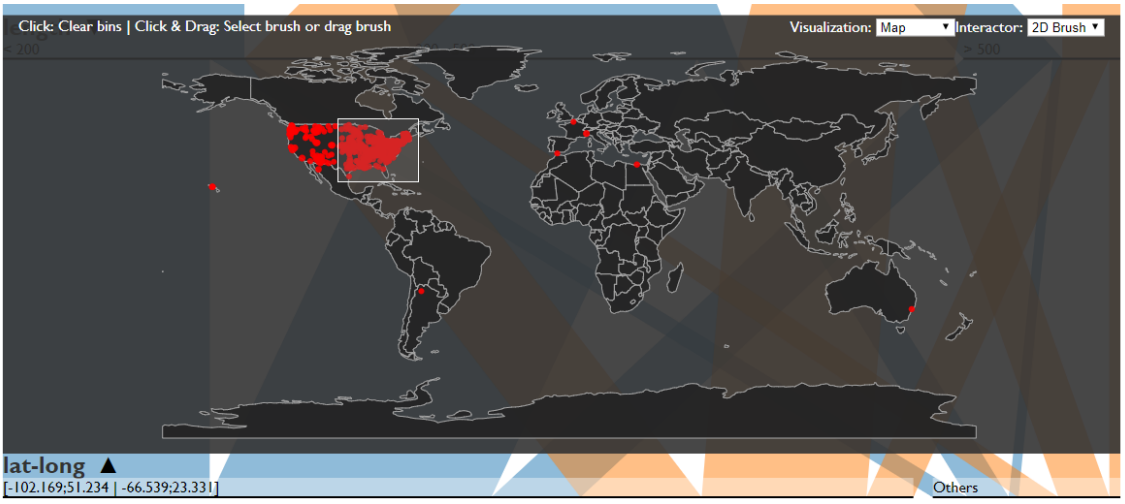

Abbildung 6.7: Die Einstellungen des Browsers für die **lat-long**-Dimension zur Vorbereitung der Beantwortung von Frage 2

die Fragestellung eingeschränkt wurden. Andererseits zögerte die Expertin des Instituts für Literaturwissenschaft beim großen Datensatz kaum, da sie sich mit den Daten bereits auseinandergesetzt hat und der Aufbau bekannt war.

Die Experten mussten sich bei der Arbeit mit dem großen Datensatz deutlich mehr anstrengen, die Werte beim Balkendiagramm des Browsers zu erkennen. Daher wurde die Arbeit bei Fragen des großen Datensatzes bezüglich Dimensionen, die nur die Balkendiagramm-Ansicht unterstützen, deutlich langsamer. Die Interaktion mit dem Balkendiagramm wird daher nur durch die natürliche Ordnung der Daten durchführbar. Bei der Karte und dem Kalender hatten die Experten sowohl bei kleinen, als auch großen Datensätzen keine Probleme.

Alle Funktionen, die für die Fragen benötigt wurden, wurden von beiden Experten unabhängig von der Datensatzgröße schnell gefunden. Auch die Bedienung des Browsers und der Parallel Sets ging für die Experten leicht von der Hand.

6.4 Reflexion

Durch das Aufzeigen der Lösung der Anwendungsfälle und die Expertenbefragung wurde klar, dass der Prototyp in seinem jetzigen Zustand in die passende Richtung entwickelt wurde. Allerdings sind auch Probleme oder Schwierigkeiten aufgefallen, die in diesem Abschnitt diskutiert werden.

Nach dem Durchführen der Anwendungsfälle und der Expertengespräche ist klar, dass sich der Prototyp im Allgemeinen gut für die Arbeit an multidimensionalen, multivariaten Datensätzen eignet. Die Parallel Sets eignen sich als Visualisierungsansatz hierbei sehr gut, um eine Übersicht über die Daten zu gewinnen. Durch die Erweiterung des Ansatzes zum Gruppieren verschiedener Kategorien zeigen sich die Stärken des Ansatzes deutlich mehr.

