Visualisierungsinstitut der Universität Stuttgart
Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Fachstudie Nr. 165

Evaluation verfügbarer Visual Analytics Toolkits anhand von Benchmark-Datensätzen

Fabian Merkle
Hanna Schäfer Sebastian Zillessen

Studiengang:	Softwaretechnik
Prüfer:	Prof. Dr. Ertl
Betreuer:	Dipl.-Ling. Florian Heimerl M. Sc. Harald Bosch M. Sc. Robert Krüger
begonnen am:	27. September 2012
beendet am:	27. März 2013
CR-Klassifikation:	A.1 INTRODUCTORY AND SURVEY, H.3.3 Information Search and Retrieval H.3.4 Systems and Software, H.5.2 User Interfaces, K.8.1 Application Packages

Kurzfassung

Diese Arbeit dokumentiert die Analyse und Bewertung der drei Tools Spotfire, Tableau und QlikView. Diese sollten dabei im allgemeinen in Bezug auf ihre Eigenschaften als Datenanalysetools und im speziellen als Werkzeuge zur Analyse der VAST-Challenges 2008 und 2009 bewertet werden. Der Umfang der Datenarten teilt sich dabei in Geo-Daten, Social Networks und Text Analyse auf.

Die Analyse der drei Tools anhand des entwickelten Bewertungskataloges konnte dabei Tableau im Sinne der Analyseparameter als geeignetstes Datenanalysetool bestimmen. Mit nur geringem Abstand in der gewichteten Bewertung, ist Spotfire ein ebenfalls gute geeignetes Tool zur Analyse der gegebenen Datensätze. QlikView schnitt im Gegensatz dazu in allen Kategorien unterdurchschnittlich ab.

In der Benutzerfreundlichkeit liegen Tableau und Spotfire durch Usablity-Features wie Warnmeldungen und Redo/Undo, eine einfache Erlernbarkeit und ein intuitives Bedienkonzept im sehr guten Bereich. QlikView ist besonders als Einsteiger nur mit Hilfe des Tutorials effizient nutzbar. Zudem ist zum Einlesen und analysieren von Daten immer wieder eine Skriptsprache notwendig. Zwar sind allgemeine Usability Features auch hier umgesetzt, doch leider nicht durchgängig in den verschiedenen Anwendungsfällen.

Im Bereich der Kollaborationsmöglichkeiten haben alle Tools nur mittelmäßig abgeschnitten. Rollenverteilungen, Änderungshistorie und das einfügen von Kommentaren waren nur selten, oder über Umwege zugänglich.

Die Laufzeitumgebung der Tools wurde grundsätzlich auf den Desktop beschränkt. QlikView und Tableau verfügen zusätzlich um funktionsfähige Webviewer. Spotfire schnitt aufgrund einer notwendigen Internetverbindung zum Starten der Desktopversion und dem fehlenden Webviewer schlechter ab. Im Bereich der Automatisierung der Importe oder der Analyse über Makros lagen alle drei Tools im schlechten Bereich.

Die Belastbarkeit war in allen Analysen sehr gut. Ladezeiten wurden nicht bemerkt und auch die Analysen waren in Bezug auf Anzahl der Visualisierungen, Platz bedarf und ähnliche Kriterien nicht beschränkt. Größere Datensätze als die zu untersuchenden Challenges wurde nicht getestet.

Im Bereich der Textanalyse haben alle drei Tool keinerlei Features geboten. QlikView konnte über eine Erweiterung Tag-Clouds erstellen, doch abgesehen davon wurden unstrukturierte Datensätze nicht unterstützt. Trotz der vielversprechenden Kooperation von Attivio mit QlikView und Spotfire, war eine Einbindung von Attivio als Erweiterung in die beiden Tools

nicht möglich.

In der Analyse von Geodaten bieten alle drei Tools die Möglichkeit über Scatterplots bekannte Kartendatensätze einzubinden. Allein Tableau verfügt über eine integrierte Kartensammlung, welche die Geoanalyse um einiges vereinfacht. Dadurch wird eine verstärkte Interaktion mit der Visualisierung möglich und Parameter wie die Höhenlage zugänglich. Auf QlikView ist das Kartenmaterial über die Erweiterung CloudMadeMaps integrierbar.

In der Analyse von Netzwerken bietet nur Spotfire eine Möglichkeit zur Visualisierung von Graphen. In QlikView und Tableau können solche Datensätze nur über Tabellen und Filter analysiert werden. Auch die Umsetzung in Spotfire bietet nur eine Eingeschränkte Graph-Visualisierung von Netzwerkdaten. Es besteht nirgends die Möglichkeit zur Erstellung von Treemaps oder anderen parametrisierten Graph-Visualisierungen.

Im Bereich der allgemeinen statistischen Visualisierungen bieten alle Tools die meisten Visualisierungstypen. Im Bereich der Multivariaten Daten sind alle Tools weniger gut ausgestattet, aber auch hier sind Basis-Visualisierungen möglich.

Die Importmöglichkeiten sind bei Tableau durch den möglichen Anschluss an Datenbanksysteme zusätzlich zu den gängigen Datensatzfiles am ausgeprägtesten. In Spotfire werden alle gängigen File-Typen unterstützt, während QlikView auf Excel-Dokumente beschränkt ist.

Die Interaktionsmöglichkeiten innerhalb des Tools sind sowohl bei Spotfire, als auch bei Tableau sehr gut ausgearbeitet. QlikView bietet hingegen nur sehr eingeschränkte Interaktionen und löst die meisten Eingaben über Menüfelder.

Der Export ist in allen Tools gleichermaßen eingeschränkt. Bilder, PDFs und die allgemeine Speicherung des Dokumentes werden von allen unterstützt. Tableau bietet zusätzlich den Export als interaktives HTML-Dokument.

Insgesamt ist keines der Tools zu einer vollständigen Analyse der drei Challenges in der Lage. Oftmals muss auf externe Tools zurückgegriffen werden. Dennoch ist ein erster Überblick über die Datensätze mit den drei Tools schnell erreichbar und auch erste Ergebnisse können so schnell vorweg gegriffen werden.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einführung

Seit 2006 findet alljährlich die sogenannte *VAST* Challenge statt. Bei diesem Wettbewerb geht es in der Regel darum, aus einem großen Datensatz versteckte Informationen zu extrahieren. Damit können Forschungsgruppen ihre Systeme mit den Datensätzen testen und evaluieren. Außerdem entstehen durch die Challenges oft neue interessante Analysemethoden für große Datensätze. Eine unabhängige Jury bewertet am Ende die Resultate und prämiert die Gewinner. Es nehmen viele Universitäten an den Challenges Teil.

Zur Lösung der Aufgaben werden häufig eigene Implementierungen verwendet oder die Daten zunächst durch eigene Skripte aufbereitet um diese dann anschließend zu analysieren.

Ziel der ausgeschriebenen Fachstudie war es deshalb zu analysieren, ob auch herkömmliche (kommerzielle und Open-Source) Anwendungen für die Analyse der Datensätze geeignet sind und auf welche Probleme man dabei stößt:

"Inhalt der Fachstudie soll sein, eine Auswahl bestehender „of-the-shelf" VA-Lösungen auf vereinzelte Datensätze der oben genannten Sammlung anzuwenden und dabei die Produkte in folgenden Punkten zu untersuchen:

- allgemeine Generizität (d.h. Unterstützung verschiedenster Datenarten und Analysevorgehen) und Erweiterbarkeit
- Möglichkeit Daten ad-hoc zu verknüpfen
- Aufzeichnung des Analyseprozesses
- Einsatz von Visualisierung, Interaktion und maschinellen Verfahren, insb. aus der maschinellen Sprachverarbeitung" ([Bos12])

2 Themenverwandte Arbeiten

Bereits vor der hier präsentierten Studie haben verschiedene Analysen von Tools zur Visualisierung von Datensätzen stattgefunden. Dabei lag der Fokus teil weise auf der allgemeinen Analyse von Tools und andererseits auf der allgemeinen Visualisierung von Datensätzen. Für uns relevante Arbeiten waren 'Visual Analytics for the Big Data Era – A Comparative Review of State-of-the-Art Commercial Systems' und 'Mastering the Information Age Solving Problems with Visual Analytics'. Das Paper über die Toolanalyse ([Kei02a]) kommt der hier geführten Analyse sehr nahe. Allerdings werden hier die allgemeinen Werte der Tools verglichen während diese Arbeit sich auf die Bewertung in Bezug auf die Lösbarkeit von VAST-Challenges konzentriert. Dennoch war die Vorarbeit sehr hilfreich in der Wahl der Kategorien für den Bewertungskatalog und als Anhaltspunkt für die Bewertung der Kategorien Allgemeine Analyse und Benutzerfreundlichkeit. Der Artikel über Problemlösung mit Visual-Analytis [DK10] wurde von uns als Hilfestellung für die Lösung der Vast-Challenges. Hier konnten Ansätze und eine Vorstellung der Denkweise bei der visuellen Analyse abgelesen werden.

3 Auswahl der Tools

Zu Beginn der Arbeit stand die Auswahl der drei Tools die es zu analysieren galt im Raum. Grundlage für diese Entscheidung war eine intensive Recherche verschiedener Tools. Dabei sollten drei Tools ausgewählt werden.

3.1 Auswahlkriterien

Bei der Auswahl der Tools wurde auf folgende Kriterien Wert gelegt:

Textuelle Analyse

Ein Kernbereich des Bewertungskataloges (siehe Kapitel 3.3 ab Seite 20) ist die Analyse von Textdaten. Diese ist daher von besonderer Bedeutung, da sie die unstrukturierten Datensätze repräsentiert. Diese Kernkompetenz wurde bereits bei der ersten Begutachtung der Programme begutachtet, doch die dort versprochenen Erweiterungen zur Analyse unstrukturierter Datensätze waren, wie bei der Analyse festgestellt wurde, noch nicht einsatzfähig. Die Spalte `Text` (siehe Kapitel 3.3 ab Seite 20) gibt eine grobe Aussage darüber, ob das Tool sich zur Textanalyse eignet. Die Bewertungsstufen sind: `ja (Erw.)` falls sehr gut geeignet, `ja` falls geeignet und `nein` falls gar nicht oder nur geschränkt dafür geeignet. Die Bewertung `ja (Erw.)` kam in dieser Kategorie bei keinem der Tools vor.

Netzwerkanalyse

Bei den Datensätzen der VAST-Challenges, die dieser Analyse zugrunde liegen, ist in zwei Fällen die Darstellung und Analyse von Knoten-Kanten-Diagrammen notwendig. Programme welche diese Funktion bereits öut-of-the-boxünterstützen sollten bevorzugt untersucht werden. Aus diesem Grund wurde diese Komponente bei der ersten Recherche ebenfalls aufgenommen. Die Spalte `Netzwerk` (siehe Kapitel 3.3 ab Seite 20) gibt grobe Aussage darüber, ob das Tool sich zur Analyse von Netzwerken im ersten Eindruck eignet. Die Bewertungsstufen sind: `ja (Erw.)` falls sehr gut geeignet, `ja` falls geeignet und `nein` falls gar nicht oder nur geschränkt dafür geeignet.

Geodaten-Analyse

Die Analyse von Geologischen Datensätzen gewinnt durch die zunehmende Vernetzung und Globalisierung immer mehr an Bedeutung. Durch Technologien wie Smartphones und darin enthaltende *GPS*-Sensoren können immer detailliertere Bewegungsprofile und -muster erstellt werden. Um solche Daten zu analysieren bedarf es gut skalierender Anwendungen und intelligenten Umgang mit großen Datenmengen und deren Visualisierung. Außerdem ist die Analyse von Geo-Daten eine der Aufgaben in den VAST-Challenges. Aus diesem Grund fließt dieser Bereich in die Auswahl der Programme ein. Die Spalte `Geo` (siehe Kapitel 3.3 ab Seite 20) gibt grobe Aussage darüber ob das Tool sich zur Analyse von Geodaten eignet. Die Bewertungsstufen sind: `ja (Erw.)` falls sehr gut geeignet, `ja` falls geeignet und `nein` falls gar nicht oder nur geschränkt dafür geeignet.

Interaktion

Die Interaktion mit den Daten ist ein wichtiges Qualitätsmerkmal bei Visual-Analytics-Tools. Erst dadurch kann bei der Exploration von Daten die Kombination aus Menschlicher Einsicht in die Hintergründe der Daten und Darstellungs-/ Rechenleistung der Programme dazu verwendet werden effektiv Informationen zu entnehmen. Aus diesem Grund ist diese Kategorie in der ersten Auswahl bereits mit berücksichtigt. Die Spalte `Interaktion` (siehe Kapitel 3.3 ab Seite 20) zeigt wie stark die Interaktion mit den Daten in dem Tool möglich ist. Die Werte dabei sind: `viel` falls die Interaktion besonders ausgeprägt, `mittel` falls Interaktion nur durchschnittlich oder `wenig` falls die Interaktion sehr begrenzt ist.

Visualisierung

Bei der ersten Auswahl der Tools sollten die verschiedenen Möglichkeiten Daten mit diesem Tool zu visualisieren Einfluss nehmen. Insbesondere wurde hierbei auf die verschiedenen Typen und deren Integration in die Anwendung Bezug genommen. Die Spalte `Visualisierung` (siehe Kapitel 3.3 ab Seite 20) enthält somit einige Kommentare über besondere Darstellungsformen der Tools. Der `hauptsächlich Business Intelligence` deutet darauf hin, dass das Tool eher für die Visualisierung von Geschäftsdaten konzipiert wurde. Diese sind damit weniger für die Visualisierung von VAST Challanges geeignet, werden aber der Vollständigkeit halber hier mit aufgeführt.

Verfügbarkeit

Die Verfügbarkeit spielt bei der ersten Analyse selbstverständlich auch eine große Rolle. Es sollten gezielt kommerzielle und Open-Source Lösungen untersucht und getestet werden. Die Spalte `Verfügbarkeit` (siehe Kapitel 3.3 ab Seite 20) gibt an was ein Tool kostet und ob es eine Testversion gibt. Häufig lässt sich der Preis leider nur durch Anfrage ermitteln. Dies ist dann entsprechend aufgeführt.

3.2 Mögliche Tools

Name	Text	Netzwerk	Geo	Interaktion	Visualisierung	Verfügbarkeit
AVS OpenViz[1]	nein	nein	ja	wenig	Heat-Maps, Fluid Simulationen	Testversion, Preis auf Anfrage
Gephi[2]	nein	ja	nein	mittel	Graph Visualisierung sehr ausgereift und gut	Frei
GGobi[3]	nein	ja	nein	mittel	Tours, Scatterplot, Barchart and Parallel Coordinates Plot	Frei
IBM Cognos Insight/Express[4]	nein	nein	nein	mittel	hauptsächlich Business Intelligence	Testversion, Preis auf Anfrage
IBM SPSS[5]	nein	ja	nein	viel	hauptsächlich statisitsch	Testversion, Preis auf Anfrage
InfoZoom[6]	nein	nein	nein	viel	nur statistische Auswertungen	Testversion, Preis auf Anfrage
INSpire[7]	ja	nein	nein	viel	hauptsächlich Business Intelligence	Preis auf Anfrage

[1] http://www.avs.com/research.html
[2] http://gephi.org/
[3] http://www.ggobi.org/
[4] http://www-142.ibm.com/software/products/us/en/subcategory/SWQ10
[5] http://www-01.ibm.com/software/analytics/spss/solutions.html
[6] http://www.infozoom.com/
[7] http://in-spire.pnnl.gov/

Name	Text	Netzwerk	Geo	Interaktion	Visualisierung	Verfügbarkeit
Knime[8]	ja	ja	nein	mittel	Scatter, Plots, R-Export, TagCloud, Interactive Table, Histogram	Frei
Miner3D Enterprise[9]	nein	nein	nein	mittel	2D/3D Trellis Charts, Visual Clustering, hauptsächlich Business Intelligence	990 Euro
Mondrian [10]	nein	nein	ja	mittel	Histograms, Scatterplots, Barcharts, Mosaicplots, Missing Value Plots, Parallel Coordinates/Boxplots, SPLOMs	Frei
MS BI Stack[11]	nein	nein	nein	viel	hauptsächlich Business Intelligence	Testversion, Preis auf Anfrage
MS Excel 2013 (64-bit)[12]	nein	nein	nein	mittel	Pivot Graph	98 Euro
NetChartsPro[13]	nein	nein	ja	wenig	ausführliche Dashboards	Testversion, Preis auf Anfrage

[8] http://www.knime.org/
[9] http://www.miner3d.com/
[10] http://rosuda.org/mondrian/
[11] http://www.microsoft.com/en-us/bi/default.aspx
[12] http://www.microsoft.com/en-us/bi/Products/OfficePreview.aspx
[13] http://www.visualmining.com/

Name	Text	Netzwerk	Geo	Interaktion	Visualisierung	Verfügbarkeit
Occulus[14]	ja	ja	ja	viel	sehr Anwendungsspezifisch	Keine Testversion, Preis auf Anfrage
Palantir[15]	nein	ja (Erw.)	ja (Erw.)	viel	umfangreiche Netzwerkanalysen, Heat-Maps	keine Testversion, Preis auf Anfrage
Panopticon[16]	nein	nein	nein	wenig	hauptsächlich Business Intelligence	Testversion, Preis auf Anfrage
Qlikview[17]	nein	nein	nein	viel	Dashboards, viele Erweiterungen	Einzellizens frei (sonst 1350 Euro)
R[18]	nein	nein	nein	viel	viele Erweiterungen	Frei
SAS[19]	ja	ja (Erw.)	ja (Erw.)	viel	hauptsächlich Business Intelligence	Testversion, Preis auf Anfrage
SMS JMP[20]	nein	nein	ja	viel	hauptsächlich Business Intelligence	Testversion, Preis auf Anfrage
Spotfire[21]	ja	ja	ja	viel	Dashboards, Plot Diagram, Gantt	Testversion, 99 Euro pro Jahr
Tableau[22]	nein	nein	ja	viel	Dashboards, Node-Link Diagram, Parallele Koordinaten	Testversion, Public Lizenz ist frei (sonst 1,999 Euro)

Tabelle 3.2: Die Liste der recherchierten Tools mit ihren Features

[14] http://www.oculusinfo.com/
[15] http://www.palantir.com/
[16] http://panopticon.com/
[17] http://www.qlikview.com/de
[18] http://www.r-project.org/
[19] http://www.sas.com/whitepapers/
[20] http://www.jmp.com/software/
[21] http://spotfire.tibco.com/
[22] http://www.tableausoftware.com

3.3 Ausgewählte Tools

Nach Abschluss der Sammlung an verfügbaren Datenanalysetools im ersten Schritt, wurden diese nach den Kriterien der Analysemöglichkeiten gefiltert. Es wurde in diesem zweiten Schritt einerseits Wert darauf gelegt vergleichbare Orientierungen und Zielgruppen abzudecken und andererseits darauf alle Datentypen (Geo/Netzwerk/Text) analysieren zu können. Ergebnis der zweiten Auswahlanalyse waren die folgenden Datenanalysetools:

- SAS
- Occulus
- Tableau
- Knime
- SPSS
- Spotfire
- INSpire
- Gephi
- JMP

Diese Vorauswahl kann in den zentralen Punkten des Bewertungskatalogs und in den Datenarten Geo, Netzwerk und Text sinnvoll bewertet werden. Zudem hatten wurde auf gute Wertungen, was die Usability und die Bekanntheit angeht, geachtet. Die letztendliche Entscheidung für die Analyse von Qlik, Spotfire und Tableau im dritten Schritt, hing unter anderem von der Verfügbarkeit einer Trial-Version sowie von der Spezialisierung einiger der Alternativen auf Business-Analyse ab. Anhand der bei der Sammlung verfügbaren Daten konnte davon ausgegangen werden, dass die drei Tools mit den Geo-, Netzwerk-Datensätzen umgehen können und zumindest zwei davon über Erweiterungen Text-Datensätzen abdecken können. Zudem war in allen drei Fällen eine Testversion verfügbar. Während Spotfire und Tableau kommerziell sind, deckt Qlik zusätzlich ein OpenSource Produkt ab.

4 Bewertungskatalog

Um eine Bewertung über die ausgewählten Programme (siehe Kapitel 3 ab Seite 15) abgeben zu können haben wir zu Beginn unserer Analyse einen **Bewertungskatalog** erstellt. Mittels diesem sollen die drei gewählten Programme (siehe Abschnitt 3.3 auf Seite 20) möglichst objektiv und nachvollziehbar untersucht und klassifiziert werden.

4.1 Übersicht

Um den Bewertungskatalog zu erstellen haben wir zunächst das Ziel dieser Analyse konkretisiert:

Ziel dieser Arbeit ist es Visual-Analytics-Tools zu vergleichen und zu bewerten. Als Bewertungsgrundlage wird hierbei ihre Fähigkeit zur Analyse von großen Datensätzen am Beispiel der VAST-Challange Daten von 2008 und 2009 herangezogen. Die Ergebnisse und Erkenntnisse werden abschließend zusammengefasst.

4.2 Bewertungskategorien

Nachfolgend soll kurz erklärt werden, unter welchen Gesichtspunkten die einzelnen Programme betrachtet und untersucht wurden und was unter diesen Punkten zu verstehen ist.

4.2.1 Analysefähigkeit - Text

Dieser Aspekt *(Gewichtung: 20%)* soll die Fähigkeit der Programme bewerten Textdaten einzulesen und diese zu visualisieren. Dabei sollen unter anderem Folgendes beachtet werden:

- *Überblick über den Text* (Identifikation von Schlüsselworten, Worthäufigkeit (Minimum / Maximum), Tagclouds)
- *Inhaltliche Analyse* (Stimmung des Textes, Betreffende Personen in Relationen setzen)
- *Mehrsprachigkeit* (Analyse von verschiedenen Sprachen, Erkennung der auftretenden Sprachen)

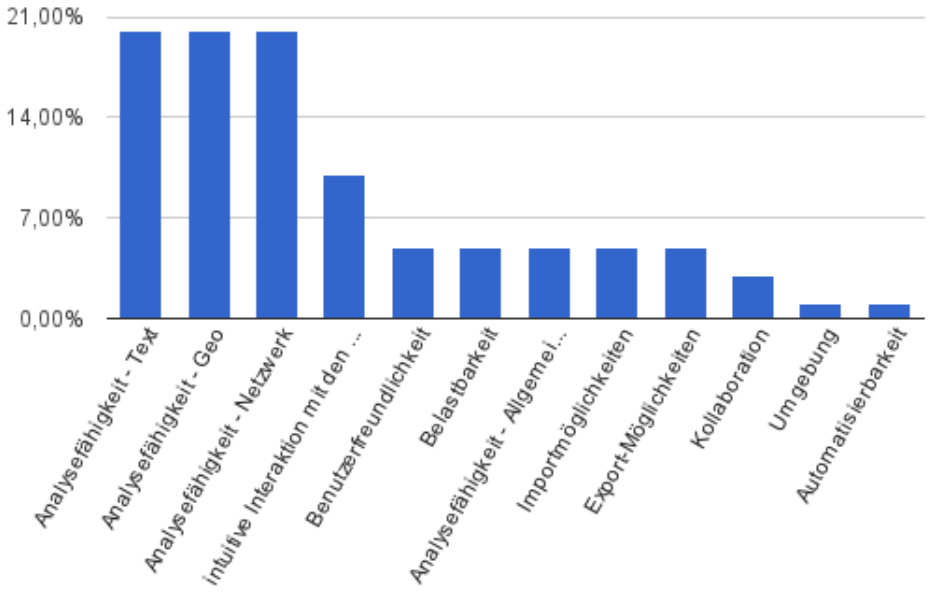

Abbildung 4.1: Bewertungskatalog - Gewichtung der Bewertung

- *Relationale Identifikation* (gleiche Autoren verschiedener Texte erkennen, Beziehung zwischen Texten herstellen, ähnliche Satzstruktur)
- *Mustererkennung* (Tabellen, Aufzählungen)

4.2.2 Analysefähigkeit - Geo

Die Analyse von Geodaten *(Gewichtung: 20%)* ist unter anderem bei der Untersuchung von Bewegungsmustern oder der Analyse von verorteten Daten von großer Bedeutung. Dabei werden folgende Aspekte betrachtet:

- *Mögliche Eingaben* (Geokoordinaten, Postleitzahlen, Städtenamen, Länder)
- *Mögliche Ausgaben* (Karte, Parameter wie Höhenlage, Bevölkerungsdichte usw., 2D, 3D, Interaktive Ausgabe)
- *Interaktion* (Verschieben, Zoomen, Filtern der Karte)
- *Visualisierung* (Auswahl an Farbmuster, Darstellung, Symbolen, verschiedene Kartenlayer)

4.2.3 Analysefähigkeit - Netzwerk

Große Netzwerke *(Gewichtung: 20%)* zu analysieren wird in der Zeit von Facebook, Twitter und Co. immer wichtiger. Es entstehen hier in den letzten Jahren beachtliche Datenmengen, die zuvor niemals vorstellbar waren (z.B. bei Facebook täglich 2,5 Milliarden Inhalte wie Statusupdates oder Ortsmarken, [Ler12]). Neben solchen sozialen Netzwerken können auch relationale Daten als Netzwerk aufgefasst werden. Um solche "Big-Data" zu analysieren bedarf es bei der Analyse von Netzwerken folgender Möglichkeiten:

- *Überblick* (Gruppenbildung, Leader-Eigenschaft, Hierarchie-Erkennung, Zusammenhang)
- *Filterung* (Dichte, Richtung/Art der Kante, Knoteneigenschaften)
- *Datengröße* (Anzahl Kanten, Anzahl Knoten, Dichte des Graphs)
- *Interaktion* (Zoom, Hervorheben, Verschieben, Daten auf Abfrage, Veränderung der Visualisierung),
- *Ausgabemöglichkeiten* (Layout-Algorithmen, Interaktive Formate)
- *Reduktion von Visual Clutter* (Edge Bundling, Ein/Ausklappen von Bereichen, Clustern von Daten)

4.2.4 Interaktion mit den Daten

Bei der Interaktion mit den Daten *(Gewichtung: 10%)* spielt in erster Linie der Umgang des Tools mit großen Datenmengen eine Rolle. Dabei geht es darum, wie der Nutzer mit den Daten agieren kann, ob Daten getaggt werden können und ob die Darstellung der Daten verändert und personalisiert werden kann.

- *Visualisierungsmantra* (Overview first, zoom and filter, details on demand. [Shn96])
- *Freie Gestaltbarkeit* (Kommentare anlegen, Texte hinzufügen, Symbole und Farben ändern, Legenden verschieben)
- *Filterung* (Teilelemente ausschließen, Selektion auf alle Diagramme übertragen, Selektion mehrerer Datensätze und Auswertung)
- *Workflow* (Nachvollziehbarkeit, Flexibilität der Reihenfolge, ggf. Hinweise, keine Störungen durch das Programm)

4.2.5 Benutzerfreundlichkeit

Bei der Benutzerfreundlichkeit *(Gewichtung: 5%)* geht es darum, wie der Benutzer es empfindet die Anwendung zu bedienen. Dabei werden neben den in der EN ISO 9241 beschriebenen Kriterien auch die "Acht Goldenen Regeln" von Shneiderman (vgl. [Shn]) zu Rate gezogen.

4.2.6 Belastbarkeit

Dieser Punkt *(Gewichtung: 5%)* behandelt in erster Linie die Belastbarkeit der Anwendungen im Bezug auf die Datengröße.

- *Umgang mit großen Daten* (Datengröße, Analysegeschwindigkeit)
- *Umgang mit kleinen Daten* (Schnelligkeit, Genauigkeit)
- *Umgang mit mehreren Datensätzen* (parallele Bearbeitung, Kombination von Bearbeitungen)
- *Verbindungsauslastung* (falls Netzwerkzugang erforderlich)

4.2.7 Analysefähigkeit - Allgemeine Daten

Dieser Punkt *(Gewichtung: 5%)* soll die allgemeinen Fähigkeiten der Anwendung zur Analyse von Daten beschreiben. Dabei geht es darum wie Standard-Datensätze ausgelesen und visualisiert werden können.

4.2.8 Importmöglichkeiten

Die Möglichkeiten für den Import *(Gewichtung: 5%)* spielen selbstverständlich eine Rolle. Da dieser Prozess jedoch nur einmal beim Visualisieren von Daten stattfindet ist dieser Punk weniger stark gewichtet.

