Sebastian Bremm, Sebastian Maier, Tatiana von Landesberger, Tobias Schreck

Fachgebiet Graphisch-Interaktive Systeme, Technische Universität Darmstadt

Ein flexibles System für die explorative visuelle Sequenz-Analyse

Zusammenfassung

Visual Analytics ist eine junge Disziplin die automatische Datenanalyse mit Visualisierung und Mensch-Computer-Interaktion verbindet. Diese Kombination hat sich in vielen Fällen bereits als besonders effektiv bei der explorativen Analyse großer und komplexer Daten erwiesen.

Ein Beispiel hierfür sind große Mengen von Sequenzdaten, also eine lineare Folge von Objekten. Diese treten in vielen Arbeitsgebieten wie z.B. der Biologie als Folge von Aminosäuren, Digitaler Bibliotheken als Folge von Zeichen und Worten und dem Finanzsektor als Folge von Kurswerten auf.

Wir stellen in dieser Arbeit ein System zur explorativen visuellen Analyse diskretisierter Sequenzen vor. Die neuen Werkzeuge sind in eine Pipeline zur Analyse großer Datenmengen eingebunden. Wir bieten dem Benutzer viele interaktive Möglichkeiten, Bestandteile der Pipeline zu kombinieren und Optionen der verwendeten Algorithmen zu beeinflussen, um der Diversität der Daten gerecht zu werden. Diese umfassenden Möglichkeiten kombiniert mit entsprechenden Visualisierungen führen zu einem Verständnis des Datensatzes und den darin enthaltenen Informationen durch den Analysten. Diese Erkenntnisse können dann sowohl als Grundlage von Entscheidungen als auch als Ausgangspunkt für weitere Analyseschritte dienen.

1 Einleitung

Die Analyse von sequenziellen Daten spielt in vielen Anwendungsgebieten eine Rolle. Die Proteinsequenzanalyse in der Biologie, die Chartanalyse im Finanzsektor oder die Analyse von Texten sind nur einige Beispiele. Gesucht werden u.a. bestimmte (Sub)Sequenzen, korrelierende Abschnitte, zyklische Wiederholungen oder besonders untypische Sequenzabschnitte. Ein Typisches Beispiel aus der Finanzindustrie wäre es, verschiedene Firmen zu finden deren Kursverhalten am Finanzmarkt sich ähnelt. Dadurch lassen sich Ursachen für das Verhalten finden und sich eventuell sogar Voraussagen treffen. Wir werden auf dieses Beispiel im Anwendungsteil zurück kommen. Eine erfolgreiche Datenanalyse ohne Hilfsmittel gelingt nur bei sehr kleinen Datenmengen. Der Umfang der Daten steigt jedoch stetig, begünstigt durch immer besser werdende technische Mittel zur Datenerhebung, Übermittlung und Speicherung.

Ein weiteres Problem sind oft wechselnde Szenarien und damit einhergehend, wechselnde Anforderungen an die Suche. Die verwendeten Analyseverfahren müssen deshalb schnell an die jeweilige Aufgabenstellung angepasst werden können. Dies trifft besonders auf die explorative Datenanalyse zu. Zum einen liegt dabei zu Beginn der Analyse oft nur sehr wenig Wissen über den internen Aufbau und die Struktur der Daten vor, zum anderen muss keine konkrete Aufgabenstelllung, die über das Finden „interessanter“ Merkmale hinausgeht, gegeben sein.

Bei der Analyse der Daten sollten zwei Aspekte betrachtet werden. Einerseits die algorithmische Analyse durch den Computer und andererseits eine visuelle Analyse durch den Menschen. Sowohl Mensch als auch Computer haben ihre spezifischen Stärken. Computer können gut und schnell bekannte Probleme anhand eines vorgegebenen Algorithmuses lösen. Der Mensch kann sich sehr schnell auf neue Anforderungen einstellen und seine Stärken bei der Erkennung von Mustern ausspielen wenn die Daten entsprechend präsentiert werden. Die Kombination dieser unterschiedlichen Methoden führt im angestrebten Idealfall schnell zu sehr erfolgreichen Ergebnissen.

Die Sequenzanalyse gliedert sich in unterschiedliche Abschnitte. Zunächst müssen die Rohdaten z.B. durch Normalisierung aufbereitet werden. Anschließend werden die kontinuierlichen Rohdaten durch eine Berechnung von prototypischen Repräsentanten diskretisiert. Auf die Definition der paarweisen Ähnlichkeiten unter den Repräsentanten folgt die eigentliche explorative visuelle Mustersuche. In all diesen Schritten wird der Benutzer so gut wie möglich durch Automatismen unterstützt, hat aber jederzeit die Möglichkeit in den Prozess einzugreifen. Um ihm diese Möglichkeit sinnvoll bieten zu können, ist es nötig die einzelnen Schritte und deren Ergebnisse zu visualisieren.