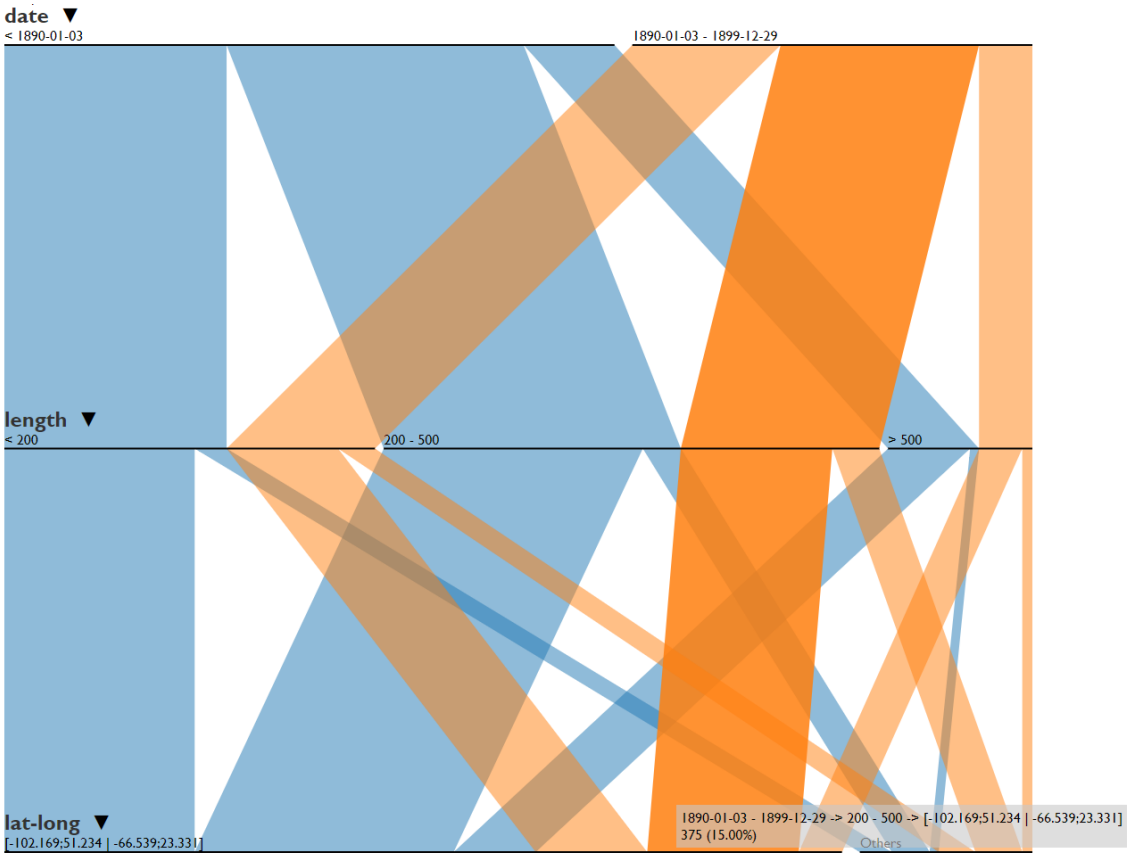

Abbildung 6.8: Das Ergebnis der Gruppierungen aller drei Dimensionen und das Ergebnis von Frage 2

Das initiale Binning hilft, bei großen Datensätzen einen Überblick zu bekommen, wie die Daten aufgebaut sind, allerdings wird es meist bei Beantwortung der Fragestellungen durch vom Nutzer festgelegtes Binning ersetzt. Trotzdem bietet es, selbst wenn es vom Nutzer nicht zum Analysieren der Daten verwendet wird, ein besseres Verständnis der Daten als eine überladene Ansicht mit Tausenden von Kategorien, wie in Abb. 4.1 zu sehen war.

Der Ansatz, das Gruppieren der Kategorien als festen Bestandteil der Visualisierung einzubauen, hat sich als richtige Entscheidung erwiesen und ist auch laut den Experten eine intuitive Art, mit den Daten zu arbeiten und die Gruppen zu erstellen. Die Möglichkeit, mehrere Ansichten für dieselben Daten verwenden zu können und mehrere Möglichkeiten zu haben, mit diesen zu interagieren und verschiedene Gruppierungen vornehmen zu können, bietet Potential, viele Fragen an die Daten beantworten zu können.