- *Einfache Datenformate* (CSV, Txt, XLS, ...)
- *Komplexe Datenformate* (SQL, Access, SAP, ...)
- *Relationen erkennen* (Verknüpfung zu bestehenden Datensätzen, Darstellung dieser, Hervorheben von Dublikaten, Warnung vor Inkonsistenzen)
- *Datenquellen* (verwalten, verändern, Verbindungen speichern, Einstellungen/Vorkonfigurationen behalten)
- *Selektion der Daten* (Einschränkung bereits beim Import, Priorisieren, Benennung, Datenformate anpassen, Trennung festlegen)
- *Erweiterte Dateneingabe* (Copy & Paste, Import von Webseiten)

4.2.9 Export-Möglichkeiten

Der Export von Daten *(Gewichtung: 5%)* stellt eine wichtige Grundlage für die Kommunikation der Daten dar. Diese Arbeit behandelt jedoch hauptsächlich die Beurteilung der Daten auf ihre Fähigkeit Daten zu analysieren. Deshalb wird dieser Punkt weniger stark gewichtet.

- *Diagramme* (3D/2D, dynamisch, Animation, "What you see is what you get"-Export, Interaktive Export-Formate)
- *Veröffentlichung* (PDF, HTML, FTP-Upload, EMail, Powerpoint, Facebook, Twitter)

4.2.10 Kollaboration

Die Zusammenarbeit mit anderen *(Gewichtung: 3%)* ist bei der Bearbeitung von Aufgaben im Team wichtig. Bei der Analyse der Anwendungen hat dieser Punkt jedoch keine hohe Priorität.

- *Kommentare* (Kommentieren von Daten/Diagrammen, Gemeinsame Interaktion)
- *Datenbestandsverwaltung* (Ablage auf Server, Gemeinsames Nutzen)
- *Programm* (Ist das Programm nötig um die Daten betrachten zu können, Gibt es eine Webversion davon)
- *Konfigurationsmanagement* (Historie, Änderungsverfolgung, Zurücksetzen)
- *Verwaltung* (Rollen-Verteilung, Benutzereinschränkungen, Teilen-Funktion Makros, TODOs)

4.2.11 Umgebung

Die Umgebung in der die verschiedenen Anwendungen laufen *(Gewichtung: 1%)* spielt bei der Analyse der Tools eine untergeordnete Rolle. In erster Linie geht es darum die Tools auf ihre Funktionalität zu untersuchen, nicht auf ihre Umgebungseigenschaften.

- *Architektur* (Standalone/ Client-Server / Cloud)
- *Betrachtung* (Notwendige Programme zum Betrachten)
- *Scalability*
- *Datensicherheit*

4.2.12 Automatisierbarkeit

Wiederkehrende Aufgaben automatisiert auszuführen *(Gewichtung: 1%)* spielt bei einigen Aufgabenbereichen (z.B. der Analyse von Unternehmensdaten) eine größere Rolle. Bei der Analyse der VAST Challenges ist dies jedoch nicht notwendig. Deswegen wird dieser Punkt weniger stark gewichtet.

- *Makros*
- *Automatischer Export / Import*
- *Benutzerdefinierte Analysen*

4.3 Begründung der Gewichtung

Die gewählte Gewichtung der verschiedenen Kriterien (siehe Abb. 4.1) lässt sich aus unserer Aufgabenbeschreibung und der Fragestellung dieser Arbeit ableiten. Die eigentlichen Daten der VAST-Challenges dienen uns als Grundlage für die Analyse der Programme. Mit ihnen wird die Eignung der verschiedenen Tools auf den Umgang mit großen Daten hin überprüft. In erster Linie geht es bei dieser Untersuchung um die Feststellung, ob verfügbare Anwendung für die Analyse der VAST-Challenge Daten der letzten Jahre geeignet sind. Diese Daten geben deshalb die drei zu untersuchenden Analysefähigkeiten vor:

- Text-Analyse
- Geodaten-Analyse
- (Soziale-)Netzwerk-Analyse

Die weniger wichtigen Aspekte dieser Untersuchung (Automatisierbarkeit, Umgebung und Kollaboration) lassen sich durch die geringe Relevanz dieser Punkte bei der einmaligen Analyse von Daten erklären.

4.4 Bewertungsstrategie

Im Anschluss an die Analyse der verschiedenen Tools von jeweils einem Teammitglied, werden die Tools in gemeinsamer Absprache verglichen und bewertet. Es handelt sich dabei um eine subjektive, qualitative Wertung mit Punkten von 0 (schlecht/nicht vorhanden) bis 5 (sehr gut). Um den Spielraum der Wertung einzuschränken werden alle Unterkategorien einzeln bewertet und der Durchschnittswert als Gesamtwertung berechnet. Jede Unterkategorie verfügt zusätzlich über Stichpunkte, anhand derer die Bewertung durchgeführt werden soll. Beispielsweise soll der Überblick über Text anhand des Vorhandenseins einer Keywords-Identifizierung, einer Bestimmung des am meisten genutzten und am seltensten genutzten Wortes sowie der Erstellung einer Tagcloud bewertet werden. Am Ende werden die

Durchschnittsbewertungen jeder Kategorie mit der Gewichtung dieser Kategorie verrechnet und für die Gesamtwertung des Tools ein gewichteter Durchschnitt berechnet.

5 Analysevorgehen

Um möglichst schnell eine gute Bewertung der Tools zu erhalten, wurde die Analyse der Tools zu Beginn auf die einzelnen Teammitglieder aufgeteilt. Dabei musste zunächst ein Kriterium zur Aufteilung der Analyseschritte festgelegt werden und anschließend eine Strategie zur Verknüpfung der Ergebnisse entworfen werden.

5.1 Aufteilung

Es wurde entschieden die Analyse anhand der Tools aufzuteilen. Auch eine Aufteilung nach Bewertungskriterien wurde in Erwägung gezogen, doch durch die Aufteilung anhand der Tools ist eine geringere Einarbeitungszeit notwendig. Zudem kann das Tool detaillierter erfasst werden. Ein Vergleich der verschiedenen Tools bietet durch die notwendige Interaktion der Teammitglieder eine bessere Objektivität in der Wertung, als eine Einzelwertung über alle Tools hinweg. Die Tools wurden wie folgt aufgeteilt:

- Fabian Merkle: Tableau
- Sebastian Zillessen: Spotfire
- Hanna Schäfer: Qlik

5.2 Stufen der Analyse

Die Analyse gliedert sich in mehrere Stufen. Dabei wird zunächst eine Aufteilung in zwei Durchläufe vorgesehen, um bei der Analyse bereits auf die ausgetauschten Vergleiche aufbauen zu können. Zudem wird jeder Durchlauf in die drei vorhandenen Datensätze aufgeteilt, sodass alle Teammitglieder zur gleichen Zeit an einem Datensatz arbeiten und sich gegenseitig austauschen können. Der gesamte Ablauf der Analyse sieht wie folgt aus:

- Analyse erster Durchlauf (Überblick)
  - Analyse VAST Challenge 2008: Text
  - Analyse VAST Challenge 2009: Geo
  - Analyse VAST Challenge 2009: Netzwerk
  - Vergleich Analyseergebnisse

- Analyse zweiter Durchlauf (Detail)
  - Analyse VAST Challenge 2008: Text
  - Analyse VAST Challenge 2009: Geo
  - Analyse VAST Challenge 2009: Netzwerk
  - Vergleich Analyseergebnisse
- Gemeinsame Bewertung im Bewertungskatalog

6 Toolvorstellung

6.1 Qlik

Qlik ist eine Business-Discovery-Platform zu schnellen und interaktive Analyse von Betriebsdaten. Qlik ist ein OpenSource Tool, das in einer eingeschränkten Version kostenlos zur Verfügung steht. Die Vollversion des Produktes ist nur beim Kauf erhältlich. Unter den ca. 26.000 Kunden befinden sich hauptsächlich Mittelständische Unternehmen, aber auch größere Betriebe wie Edeka und Toshiba.

6.1.1 Plattformen, Verfügbarkeit, Installation

Die Qlik Anwendung steht als Web- und Desktopanwendung zur Verfügung. Unter der Desktopsystemen werden Windows 32bit und 64bit unterstützt. Zusätzlich steht die Kaufversion von Qlik als Mobile App zur Verfügung. Ein besonderer Aspekt ist, das neben den eingebauten Funktionen mehrere Erweiterungen von verschiedensten Analyserichtungen angeboten werden. Diese sind von anderen Nutzern, oder Mitgliedern der Community geschrieben worden und können teilweise auch kostenfrei heruntergeladen werden. Zur Nutzung der Erweiterungen wird jedoch das Einschalten des *Webviews* verlangt und somit eine bestehende Internetverbindung vorausgesetzt.

6.1.2 Erster Eindruck

Anhand der Demonstrationsvideos und Beispieldatensätze sind die Erwartungen an eine einfache Bedienung und präzise Ergebnisse sehr hoch. Dennoch werden je nach System bereits beim Download Mängel deutlich, da erst nach mehrmaligem registrieren und fehlerhaften Links das Programm installiert werden kann. Ein positiver Aspekt ist die Startseite 6.1 des Tools, welche insbesondere über Beispiele und Anleitungen den Einstieg erleichtern soll.

Bei genauem Hinschauen ist jedoch schon die Beschriftung der Menüpunkte wenig intuitiv und oftmals sogar irreführend 6.2. Viele Funktionen können hier erst nach Lesen der Anleitung effektiv genutzt werden.

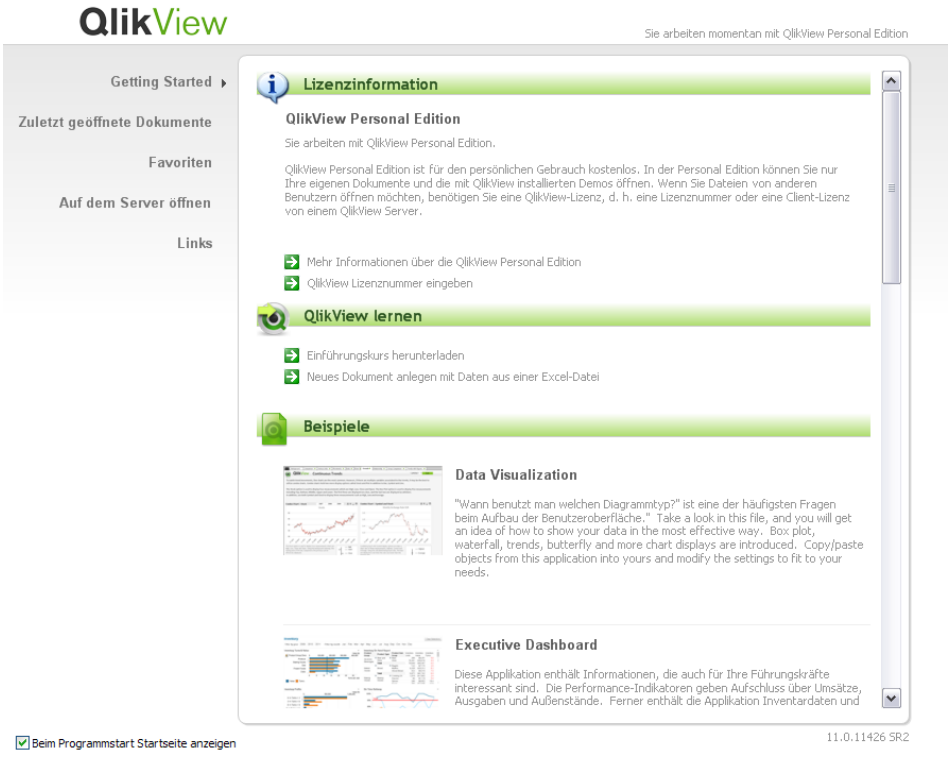

Abbildung 6.1: Qlik - Startseite mit Einführungstutorial und Beispieldatensätzen ermöglicht einen schnellen Einstig in die Funktionen des Programms und bietet Einsteigern alle notwendigen Links zur weiteren Nutzung

Abbildung 6.2: Qlik - Menüleiste in einem neuen Dokument, erschwert das genaue Zuordnen von gewünschten Aktionen zu Menüpunkten

6.1.3 Dateneingabe

Das Einlesen von Daten wird in Qlik durch einen Assistenten durchgeführt. Dabei fällt zunächst die Einschränkung der Datensätze auf Excel-Dateien auf. Andere Datensätze können nur durch manuelles Schreiben eines alternativen Skriptes eingelesen werden. In diesem Fall muss das Einlesen ohne einen Assistenten 6.3 durchgeführt werden. Durch die Einträge in Foren der Community ist es meist möglich ein fertig konfiguriertes Skript

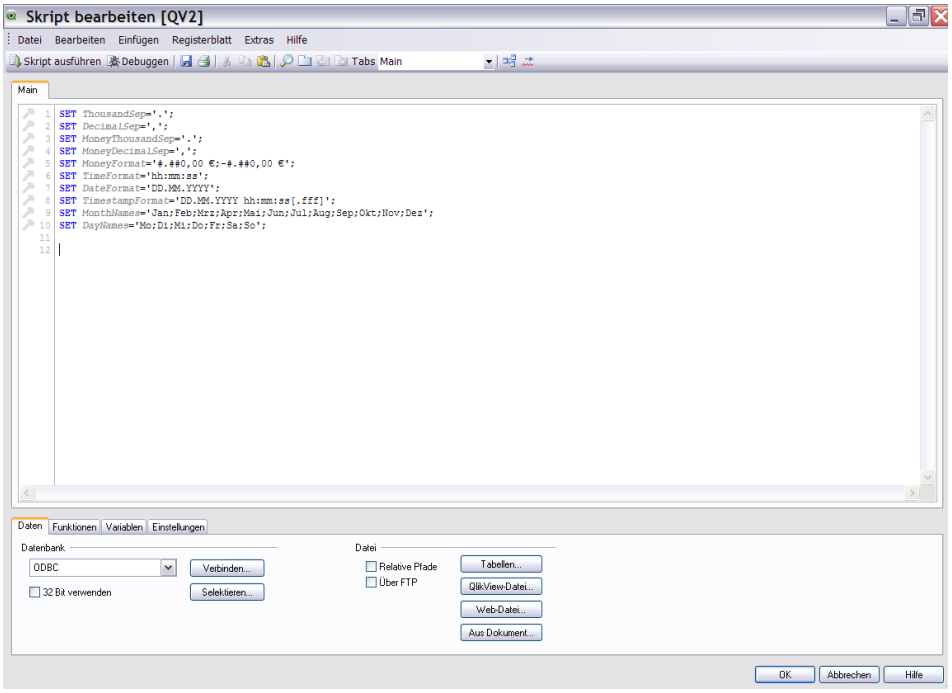

Abbildung 6.3: Qlik - Skript zur Datenerfassung im Editor geöffnet. Zwar werden Eingabehilfen gestellt, doch das ändern des Skriptes ist für Neunutzer nur langsam erlernbar

zu finden, doch da für die Analyse meist Excel-Dateien zur Verfügung stehen, oder leicht konvertiert werden können, werden wir uns im Weiteren auf die Dateneingabe über den Editor beschränken. Dabei werden die Spaltenüberschriften optional als Datenüberschriften gewählt. Zudem kann schon innerhalb des Assistenten ein Einstiegsdiagramm erstellt werden.

6.1.4 Erste Visualisierung

Beim ersten Einlesen der Daten kann im Assistent eine Startvisualisierung gewählt werden. Dabei stehen die im Bild gezeigten Diagrammtypen 6.4 zur Auswahl. Das gewählte Diagramm wird anschließend im Assistenten modifiziert und beim Öffnen des neuen Dokumentes auf der Visualisierungsfläche angezeigt. Das Diagramm kann später jederzeit über die Einstellungen modifiziert oder gelöscht werden. In unserem Beispiel wurde das Balkendiagramm als erste Visualisierung gewählt. Es wird nun ein Balkendiagramm aus den gewählten Daten und berechneten Daten angelegt 6.5. Zusätzlich legt Qlik sofort eine Liste an, in welcher die Parameter der Visualisierung stehen. Durch die Auswahl von Daten in der Liste kann mit der Visualisierung interagiert werden 6.6. Innerhalb des "Visualisierungssheets"können nun weitere Objekte hinzugefügt werden. Dabei stehen zunächst folgende Steuerelemente und Anzeigeformen zur Auswahl:

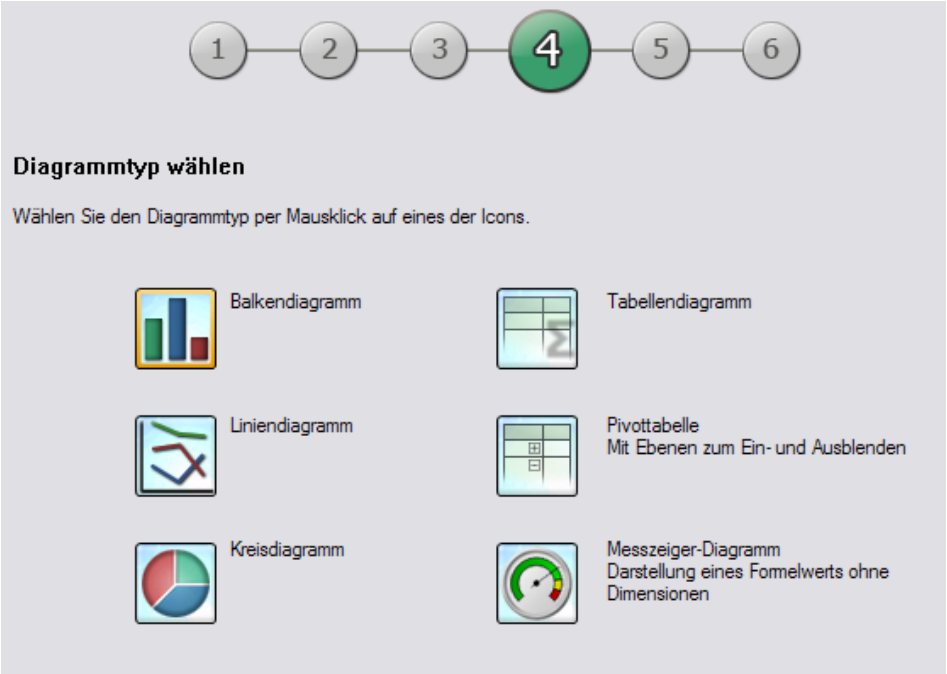

Abbildung 6.4: Qlik - Vorschläge an Standarddiagrammen nach der Dateneingabe

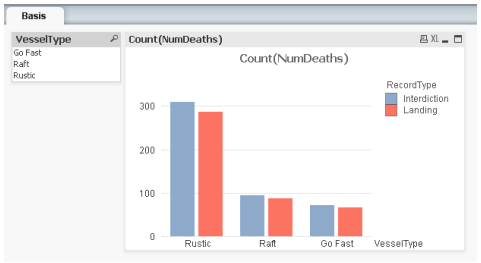

Abbildung 6.5: Qlik - Erstes Diagramm mit Listbox nach Erstellung. Das Diagramm zeigt ohne Verwendung der Listbox alle Daten der Tabelle an.

- Listbox
- Statistikbox
- Multibox
- Tabellenbox
- Diagramm
- Inputbox
- Statusbox
- Schaltfäche
- Textbox
- Linienobjekt
- Schieberegler
- Lesezeichenbox
- Suchbox
- Sammelbox
- Containerbox

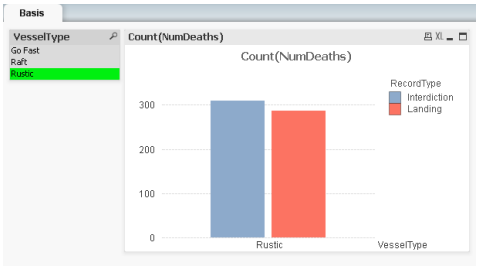

Abbildung 6.6: Qlik - Erstes Diagramm mit Listbox nach Markierung. Sobald die Listboxauswahl eingeschränkt wird, verändert sich auch der zugehörige Datensatz in der Visualisierung

Interessant für die Interaktion ist zu Beginn die Listbox, welche alle Werte einer Datenspalte auflistet und so eine Auswahl ermöglicht. Diese Auswahl wird dann auf allen anderen Objekten synchronisiert. Für die Visualisierung sind zunächst die Steuerelemente von Bedeutung, welche eine visuelle Repräsentation der Daten bietet, wie es im Falle von Qlik nur die Diagramme tun. Die Statistikbox enthält dagegen keine Visualisierung statistischer Daten, sondern Felder, in welchen statistisch relevante Werte wie Median, Summe usw. angezeigt werden. Als Diagrammtypen stehen folgende im Assistenten bildlich veranschaulichte Diagramme 6.7 zur Auswahl. Die Symbole repräsentieren:

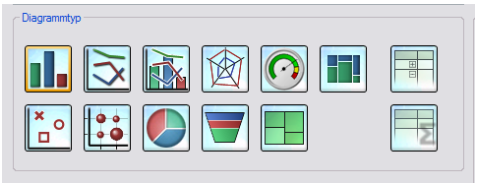

Abbildung 6.7: Qlik - Auswahl an Diagrammen zur Visualisierung

- Balkendiagramm,
- Liniendiagramm,
- Kombi-Diagramm,
- Punktdiagramm,
- Kreisdiagramm,
- Pivottabelle,
- Tabellendiagramm,
- Matrixdiagramm,
- Netzdiagramm,
- Messzeiger-Diagramm,
- Blockdiagramm,
- Trichterdiagramm,
- Marimekko-Diagramm

Wie an der Aufzählung zu sehen ist, handelt es sich bei der Auswahl eher um statistische Standarddarstellungen welche nur eine eingeschränkte Art der Visualisierung bieten.

6.1.5 Einschränkungen

Qlik ist auf den ersten Blick ein sehr einfaches Tool, das hauptsächlich Basisdiagramme zur statistischen Visualisierung bietet. Eine besondere Einschränkung ist dabei die Eingrenzung des Dateiformates auf Excel-Formate, wodurch das Einlesen von unstrukturierten Datensätzen grundsätzlich verhindert wird. Eine Stärke, welche die Einschränkungen ausgleicht ist die starke Individualisierbarkeit des Tools über Erweiterungen und Veränderungen im Skript. Insbesondere durch die Erweiterungen wird eine große Basis an Visualisierungsmöglichkeiten geschaffen. Über eine Kooperation mit Attivio[1] kann beispielsweise auch freier Text innerhalb von Qlik analysiert werden. Eine Beispieldatei kann hierzu bereits getestet werden. Leider ist es bisher nicht möglich Attivio als Erweiterung in die Desktop-Anwendung einzubinden, um eigene Daten zu Analysieren. Eine weitere für unsere Analyse relevante Einschränkung ist das Erstellen von Graphen anhand eines Node-Link-Datensatzes. Hierzu konnten wir leider keine Erweiterungen finden.

6.2 Spotfire

Spotfire[2] wird von der Firma TIBCO[3] hergestellt.

Die Wurzeln von Spotfire reichen bis zur Universität Maryland zurück. Dort arbeitete Christoph Ahlberg in den frühen 90-er Jahren gemeinsam mit Ben Shneiderman an Anwendungen, die dynamische Abfragen und Visualisierungen ermöglichen sollten. Daraus entstand später das "Information Visualization and Exploration Environment" (IVEE). Spotfire wurde 1996 veröffentlicht und 2007 von Tibco aufgekauft (vgl. [Shn99]).

Spotfire wird in vielen verschiedenen Industriezweigen verwendet. Sowohl Telekommunikationsunternehmen, Regierungen als auch Finanzdienstleister setzen Spotfire als Visual-Analytics-Tool ein. Es existieren auch einige wissenschaftliche Arbeiten zu Spotfire (z.B. [Ahl96], [PS08]).

6.2.1 Plattformen, Verfügbarkeit, Installation

Nach einer Registrierung zur 30-tägigen kostenlosen Testversion (e-Mail Adresse erforderlich) kann man sich Spotfire für Microsoft Windows herunterladen. Es werden dabei Microsoft Windows 7, Vista und XP unterstützt. Zum Verwenden des Visual-Analytics-Tools muss man sich auf dem Server von Tibco mit der installierten Anwendung registrieren.

[1] http://qlikmarket.qlikview.com/QvAJAXZfc/opendoc.htm?document=docs/QlikView%20Attivio%20Demo%20App.qvw&host=qlikmarket&anonymous=true&utm_source=attivio&utm_medium=link&utm_campaign=qlikview+partner+demo
[2] http://www.spotfire.tibco.com
[3] http://www.tibco.com

Die ebenfalls zur Verfügung stehende Web-Player Versionen sind plattformunabhängig. Diese wurden im Rahmen dieser Arbeit jedoch nicht weiter untersucht, da der Aufwand für die Einrichtung und Konfiguration nicht Inhalt dieser Arbeit war.

6.2.2 Erster Eindruck

Die Oberfläche von Spotfire macht einen sehr aufgeräumten und übersichtlichen Eindruck, was sich auch beim späteren Verwenden der Anwendung bemerkbar macht. Es ist ersichtlich, dass großen Wert auf möglichst schnelle und einfache Benutzung gelegt wurde (siehe Abb. 6.8).

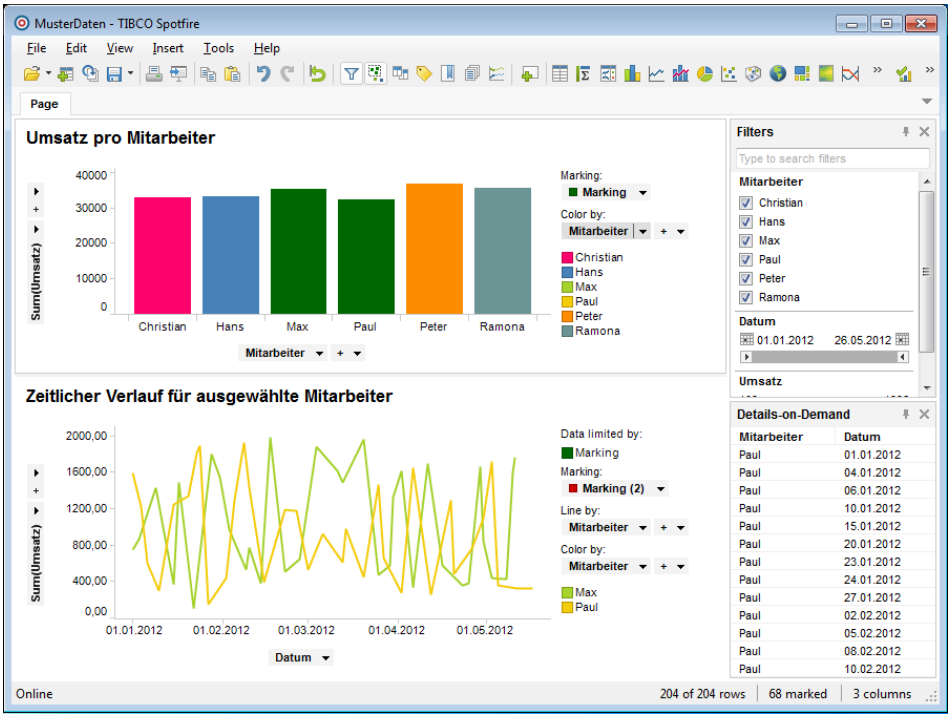

Abbildung 6.8: Erste Visualisierung. Hier ein Bar-Chart mit einer Detail-Ansicht. Der zeitliche Verlauf der Umsätze wird nur für die beiden selektierten Mitarbeiter dargestellt.

Der direkte Verweis auf eine Einführung in die Bedienung des Programmes [4] ermöglicht es schnell einen Überblick über die Funktionalität und das Einlesen von Daten zu erhalten.

[4] http://stn.spotfire.com/gettingstarted/spotfire/400/en/Default.aspx

6.2.3 Dateneingabe

Spotfire unterstützt verschiedenste Dateiformate. Neben den Standards (Excel- und CSV-Dateien) werden neben den hauseigenen Speicherformaten auch *ESRI-Shapes* und Microsoft Access Datenbanken unterstützt. Über das *UDL*-Format können außerdem anderen Datenbanken angebunden werden (z.B. über *ODBC* oder *ADO*(*-.NET*)). Nach der Auswahl der gewünschten Datei hilft ein Assistent (siehe Abb. 6.9) dabei, die Daten korrekt einzulesen. Neben den Einstellungen zu Datentypen, Zeichensätzen und Trennzeichen findet man auch die Möglichkeit Teile des Datensatzes zu ignorieren und Namen und Werte festzulegen.