Um dies zu erreichen kombinieren wir verschiedene, auf die einzelnen Schritte und deren Darstellung spezialisierte Programme. Dadurch wird sichergestellt, dass jeder Teilprozess mit der nötigen Qualität bearbeitet werden kann und so Folgefehler minimiert werden. Gleichzeitig ist durch die Kombination verschiedener Programme eine größtmögliche Flexibilität gewährleistet und je nach Anwendungsfall kann die Bearbeitungspipeline angepasst werden.

Ein Schwerpunkt liegt dabei auf der Visualisierung der einzelnen Datenanalyse Schritte. Nur wenn der Benutzer die einzelnen Vorgänge versteht und die Auswirkungen der Algorithmen und seiner Einstellungen sieht, kann eine zielführende und erfolgreiche Analyse der Daten gewährleistet werden.

Um eine hohe Flexibilität zu gewährleisten, bieten wir dem Benutzer verschiedene Visualisierungsansätze die effektiv miteinander kombiniert und vernetzt sind.

Ein weiterer wichtiger Aspekt in diesem Analyseszenario ist die interaktive Definition der paarweisen Ähnlichkeit von Objekten. Dadurch kann der Nutzer auch an diesem Punkt eigenes und problemspezifisches Wissen unkompliziert in den Analyseprozess einbringen.

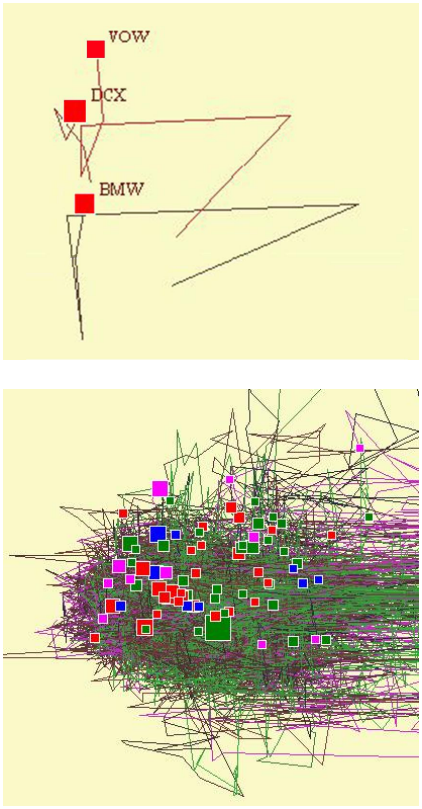

Abb. 1: Die verwendeten Daten stellen die Werte zu Risiko und Ertrag der Wertpapierkurse verschiedener Firmen über die Zeit dar. Die einzelnen Zeipunkt werden chronologisch durch Graden verbunden. Bei zu vielen Zeitpunkten und Unternehmen wird diese Form der Visualisierung schnell unübersichtlich.

Die nachfolgende Arbeit gliedert sich wie folgt:

- Kapitel 2 gibt einen kurzen Überblick über dieser Arbeit zugrunde liegende Literatur und alternative Ansätze zur Sequenzanalyse.
- Kapitel 3 widmet sich den Hintergründen der Arbeit und des Datensatzes.
- Kapitel 4 behandelt Details zur Datenaufbereitungspipeline.
- In Kapitel 5 widmet sich der Sequenzanalyse.
- In Kapitel 6 wird unser System zur visuellen Sequenzanalyse vorgestellt und an einem Beispiel demonstriert.
- Kapitel 7 fasst die gewonnenen Erkenntnisse zusammen und bietet einen Ausblick auf zukünftige Arbeiten.