Die Schwachstelle des Browsers ist seine Skalierbarkeit bei bestimmten Ansichten. Den Experten nach fiel die Arbeit mit der Ansicht des Säulendiagramms bei vielen Kategorien schwer, da die Balken sehr schmal werden, man die Bezeichnung der Säulen nicht mehr erkennen kann und so abschätzen muss, an welchem Punkt die erwarteten Daten liegen. Die Bereichsauswahl bietet hier eine Hilfestellung mit dem Anzeigen der Randwerte, allerdings ist bei dem Partitionieren keinerlei Information vorhanden. Auch die Karte wird bei vielen angezeigten Dimensionen sehr klein, wodurch die Bereichsauswahl bei präzisen Fragen

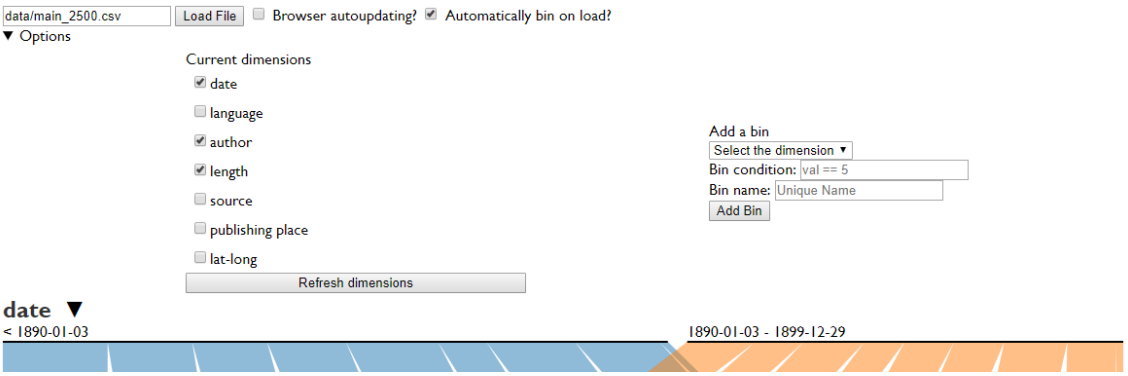

Abbildung 6.9: Die Einschränkung der sichtbaren Dimensionen zur Vorbereitung der Beantwortung von Frage 3

Abbildung 6.10: Die Einstellungen des Browsers für die **date**-Dimension zur Vorbereitung der Beantwortung von Frage 3

schwierig wird.

Die Verbesserungsvorschläge der Experten bieten bereits eine Grundlage für zukünftige Arbeit. Dem Experten des Instituts für Visualisierung ist aufgefallen, dass die Übergänge zwischen den Kategorien bei späteren Dimensionen schwer abzugrenzen waren. Im Vergleich zum Konzept wurden die Balken der Kategorien auf einfache Linien reduziert, was bei vielen Kategorien zu Verwirrung sorgen kann. Die Expertin des Instituts für Literaturwissenschaft schlug außerdem eine genauere Bezeichnung der einzelnen Dimensionen vor. Wenn die Namen der rohen Daten verwendet werden, ist meist nicht klar, um welche Daten es sich genau handelt. Als Beispiel war die Dimension **length** der in Tabelle 6.1 aufgezeigten Daten, welche ohne Kontext für die Länge des Artikels in Zeichen oder in Wörtern stehen kann. Für solche Ungenauigkeiten würde es helfen, einen eigenen Namen für Dimensionen festlegen zu können. Abschließend schlugen beide Experten mehr Ansichten und Interaktoren für den Browser vor, da dieser bisher auf drei Ansichten und drei Interaktoren beschränkt ist.