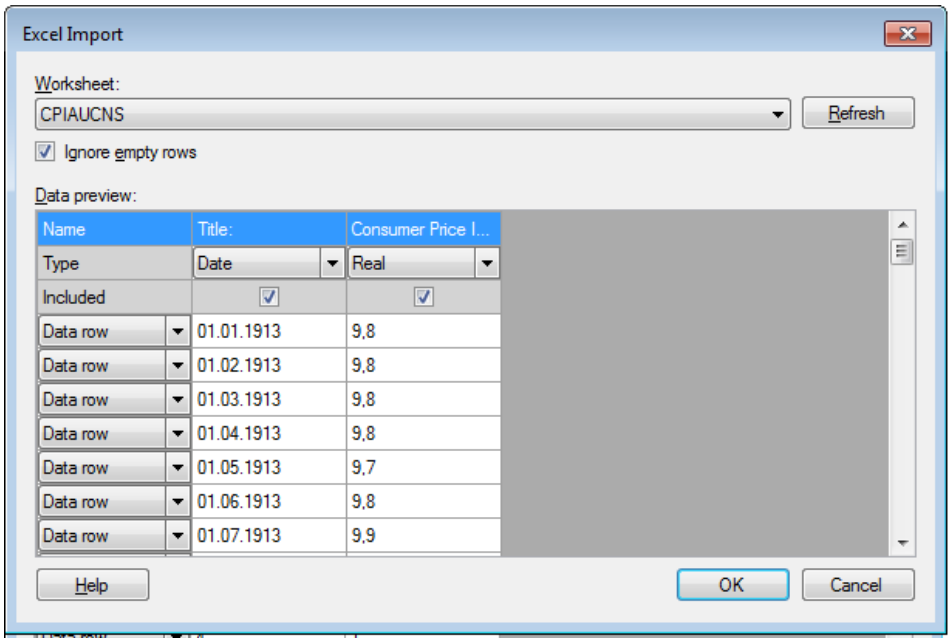

Abbildung 6.9: Import einer zweispaltigen Excel-Tabelle. Die Namen der Spalten wurden bereits automatisch erkannt. Beide Spalten wurden für den Import vorgemerkt.

6.2.4 Erste Visualisierung

Nach dem Einlesen der Daten versucht Spotfire immer ein Bar-Chart zu erstellen. Dies ist nicht immer sinnvoll. Eine schnell verständliche und intuitive Darstellung der Inhalte ermöglicht jedoch rasch das Erstellen einer alternativen Visualisierung. In dem eingelesenen Beispiel[5] handelt es sich um den "Consumer Price Index for All Urban Consumers". Ein Bar-Chart über den zeitlichen Verlauf macht hier durchaus Sinn. Mit wenigen Maus-Klicks

[5]http://research.stlouisfed.org/fred2/series/CPIAUCNS/downloaddata?cid=9

kann man sich hier den Verlauf des Kaufverhaltens über die Zeit darstellen (siehe Abb. 6.10).

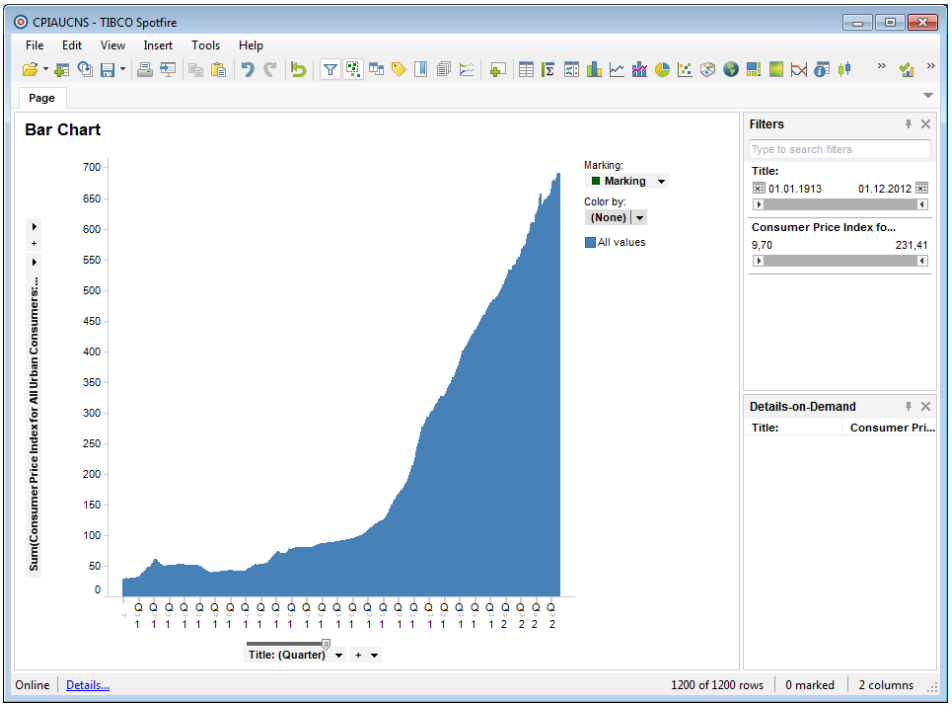

Abbildung 6.10: Erste Visualisierung einers Datensatzes

Besonders hilfreich ist dabei, dass neue Visualisierungen des Datensatzes schnell und intuitiv aus der Toolbar hinzugefügt werden können. Dabei kann die Visualisierung in unterschiedliche Seiten aufgeteilt werden, sodass immer ausreichen Platz vorhanden ist. Neue und bereits bestehende Visualisierungen werden dabei mittels *brushing and linking* verbunden. Außerdem verfolgt Spotfire das "Visual Information-Seeking Mantra"(vgl. [Shn96]): Übersicht, Filter und Zoom, Details auf Nachfrage. Dieses Konzept wird unter anderem gut durch die Sidebar realisiert. Dort kann der Datensatz schnell gefiltert werden und selektierte Datensätze werden zur genaueren Betrachtung im Detailfenster angezeigt.

Jede Visualisierung hat bei Spotfire eine Titelleiste. In dieser können die Eigenschaften der Visualisierung (wie Datenanbindung, Farben, Skalierung, Zoom, etc.) konfiguriert werden. Das dazu verwendete Fenster ist gut gegliedert und eindeutig beschriftet. Man findet sich schnell zurecht und kann gewünschte Änderungen gezielt vornehmen. Alle Diagramme sind außerdem mit Legenden und Beschriftungen versehen. In diesen kann meistens auch die Datenquelle direkt verändert werden. So ist es z.B. möglich in einem Diagramm die Achsen schnell auszutauschen und so andere Daten zu visualisieren.

Wie aus der Menüleiste direkt hervorgeht unterstützt Spotfire folgende Visualisierungen:

- Bar-Charts
- Scatter-Plots
- Line-Charts
- Pie-Charts
- 3D-Scatter-Plots
- Heat-Maps
- Tree-Maps
- Parallel-Coordinates
- Box-Plots
- Node-Link-Diagramme

Diese verschiedenen Diagramme können auf unterschiedlichen Seiten angelegt werden um so eine optimale Gliederung der Daten zu erhalten. Darüber hinaus unterstützt Spotfire die Verlinkung von Daten. Selektionen in einem Datensatz führen automatisch auch dazu, dass dieser Datensatz in anderen Visualisierungen hervorgehoben wird (siehe Abb. 6.11).

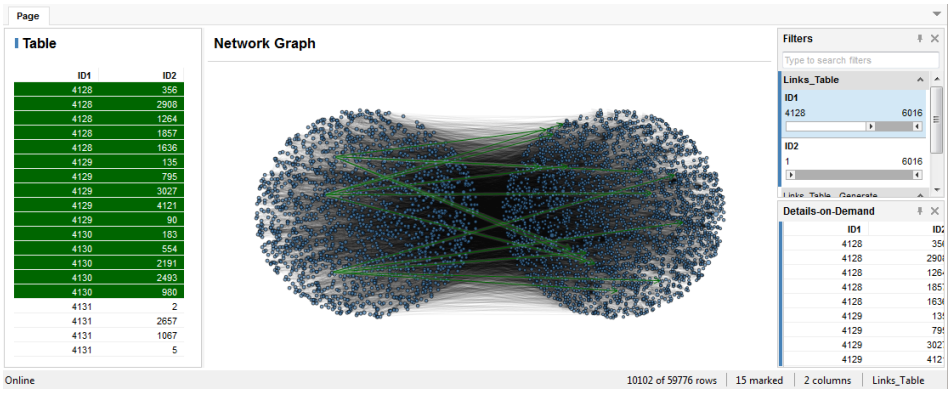

Abbildung 6.11: Hervorheben von selektierten Datensätzen in allen Visualisierungen. Links: Darstellung der Kanten als Liste, Rechts: Darstellung des Graphen als Knoten-Kanten-Diagramm, Knoten : blau, nicht markierte Kanten : schwarz, markierte Kanten : grün

Neben dem reinen Hervorheben von Datensätzen werden auch abhängige Datensätze unterstützt. So können für selektierte Datenpunkte in einer anderen Visualisierung andere Eigenschaften dargestellt werden. Ein denkbares Szenario hierfür ist z.B. die Visualisierung von Umsatz in Abhängigkeit von Verkäufern. In der ersten Visualisierung sieht man in einem Bar-Chart die Verteilung der Umsätze pro Mitarbeiter. Sobald man einen oder einige Mitarbeiter auswählt sieht man in einem Line-Plot den zeitlichen Verlauf des Umsatzes (siehe Abb. 6.8). In Spotfire nennt sich dies Detail-Visualisierung und kann für alle Diagramme verwendet werden.

6.2.5 Einschränkungen

Spotfire macht auf den ersten Blick einen sehr guten Eindruck und vermittelt große Stärken in der Visualisierung von strukturierten Datensätzen (Tabellen, etc.). Sobald es jedoch um

unstrukturierte Datensätze (z.B. Text) geht, scheitert das Programm. Weder Word-Clouds noch sonstige textbasierte Visualisierungen sind möglich.

Attivio[6] ist ein mächtiges Text-Analyse-Tool. Gemäß der Produktinformationen von Attivio[7] existiert eine Kooperation zwischen Attivio und Spotfire. Leider war es jedoch nicht mögliche Spotfire um dieses Plugin zu erweitern, da es keinerlei Informationen zur Installation gibt. Auf unsere Nachfrage nach einer solchen Unterstützung wurde uns von Spotfire nur mitgeteilt, dass Spotfire nur für die Verwendung von strukturierten Daten geeignet sei.

6.3 Tableau

Tableau ist ein Datenvisualiserungs-Tool von Tableau Software[8] aus Seattle, USA welches die Zentrale[9] in Standford University hat. Als wissenschaftlich Vorgeschichte hat Tableau dabei einem Fachartikel names "Polaris"[10] . Es wird geschätzt[11], dass Tablau 2010 einen Umsatz von bis zu 40 Millionen US Dollar erwirtschaftet hat.

6.3.1 Plattformen, Verfügbarkeit, Installation

Es gibt folgende Versionen des Produkts: die Standardsoftware **Tableau Desktop** (Personal oder Professional Edition), die Business Intelligent Lösung **Tableau Server** und die kostenlose **Tableau Public** zur Erstellung von Inhalten für Webseiten. Eine 14 Tage Testversion ist verfügbar. Wir werden uns nur mit der Version Tableau Desktop Professional 7.0 mit deutschem Interface beschäftigen, für die uns eine akademische Lizenz zur Verfügung steht. Alle Produkte benötigen Microsoft Windows XP oder höher.

Inhalte können für Nutzer ohne Softwarelizenz auch als html-Seite exportiert werden.

6.3.2 Erster Eindruck

Tableau animiert zum Herumspielen mit den Daten. Das Programm ist klar strukturiert und die Grundfunktionalitäten sind auch ohne Tutorial schnell zu erlernen. Die Arbeitsweise spiegelt sich in dem Hauptfenster wieder (siehe Abb. 6.12):

[6] www.attivio.com
[7] http://www.attivio.com/partners/partner-showcase/technology-alliances/1110-tibco-software.html
[8] http://www.tableausoftware.com/about
[9] http://en.wikipedia.org/wiki/Tableau_Software
[10] Chris Stolte, Diane Tang, and Pat Hanrahan - "Polaris: A System for Query, Analysis, and Visualization of Multidimensional Databasesërschienen 2006 in "communications of the acm Vol 51"http://mkt.tableausoftware.com/files/Tableau-CACM-Nov-2008-Polaris-Article-by-Stolte-Tang-Hanrahan.pdf
[11] http://www.bizjournals.com/seattle/blog/techflash/2010/07/tableau_software_grows_like_gangbusters_plans_to_hire_100.html

- **Datenzugang**: in der linken Spalte werden die Spalten des Datensatzes (Dimensionen) und automatisch erkannte Kennzeilen angezeigt. Diese werden mit Drag-and-Drop zu einer Visualisierung hinzugefügt.
- **Visualisierungsbereich**: in der zentralen Oberfläche findet die Hauptinteraktion mit den Daten statt. Dabei platziert man die Datenüberschriften je nach Bedarf auf einem Diagramm oder benutzt sie als Filter oder Markierung.

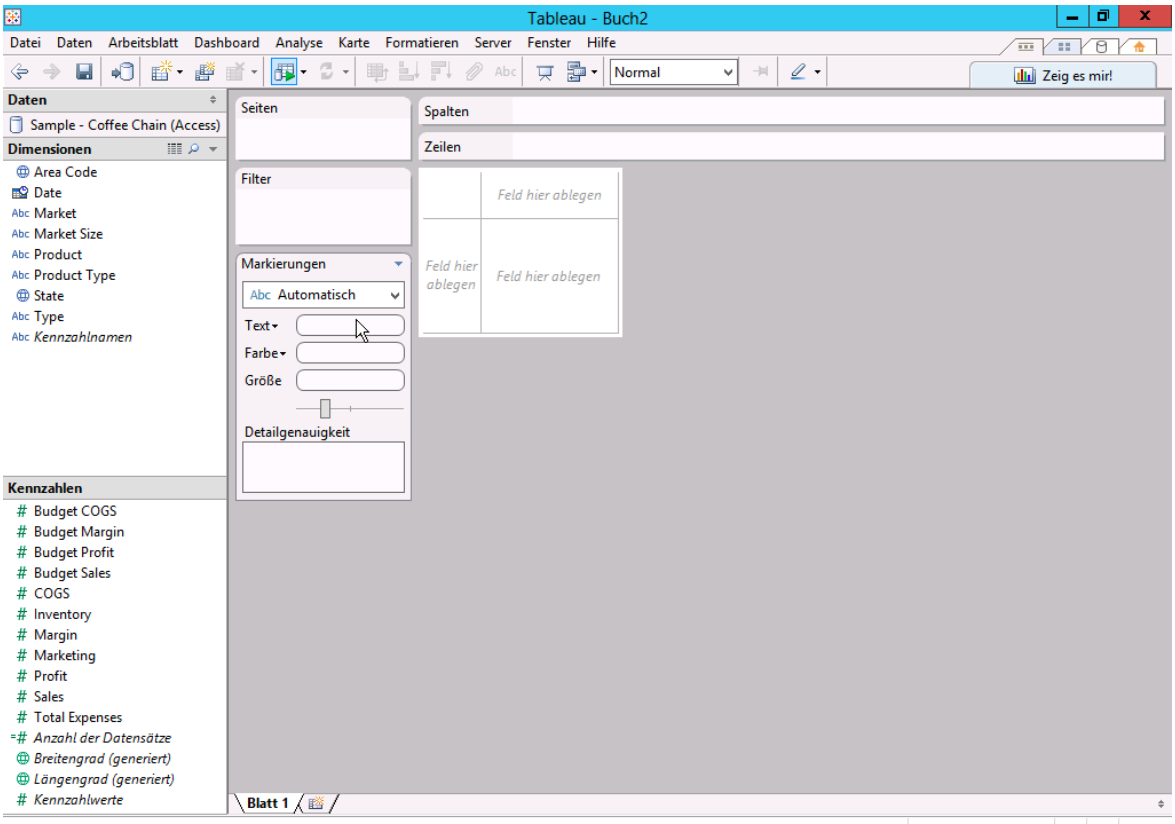

Abbildung 6.12: Das Hauptfenster von stellt bereits die Arbeitsmaske dar in der die gearbeitet werden kann

6.3.3 Dateneingabe

Tableau unterstützt sehr viele Dateiformate. Die Standardformat (Excel-, Access- und Text-Dateien) sind vertreten, genauso wie speziellere Datenbankanbindungen[12]. Außerdem können über ODBC anderen Datenbanken angebunden werden. Genauso wie bei Spotfire (siehe

[12]Tabeleau-Datenextraktion, IBM DB2, Oracle, MySQL, PosgreSQL, Microsoft PowerPoint, Microsoft Azurs, SAP NetWeaver, SAP HANA, und weitere

Abb. 6.9) hilft ein Assistent (siehe Abb. 6.13) beim Einlesen der Daten. Es kann angegeben werden welche Daten der Datei analysiert werden sollen und ob die Spalten über die ersten Zeile bezeichnet werden. Es wird dann auf eine Verbindung zu dieser Datei in Tableau abgespeichert. Dadurch werden die Dateien in Tableau bei einer Änderung der Quelldatei ebenfalls aktualisiert.

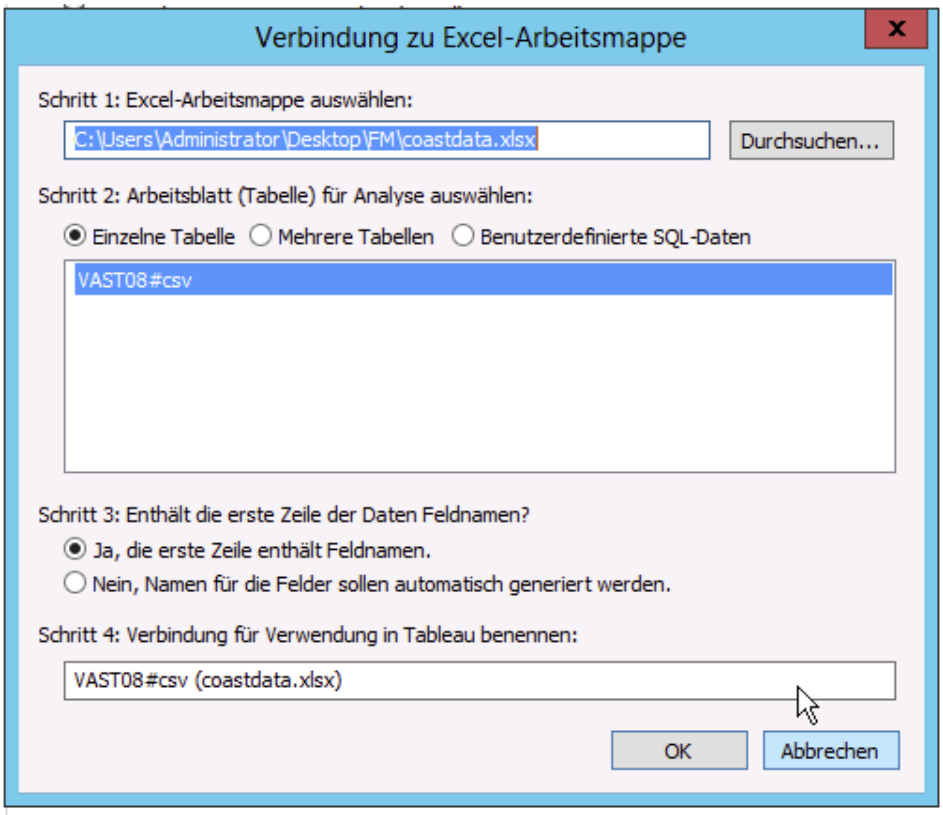

Abbildung 6.13: Verbundungsassistent zum Aufbau von Datenverbindungen

6.3.4 Erste Visualisierung

Beim Start des Programms erscheint ein Startassistent (siehe Abb. 6.14). Von dort aus können neue Datenverbindungen aufgebaut, bisherige Arbeitsmappen geöffnet oder Beispieldatensätze geladen werden.

Nun erscheint das Hauptfenster (siehe Abb. 6.12) ohne Visualisierung. Wobei sich die eben importierten Daten als Überschriften in der linken `Daten` Spalte zugänglich sind. Kleine Icons neben dem Titel zeigen den Datentyp an: Strings, Zeitliche Daten, Geo-Koordinaten und unter `Kennzahlen` aufgelistete numerische Daten. Mit folgenden Schritten erstellen wir nun eine Diagramm:

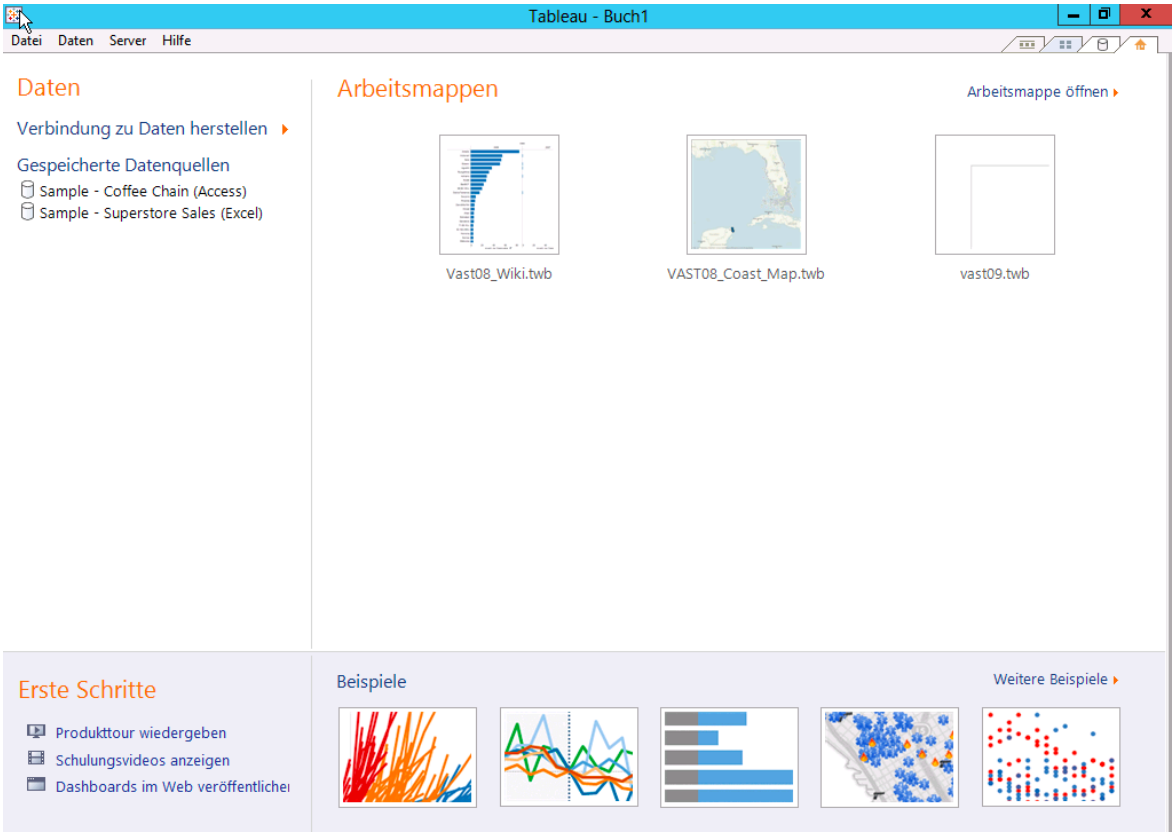

Abbildung 6.14: Startassistent zum schnellen Finden von momentanen Projekten und Beispielprojekten

1. **Festlegen der Daten**: Wir ziehen aus dem linken Datenbereich `Dimensionen` oder `Kennzahlen` auf die Felder `Spalten`, `Zeilen` und `Markierungen`. Optional stehen auch die Felder `Seiten` und `Filter` zur Verfügung um die angezeigten Daten interaktiv zu bestimmen.
2. **Diagrammauswahl**: Je nachdem welchen Typ unsere ausgewählten Daten haben, wählt das Programm bereits einen sinnvollen Datentyp. Wir können diesen anpassen indem wir oben rechts im Eck des Hauptfenster (siehe Abb. 6.12) auf den Button `Zeig es mir!`(siehe Abb. 6.15) klicken. Hier erhält man eine Übersicht über alle verfügbaren Visualiserungsweisen. Tableau erkennt dabei anhand der zur Darstellung gewählten Daten welche Diagrammtypen sinnvoll wären und markiert diese farbig. Auf diese Weise fällt die Wahl der geeigneten Darstellung sehr leicht.

Wie aus dem `Zeig es mir!` Fenster (siehe Abb. 6.15) hervorgeht, unterstützt Tableau folgende Visualisierungen:

- Texttabellen

Abbildung 6.15: Show me Fenster zur schnellen Diagrammtypbestimmung

- Heatmaps
- Balkendiagramme: horizontal, gespalten- und nebeneineanderliegend
- Streudiagramme, Kreisansichten
- Symbol- und gefüllte Karten
- Lininediagramme: fortlaufend und diskret
- Bereichsdiagramme: fortlaufend und diskret
- Kreisdiagramme
- Mehrdimensionale Diagramme: Zweifachpunkt, Zweifachlinien, Zweifachkombiniation
- Sonderformen: Bulletdiagramme: Ganttdiagramme, Histogramm

Die große Stärke eines Visiualisierungstools wie Tableau ist die Interaktion mit den Daten, insbesondere *brushing and linking*. Man wählt geeignete Diagrammtypen für mehrere Perspektiven auf die Daten und beobachtet wie sie zusammenhängen. Zum Beispiel hat man eine Karten auf der die Umsätze einer Firma nach Region angezeigt werden. Man kann nun mit der Maus bestimmte Regionen nach Belieben markieren wodurch sofort die Umsätze nur dieser Regionen in einer Zeitlinie anzeigt werden.

Tableau übernimmt diese Konzeption indem Diagramme in **Dashboards** zusammengefasst werden: Jedes Diagramm und Dashboard in Tableau wird in einer Arbeitsmappe dargestellt und bearbeitet. Arbeitsmappen werden wie die Tabellenblätter bei Microsoft Excel am unteren Bildschirmrand als Reiter dargestellt (vgl. siehe Abb. 6.12). Hat man einige Diagramme fertiggestellt kann ein Dashboard erstellen und dort alle gewünschten Diagramme per Drag-and-Drop platzieren. Nun legt man fest wie sich das Markieren von Datensätzen auf die anderen Diagramme auswirkt. Dies erreicht man durch Klicken der Option `Als Filter verwenden` (Rechten Maustaste auf ein Diagramm im Dashboard) oder durch Erstellen von benutzerdefinierten `Aktionen` im Menü `Dashboard`. Dashboards lassen sich durch Überschriften, Filter- und Textfelder sehr ansprechend gestalten. Für ein Beispiel siehe Abb. 6.16.