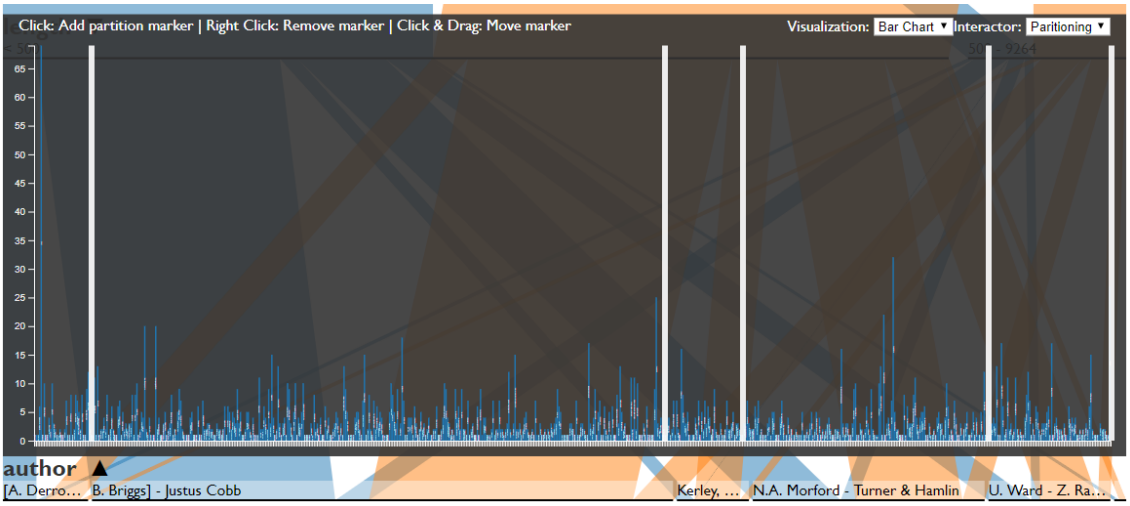

Abbildung 6.11: Die Einstellungen des Browsers für die **author**-Dimension zur Vorbereitung der Beantwortung von Frage 3

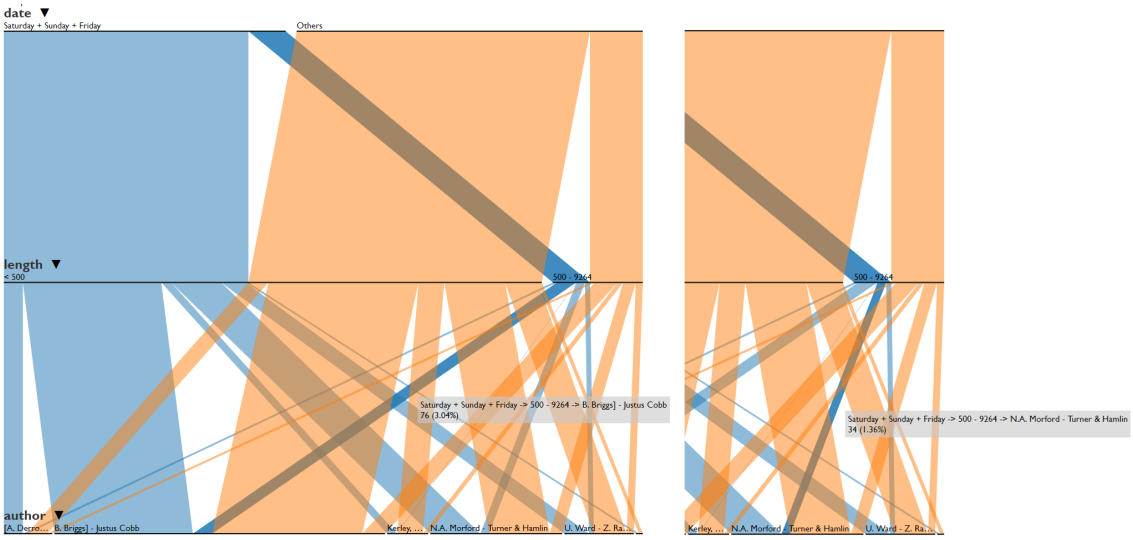

Abbildung 6.12: Das Ergebnis der Gruppierungen der Dimensionen und das Ergebnis von Frage 3

7 Zusammenfassung und Ausblick

7.1 Zusammenfassung

Das Ziel dieser Arbeit war es, einen Prototypen für das Untersuchen historischer Zeitungsartikel im Rahmen des Oceanic Exchanges-Projekts zu erstellen, der sich der Parallel Sets als Grundlage bedient und diese erweitert, um multivariate Datensätze darstellen zu können. Außerdem sollen diese mithilfe passender Visualisierungen in Partitionen unterteilbar sein.

Zunächst wurden die Grundlagen, die für das Erarbeiten des Prototyps nötig sind, zusammengefasst und erklärt. Dann wurden verwandte Arbeiten untersucht, um festzustellen, ob es schon ähnliche Ansätze gibt, die sich dieses Problems annehmen und diese im Vergleich zu den Anforderungen analysiert.