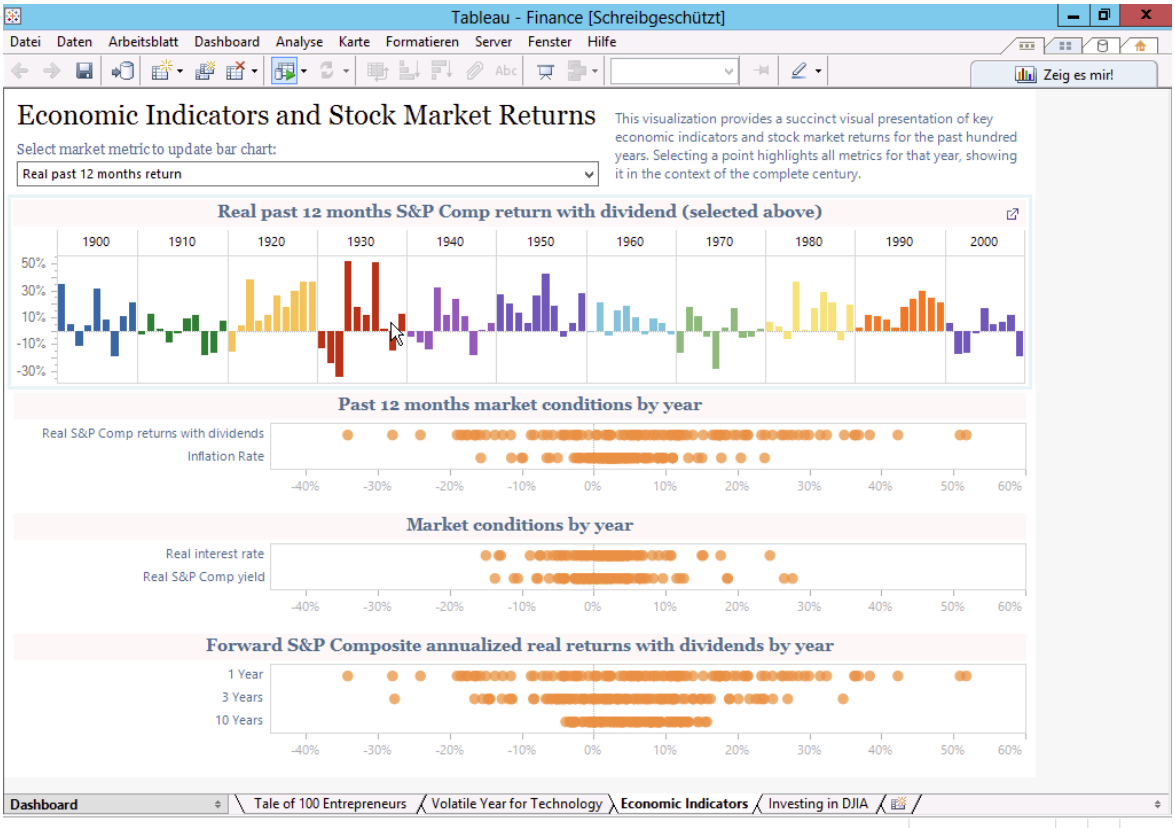

Abbildung 6.16: Beispiel einer Aufbereitung von Zusammenhängen in einem Dashboard

6.3.5 Einschränkungen

Wie auch Spotfire hat Tableau hat keine Möglichkeit textbasierte Visualisierung durchzuführen. Zusätzlich ist es in Tableau von Haus aus nicht möglich Netzwerke in Node-Link-Diagrammen darzustellen.

7 Festgestellte Ergebnisse der Challenges

7.1 VAST-Challenge 2008 - Migrants (Geo)

Bei diesem Datensatz handelt es sich um eine Aufzeichnung der Überfahrtversuche zwischen 2005 und 2007 von der "Isla Del Sueño" in die USA. Der Datensatz enthält Festnahmen von Flüchtlingen durch die Küstenwache der Vereinigten Staaten, sowie Informationen über deren (vermeindlich) illegale Ankünfte an der Küste der USA.

Die Fragestellung bezüglich des Datensatzes lautet[1]:

- Beschreiben Sie die Wahl der Landungsplätze und deren Veränderung während den drei Jahren?
- Beschreiben Sie die geografische Struktur und deren Veränderung der Festnahmen während den drei Jahre?
- Welchen durchschnittlichen Erfolg (Erfolgsrate) haben die Flüchtlinge während den drei Jahre?

Bei dieser Challenge lag der Fokus der Analyse also auf der Untersuchung von geobasierten Daten und der Visualisierung dieser. Außerdem sollten verschiedene zeitliche Verläufe und Muster erkannt werden.

7.1.1 Qlik

Einlesen der Daten

Da der Datensatz zur VAST-Challenge 2009 nur als CSV vorliegt, muss dieser zunächst korrekt in einer Excel-Datei eingelesen werden. Dabei ist insbesondere darauf zu achten, dass die Spalten richtig formatiert sind und beispielsweise nicht Datumsangaben in Zahlen umgewandelt werden. Im Anschluss an die Konvertierung, kann die Excel-Tabelle in Qlik eingelesen werden. Die Spaltenüberschriften werden automatisch erkannt. Um die Koordinaten zu nutzen, müssen diese beim Einlesen und Bearbeiten der Exceltabelle in zwei Spalten aufgeteilt werden. Das kann über das ';' Trennzeichen einfach erreicht werden. Schlussendlich muss in der neu erstellten Qlik-Datei das Trennzeichen innerhalb einer Zahl von '.' auf ',' umgestellt werden, damit die Koordinaten als Komma-Zahl erkannt werden.

[1] vgl. http://www.cs.umd.edu/hcil/VASTchallenge08/tasks.html

Weg zur Lösung und Ergebnis

Die VAST Challenge 2009 bestand aus drei Fragestellungen. Zunächst sollte die Erfolgsrate der Einwanderungen bestimmt werden. Leider bietet Qlik keine Möglichkeit gestapelte Balkendiagramme darzustellen. Daher wurde die Erfolgsrate über berechnete Felder bestimmt. Anhand der Summe von Einwanderungsversuchen, sowie der Summe erfolgreicher Einwanderungen lässt sich der Quotient leicht berechnen. In der folgenden Abbildung sind die dazu verwendeten berechneten Felder in den verschiedenen Markierungen zu sehen. Im Falle der hier verlangten Daten, ist dazu kein Skript notwendig, da die Statistikbox 7.1 bereits alles notwendige bereitstellt.

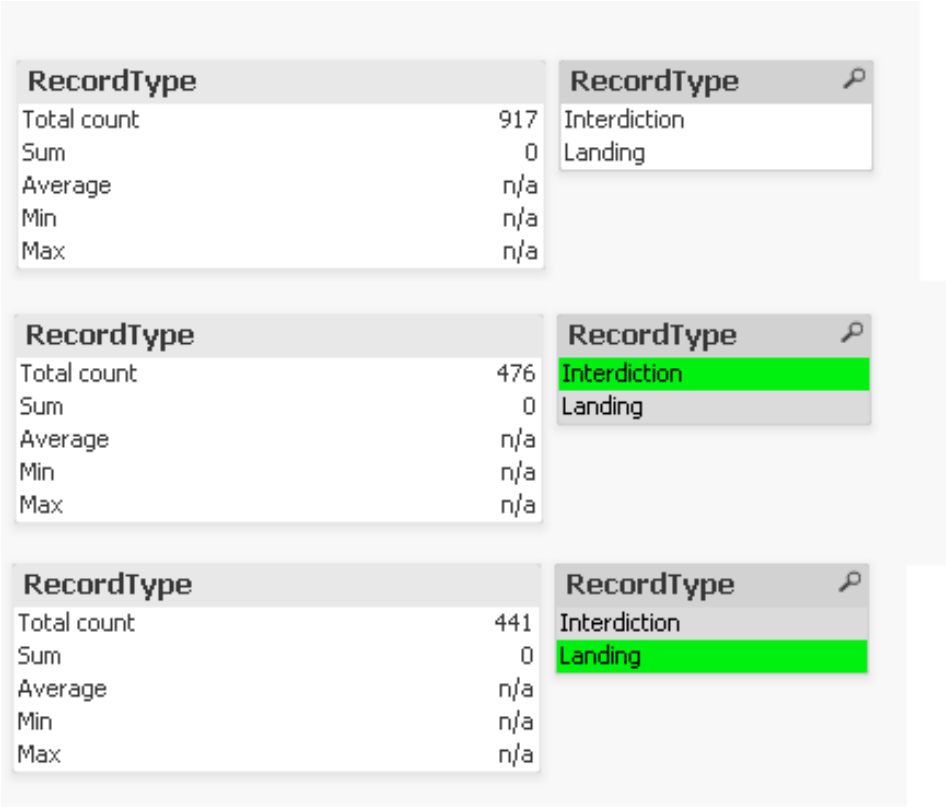

Abbildung 7.1: Qlik - Berechnete Felder zur Landungsrate. Aus der Gesamtzahlt und den jeweils gefilterten Zahlen lässt sich die Erfolgsrate bestimmen.

Aus diesen Feldern berechnet sich nun eine Erfolgsrate von 0.481 Prozent. Da Qlik außerdem die Datumsangaben nicht als Datum erkennt, ist es schwierig diese Wahrscheinlichkeit nach den Jahren zu Filtern. Daher kann nur die gesamte Erfolgswahrscheinlichkeit angegeben werden.

Die zweite und dritte Frage beziehen sich auf die Geo-Analyse. Hierbei sollen die gegebenen Koordinaten der Landestellen auf deren Entwicklung über die Zeit untersucht werden. Dabei

werden zwei Fälle unterschieden. Die erfolgreichen Einwanderungen sowie die abgefangenen Boote.

Die Landungen, welche erfolgreich waren, sind in der Abbildung 7.2 sichtbar. Das Filtern dieser Werte ist in Qlik sehr einfach durch Auswahl des Wertes Landing im Recordtype möglich.

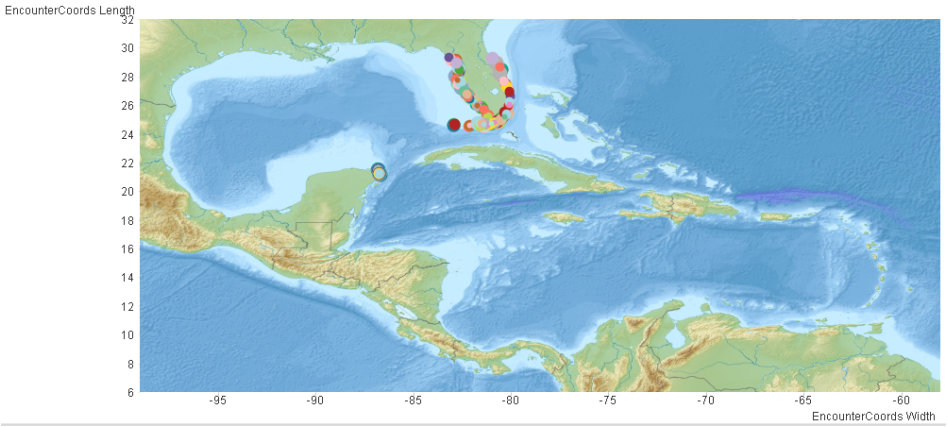

Abbildung 7.2: Qlik - Karte mit eingezeichneten Landepunkten. Jede Koordinate entspricht dabei einer anderen Farbe, um die einzelnen Punkte leichter erkennbar zu machen.

Um nun die Entwicklung über die Jahre zu sehen, muss jedes Datum in der EncounterDate Tabelle einzeln ausgewählt werden. Leider erkennt Qlik das Datum nicht als Datum und kann daher nicht direkt nach der Jahreszahl filtern.

Dabei lässt sich folgende Entwicklung bemerken: Im Jahr 2005 waren die meisten Landungspunkte auf einem kleinen Landstück unterhalb der Halbinselspitze. Zusätzlich landeten einige der Boote an der Spitze der Halbinsel. 2006 wurden diese Ladepunkte erweitert um Punkte seitlich der Halbinsel auf der linken Seite. Auch die alten Zielpunkte wurden beibehalten und gleichermaßen angesteuert. Es gab keine Muster in der zeitlichen Abfolge der Landungspunkte. Zusätzlich zur Landung auf der Halbinsel wurde auch die Spitze des linken Landabschnittes vereinzelt angesteuert. Im Jahr 2007 wurde dieser Zielpunkt auf trotz der weiteren Weges über den Meeresabschnitt zum linken Landabschnitt öfter angesteuert. Zusätzlich wurde die Rechte Hälfte der Halbinsel zu den Zielpunkten hinzugefügt. Allgemein ist 207 eine Zunahme der gleichzeitigen Landings zu bemerken, welche auf eine insgesamt erhöhte Ausreiserate schließen lässt.

Die Landungen, welche abgefangen wurden, sind in der Abbildung 7.3 sichtbar. Auch hier wird gefiltert, diesmal nach den Interdictions, und die Werte der Datumstabelle durchwandert.

Dabei lässt sich folgende Entwicklung bemerken: Im Jahre 2005 wurden viele Boote unterhalb der Halbinselspitze oder rund um den kleinen Landabschnitt links von der Halbinsel abgefangen. Die meisten Punkte sind im Wasser, aber auch auf dem kleinen Landabschnitt

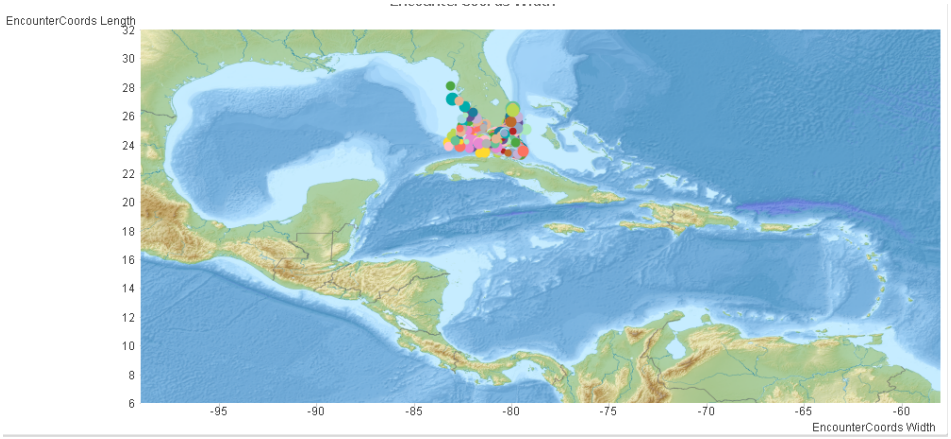

Abbildung 7.3: Qlik - Karte mit eingezeichneten Abfangpunkten

wurden noch Passagiere abgefangen. Zusätzlich wurden einige Schiffe rechts von der Halbinselspitze abgefangen. Da dies nicht den allgemein üblichen Landungszielen entspricht, lässt sich vermuten, dass diese Boote aufgrund eines falschen Kurses abgefangen werden konnten. Im Jahre 2006 werden die Boote immer weiter außen in den umgebenden Meeren abgefangen. Dass lässt vermuten, dass die Küstenwache ihre Kontrollen verschärft hat. Auch kurz vor dem linken Landabschnitt werden nun Boote abgefangen. Diese sind jedoch ebenso wie die dort ankommenden Boote in geringer Anzahl. 2007 werden zusätzlich auf der rechten Seite der Halbinsel sowie unterhalb der Halbinselspitze Boote abgefangen. Ebenso wie die Anzahl der Einwanderungen nimmt 2007 auch die Anzahl der abgefangenen Boote zu.

Hindernisse / Erkenntnisse

In der Geoanalyse gab es kaum Hindernisse bei der Datenverarbeitung. Die Darstellung der Karten als Scatterplot war umständlich aufgrund der manuellen Festlegung der Grenzkoordinaten der Bilddatei. Mit der Erweiterung CloudMadeMaps 7.4 war dies jedoch kein Problem.

Die berechneten Felder sind keine schöne Lösung und es wäre schön gewesen, diese auch gefiltert auszuwerten, statt nur über Markierungen die gewünschten Ergebnisse auszuwählen. Dennoch war dies kein Hindernis um an das Ergebnis der Analyse zu kommen. Eine starke Einschränkung war außerdem das Fehlen der Datumserkennung, um die Landepunkte nach Jahre zu filtern, statt die einzelnen Datumsangaben der Reihe nach durchzugehen.

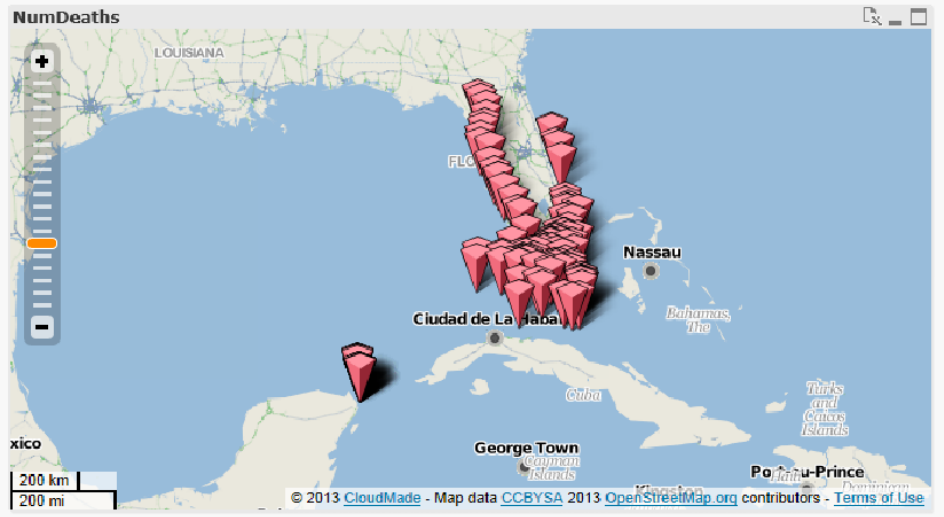

Abbildung 7.4: Qlik - Karte mit eingezeichneten Landepunkten

7.1.2 Spotfire

Einlesen der Daten

Dieses Dateiformat (*CSV*), indem der Datensatz vorliegt, wird bei Spotfire standardmäßig unterstützt und kann sofort ohne weitere Probleme eingelesen werden. Der Datensatz enthält sowohl die GPS-Position der Begegnungen mit der Küstenwache als auch die Start-Koordinaten in einer Spalte.

Um mit diesem Problem umzugehen unterstützt Spotfire sog. berechnete Spalten. Mit einem einfachen Befehl konnte so aus den GPS-Spalten der Längen- und Breitengrad extrahiert werden (`Real(left([EncounterCoords],Find(",",[EncounterCoords]) - 1))`). Für das Anlegen solcher berechneter Spalten gibt es einen eigenen Assistenten der einem neben der Übersicht über die zur Verfügung stehenden Spalten auch die möglichen Funktionen mit einer kurzen Erklärung anzeigt. Mithilfe dieses Assistenten ist es möglich schnell möglich einfache berechnete Spalten hinzuzufügen (siehe Abb. 7.5).

Auf diese Weise wurden die beiden GPS-Positionen in einzelne Spalten aufgeteilt.

Weg zur Lösung und die Ergebnisse

Die drei Fragen, die es bei dieser Challenge zu beantworten galt werden nun nachfolgend beantwortet.

Beginnend stand die durchschnittliche Erfolgsrate der Flüchtlinge im Mittelpunkt. Hierzu wurde der Datensatz mittels eines Balkendiagramms ausgewertet (siehe Abb. 7.6). Die hier aufgetragenen Daten (monatsweise) entsprechen der Anzahl der Passagiere. Die pink

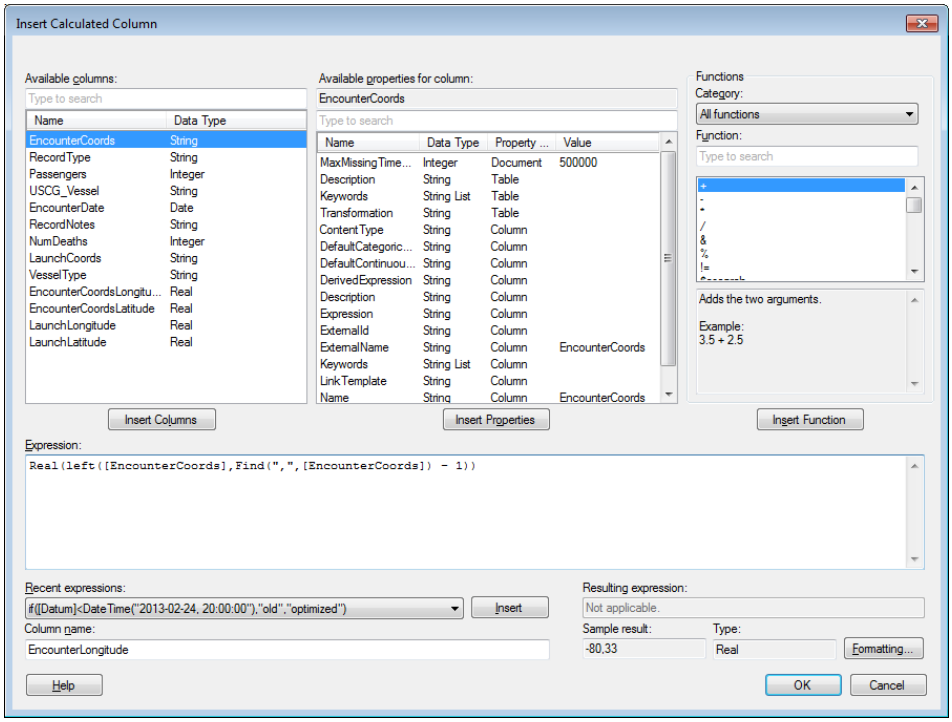

Abbildung 7.5: Spotfire - Migranten - Hinzufügen einer berechneten Spalte im dafür vorgesehenen Wizard.

gefärbten Bereiche sind die Festnahmen durch die Amerikanische Küstenwache. Die blau gefärbten Bereiche entsprechen den Landungen. Wie man gut erkennen kann steigt die Zahl der Flüchtlinge auf das Jahr gesehen kontinuierlich an. Insbesondere in den Sommermonaten (April bis September) ist die Anzahl der Überfahrten besonders hoch. Die abgebildeten Grafiken wurden in Spotfire mittels einfacher Bar-Charts realisiert. Wie bei den meisten Visualisierungen in Spotfire können auch für diese Aggregations-Funktionen (Anzahl der Passagiere) und Gruppierungen (Datum) angewendet werden.

Wichtiger jedoch ist die Erfolgsquote der Passagiere. In der Grafik siehe Abb. 7.7 wurde das oben verwendete Bar-Chart deshalb in eine prozentuale Ansicht transferiert. Leider unterstützt Spotfire es nicht, aus verschiedenen Zeilen eines Datensatzes händisch neue Werte zu berechnen (hier wäre z.B. der Quotient aus den erfolgreichen und misslungenen Überfahrten interessant).

Jedoch unterstützt Spotfire sog. "100% Stacked Bars". Bei diesen werden die aggregierten Werte der Y-Achse auf 100% aufgetragen. So lässt lässt sich der Verlauf der Erfolgsquote gut darstellen. Wie man erkennt ist die Erfolgsquote zu Beginn der Aufzeichnung sehr niedrig. Besonders im Monat Februar (in allen drei Jahren der Aufzeichnung) scheint die erfolgreiche Überfahrt sehr schwer zu sein. Im Jahr 2007 sind die Flüchtlinge bei Ihren Überfahrten am Erfolgreichsten und haben dabei die größte Wahrscheinlichkeit die USA zu erreichen. In den Sommermonaten zeigt sich über die drei Jahre eine klare Tendenz

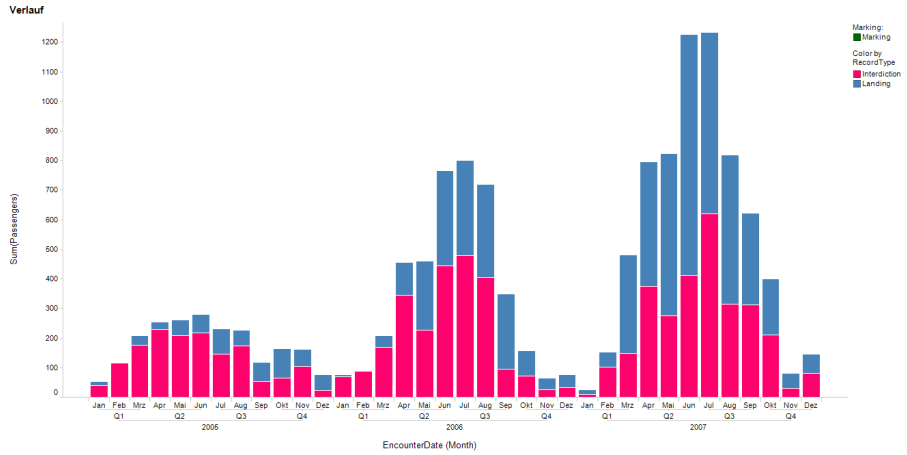

Abbildung 7.6: Spotfire - Migranten - Der zeitliche Verlauf der Passagierzahlen

zum Erfolg. Bemerkenswert ist jedoch, dass in den ersten beiden Jahren mehr als 50% der Flüchtlinge gefangen genommen werden (siehe Abb. 7.7). Erst im letzten Jahr schafften es ca. 60% der Flüchtlinge.

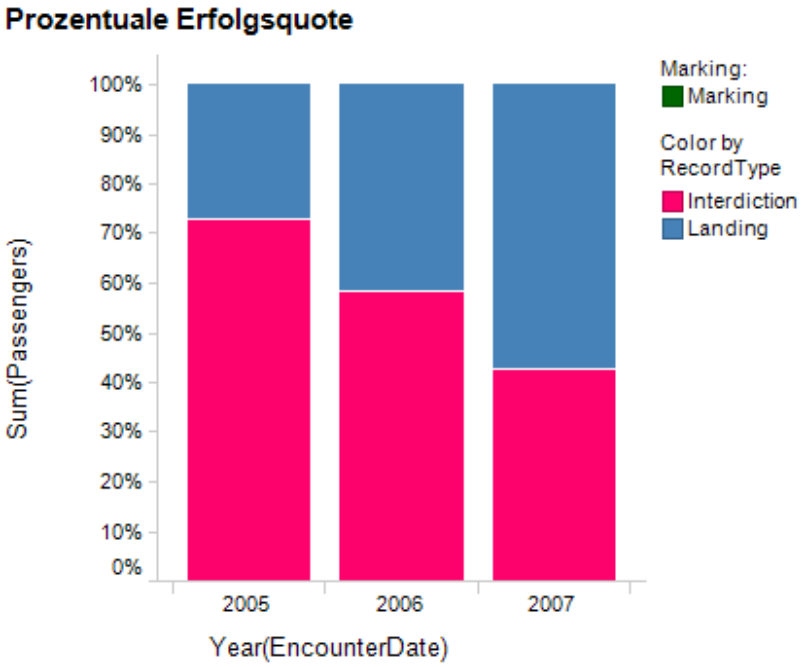

Abbildung 7.7: Spotfire - Migranten - Die Erfolgsquote pro Jahr

Um über den gesamten Datensatz und mögliche Korrelationen einen Überblick zu erhalten wurde außerdem noch ein Parallele-Koordinaten-Diagramm angefertigt (siehe Abb. 7.8). In diesem lassen sich einige Zusammenhänge zwischen den Daten (insbesondere der Anzahl

der Passagiere und des Boots-Typs) erkennen. Diese wurden jedoch für die weitere Analyse nicht benötigt und werden deshalb nicht weiter berücksichtigt.

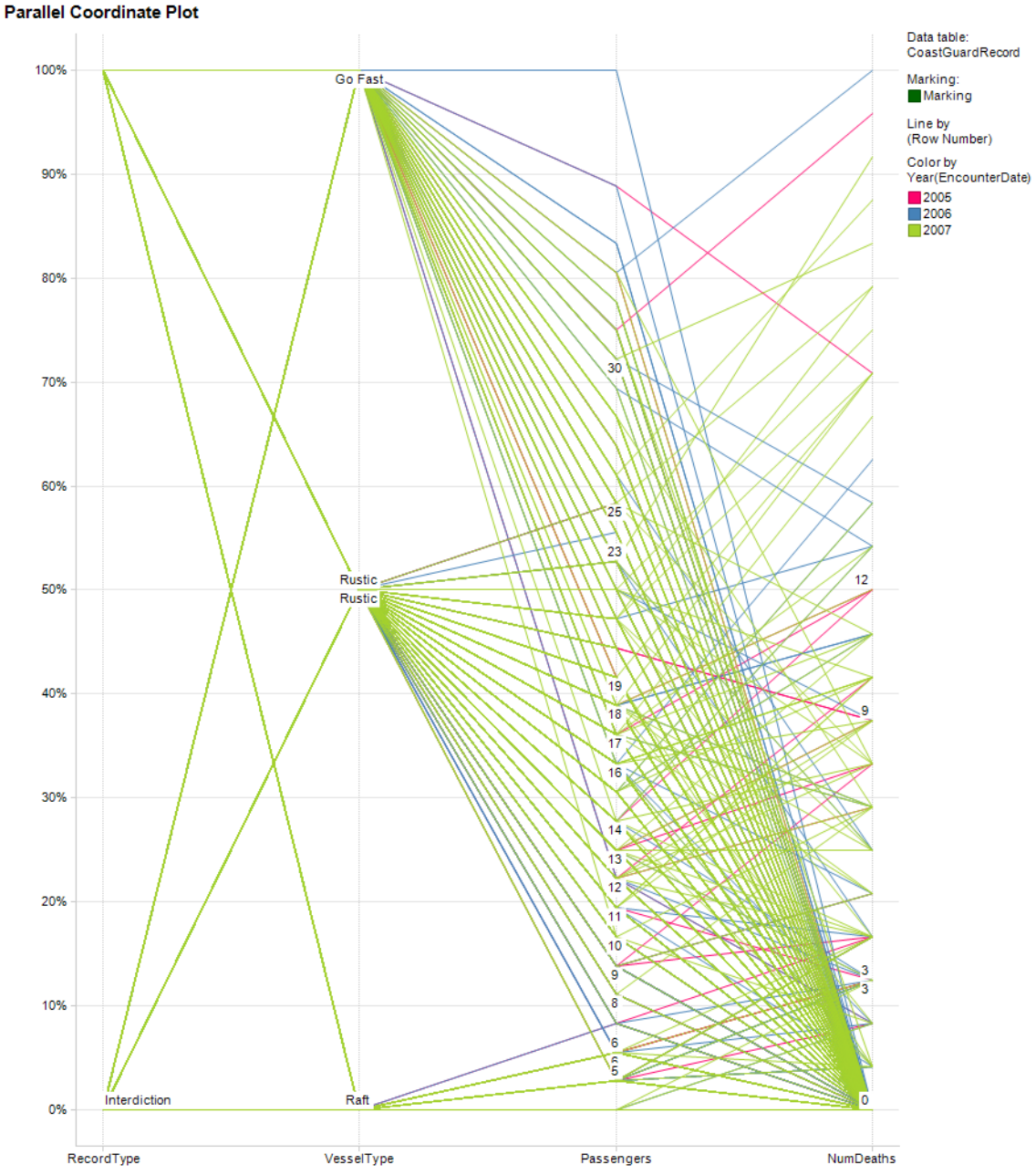

Abbildung 7.8: Spotfire - Migranten - Parallele Koordinaten zur Gewinnung eines Überblickes über den Datensatz und zur Erkennung möglicher Korrelationen

Der zweite Teil diese Challenge war die geografische Struktur der Flüchtlingsbewegung zu analysieren. Hierzu bietet Spotfire eine Visualisierung für Geografische Daten in Form einer

Karte an. Im Prinzip handelt es sich dabei um einen Scatter-Plot der GPS-Koordinaten in eine Ebene zeichnen kann. Als Hintergrund für die Visualisierung kann jedoch Kartenmaterial in Form eines Bilder hinterlegt werden. Spotfire selbst kommt jedoch nicht mit solchem Kartenmaterial einher. Teil des Datensatzes war eine *KMZ*-Datei, welche die "Isla Del Sueño" enthält.

Für eine korrekte Projektion der Karte auf die Map-Darstellung von Spotfire wird jedoch eine sog. Merkator Projektion benötigt. Bei dieser Projektion wird eine winkeltreue Ansicht der Erdoberfläche erstellt. Längen- und Breitengrade werden durch diese Projektion zu rechtwinklig zueinander stehenden Geraden.

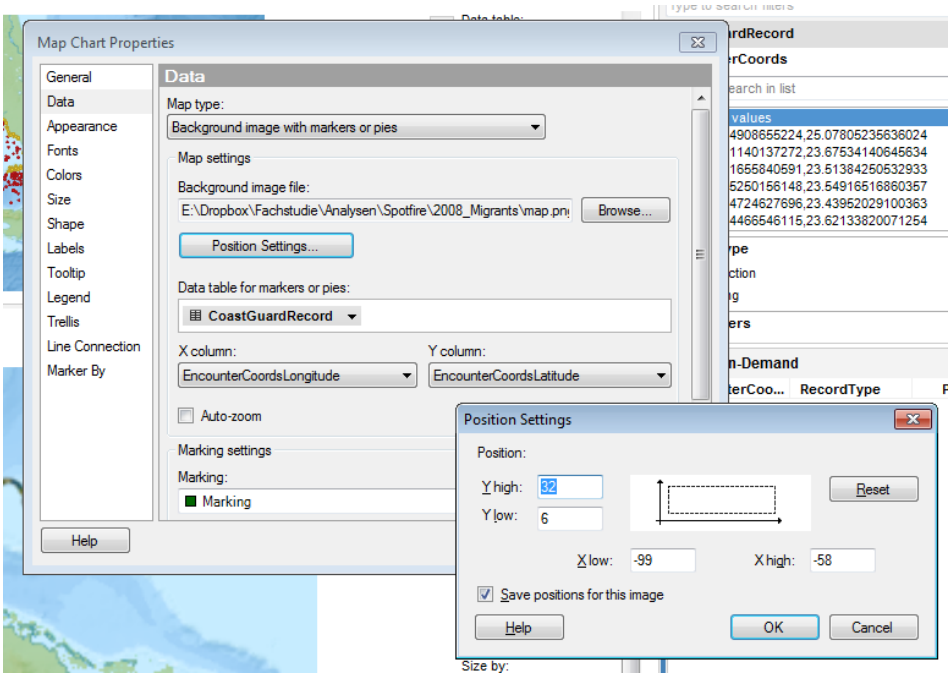

Abbildung 7.9: Konfiguration der Kartendarstellung

Als Grundlage für die Karte wurde hier Kartenmaterial ([Ded10]) aus dem Internet verwendet. Bei dieser Karte sind die Geodaten bereits mit angegeben. Um die "Isla Del Sueño" darzustellen wurde dann die *KMZ*-Datei verwendet. In ihr ist die Insel als Grafikdatei enthalten. Mittels GIMP[2] wurde das Bild entsprechend der Angaben in der *KMZ*-Datei um 73.3 Grad gedreht und an die richtige Position in der Karte gesetzt. Dieses Bild (als Grundlage u.a. für siehe Abb. 7.10) diente der Analyse unter Spotfire als auch der Analyse mit Qlik (siehe Abschnitt 7.1.1 auf Seite 49) als Grundlage. Die Konfiguration der Map-Visualisierung von Spotfire (siehe Abb. 7.9) lässt es nun zu, dass man die Hintergrund-Grafik setzt und die Position der Karte auf dem Globus angibt. Dieser Weg ist leider nicht sehr komfortabel und erfordert einiges an Vorwissen. Außerdem müssen die Koordinaten des Kartenausschnittes

[2] http://www.gimp.org/

bereits bekannt sein. Diese Visualisierung ist also nicht sofort ohne weiteres Zutun auf Daten anwendbar.

Nachdem eine so geschaffene Visualisierung für die Geo-Daten zur Verfügung stand konnte jedoch die Frage nach der geografischen Struktur der Flüchtlingsbewegung rasch beantwortet werden:

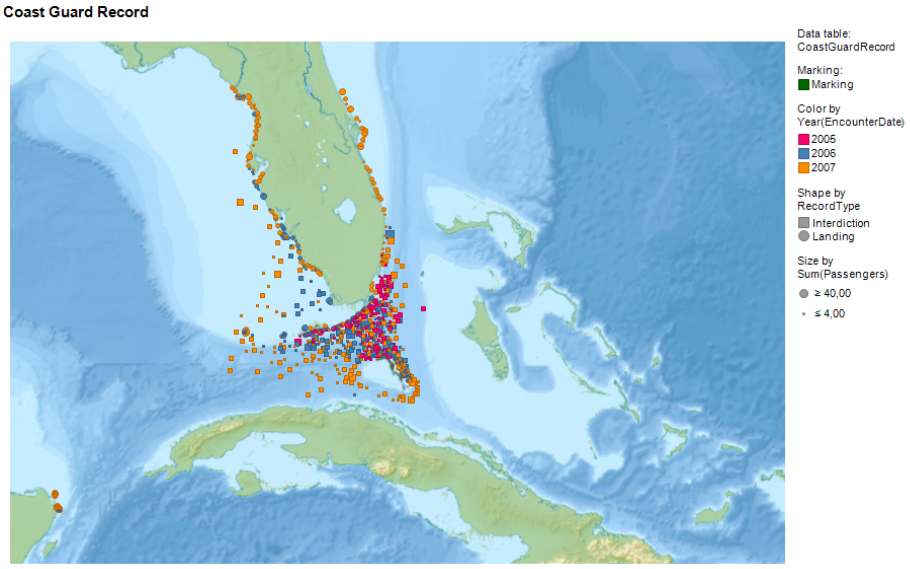

Abbildung 7.10: Der geografische Überblick

Der erste Überblick über den Datensatz (siehe Abb. 7.10) lässt anhand der eingesetzten Farbcodierung für die Jahre bereits eine deutliche Tendenz erkennen. Im Jahr 2005 (rosa) konzentrierten sich die Festnahmen oder erfolgreichen Landungen auf einen trichterförmigen Bereich, ausgehend von der "Isla Del Sueño" in Richtung Nordnordwesten. Die Flüchtlinge hatten alle die Florida Keys als Ziel (siehe Abb. 7.12).

Bereits im Jahr 2006 (blau) verteilten sich die Flüchtlinge über einen sehr viel größeren Bereich und hatten als Ziel bereits das amerikanische Festland von Florida. Um der Küstenwache zu entgehen wurden immer weiter nördlich liegende Ziele, bevorzugt auf der vom Atlantik abgewandten Seite, angestrebt. Überraschend ist hier auch, dass es bereits erste Landungen an der Küste Mexikos (bei Cancún) gab (siehe Abb. 7.12). Anscheinend versuchen die Flüchtlinge hier einen indirekten Weg in die USA zu finden. Im Jahr 2005 (orange) verteilen sich die zunehmende Zahl der Flüchtlinge immer mehr. Sie dringen mit ihren Booten bis auf die Höhe von Jacksonville vor und riskieren auch die deutlich rauere, dem Atlantik zugewandte Seite Floridas um in den USA landen zu können. Um der Küstenwache zu entgehen werden größere Umwege in Kauf genommen. Dies ist an der zunehmenden Anzahl der Festnahmen auf hoher See (siehe Abb. 7.11) zu erkennen. Die Zahl der Landungen in Mexiko steigt ebenfalls.

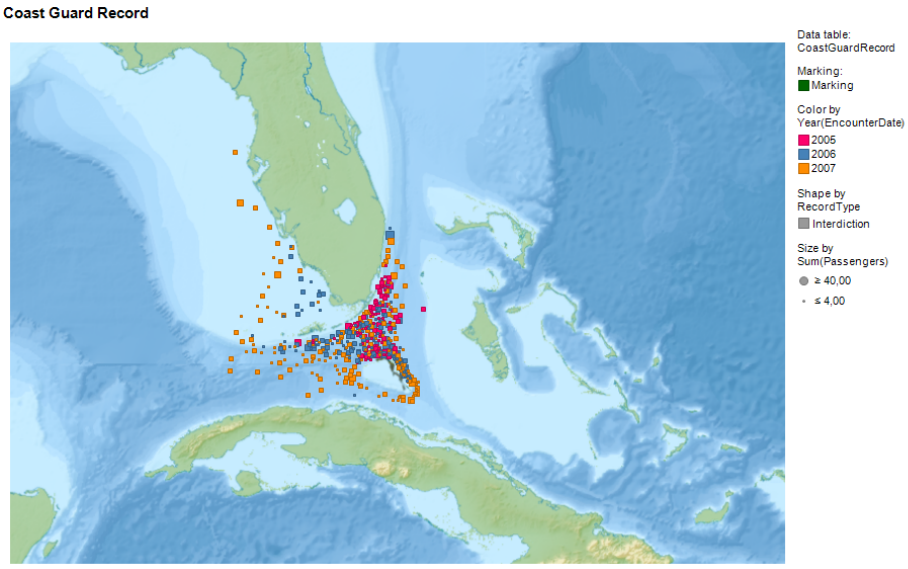

Abbildung 7.11: Der Verlauf der Festnahmen

Abbildung 7.12: Der Verlauf der Landungen

Hindernisse / Erkenntnisse

Spotfires Unterstützung bei der Analyse von Geo-Daten ist nicht wirklich gut. Es gibt ein Darstellung für Geodaten, diese beschränkt sich jedoch auf die um wenige Features (Karte als Hintergrund, Positionierung dieser) erweiterte Visualisierung eines Scatter-Plots. Eine

automatische Integration von Kartendaten anhand der Geodaten wäre durchaus sinnvoll und kann in der heutigen Zeit von Google Maps und OpenStreetMap eigentlich auch nicht zu viel verlangt sein.

Die Interaktion und Filterung der Daten, sowie die Fähigkeit für berechnete Spalten, ermöglichen jedoch schnelle und informative Ergebnisse die durch eine gute und stabile Visualisierung unterstützt werden. Lediglich manche Berechnungen (z.B. Distanz, Erfolgsrate, ...) fehlen und könnten noch integriert werden.

Die Funktion für Spalten benutzerdefinierte Sortierreihenfolgen anzulegen ist insbesondere bei der Verwendung von parallelen Koordinaten ein nützliches Feature, so kann Visual Clutter reduziert werden.

Insgesamt ließ sich die Fragestellung der VAST-Challenge 2008 hiermit jedoch gut beantworten.

7.1.3 Tableau

Einlesen der Daten

Vordem einlesen der csv-Datei müssen die Geo-Koordinaten in Längen- und Breitengrade getrennt werden. Dies funktioniert Mithilfe eines Assistenten, welcher ähnlich zu Spotfire funktioniert (vgl. Unterkapitel Einlesen der Daten des vorherigen Kapitels), oder auch in MS Excel durch die Funktion `Text in Spalte`.

Weg zur Lösung und Ergebnisse

Die VAST-Challenge 2008 dreht sich hauptsächlich um Geo-Daten. Diese zu Analysieren ist dank dem automatischen Zugriffs eingebauten Kartendaten sehr komfortabel. Die Landungsplätze lassen sich auf einer Karte sehr gut darstellen. Wie gefordert soll dabei zeitlich unterschieden werden. Dies ist in Tableau sehr gut möglich in dem man das Datumsfeld in den Seitenfilter zieht. Daraufhin kann man schrittweise durch die Zeitintervalle (Tag, Monat, Quartal oder Jahr) durchgehen. Dies lässt sich auch automatisch in einer Endlosschleife durchführen. Falls gewünscht ist es auch möglich die Änderung zu den Vorperioden eingegraut oder mit einer Verbindung beizubehalten. In dem hier verwendeten Anwendungsfall trägt dieses Feature nicht zur Übersichtlichkeit bei und wurde deshalb deshalb weggelassen.

Nun folgenden die Grafiken die sich mit Tableau aus den Geo-Daten der VAST Challenge 2008 erstellen lassen.

Abbildung 7.13 zeigt die Lösung zu der Frage der Veränderung der Landungen über Zeit. Die Farbe bestimmt dabei wie viele Passagiere der jeweilige Datenpunkt beinhaltet. Je dunkler ein Punkt, desto mehr Passagiere sind betroffen. Die Anzahl schwankt hier dabei zwischen 4 (hell) und 34 (dunkelrot).

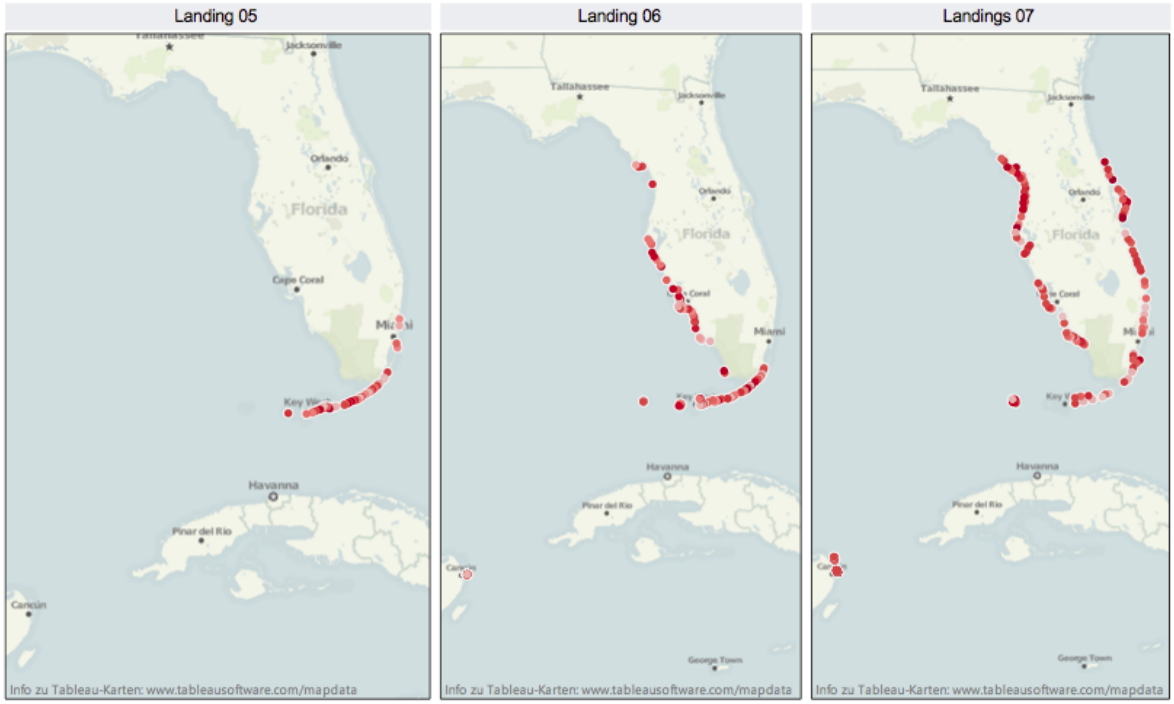

Abbildung 7.13: Tableau - Migranten - Darstellung der Landings nach Jahren

Es ist zu erkennen wie sich die Landings verhalten haben (siehe Abb. 7.13):

1. **2005**: Die Einwanderer konzentrieren sich auf den Süden Floridas. Die größten Boote sind dabei hauptsächlich bei den Florida Keys Inseln gelandet. Vereinzelten, kleinere Boote landeten aber auch am südöstlichen Küstenstreifen von Florida.
2. **2006**: Die Einwanderung steigt um ein vielfaches an. Südlich verhält sie sich wie im Jahr 2005. Jedoch gibt es nun zahlreiche zusätzliche Landungen der gesamten Westküste von Florida entlang. An der südöstlichen Spitze Floridas gab es wendiger Landungen, dafür erste Landungen vor Lancún, Mexiko.
3. **2007**: Wieder steigen die Einwanderungen stark. Die Landungen finden jetzt noch weiter nördlich an der Westküste von Florida statt. Des weiteren ist dem gegenüberliegend jetzt auch ein Großteil der Ostküste von Landungen betroffen. Rund um Key West im Süden gibt es weniger Landungen. Dies könnte mit den verstärkten Interdictions in dieser Region in 2006 zusammenhängen (siehe Abb. 7.14). Die Landungen vor Lancún, Mexiko haben auch weiter zugenommen.

Ähnlich zu der vorherigen Aufgabe, war im nächsten Schritt die Veränderung der Interdictions der Küstenwache zu untersuchen. Die Grafik siehe Abb. 7.14 ist genauso aufgebaut wie siehe Abb. 7.13 und die Farbe bestimmt wieder die Anzahl der Passagiere, welche jeweils von der Küstenwache an den Koordinaten abgefangen wurden. Diese Zahl schwankt hier dabei zwischen 4 (hell) und 40 (dunkelrot).

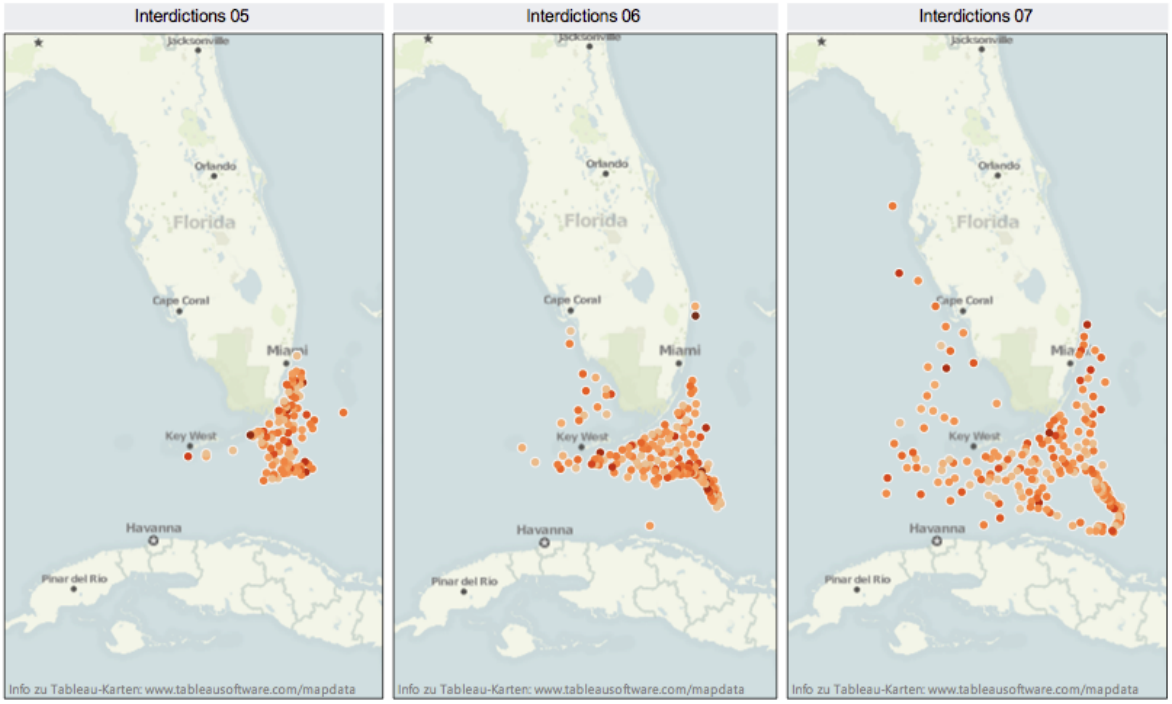

Abbildung 7.14: Tableau - Migranten - Darstellung der Interdictions nach Jahren

Nun lässt sich das Verhalten der Interdictions, also die von der Küstenwache abgefangen Landungsversuche, herausfinden(siehe Abb. 7.14):

1. **2005**: Die Interdictions konzentrieren sich auf den Bereich zwischen dem südlichen Festland Floridas und der südlich davon gelegenen ïmaginärenÏnsel welche als Startposition für die Einwanderer dient.

2. **2006**: Genauso wie die Einwanderung, steigen auch die Interdictions auf ein vielfaches an. Die Küstenwache verzeichnet, zusätzlich zum bisherigen Abfanggebiet, nun auch etwas weiter westlich also südlich von Key West Erfolge. Der Grund liegt wahrscheinlich daran, dass dieses Gebiet im Vorjahr vernachlässigt wurde und viele Landungen zu verzeichnen hatte. Die Interdictions verlaufen auch etwas ausgedehnter der Startinsel und der Westküste Floridas entlang.

3. **2007**: Die Interdiciton nehmen noch weiter zu. Der Trend der Verlagerung in Richtung Isla nimmt weiter zu und er sind Interdictions um die gesamte Insel zu verzeichnen. Die Erfolge werden außerdem weiter von der Küste entfernt verzeichnet. Vor allem die Interdictions im Osten des Golf lassen darauf schließen, dass auch Boote abgefangen wurde, die nach Mexiko unterwegs waren. Der Landungen entsprechend (siehe Abb. 7.13), sind 2007 auch viel Aktivitäten weiter im Norden des Festlandes von Florida zu beobachten.

Abschließend war in der Challenge nach der Erfolgsquote der Migranten gefragt. Diese lässt sich in dem folgenden Bereichsdiagramm siehe Abb. 7.15 erkennen. Blau sind ist die Anzahl der Interdictions und Orange die Anzahl der Landings.

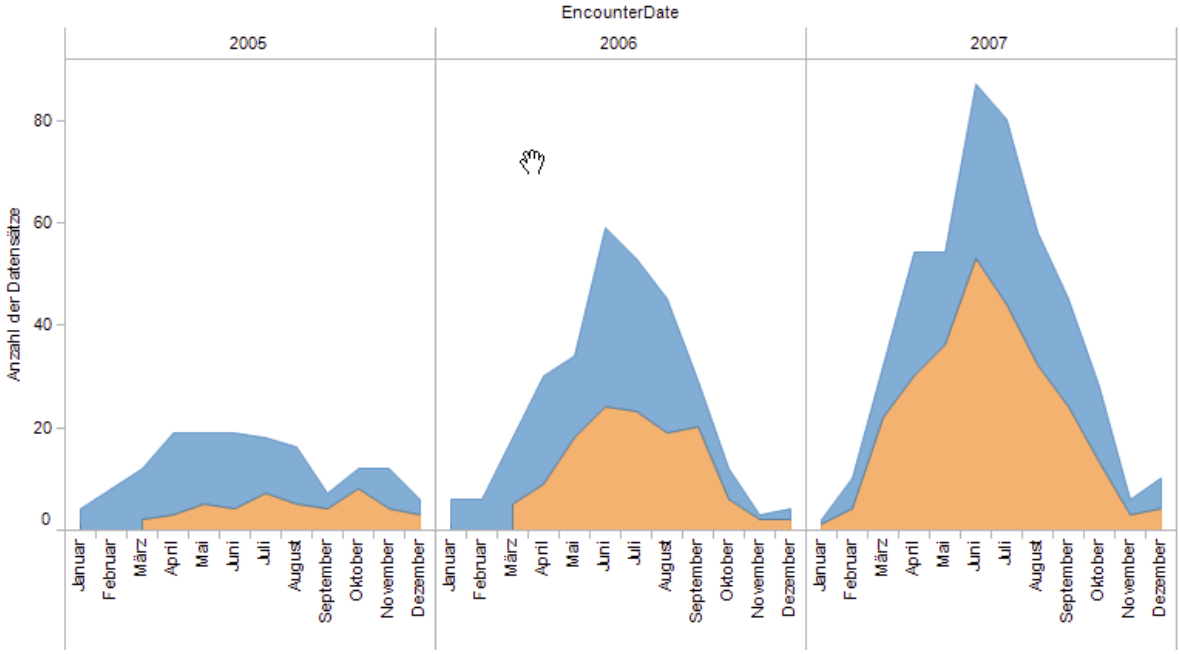

Abbildung 7.15: Tableau - Migranten - Darstellung Anzahl Landings (orange) und Interdictions (blau) nach Jahren

Die gefragte Erfolgsquote der Einwanderer liegt bei **48,1%**. Dies lässt sich ermitteln indem man ein Kuchendiagramm mit `Record Type` als Inhalt wählt und dann unter `Arbeitblatt - Zusammenfassung anzeigen` die statistischen Kennzahlen einsieht. Diese Zusammenfassung ist sehr hilfreich um schnell einen Überblick über den Datensatz zu erhalten.
In dem benannten Bereichsdiagramm lässt sich erkennen, dass die Einwanderung von Jahr zu Jahr um ein Vielfaches zunimmt (siehe Abb. 7.15). Im Sommer sind die Hauptaktivitäten zu verzeichnen, wobei der Juni den Höhepunkt bildet. Während 2005 die Verteilung fast gleichmäßig war bildet sich bis 2007 diese sehr starke Spitze heraus. Sowohl bei den Landings (orange) also auch bei den Inderdictions (blau). Von November bis Februar setzt die Einwanderung aber in jedem Jahr fast vollständig aus.

Hindernisse / Erkenntnisse

Bei dieser Challenge wird hauptsächlich mit Geo-Daten gearbeitet. In diesem Bereich spielt Tableau seine Stärken voll aus: Das Arbeiten mit der Karte ist einfach und fehlerresistent, da das Mapping von Tableau übernommen wird. Der abspielbare Filter ermöglicht es einfach zwischen Jahren zu wechseln und die Veränderungen zuermittlen. Auch das Ablesen der Erfolgsquote ist durch den Einsatz der automatisch berechneten Zusammenfassung sehr

einfach. Einziges, aber einfach zu überwindendes Hindernis, war der Import der Daten, da die Breiten- und Längengrade getrennt werden müssen. Dies würde zu Beginn des Kapitels beschreiben.

7.2 VAST-Challenge 2008 - Wiki Edits (Textanalyse)

Beim zweiten Teil der VAST-Challenge 2008 geht es um die Analyse von Bearbeitungen einer (fiktiven) Wikipedia-Seite zur "Paraiso Bewegung". Diese Bewegung hat einen großen sozialen Einfluss auf diese Region der Welt und wird auch für die wachsenden Flüchtlingszahlen (siehe Abschnitt 7.1 auf Seite 49) verantwortlich gemacht. Es liegt ein Teil der Bearbeitungsvorgänge der zugehörigen Wikipedia-Seite als Datensatz vor.