Mit diesen Grundlagen wurde ein Konzept für den Prototypen entwickelt. Dieser basiert auf den erstmalig vorgeschlagenen Parallel Sets und erweitert diese um ein Binning und das Konzept eines Browsers zur Interaktion mit dem Nutzer. Dieser Browser bietet für jede Dimension des Datensatzes angepasste Ansichten, die sich den Datentypen anpassen. Jede dieser Ansichten bietet dann ebenfalls angepasste Interaktoren an, die die Schnittstelle zwischen dem Erstellen der Partitionen und dem Nutzer bieten. Dieses Konzept wurde anhand des Prototypen implementiert und im Vergleich zum Konzept beschrieben.

Anschließend wurden Anwendungsfälle beschrieben und erläutert, die die Möglichkeiten des Prototypen aufzeigen, ähnliche Problemstellungen zu den in der Anforderung gestellten Problemen zu lösen. Die darauffolgenden Expertengespräche, welche mit Auszügen des realen Datensatzes durchgeführt wurden, und deren Diskussion beschrieb dann die Aspekte des Prototypen, die gut gelungen sind, seine Schwächen und mögliche Verbesserungen, die die Experten für brauchbar hielten.

7.2 Ausblick

Mithilfe der Ergebnisse der Anwendungsfälle und Expertengespräche ließen sich Erkenntnisse erarbeiten, die eine Basis für spätere Erweiterungen des Prototypen darstellen.

Das Potential des Prototypen besteht aus der Anzahl der vorhandenen Ansichten. Diese sind im Rahmen der Arbeit auf das Säulendiagramm, die Karte und den Kalender beschränkt. Allerdings lassen sich noch deutlich mehr Ansichten für verschiedene Datentypen entwickeln, beispielsweise eine Tagcloud für Zeichenketten oder ein Zahlenstrahl für Zahlenwerte, statt diese als Säulendiagramm darzustellen.

Neben den Ansichten ist auch die Auswahl an Interaktoren sehr beschränkt. Für die Karte existiert beispielsweise nur die zweidimensionale Auswahl eines rechteckigen Bereichs, welcher die Auswahlsmöglichkeiten bei präziseren Untersuchungen sehr einschränkt. Hierfür ließe sich auch eine konkrete Auswahl mehrerer Länder oder Kontinente realisieren.

Ein Implementierungsdetail ist die native Unterstützung großer Datensätze. Bisher müssen Datensätze, die untersucht werden sollen, auf kleine Ausschnitte reduziert werden, damit sie korrekt verarbeitet werden können. Allerdings sollte dies vom Prototypen selbst erkannt und dem Nutzer mitgeteilt werden können.

Literaturverzeichnis

[Car99] M. Card. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, 1999 (zitiert auf S. 4).

[CMS09] S. Card, J. Mackinlay und B. Shneiderman. "Information Visualization". In: *Human-computer interaction: Design issues, solutions, and applications* 181 (2009) (zitiert auf S. 3).

[FMH08] W. Freiler, K. Matkovic und H. Hauser. "Interactive Visual Analysis of Set-Typed Data". In: *IEEE Transactions on Visualization and Computer Graphics* 14.6 (2008), S. 1340–1347 (zitiert auf S. 10).

[ID90] A. Inselberg und B. Dimsdale. "Parallel Coordinates: A Tool for Visualizing Multi-dimensional Geometry". In: *Proceedings of the 1st Conference on Visualization '90*. VIS '90. San Francisco, California: IEEE Computer Society Press, 1990, S. 361–378 (zitiert auf S. 7).

[KBH06] R. Kosara, F. Bendix und H. Hauser. "Parallel Sets: Interactive Exploration and Visual Analysis of Categorical Data". In: *IEEE Transactions on Visualization and Computer Graphics* 12.4 (Juli 2006), S. 558–568 (zitiert auf S. 9, 13, 18).

[LSP+10] A. Lex, M. Streit, C. Partl, K. Kashofer und D. Schmalstieg. "Comparative Analysis of Multidimensional, Quantitative Data". In: *IEEE Transactions on Visualization and Computer Graphics* 16.6 (2010), S. 1027–1035 (zitiert auf S. 10).

[Ste+46] S. S. Stevens et al. "On the Theory of Scales of Measurement". In: *SCIENCE* 103.2684 (1946) (zitiert auf S. 4, 5).

Alle URLs wurden zuletzt am 19. 02. 2019 geprüft.

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