Durch die Verwendung von Visual-Analytic-Tools soll die soziale Struktur der Bearbeiter und deren Zugehörigkeit zu verschiedenen Gruppen analysiert werden.

Die Fragestellung bezüglich des Datensatzes lautet[3]

- Welche verschiedenen Fraktionen editieren den Eintrag und wer sind ihre Mitglieder? Beschreiben Sie die Gruppen und ihre Mitglieder aufgrund der Bearbeitungsvorgänge.
- Ist die ”Paraiso Bewegung” in gewalttätige Handlungen verstrickt?

Der Fokus dieser Challenge lag auf der Analyse der gegebenen Bearbeitungsvorgänge und der Extraktion von textuellen Attributen aus den Bearbeitungs-Kommentaren. Außerdem sollte aus den so extrahierten Daten ein Netzwerk der Mitglieder erstellt werden um Gruppen ausfindig machen zu können.

7.2.1 Qlik

Einlesen der Daten

Das Einlesen der Daten des Wikipedia Challenges war insofern schwierig, dass Qlik kein Einlesen von Textdaten unterstützt. Daher beschränkt sich die Analyse hauptsächlich auf die Paraiso Edits Datei, welche über Trennzeichen zu einer Exceldatei gemacht werden muss. Beim Verwenden dieser Datei stellt sich jedoch schnell heraus, dass die wichtigen Informationen zu den Einträgen im Textbereich stehen und daher durch einfaches Auslagern in eine Spalte keine Analyse möglich ist. Daher muss die Datei manuell in eine Form umgewandelt werden, welche die Stichpunkte bereits als Tabellenspalten enthält. Dabei kann der Kommentar nach Stichworten wie ’Undid’, ’Revertet’ und ’to revision of’ gefiltert werden. Dadurch ergibt sich eine Datentabelle, die es ermöglicht grobe Gruppierungen herauszulesen. Qlik bietet außerdem zur Lösung des Textteiles eine Erweiterung namens WordMap, welche eine TagCloud aus eingegebenem Text erstellt. Hierbei muss jedoch von

[3] vgl. http://www.cs.umd.edu/hcil/VASTchallenge08/tasks.html

einem Einlesen des Datensatzes unterschieden werden, da nicht Qlik den Text verarbeitet, sondern in der Erweiterung selbst ein Textfeld zu Eingabe bereitsteht. Dies ist im Sinne der Analyse nur von geringem Vorteil, da beide Gruppierungen einen sehr ähnliche Wortschatz verwenden, doch es wurde genutzt um einen Überblick über die gegebene Wikipedia-Seite an sich zu gewinnen.

Weg zur Lösung und Ergebnisse

Zunächst galt es einen Überblick über das Thema zu erreichen. Dies wurde in Qlik über die TagCloud des Wiki-Artikels gelöst. In der folgenden Abbildung ist eine solche TagCloud mit der Beschränkung auf die 100 häufigsten Wörter abgebildet.

Abbildung 7.16: Qlik - Wiki Daten - Meiste Wörter in Artikel

Leider lässt sich durch diese Art von Darstellung nur der Inhalt und nicht die emotionale oder gewalttätige Stimmung ableiten. Auch bei einer Analyse der Kommentare auf diese Art, kam ein sehr einheitlich verwendetes Vokabular zum Vorschein. Daher kann mithilfe von

Qlik keine weitere Textanalyse geboten werden. Die versprochene Kooperation mit Attivio kann bisher nicht selbst genutzt sondern nur ein einer Beispieldatei getestet werden.

Die Art der Gruppierungen ist anhand der selbst Erstellten Tabelle über die Aktionen in den Paraiso Edits möglich. In der Abbildung 7.17 ist zu sehen, wie in einem Balkendiagramm über die Anzahl an Aktionen pro Mitglied die aktivsten Mitglieder ausgewählt wurden. In den in der Abbildung unten liegenden Balkendiagrammen ist zeitgleich zu sehen, wessen Beiträge von dieser Nutzergruppe editiert oder gelöscht wurden.

Abbildung 7.17: Qlik - Wiki Daten - Aktive Nutzer

Durch einzelnes Auswählen dieser aktiven Nutzer, kann verglichen werden, wer wessen Beiträge editiert/löscht und zu welchen Beiträgen zurückgesprungen wird. Dabei entstehen 3 Gruppierungen. Die Gruppierung der Befürworter des Paraiso Movements, die Gegner und eine Gruppe an Teilnehmern, die hauptsächlich Einträge löschen, jedoch auf beide Gruppen bezogen aktiv sind.

Die erste Gruppe gründet sich um Victoria V als aktivste Nutzerin. Die anderen Teilhaber dieser Gruppe sind schwer zu definieren, da es sich um Einzelbeiträge handelt und selten um sehr aktive Nutzer.

Die zweiter Gruppe identifiziert sich über die Beziehung zu Victoria, indem diese Leute oftmals deren Beiträge löschen und umgekehrt auch oftmals von ihr editiert werden. Diese Gruppe enthält die Nutzer Edemir, Augustin, Rm99 und Dailos Salamanca als aktivste Nutzer. Schön zu sehen ist dies an Abbildung7.18, in welcher deutlich wird, dass VictoriaV immer wieder die Versionen dieser 3 Nutzer rückgängig gemacht oder gelöscht hat.

Die dritte Gruppe wird geführt von BakBot und enthält unter den aktiven Nutzern Amado und Sara. Diese Gruppe wird unter Einbezug der weniger aktiven Nutzern die größte sein, da sie auch die unbeteiligten Nutzer oder diejenigen Nutzer, welche Rechtschreibung korrigieren enthält. Schön zu sehen ist dies an Abbildung7.19, in welcher deutlich wird, dass BakBot viele Änderungen macht, diese aber nie zweimal auf einen Nutzer bezieht. insgesamt sind die betroffenen Nutzer eher unwichtig und dafür vielzählig.

Anhand der Verteilung von Edits kann nicht absolut festgestellt werden, welche der Gruppen 1 und 2 die Befürworter und welche die Gegner des Paraiso Movements sind. Anhand der Art

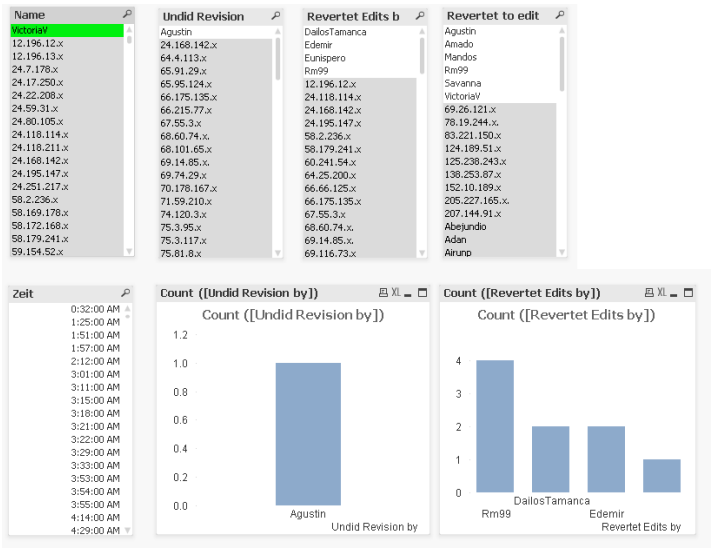

Abbildung 7.18: Qlik - Wiki Daten - Daten über VictoriaV. Hier ist deutlich zu sehen, dass Victoriav die Beiträge von Augustin gelöscht hat, sowie die Versionen von Edemir, Rm99 und Dailos Salamanca rückgängig gemacht hat

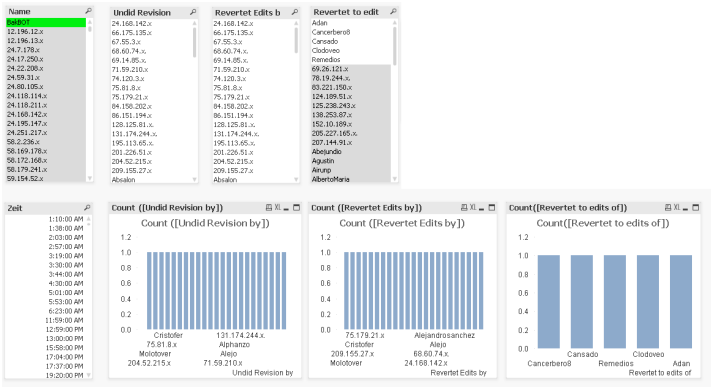

Abbildung 7.19: Qlik - Wiki Daten - Daten über BakBot. Hier ist deutlich zu sehen, dass Bakbot die Beiträge von vielen verschiedenen, aber eher unwichtigen Nutzern gelöscht oder rückgängig gemacht hat. Er hat dabei keinen Nutzer zweimal bearbeitet.

von Gruppierung lässt sich jedoch vermuten, dass die Gruppe von mehreren aktiven Nutzern um Edemir das Paraiso Movement unterstützen und die Gruppe an vereinzelten Beiträgen um die sehr aktive Victoria V sich gegen die gewaltbereite Einstellung der Befürworter wehrt.

Hindernisse / Erkenntnisse

Ein großes Hindernis in der Analyse des Wiki Challenges war die felhlende Möglichkeit die Datensätze einzulesen. Da keine Texte verarbeitet werden konnten, fehlen bei der Analyse wichtige Informationen über die Einstellung der Nutzer und die Arte der Wortwahl. Auch die nun analysierten Daten konnten nur durch aufwändige Vorarbeit an der Edits Datei erreicht werden. Das Tool hat in diesem Fall nur einen geringen Beitrag zum Erfolg des Challenges beigetragen.

7.2.2 Spotfire

Die Analyse der Wiki-Edits der Paraiso Bewegung stellte für Spotfire die größte Herausforderung da. Unstrukturierte Daten kann Spotfire nicht einlesen. Selbst wenn die Daten vorher mit anderen Programmen aufbereitet werden, tut sich Spotfire mit diesen Daten schwer.

Einlesen der Daten

Um die Daten überhaupt analysieren zu können wurde das zur Verfügung gestellte Bearbeitungsprotokoll mittels Excel in eine tabellarische Struktur gebracht. Dabei handelt es sich um eine Arbeitsmappe aus drei Tabellenblättern. Das Tabellenblatt *Edits* hat folgende Struktur:

- Name
- Name ID
- Aktion
- Talk
- contribs
- m
- Größe
- DateTime
- Undid Revision by
- Revertet Edits by
- Revertet to edits of
- Using

Das Tabellenblatt *Node* besteht aus den Spalten

- Name ID
- Name

Das Tabellenblatt *Links* besteht aus den Spalten

- ID1
- ID2
- Action

Spotfire benötigt diese Aufteilung um für die Knoten-Kanten-Diagramme eine vernünftige Datenbasis zu haben. Nur so können Beziehungen richtig dargestellt werden. Um diese Daten

zu generieren wurden verschiedene Excel-Befehle (wie z.B. `SVERWEIS`,`VERKETTEN`, `LINKS` und `FINDEN`) verwendet. Die genaue Zerlegung der Daten wird hier nicht weiter behandelt.

Weg zur Lösung und Ergebnisse

Die so generierten Daten konnten dann als drei Datenquellen in Spotfire geladen werden und dort über die Relationen miteinander verknüpft werden (siehe Abb. 7.20).

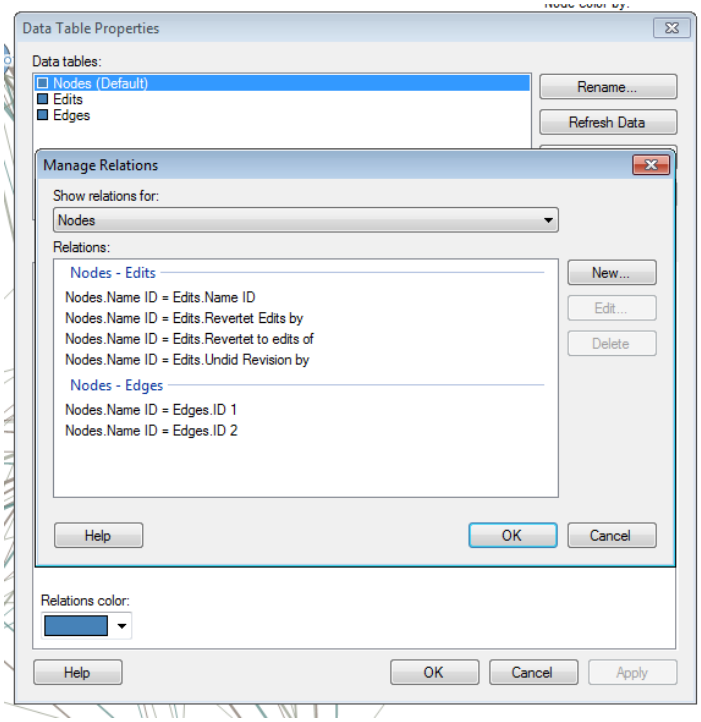

Abbildung 7.20: Herstellen der Relation zwischen den drei Tabellenblättern

Während der Analyse wurde jedoch festgestellt, dass diese Relationen nicht immer so funktionieren wie man sich das vorstellt. Hin und wieder werden diese einfach nicht berücksichtigt. Die einzige Visualisierung die mit den Relationen wirklich gut klar zu kommen scheint ist die Knoten-Kanten-Darstellung.

Um einen ersten Überblick über die Aktivitäten am Paraiso-Manifest zu erhalten wurden die Aktivitäten der einzelnen Personen zunächst in einer Bar-Chart visualisiert (siehe Abb. 7.21).

Da stark davon ausgegangen werden kann, dass es Befürworter und Gegner des Paraiso-Manifestes gibt sind diese enthusiastisch partizipierenden Bearbeiter der erste Einstiegspunkt für eine Analyse. Bei diesen Schritten halfen die vielen kleinen Features von Spotfire, wie z.B. das Sortieren von Bar Charts, schnelle Filterung der Daten durch Achsen-Selektionen (Slider am Rand) und die Aggregation von Daten.

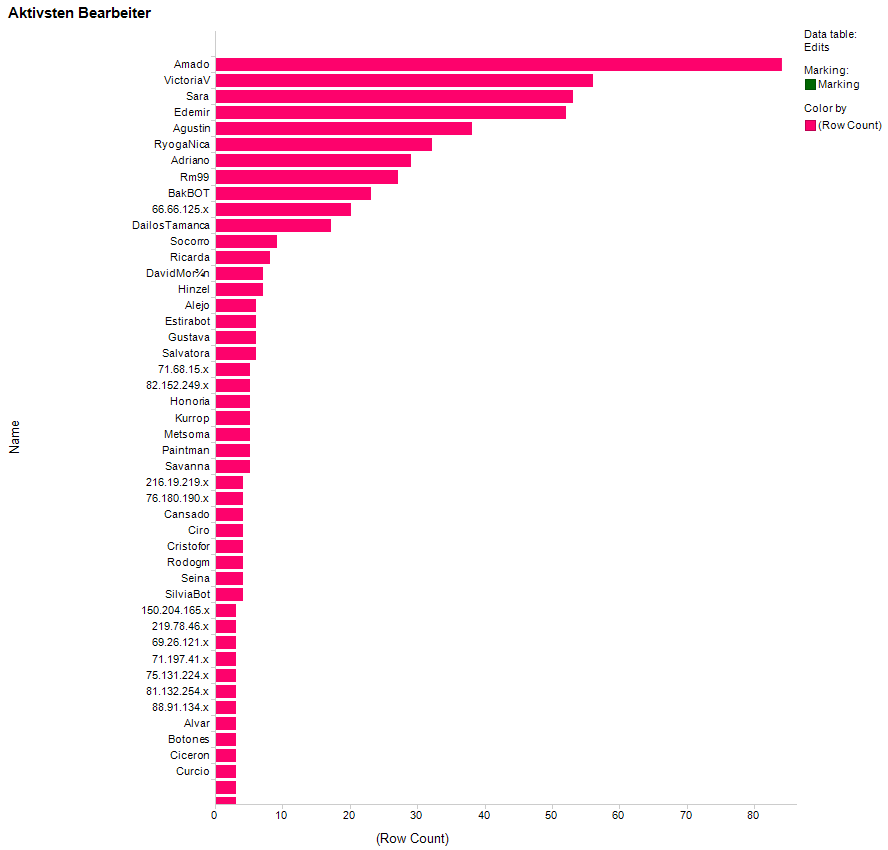

Abbildung 7.21: Aktivste Bearbeiter des Paraiso Manifests

Da es in erster Linie um die Verbindung zwischen den einzelnen Bearbeitern ging wurde in einem zweiten Schritt eine Aufstellung der Aktionen auf andere Versionen des Paraiso Manifests pro User durchgeführt (siehe Abb. 7.22).

Besonders interessant sind hierbei die Nutzer BakBOT, Edemir, VictoriaV, Sara und Agustin. Diese tauchen unter den Top 10 in beiden Anordnungen auf. Sie scheinen am aktivsten zu sein.

Anschließend wurden die Daten mit einem Knoten-Kanten-Diagramm visualisiert. Hier kam die Aufteilung der Daten in die drei Tabellenblätter zum Einsatz. Nach den ersten erfolglosen Versuchen mittels eines Layouts und geschickter Anordnung zu einem guten Ergebnis zu kommen wurde ein kleiner Trick angewandt: Spotfire bietet eine Gruppierung der Knoten an. Diese kann auch anhand berechneter Spalten stattfinden. Das Knoten-Kanten-Diagramm bietet eine Berechnung des Grades eines Knotens an. Da die relevanten oben genannten Knoten einen unterschiedlichen Grad aufwiesen wurde der Graph danach gruppiert. Das Ergebnis (siehe Abb. 7.23) ermöglicht bereits einen guten Einblick in die Abhängigkeiten der Bearbeiter. Neben den bereits entdeckten Bearbeitern fällt hier noch zusätzlich Rm99 ins Auge. Dieser scheint große Meinungsverschiedenheiten mit VictoriaV zu haben.

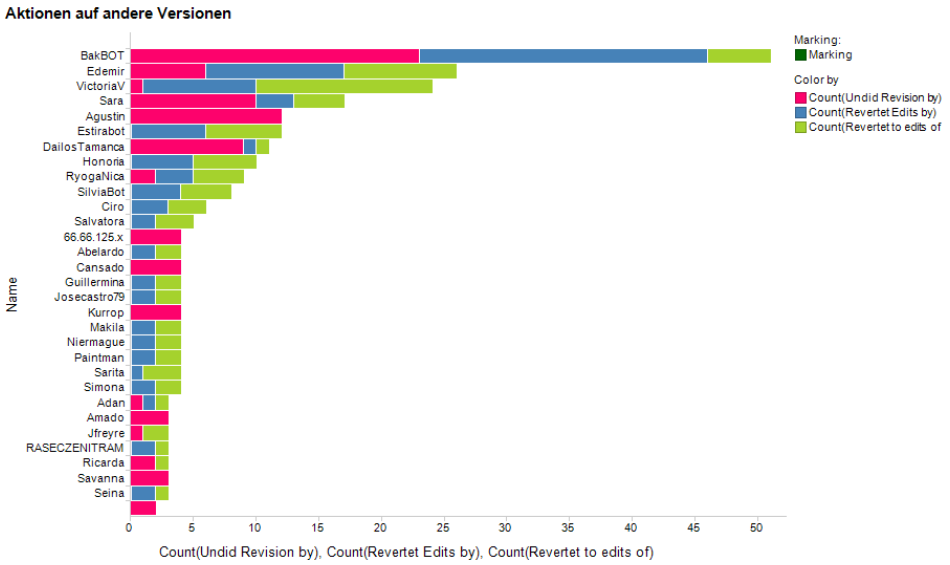

Abbildung 7.22: Die Bearbeiter die Aktionen auf Versionen von anderen Bearbeitern durchgeführt haben

Die drei Bearbeiter Edemir, Agustin und Rm99 (Gruppe 1) revidieren nie Änderungen der anderen Gruppenmitglieder. Jedoch boykottieren sie die Änderungen von den beiden Bearbeitern VictoriaV und Sara (Gruppe 2). Da die beiden Bots Estirabot und BakBot regelmäßig die Änderungen der Gruppe 1 revidieren scheint es so, als ob diese Gruppe Änderungen durchführt, die kontrovers sind. Es liegt also nahe zu vermuten, dass diese Gruppe 1 (ohne die Bots) Gegner der Paraiso Bewegung sind.

Der Bearbeiter Cansado könnte aufgrund seiner fehlenden Interaktion mit Sara und VictoriaV und seines regelmäßigen Einflusses auf Edemir, Agustin und Rm99 ebenfalls zu der Befürworter-Gruppe gehören. Dieser Anhaltspunkt ist jedoch nur sehr vage formulierbar.

Neben diesen bereits erkannten Gruppen (Gegner, Befürworter und den Bots) gibt es noch eine Ansammlung von unbeteiligten Personen, die den Eintrag offensichtlich nur editieren und korrigieren.

Um die Frage nach den gewalttätigen Handlungen zu beantworten müssen nun zusätzlich noch die Kommentare der Bearbeitungen zu Rate gezogen werden.

Hier wurden die Kommentare einfach nach Begriffen wie ”Health”, ”Blood”, ”Gun”, ”Shot” und ”Fight” durchsucht (siehe Abb. 7.24). In dem Datensatz gibt es offenbar acht Treffer für Gewalt. Unter anderem auch Kommentare von Edemir und Sara. Es lässt sich vermuten, dass es zu gewalttätigen Auseinandersetzungen kam.

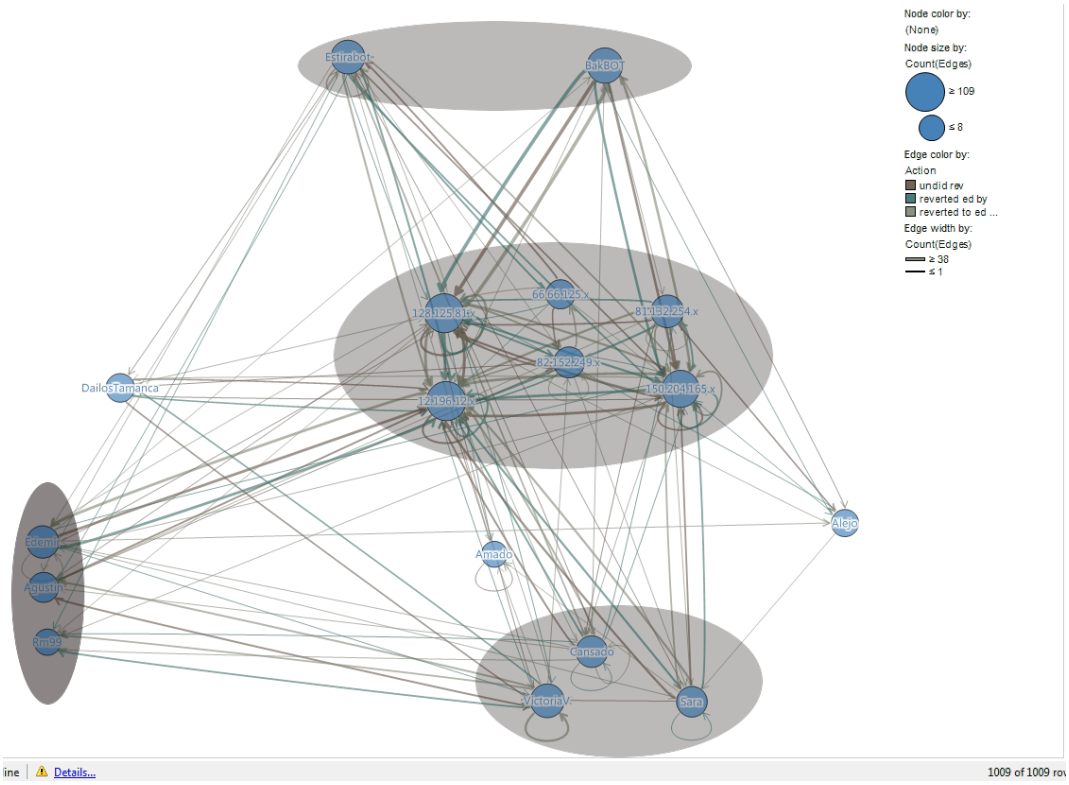

Abbildung 7.23: Der gruppierte Knoten-Kanten-Graph

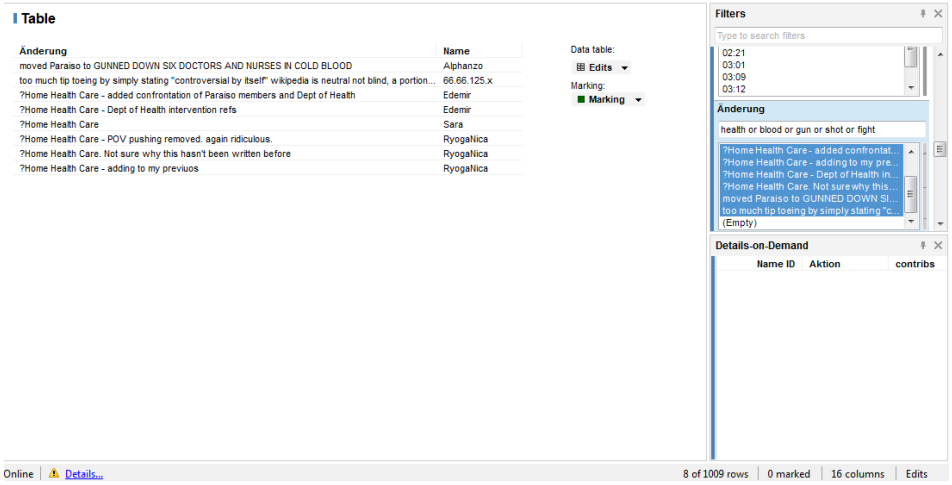

Abbildung 7.24: Suche nach gewalttätigen Begriffen im Datensatz

Hindernisse / Erkenntnisse

Die Aufteilung die notwendig war um die Daten für das Knoten-Kanten-Diagramm verwenden zu können war bei der weiteren Analyse der Daten eher hinderlich. Keine der detailierten Visualisierungen konnte mit einer referenzierten Datenquelle umgehen. Hier scheint es so, als ob Spotfire eine Lücke aufweist. Darüber hinaus ist die Analyse der Texte nur schwer mit Spotfire möglich. Eine vorherige Zerlegung der Bearbeitungsprotokolle mittels Excel ist nicht wirklich intuitiv und sollte bei der Analyse eher vermieden werden.

Die Visualisierung als Knoten-Kanten-Diagramm hingegen hat die Analyse der Daten wieder ungemein unterstützt. Insbesondere auch die Gruppierung der Daten nach ihrem Grad der Vernetzung ermöglichte eine Darstellung des Datensatzes, die alle unwichtigen Bearbeiter versteckte und so die wichtigsten Personen in den Vordergrund rückte.

Die Textsuche in Spotfire verträgt sich mit boolschen-Ausdrücken (so wie z.B. "text1 or text2"), was beim Suchen in den Kommentaren hilfreich war.

7.2.3 Tableau

Einlesen der Daten

Das Tableau nicht für die Textanalyse geeignet ist merkt man daran, dass es schon bei dem Import nichts mit den von der VAST Challenge zur Verfügung gestellten Tabellen anfangen kann. Um wenigsten die erste der gestellten Fragen beantworten zu können, wird hier die vorbearbeitete Tabelle importiert, wie bereits im Kapitel zum Einlesen von Daten bei Spotfire beschrieben.

Weg zur Lösung und Ergebnisse

Um herauszufinden, welche Fraktionen es gibt lässt sich ein Dashboard erstellen, von welchem aus die einzelnen Benutzer durchgegangen werden können, Abbildung 7.25.

Es gibt dabei 3 wichtige Elemente in diesem Dashboard (Abbildung 7.25):

- **Tablelle** `reverted from`: Zeigt an welcher Benutzer (Spalte **name**) wie viele Eintrage eines anderen Benutzers (Spalte **Revertet Edits by**) geändert hat. Es ist somit ein Indikator gegen welche Fraktion ein Benutzer arbeitet.
- **Tablelle** `reverted to`: Zeigt an welcher Benutzer (Spalte **name**) wie viele Eintrage in die Eintrage eines andere Benutzers (Spalte **Revertet Edits by**) zurückgeändert hat. Es ist somit ein Indikator für welche Fraktion ein Benutzer arbeitet.
- **Filterbereich** `name` (rechte Spalte unten): Mit diesem Bereich kann die Auswahl auf bestimmte Benutzer oder mehrer Benutzer (z.B. alle einer Fraktion) eingeschränkt werden. Dadurch können die Aktivitäten einer einzelnen Fraktion genauer untersucht werden. Tableau bietet bei dem Ausschließen ein Freitext-Suchfeld an.

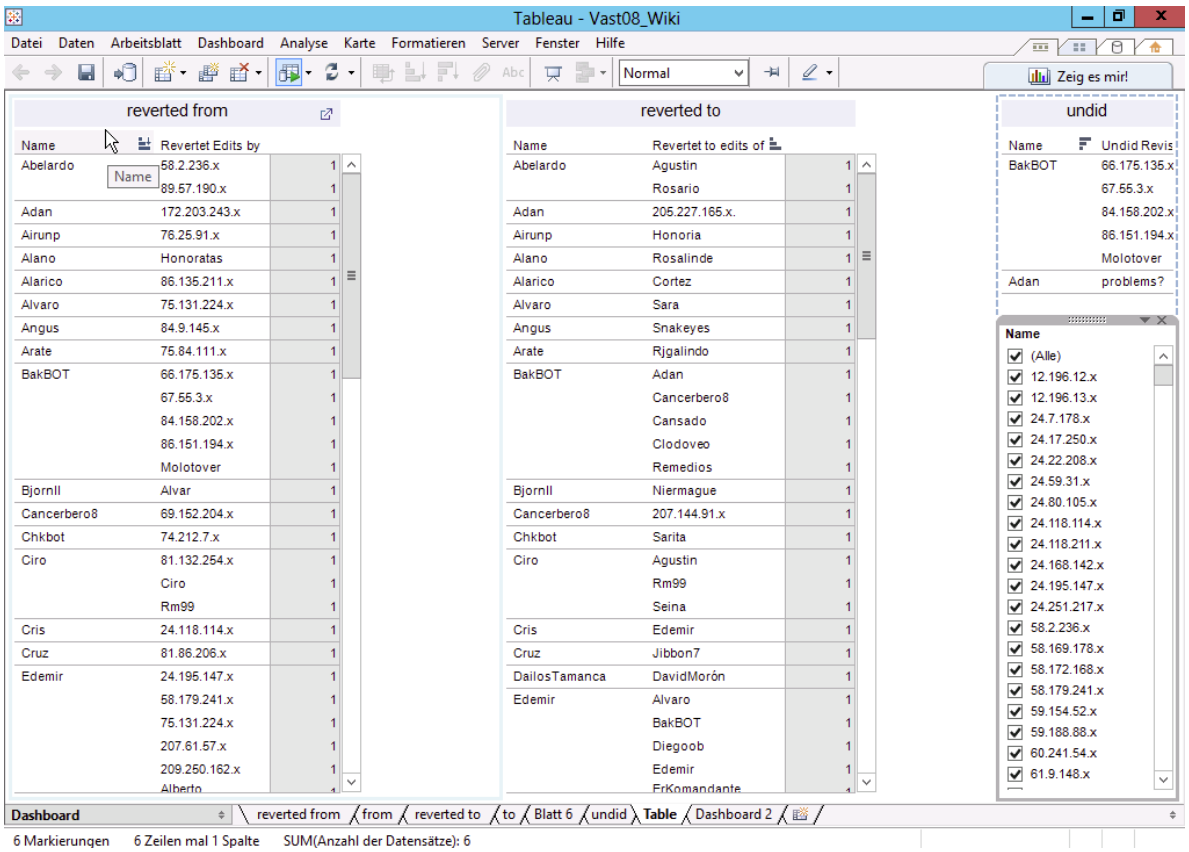

Abbildung 7.25: Dashboard zum Herausfinden der Fraktionen mittels der Anzahl an Änderungen

Die aktivsten Benutzer zu ermitteln kann ein Balkendiagramm erstellt werden (siehe Abbildung 7.26)

Mittels des Dashboards (siehe Abbildung 7.25) und des Balkendiagramms (siehe Abbildung 7.26) können nun folgendes Ergebniss ermittelt werden:

1. Die Hauptnutzer sind BakBOT, Edemir, VictoriaV, Honoria, SilviaBot, Cira und Sara.
2. Es gibt 2 Fraktionen unter den Benutzern, die durch überdurchschnittlich viele Ediths herausstechen: Fraktion 1 wird angeführt von Victoria und Sara und enthält unter anderem Cira. Fraktion 2 wird angeführt von Edemir und Estirabot und enthält unter anderem Rm99 und BakBOT. Dies wurde durch das Heraussuchen von Mustern aus Dashboards (siehe Abbildung 7.25) ermittelt. Die Muster sind dabei wenn ein Mitglied häufig die Einträge von einer Gruppe auf die Einträge von eignen Mitglieder ändert.

Weitere Ergebnisse in Tableau zu der Challenge zu erhalten ist mühsam und kommt einem manuellen Durchgehen der Daten gleich.

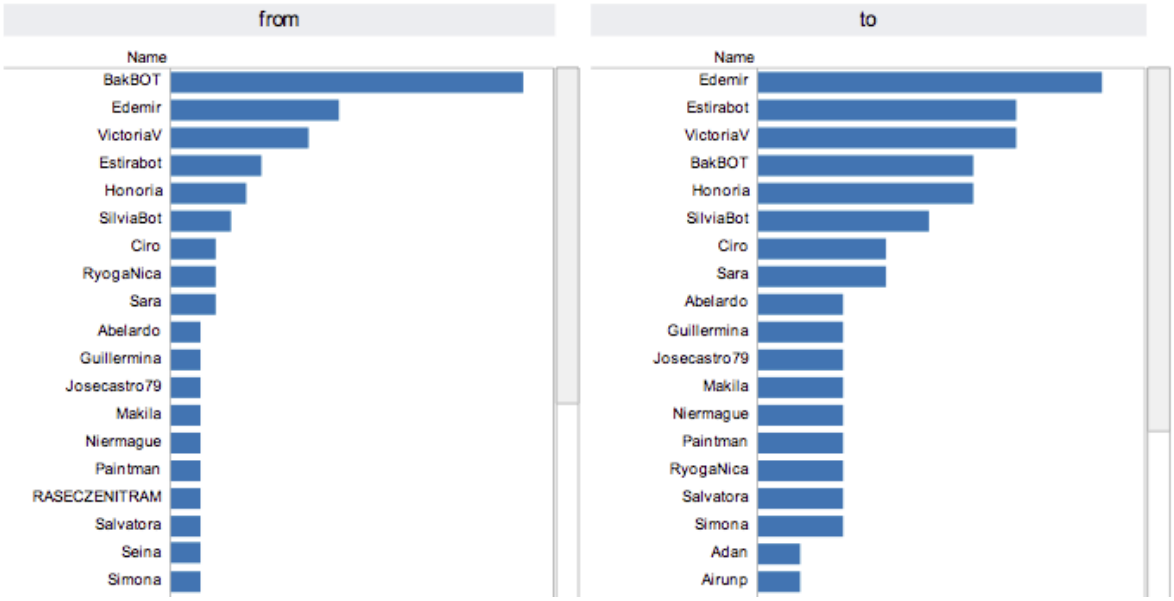

Abbildung 7.26: Balkendiagramm mit der Anzahl der Aktivitäten zur Identifizierung der wichtigsten Benutzer

Hindernisse / Erkenntnisse

Wie in der Einleitung des Kapitels erwähnt, ist Tableau nicht für die Analyse von Texten geeignet. Das Vorbearbeiten der Daten ist sehr aufwendig und ermöglicht es nicht semantische Informationen aus dem Text zu lesen. Wie bei der VAST Challange 2009 zu Social Networks ist das Fehlen von Knoten-Kanten-Diagrammen ungeschickt, da man dadurch Gruppierungen besser darstellten könnte. Es ist auch von Nachteil, dass sich keine horizontale mehrfarbige Balkendiagramme erstellen lassen. Wagerechte Diagiagram sind möglich, werden im Allgemeinen aber für zeitliche Zusammenhänge verwendet.

7.3 VAST-Challenge 2009 - Social Network (Netzwerkanalyse)

Die VAST-Challenge 2009 erfordert die Analyse des fiktiven sozialen Netzwerkes "Flitter". Dieses wird von Arbeitskollegen und Freunden verwendet um untereinander zu kommunizieren. Das 'Flitter'-Netzwerk unterhält eine Verbindung zu einem kriminellen Ring, der möglicherweise einen Angestellten für seine Zwecke rekrutiert hat. Es wurden Daten von Flitter zur Verfügung gestellt um dies zu Analysieren. Die vorliegenden Daten beinhalten eine Relations- und eine Nutzer-Tabelle über die der zugrunde liegende Kommunikationsgraph aufgebaut werden kann.

Vorab gibt es eine Beschreibung von zwei möglichen Szenarien, wie das neue Mitglied des kriminellen Rings rekrutiert wurde und über welche Verbindungen es zu diesem verfügt. Dabei spielen Handlanger, Mittelsmänner und der Boss der Organisation eine Rolle.

Die Fragestellungen bezüglich dieses Datensatzes lauten[4]

- Welche der beiden beschriebenen Szenarien liegt laut des Datensatzes der Aufgabe zu Grunde?
- Identifizieren Sie den Angestellten, die Handlanger, den oder die Mittelsmänner, den Anführer und des Anführers internationale Kontakte sowie andere relevante Personen.

Bei dieser Aufgabe geht es also in erster Linie um die Analyse eines großen Netzwerkes (ca. 6000 Knoten mit 30.000 Kanten) mittels Visual-Analytic-Tools. Die beschriebenen Szenarien galt es dabei in Eigenschaften eines Graphs zu transferieren und diesen danach zu untersuchen.

7.3.1 Qlik

Einlesen der Daten

Bereits das Einlesen der Daten ist in Qlik sehr schwer zu erreichen, da in der Social Network Challenge die Daten in verschiedenen Quellen verteilt sind. Daher muss zunächst eine Exceltabelle erstellt werden, welche alle vorhandenen Daten beinhaltet. Die Informationen des Knoten Kantendiagramms sind hierbei schwer zu integrieren, da zu jedem Eintrag mehrere Konten und Kanten existieren. Daher werden die Graphdaten in einer zweiten Qlik-Datei analysiert. Zugehörige Namen werden demnach erst nach der Graphanalyse den IDs zugeordnet. Damit der Graph ungerichtet dargestellt werden kann, müssen alle Kanteneinträge in umgekehrter Richtung in der Tabelle angefügt werden. Andernfalls, kommt es zu Unstimmigkeiten in der Anzahl der Kanten von und zu einem Knoten.

Lösungsansatz und Hindernisse

Da Qlik keine integrierte Lösung und auch keine Erweiterung für die Darstellung von Graphen bietet ist es kaum möglich eine sinnvolle Analyse des Datensatzes zu bewerkstelligen. Es können in Balkendiagrammen beispielsweise die Nutzer mit den meisten Kanten ermittelt werden. Dabei muss jedoch auf die doppelte Anzahl jeder Kante geachtet werden. Diese Graphik in Abbildung 7.27 zu sehen.

Von dort aus, können nun die potenziellen Bosse (mehr als 100 Kontakte) gefiltert werden und die Eigenschaften der mit Ihnen verbundenen Mittelsmänner gezeigt werden. Das Ermitteln der letztendlichen Struktur ist dabei sehr umständlich, da die gegebene Auswahl zwischengespeichert werden muss und die Analyse nahe an eine manuelle Auswertung kommt. Das Tool ist dabei kaum mehr als eine Zählhilfe.

[4] vgl. http://hcil.cs.umd.edu/localphp/hcil/vast/index.php/taskdesc/index

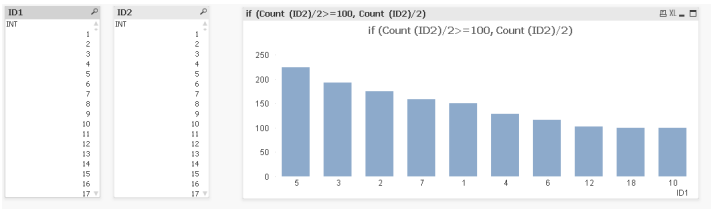

Abbildung 7.27: Qlik - Social Network - Potenzielle Bosse des Netzwerkes berechnet über die Anzahl der Verbindungen über 100. Da die Verbindungen verdoppelt wurden, muss diese Anzahl halbiert werden oder mit 200 verglichen werden.

Ein weiterer Versuch der Visualisierung war das Eintragen der Kanten in einen Scatterplot 7.28. Aufgrund der Anzahl an Kanten kam es dabei jedoch zu starken Performance-Problemen, bei gleichzeitig geringem Mehrwert der Darstellung.

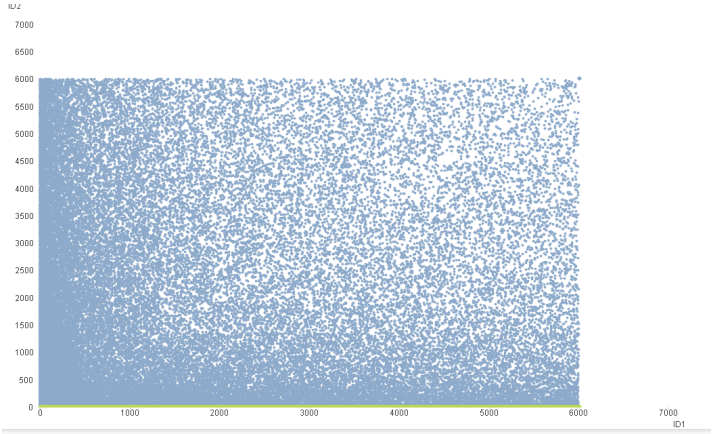

Abbildung 7.28: Qlik - Social Network - Darstellng des Netzwerkes als Scatterplot. Aufgrund der hohen Anzahl Kanten ist dies wenig übersichtlich.

7.3.2 Spotfire

Einlesen der Daten

Das Einlesen der Daten war mittels Spotfire schnell erfolgreich. Die Kanten des Netzwerks und die Knoteneigenschaften standen in zwei separaten Dateien zur Verfügung. Bei den Einträgen handelt es sich um Hin- und Rückkanten. Leider unterstützt Spotfire jedoch keine ungerichteten Graphen. Aus diesem Grund musste die Kantendaten manipuliert werden. Der Datensatz wurde einfach um die gespiegelten Einträge erweitert ($E' = E \cup \{(x,y)|\forall(y,x) \in E\}$.

Anschließend konnten beide Datensätze eingelesen und mittels der Relationen verknüpft werden.

Weg zur Lösung

Zur Analyse der Daten dient natürlich ein Node-Link-Diagramm welches von Spotfire standardmäßig unterstützt wird.

ID	Name	Type	Degree
38	@krintz	person	80
60	@roark	person	78
63	@letelier	person	82
68	@saraswat	person	80
79	@terekhov	person	78
92	@tolbert	person	82
100	@schaffter	person	80
115	@classe	person	82
140	@bailey	person	78
142	@lafouge	person	80
171	@supornpaibul	person	82
175	@akbar	person	78
227	@inamori	person	80
228	@anastasakis	person	80

Tabelle 7.1: Mögliche Angestellte des Datensatzes (der Grad der Knoten muss halbiert werden, da wie in siehe Abschnitt 7.3.2 auf Seite 77 beschrieben die Kanten durch das hinzufügen der Rückkanten verdoppelt wurden)

Der Weg zu Lösung begann mit der Suche nach möglichen Kandidaten für den Angestellten. Hierbei nutzen wir die Information aus beiden Szenarien, dass der Angestellte um die 40 Kontakte haben soll. Spotfire bietet hierzu bei Node-Link-Diagrammen direkt die Möglichkeit zur Berechnung des Grades der Knoten. Eine anschließende Filterung der Daten nach Knoten die einen Grad zwischen 39 und 41 haben. Die daraus resultierenden möglichen Angestellten sind in in der Tabelle (siehe Abb. 7.29) dargestellt.

Anschließend wurden die möglichen Handlanger gesucht. Diese wurden ebenfalls über den Grad der Knoten identifiziert. Gemeinsam mit den möglichen Angestellten lässt sich so ein Netzwerk der Handlanger und Angestellten erstellen. Das sind alle Knoten mit den Graden 30-41 (siehe Abb. 7.29). In diesem Netzwerk sind die möglichen Angestellten grün markiert. Da der verdächtigte Angestellte drei solche Handlanger haben soll, lässt sich mit dieser Visualisierung die Anzahl der verdächtigen bereits stark reduzieren (siehe Abb. 7.2).

Nun folgt die Suche nach dem Mittelsmann in der Organisation. Dieser wurde ebenfalls mittels des Knotengrades ermittelt. In beiden Szenarien müssen die Mittelsmänner vier

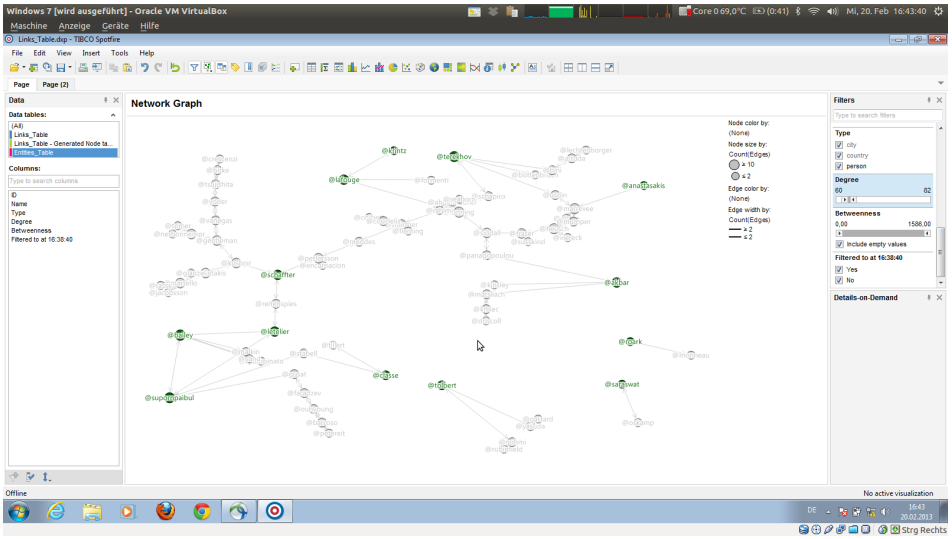

Abbildung 7.29: Die verdächtigten Angestellten (grün) und die mögliche Handlanger in einem Node-Link-Diagramm dargestellt.

ID	Name	Type	Degree
63	@letelier	person	82
79	@terekhov	person	78
92	@tolbert	person	82
100	@schaffter	person	80
140	@bailey	person	78
142	@lafouge	person	80
171	@supornpaibul	person	82
175	@akbar	person	78
227	@inamori	person	80

Tabelle 7.2: Die möglichen Angestellten (reduziert)

bis fünf (die drei Handlanger und ein bis zwei weiter) Kontakte haben, also vom Grad acht bis zehn im modifizierten Datensatz. Es kommen 1956 Mitglieder des Netzwerkes als Mittelsmänner in Frage.

Anschließend wurde für alle diese möglichen Angestellten ermittelt ob die drei in Frage kommenden Handlanger untereinander Kontakt haben, denn dies ist ebenfalls in beiden Szenarien nicht erlaubt. Dies ist in Spotfire nicht automatisierbar. Aus diesem Grund wurde für jede der Konstellationen der Graph betrachtet, indem die in Frage kommenden Mittelsmänner ignoriert wurden. Dazu wurden alle Knoten radial angeordnet und über einen Filter die 1956 Knoten mit dem Grad 8-10 ausgefiltert. Die Handlanger wurden dann in die Mitte des Kreises geschoben. So kann schnell festgestellt werden ob diese gemeinsame Nachbarn

auf dem äußeren Kreis haben(siehe Abb. 7.30). In diesem Fall wurde nur für die drei Knoten 194, 261 und 563 festgestellt, dass keine weiteren Verbindungen unter den Handlangern bestehen. Damit ist der verdächtige Angestellte @schaffter.

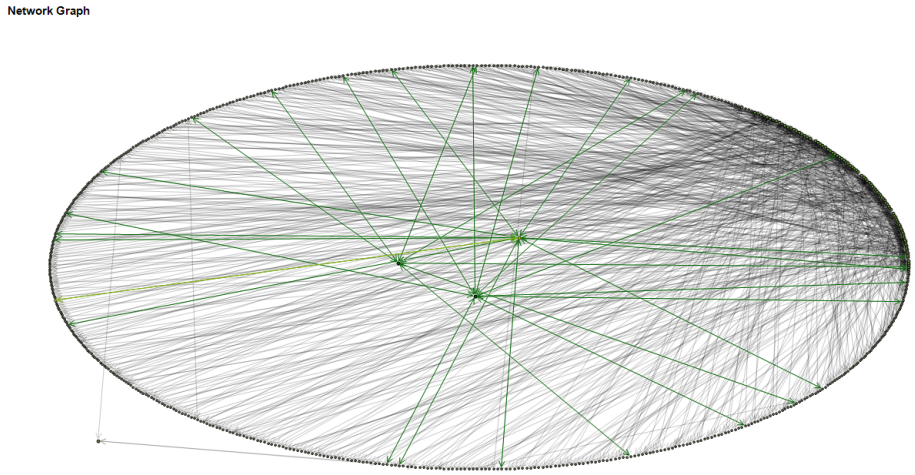

Abbildung 7.30: Die drei Handlanger 194, 261 und 563 in der Mitte und ihre Verbindungen zu allen anderen Flitter-Mitgliedern. Es gibt keine gemeinsamen Kontakte.

Eine Analyse der so ermittelten Handlanger unter Verbindung der bereits gefilterten Mittelsmänner hat dann ergeben, dass der Mittelsmann 4994 (@good) ist. Szenario B kann nicht eintreten, da die ermittelten Handlanger nicht jeweils zu einem Mittelsmann eine Verbindung unterhalten.

Da Borris (@good) nur 5 Kontakte hat ist die Ermittlung des Fearless Leaders nun einfach: die einzigen beiden infrage kommenden Kontakte sind 4(@szemeredi) und 1612(@moilanen). Letzterer hat jedoch nur 15 Kontakte und kommt damit nicht als Boss in Frage. Somit ist @szemeredi der Boss des Netwerkes.

Ergebnis

Durch die obige Analyse konnte festgestellt werden, dass das Szenario A plausibel ist. Der Boss der kriminellen Organisation ist @szemeredi. Sein "Borris" ist @good. Dessen drei Handlanger @kushnir, @pettersson und @reitenspies den Angestellten @schafftler umgedreht und für sich gewonnen haben (siehe Abb. 7.31).

Hindernisse / Erkenntnisse

Während der Analyse waren verschiedene Fähigkeiten von Spotfire sehr hilfreich. Zum einen kann bei Node-Link-Diagrammen der Grad der Knoten automatisch berechnet werden.

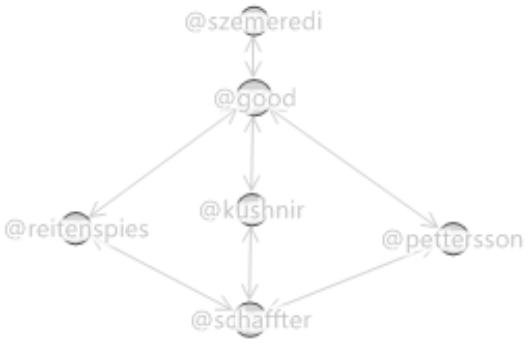

Abbildung 7.31: Das strukturelle Ergebnis der Analyse. Oben ist @szemeredi, der Boss, unten befindet sich der verdächtige Angestellte.

Die verfügbare Filterung im Programm lässt es außerdem zu, dass man jeden Datensatz (auch diesen mit 6000 Knoten und 20000 Kanten) schnell filtern kann. Anfragen und Veränderungen in der Visualisierung wurden schnell und ohne ungewöhnliche Wartezeiten realisiert. Markierungen übertragen sich von einer Darstellung in die andere, was zum Verständnis enorm beigetragen hat. Markierungen können außerdem als Filter übernommen werden. Dieser kann dann ausgewählt oder ausgeschlossen werden. Dadurch ließ sich dieser Datensatz gezielt durchforsten. Die verschiedenen Layout-Algorithmen (vor allem der radiale und das *Forced-Directed-Layout*) trugen zum Verständnis ebenso bei.

Die Verbindung von zwei Tabellen miteinander (hier Knoten- und Kantentabelle) erfolgte problemlos. Lediglich die fehlende Unterstützung von ungerichteten Node-Link-Diagrammen ist hier anzumerken.

7.3.3 Tableau

Um diese Challenge sinnvoll bearbeiten zu können, ist es unabdingbar die Daten in eine Knoten-Kanten-Diagramm darzustellen. Dies ist in Tableau nicht ohne weiteres möglich.

Es gibt in einem offiziellen Foreneintrag auf der Tableau Homepage zwar die Nennung von zahlreichen Tools[5] , diese lösen das Problem allerdings nur unzureichend. Bei den Tools werden die Datensätze um Koordinaten ergänzt. Diese lassen sich dann in Tableau einlesen und in einem zwei dimensionalen Koordinatensystem eines Streudiagramms darstellen.

[5] http://community.tableausoftware.com/thread/109050

Daraufhin wird umständlich an der Formatierung gearbeitet, bis das Netzwerk als Knoten-Kanten-Diagramm erkennbar ist. Dadurch geht allerdings jegliche Flexibilität verloren und ein flexibles Arbeiten mit den Daten ist nicht mehr möglich.

Auf diesem Grund wird auf die Durchführung der Challenge mit Tableau verzichtet. Informationen die man ohne eine Darstellung des Netzwerks erhalten könnte, gehen genauso gut mit einem anderen Tool wie zum Beispiel MS Excel oder dem manuellen Durchgehen der Daten.

Als hypothetisches Beispiel, wie eine Netzwerkanalyse funktionieren könnte zeigt die folgende Abbildung ein solches Streudiagram eines Netzwerks, siehe Abb. 7.15.

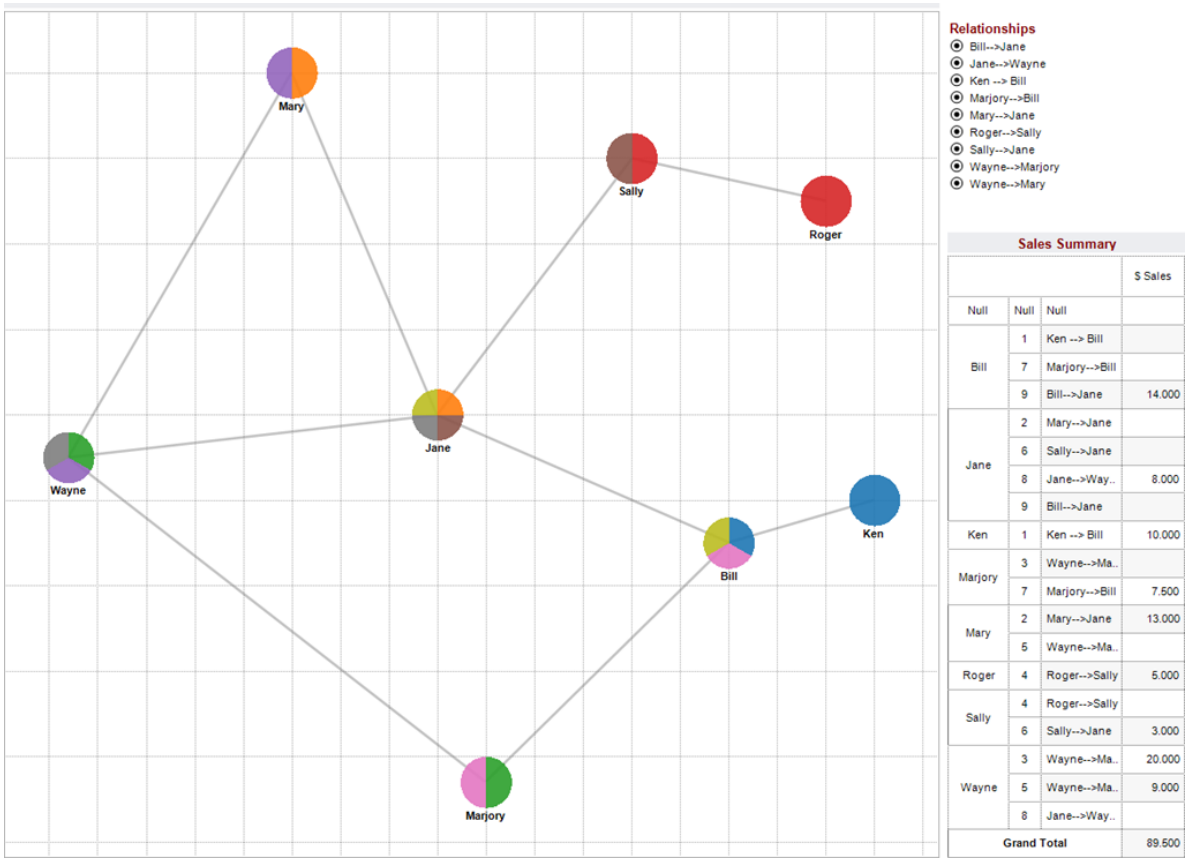

Abbildung 7.32: Beispiel eines Knoten-Kanten-Diagramms in Tableau[a]

[a] http://www.clearlyandsimply.com/clearly_and_simply/2012/12/build-network-graphs-in-tableau.html

8 Abschließende Bewertung der Tools (Bewertungskatalog)

In diesem Kapitel soll das Ergebnis der Analyse in Form des ausgefüllten Bewertungskatalogs dargestellt werden. Es wurden folgende Kategorien bewertet:

- Analysefähigkeit - Text
- Analysefähigkeit - Geo
- Analysefähigkeit - Netzwerk
- intuitive Interaktion mit den Daten
- Benutzerfreundlichkeit
- Belastbarkeit
- Analysefähigkeit - Allgemeine Daten
- Importmöglichkeiten
- Export-Möglichkeiten
- Kollaboration
- Umgebung
- Automatisierbarkeit

Für genaue Beschreibungen der Kategorien, betrachten Sie das Kapitel Bewertungskatalog. Die abschließende Punkteverteilung und Bewertung der Tools im Bewertungskatalog wird in folgender Tabelle dargestellt:

8.1 Bewertung der Tools

Kategorie	Subkategorien	Bewertungskriterium	Relevanz	Tableau	Spotfire	Qlik
Benutzerfreundlichkeit			0,05	4,5	4,5	2,17
	Erlernbarkeit	-Tutorium, -Wie lange braucht ein Benutzer um einen Datensatz zu visualisieren?		4	5	2
	Robustheit gegenüber Bedienfehlern	Warnmeldungen,Redo/Undo		5	5	4
	Wiedererkennung	Toolhilfen, einheitliches Design, einheitlicher Aufbau		4	4	4
	Interaktion mit Benutzer	Ladestatus, sinndienliche Hinweise, Übersichtlichkeit, (Feature)Suche, Visualisierungsvorschläge		5	4	0
	Bedienbarkeit	Eingabemöglichkeiten, Befehlssprache notwendig		4	4	1
	Visualisierungs-Mantra	Overview first, zoom and filter, details on demand		5	5	2
Kollaboration			0,03	2,67	1,67	1,33
	Kommentare	Kommentieren von Daten/-Diagrammen, gemeinsame Interaktion		5	4	2
	Datenbestandsverwaltung	Ablage auf Server, gemeinsames Nutzen		3	3	1
	Programm	Benötigen alle das Program, gibt es einen interaktiven Export		5	2	2
	Konfigurationsmanagement	Historie, Änderungsverfolgung, Zurücksetzen,...		1	1	1

Kategorie	Subkategorien	Bewertungskriterium	Relevanz	Tableau	Spotfire	Qlik
	Rollen	Verteilung, Benutzereinschränkungen		0	0	0
	Teilen-Funktion	Makros, Todos		2	0	2
Umgebung			0,01	5	1,5	3,5
	Stand-Alone/ Client-Server /Cloud			5	1	5
	External Viewers			5	2	2
	Scalability			0	0	0
	Memory concept			0	0	0
	BI infrastructure integration			0	0	0
	Datensicherheit	Verschlüsselung, Anonymisierung, Benutzerrechte		0	0	0
Automatisierbarkeit			0,01	2,5	1,25	1
	Makros	Erstellen, Verwalten, Verteilen, In verschiedenen Projektmappen nutzen		1	0	1
	Automatischer Export			0	0	0
	Benutzerdefinierte Analysen			4	1	1
	Automatischer Import			5	4	2
Belastbarkeit			0,05	4,6	4,2	3,6
	Umgang mit großen Daten	Datengröße, Analysegeschwindigkeit		4	3	4
	Umgang mit kleinen Daten	Schnelligkeit, Genauigkeit		4	4	4
	Umgang mit mehreren Datensätzen	Können Datensätze parallel bearbeitet werden, Können diese kombiniert werden?		5	5	0

Kategorie	Subkategorien	Bewertungskriterium	Relevanz	Tableau	Spotfire	Qlik
	Beschränkungen durch das Program	Anzahl der Statistiken, Anzahl der Diagramme		5	5	5
	Verbindungsauslastung	Wenn ein Netzwerkzugang erforderlich ist, wie sehr wird dieses dann belastet?		5	4	5
Analysefähigkeit - Text			0,2	0	0	0,5
	Überblick über Text	Keywords Identifizieren, Maximum, Minimum, Tagcloud		0	0	2
	Inhaltliche Analyse	Stimmung des Textes, Mehrsprachigkeit, Treffersicherheit, Komplexität der Sprache, Protagonisten erkennen, Betreffende Personen in Relationen setzen		0	0	0
	Relationale Erkennung	Mehrere Texte: Welche stehen in Relation, Gleiche Autoren erkennen,		0	0	0
	Mustererkennung	Ähnliche Satzstruktur, Tabellen, Aufzählungen,		0	0	0
Analysefähigkeit - Geo			0,2	5	3	3
	Mögliche Eingabe	Geokoordinaten, PLZ, Städtenamen, Länder		5	3	4
	Mögliche Ausgaben	Karte, Parameter wie Höhenlage, Bevölkerungsdichte usw. einbaubar?, 2D, 3D, Interaktive Ausgabe		5	3	3
	Interaktion	Verschieben, Zoomen, Filtern der Karte		5	4	3

Kategorie	Subkategorien	Bewertungskriterium	Relevanz	Tableau	Spotfire	Qlik
	Visualisierung	Auswahl an Farbmuster, Darstellung, Symbolen, Kartenlayer		5	1	3
	Überblick	Sind in großen Datensätzen grobe Zusammenhänge erkennbar?		5	4	2
Analysefähigkeit - Netzwerk			0,2	0	2,88	0
	Überblick	Gruppenbildung, Leader-Eigenschaft, Hierarchie-Erkennung, Zusammenhang,		0	3	0
	Filterung	Dichte, Richtung/Art der Kante, Bestimmte Knoten hervorheben,		0	3	0
	Datengröße	Anzahl Kanten, Anzahl Knoten, Dichte des Graphs		0	4	0
	Interaktion	Zoom, Hervorheben, Verschieben, herziehen, ausblenden, Daten auf Abfrage, Veränderung der Visualisierung (Farbe selbst anpassen, Eigenschaften auf Visuelles mappen, ...)		0	5	0
	Ausgabemögichkeiten	Baum, Treemap, Radiale, ...		0	3	0
	Reduktion von Visual Clutter	Edge Bundling, Ein/Ausklappen von Bereichen, ...		0	0	0
	Treemap			0	0	0
	Graphvisualisation			0	5	0
	Clustering			0	3	0

Kategorie	**Subkategorien**	**Bewertungskriterium**	**Relevanz**	**Tableau**	**Spotfire**	**Qlik**
Analysefähigkeit - Allgemeine Daten			0,05	4,23	4,46	3
	Überblick	Veränderungen, Konstanten, Oszilierend, ...		4	4	3
	Ausgabe	Animation, Filmstreifen, Diagramm,		5	4	4
	Interaktion	Zoom in Ausgabe, Veränderung der Zeitabstände, Hervorheben der einzelnen Bereiche, markieren von Objekten um diese genauer Verfolgen zu können,		5	5	3
	Eingabe	Automatische Zuordnung der Daten zu den Zeitintervallen, Einlesesformate, Hinzufügen von weiteren Datensätzen in bestehende Analyse		5	5	1
	Column calculations/column and row combinations			3	3	3
	Joints/ Joints on filtered Tables			5	4	0
	Querying functions (Group by sum, average, Count,ordering)			5	5	3
	Univariate			5	5	5
	Bivariate			5	5	5
	Multivariate			3	5	2
	Bar- line-pie-Chart Histogramm			5	5	5

Kategorie	Subkategorien	Bewertungskriterium	Relevanz	Tableau	Spotfire	Qlik	
	Scatterplot			5	5	5	
	Parallel Coordiantes			0	3	0	
Importmöglichkeiten			0,05	5	3,83	0,67	
	Einfache Datenformate	CSV, Txt, XLS,		5	5	1	
	Komplexe Datenformate	SQL, Access, SAP, ...		5	3	0	
	Verknüpfungserkennung	Verknüpfung zu bestehenden Datensätzen, Darstellung dieser, Hervorheben von Dublikaten, Warnung vor Inkonsistenzen,		5	4	0	
	Datenquellen verwalten	Verbindungen speichern, Einstellungen/Vorkonfigurationen behalten,		5	3	0	
	Selektion der Daten	Einschränkung bereits beim Import, Priorisierung, Benennung, Datenformate anpassen, Trennung festlegen,		5	5	3	
	Erweiterte Dateneingabe	Copy	Paste, Import von Webseiten		5	3	0
intuitive Interaktion mit den Daten			0,1	4,5	4,25	2	
	Visualisierungsmantra	Overview first, zoom and filter, details on demand		5	5	2	
	Freie Gestaltbarkeit	Kommentare anlegen, Texte hinzufügen, Symbole und Farben ändern, Legenden verschieben,		5	4	2	

Kategorie	Subkategorien	Bewertungskriterium	Relevanz	Tableau	Spotfire	Qlik
	Filterung	Teilelemente ausschließen, Selektion auf alle Diagramme übertragen, Selektion mehrerer Datensätze und Auswertung dieser		4	5	3
	Workflow	logisch, Flexibilität der Reihenfolge, ggf. Hinweise, keine Störungen durch das Program		4	3	1
Export-Möglichkeiten			0,05	3	2,5	2
	Diagramme	3d/2d, Dynamisch, Animation, What You see is what u get, Interkation mit Export auch ohne das Program		4	3	1
	Veröffentlichung	PDF, HTML, FTP-Upload, EMail, Powerpoint, Facebook, Twitter ...		2	2	3
Gesamtwertung			1	2,67	2,65	1,56

Tabelle 8.2: Ausgefüllter Bewertungskatalog mit Punkten zu allen 3 Tools

In folgender Graphik wird das Ergebnis der Hauptkategorien aller drei Tools als Balkendiagamm dargestellt.

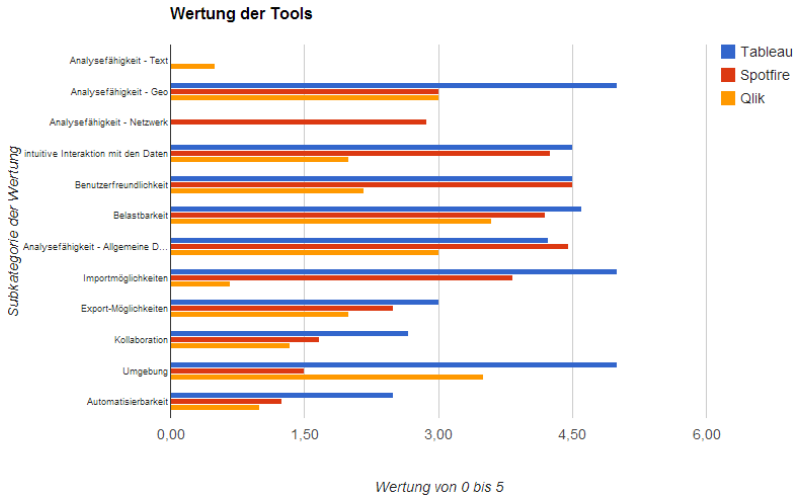

Abbildung 8.1: Bewertungskatalog - Bewertung der Tools

Die folgenden Kapitel sollen die jeweiligen Entscheidungskriterien für die Punktevergabe erläutern.

8.2 Analysefähigkeit - Text

Keines der untersuchten Tools ist für eine Analyse von unstrukturierten Datensätzen, wie Texten geeignet. Qlik bietet über die Erweiterung WordMap die Möglichkeit einkopierten Text in eine TagCloud zu verwandeln. Allerdings geschieht die Datenverarbeitung dabei innerhalb der Erweiterung und nicht über die gewöhnliche Datenschnittstelle. Durch Vorbearbeiten der Textschnipsel anhand von Trennzeichen, konnten zumindest Standard-Aktionen, wie 'Undid Revision by' + Name gefiltert werden. Diese können dann als Tabelle wie andere Datensätze innerhalb der Tools verarbeitet werden. Die Punkteverteilung im Bereich Analysefähigkeit - Text sieht wie folgt aus: Tableau- 0 Spotfire- 0 Qlik- 0,5

8.3 Analysefähigkeit - Geo

Eine einfache Geo-Analyse anhand von Koordinaten in einem Scatterplot ist in allen drei Tools möglich. In Tableau ist die Analyse von Geodaten besonders einfach, da Kartenmaterialien bereits im Programm integriert sind. In Qlik ist dieser Vorteil über die Erweiterung CloudMadeMaps erreichbar. Spotfire bietet dazu keine zusätzliche Erweiterung. Die Eingaben sind meist auf Längen- und Breitengrade eingeschränkt. In der GeoQlik Extension

können Daten auch in Form von Staaten und Counties der USA angegeben werden. Weitergehende Features zu Geo-Analyse wie 3D Darstellungen und Einbeziehen der Daten über bestimmte Kartenabschnitte sind in keinem der Tools verfügbar. Die Punkteverteilung im Bereich Analysefähigkeit - Geo sieht wie folgt aus: Tableau- 5 Spotfire- 3 Qlik- 3

8.4 Analysefähigkeit - Netzwerk

Die graphische Analyse von Netzwerkdaten wird nur von Spotfire unterstützt. Zwar können in Tableau und in Qlik anhand der allgemeinen Netzwerkdaten verschiedene Statistiken erstellt werden oder über Filtern von Daten Zusammenhänge hergestellt werden, doch ein Graph kann dabei nicht dargestellt werden. In Spotfire sind verschiedene Layout Algorithmen zu Formatierung des erstellten Graphen integriert. Allerdings können keine hierarchischen Layouts wie Beispiel die Treemap und keine radialen Layouts hergestellt werden. Auch der Umgang mit großen Graphen ist kein Problem. Beim Test von einem Graphen mit 6000 Knoten und 20000 Kanten kam es zu keinen Problemen. Die Punkteverteilung im Bereich Analysefähigkeit - Netzwerk sieht wie folgt aus: Tableau- 0 Spotfire- 2,88 Qlik- 0

8.5 Intuitive Interaktion mit den Daten

Sowohl in Spotfire als auch in Tableau ist die Umsetzung des Visualisierungsmantra, eine freie Gestaltungsmöglichkeit, einfache Filterung und ein logischer Workflow sehr gut umgesetzt. Alle Interaktionen mit den Daten und die Visualisierungen sind intuitiv möglich. Kleinere Mängel gibt es beim Setzen von Kommentaren und bei der flexiblen Verwendung von Filtern. In Qlik sind zwar die meisten Schritte ebenfalls möglich, jedoch immer nur über Assistenten und Menüs erreichbar. Die Interaktion ist hier nur selten intuitiv und zudem oftmals anhand der Vielzahl ähnlicher Menüpunkte sehr verwirrend. Die Punkteverteilung im Bereich intuitive Interaktion mit den Daten sieht wie folgt aus: Tableau- 4,5 Spotfire- 4,25 Qlik- 2

8.6 Benutzerfreundlichkeit

Sowohl Spotfire als auch Tableau sind zu Beginn sehr leicht zu erlernen und bieten einen intuitiven Umgang mit den Daten. Qlik setzt dagegen auf Assistenten um Fehler zu vermeiden und benötigt zudem eine Skriptsprache, um Daten einzulesen oder zu berechnen. Alle drei Tools bieten eine gute Robustheit, indem sie Redo/Undo anbieten und mit Warnmeldungen arbeiten. Auch der Aufbau der Tools ist schlüssig und baut auf bekannte UI-Elemente auf. Das Feedback an die Nutzer ist in Tableau besonders ausgeprägt, indem sogar gut passende Visualisierungen zu den gegebenen Daten vorgeschlagen werden. In Spotfire ist abgesehen von diesem Punkt ebenfalls eine gute Interaktion mit dem Nutzer über schnelle Reaktion des Programms und regelmäßiges Feedback gegeben. In Qlik sind derartige benutzerfreundliche

Features nicht vorhanden. Die Punkteverteilung im Bereich Benutzerfreundlichkeit sieht wie folgt aus: Tableau- 4,5 Spotfire- 4,5 Qlik- 2,17

8.7 Belastbarkeit

Die Belastbarkeit der Tool wurde anhand der vorhandenen Datensätze getestet. Darüber hinaus wurden keinen speziell großen Datensätze getestet. Bei den Analysen konnte ein guter Umgang mit den Datensätzen festgestellt werden. In der Analyse gibt es keine Beschränkungen an Platz und Anzahl der Visualisierungen. In den Datensätzen gibt es zumindest bei Tableau und Spotfire keine Begrenzung nach Anzahl oder Größe. Qlik hingegen kann nur einen Datensatz zugleich verarbeiten und benötigt beispielsweise bei Wechseln zwischen Sheets immer wieder einige Sekunden zum Rendern der Visualisierungen. Da es sich bei den Testprogrammen um Desktopanwendungen handelt, ist primär keine Wartezeit für Verbindungen notwendig. Allerdings benötigt Spotfire für die Anmeldung eine Verbindung zum Server sowie Qlik zur Verwendung von Erweiterungen. Die Punkteverteilung im Bereich Belastbarkeit sieht wie folgt aus: Tableau- 4,6 Spotfire- 4,2 Qlik- 3,6

8.8 Analysefähigkeit - Allgemeine Daten

Im Bereich der allgemeinen Analyse schnitten alle drei Tools sehr gut ab. Ein-/Ausgabe und Interaktion sind überall gut umgesetzt. Standardvisualisierungen von Uni- und Bivariaten Daten sind überall möglich. Berechnungen einzelner Felder oder berechneter Spalten ist ebenfalls möglich, aber meist umständlich umgesetzt. Joins sind nur in Spotfire und Tableau möglich. Bei allen Tools mangelt es an Visualisierungen von multivariaten Daten wie Parallele Koordinaten. Hier bietet Spotfire die beste Auswahl, aber auch diese ist sehr beschränkt. Die Punkteverteilung im Bereich Analysefähigkeit - Allgemeine Daten sieht wie folgt aus: Tableau- 4,23 Spotfire- 4,46 Qlik- 3

8.9 Importmöglichkeiten

Der Import ist von Tableau am besten umgesetzt, da hier neben den gängigen Datenformaten auch Datenbanken angeschlossen werden können. In Spotfire werden alle gängigen Datenformate unterstützt, wahrend Qlik auf Exceldateien eingeschränkt ist. Außerdem ist Qlik nicht standardmäßig dafür ausgelegt mehrere Datensätze einzulesen oder diese zu verknüpfen. Dies ist in beiden anderen Tools leicht erreichbar. Die Punkteverteilung im Bereich Importmöglichkeiten sieht wie folgt aus: Tableau- 5 Spotfire- 3,83 Qlik- 0,67

8.10 Export-Möglichkeiten

Der Export ist in allen Tools in Form von Bildern, PDFs und als gespeicherte Datei möglich. Tableau bietet zusätzlich funktionsfähige HTML Seiten als Export. Dies kann in Spotfire und Qlik nur über Öffnen der Datei in der Webversion des Programmes geschehen. Die Punkteverteilung im Bereich Export-Möglichkeiten sieht wie folgt aus: Tableau- 3 Spotfire- 2,5 Qlik- 2

8.11 Kollaboration

Alle drei Tools bieten die Möglichkeit Kommentare zu verfassen. Allerdings ist die Lösung über Workarounds in Spotfire und Qlik sind nicht zufriedenstellend. Ebenfalls können die Daten in allen drei Tools weitergegeben werden. In Tableau ist das sogar als funktionsfähiger Webview möglich. In Qlik kommt es in der Trial Version bei der Weitergabe von Dateien zu Einschränkungen. Es gibt in keinem der Tools eine Rollenverteilung und auch die Versionsverwaltung ist in allen drei Tools auf Redo/Undo Aktionen beschränkt. Die Punkteverteilung im Bereich Kollaboration sieht wie folgt aus: Tableau- 2,67 Spotfire- 1,67 Qlik- 1,33

8.12 Umgebung

Alle drei Tools bieten eine Desktopversion an. Dabei ist bei Spotfire zur Anmeldung eine Internetverbindung notwendig. Qlik bietet zusätzlich eine Webversion seines Programms oder einen Webview innerhalb der Desktopversion zur Verwendung der Erweiterungen an. Tableau bietet zusätzlich eine Cloud-basierte Programmversion an. Zur Skalierbarkeit, Datensicherheit und der internen Infrastruktur sind uns keine Informationen bekannt. Die Punkteverteilung im Bereich Umgebung sieht wie folgt aus: Tableau- 5 Spotfire- 1,5 Qlik- 3,5

8.13 Automatisierbarkeit

Die Automatisierungen von Visualisierungen wird in keinem der Tools angeboten. Allerdings lässt sich der Datenimport in Tableau über eine Live-Verbindung kontinuierlich aktualisieren. Dies ist in Spotfire nur manuell und in Qlik nur, bei nicht veränderter Datenstruktur möglich. Die Punkteverteilung im Bereich Automatisierbarkeit sieht wie folgt aus: Tableau- 2,5 Spotfire- 1,25 Qlik- 1

9 Fazit

Nach dem intensiven Vergleich der drei Visualiserungstools: Qlik, Spotfire und Tableau mittels der VAST Challenges 2008 und 2009 lässt sich folgendes feststellten:

Tableau geht als Sieger hervor. Spotfire liegt bei der durchgeführten Bewertung nur etwas unter Tableau und ist bei der Netzwerkanalyse klar überlegen. Qlik liegt weit abgeschlagen hinter den anderen Tools, ist jedoch frei erhältlich.

Es hat sich gezeigt die Eignung eines Tools stark von der Aufgabe abhängt. Geo-Analysen sind gut, Textanalyse kaum und Netzwerkanalyse nur mit Spotfire möglich:

- Alle Tools kommen mit **Geo-Daten** zurecht. Hier sind die Tools sehr wertvoll. Es ist schnell möglich die gefragten Information zu erhalten. Die Geo-Daten darzustellen ist mit Tableau am einfachsten, da die Karten im Programm integriert sind. Nach dem hinzufügen der Kartendaten ist die Analyse mit Spotfire und Qlik ebenso möglich.
- Bei der **Textanalyse** ist keines der Tools wirklich hilfreich. Die Daten mussten mühsam von Hand vorbearbeitet werden. Bei der Analyse kommen gewohnte Vorteile der flexiblen Interaktion mit Daten bei Tableau und Spotfire kaum zum tragen. Die Ergebnisse hätten mit Standardsoftware ebenso gut gelöst werden können. Einzig Qlik hatte die Möglichkeit die Wörter der Einträge in einer Tag-Cloud darzustellen.
- Die **Netzwerkanalyse** ist ausschließlich mit Spotfire möglich, da es als Einiges die Daten in einem Knoten-Kanten-Diagramm darstellen kann. Hier erweist sich Spotfire als sehr wertvoll und man kommt schnell zu einer leicht nachvollziehbaren Lösung.

Es ist empfehlenswert bei einer anderweitigen Aufgabenstellungen weitere Visialisierungstools bei der Auswahl mit ein zu beziehen.

Abschließend lässt sich sagen: Die Visualieserungstools erleichtern durch Echtzeit-Interaktion mit großen Datenmengen das Erstellten und Erarbeiten von nachvollziehbaren Lösungen auf spezielle Fragestellungen. Die Wahl des Tools spielt für das Ergebnis eine wichtige Rolle, dabei könnten die hier genannten Empfehlungen hilfreich sein.

Glossar

.NET Das .NET Framework ist eine Entwicklungsplattform, mit der Sie Anwendungen für Windows, Windows Store, Windows Phone, Windows Server und Windows Azure erstellen können. Die .NET Framework-Plattform enthält die Programmiersprachen C# und Visual Basic, die Common Language Runtime (CLR) sowie eine umfangreiche Klassenbibliothek. http://msdn.microsoft.com/de-de/vstudio/aa496123.aspx 38

ADO ActiveX Data Objects 38

CSV comma-seperated values 53

ESRI-Shapes Hierbei handelt es sich um ein Dateiformat zum Speichern von Geodaten. Entwickelt von ESRI Inc. (Environmental Systems Research Institute) ist ein US-amerikanischer Softwarehersteller von Geoinformationssystemen (GIS) 38

Forced-Directed-Layout TODO 81

GPS Global Positioning System 16

KMZ Keyhole Markup Language, wird hauptsächlich von Google Earth und Google Maps verwendet 57

ODBC Open Database Connectivity 38

UDL Universal-Data-Link 38

VAST Visual Analytics Science and Technology 11

Glossar

brushing and linking The idea of linking and brushing is to combine different visualization methods to overcome the shortcomings of single techniques. Interactive changes made in one visualization are automatically reflected in the other visualizations. Note that connecting multiple visualizations through interactive linking and brushing provides more information than considering the component visualizations independently [Kei02b] 39, 46

Literaturverzeichnis

[Ahl96] C. Ahlberg. Spotfire: an information exploration environment. *Newsletter ACM SIGMOD Record*, S. 25–29, 1996. (Zitiert auf Seite 36)

[Bos12] H. Bosch. Evaluation verfügbarer Visual Analytics Toolkits anhand von Benchmark-Datensätzen, 2012. Ausschreibung zu dieser Arbeit. (Zitiert auf Seite 11)

[Ded10] U. Dedering. Map of Middle America, including the Gulf of Mexico and the Caribbean Sea. Equirectangular projection. Strechted by 106.0URL http://commons.wikimedia.org/wiki/File:Middle_America_relief_location_map.png. Besucht am 19.01.2013. (Zitiert auf Seite 57)

[DK10] G. E. F. M. D.A. Keim, J. Kohlhammer. Solving Problems with Visual Analytics. *Mastering the Information Age*, 2010. (Zitiert auf Seite 13)

[Kei02a] D. A. Keim. Information Visualization and Visual Data Mining. *IEEE Transactions on Visualization and computer graphics*, 2002. (Zitiert auf Seite 13)

[Kei02b] D. A. Keim. Information Visualization and Visual Data Mining. *IEEE Transactions on Visualization and computer graphics*, 2002. (Zitiert auf Seite 99)

[Ler12] A. Lerg. Gigantische Datenfluten bei Facebook, 2012. URL http://computer.t-online.de/facebook-gigantische-datenfluten-pro-tag/id_58980908/index. Besucht am 17.02.2013. (Zitiert auf Seite 23)

[PS08] A. Perer, B. Shneiderman. Systematic yet flexible discovery: guiding domain experts through exploratory data analysis. In *Proceedings of the 13th international conference on Intelligent user interfaces*, IUI '08, S. 109–118. ACM, New York, NY, USA, 2008. doi: 10.1145/1378773.1378788. URL http://doi.acm.org/10.1145/1378773.1378788. (Zitiert auf Seite 36)

[Shn] B. Shneiderman. Eight Golder Rules of Interface Design. URL http://faculty.washington.edu/jtenenbg/courses/360/f04/sessions/schneidermanGoldenRules.html. Eingesehen am 24.02.2013. (Zitiert auf Seite 23)

[Shn96] B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. *Proceedings of the 1996 IEEE Symposium on Visual Languages*, S. 336–343, 1996. (Zitiert auf den Seiten 23 und 39)

[Shn99] B. Shneiderman. Dynamic queries, starfield displays, and the path to Spotfire, 1999. URL http://www.cs.umd.edu/hcil/spotfire/. überarbeitet Juli 2007, besucht am 01.03.2013. (Zitiert auf Seite 36)

Erklärung

Hiermit versichern wir, diese Arbeit
selbständig verfasst und nur die angegebenen
Quellen benutzt zu haben.

_____

(Fabian Merkle Hanna Schäfer Sebastian Zillessen)