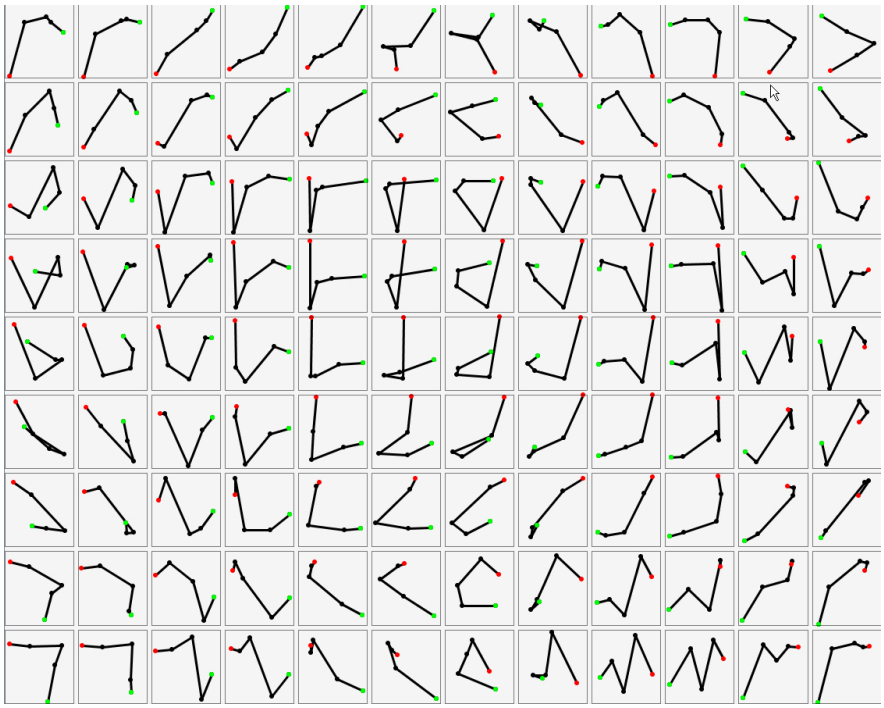

Abb. 2: Auf der Abbildung ist eine austrainierte SOM zu sehen, jedee Zelle repräsentiert ein Neuron. Die in jeweils 5 Zeitpunkte unterteilten Trajektorien wurden als Eingabe verwendet. Jede Zelle visualisiert die 10 Dimensionen ihres Vektors, wobei der Startpunkt grün und der Endpunkt rot markiert ist.

2 Literatur

Die Analyse von Sequenzdaten stellt ein breites Feld dar. Aus der Bioinformatik stammen viele Arbeiten zur Analyse von Sequenzdaten. Sequenzdaten sind dort z.B. durch eine Folge von Aminosäuren oder Nukleotiden in Form von 20 bzw. vier Buchstaben charakterisiert. Eines der bekanntesten Werkzeuge um eine Sequenz mit einer großen Menge anderer Sequenzen zu vergleichen ist BLAST (Basic Local Alignment Search Tool) [Korf et al. 2003]. Möchte man, wie in unserem Fall mehrere Sequenzen gegeneinander vergleichen, bietet sich die Clustal Algorithmenfamilie an [Larkin et al. 2007]. VISA [Posfai et al. 1994] rückt die Visualisierung zur Sequenzanalyse stärker in deren Vordergrund. Das Auftreten häufiger Subsequenzen wird durch ein Histogrammvisualisierung der einzelnen Sequenzen dargestellt. [Keim et al. 2006] kombinierten automatisierte und visuelle Sequenzanalyse in einem System. Dabei wird ein neues flexibles System zur Correlationsanalyse mit einer interaktiven Visualisierung verbunden. Viele dieser Ansätze dienten uns als Inspiration, sind allerdings aufgrund ihres spezialisierten Einsatzbereichs sehr in ihren Möglichkeiten beschränkt. So unterscheidet sich zum Beispiel die Fragestellung in manchen Punkten. Bei den von uns analysierten Sequenzen spielt Zeit eine wesentliche Rolle, wohingegen bei der biologischen Sequenzanalyse Spalten einzelner Zeilen verschoben werden, um korrespondierende Subsequenzen passend anzuordnen.

Auch die Analyse von Finanzdaten ist ein sehr aktives Forschungsfeld in dem große Mengen von Sequenzdaten anfallen. Anders als in der Biologie setzen sich die Sequenzen aus chronologischen Folgen numerischer Werte zusammen. Häufig werden diese im Zuge des Analyseprozesses diskretisiert. [Lin et al. 2004] entwickelten eine Baum-orientierte Visualisierung um häufig auftretende Teilsequenzen zu finden und darzustellen. Das System berücksichtigt Ähnlichkeiten, skaliert aber schlecht mit der Alphabetgröße.

[Chang et al. 2007] stellten ein System zur Analyse zeitabhängiger Transaktionsdaten vor. Sie verfolgen den Ansatz, verschiedene Techniken und Visualisierungen zu kombinieren, um unterschied-

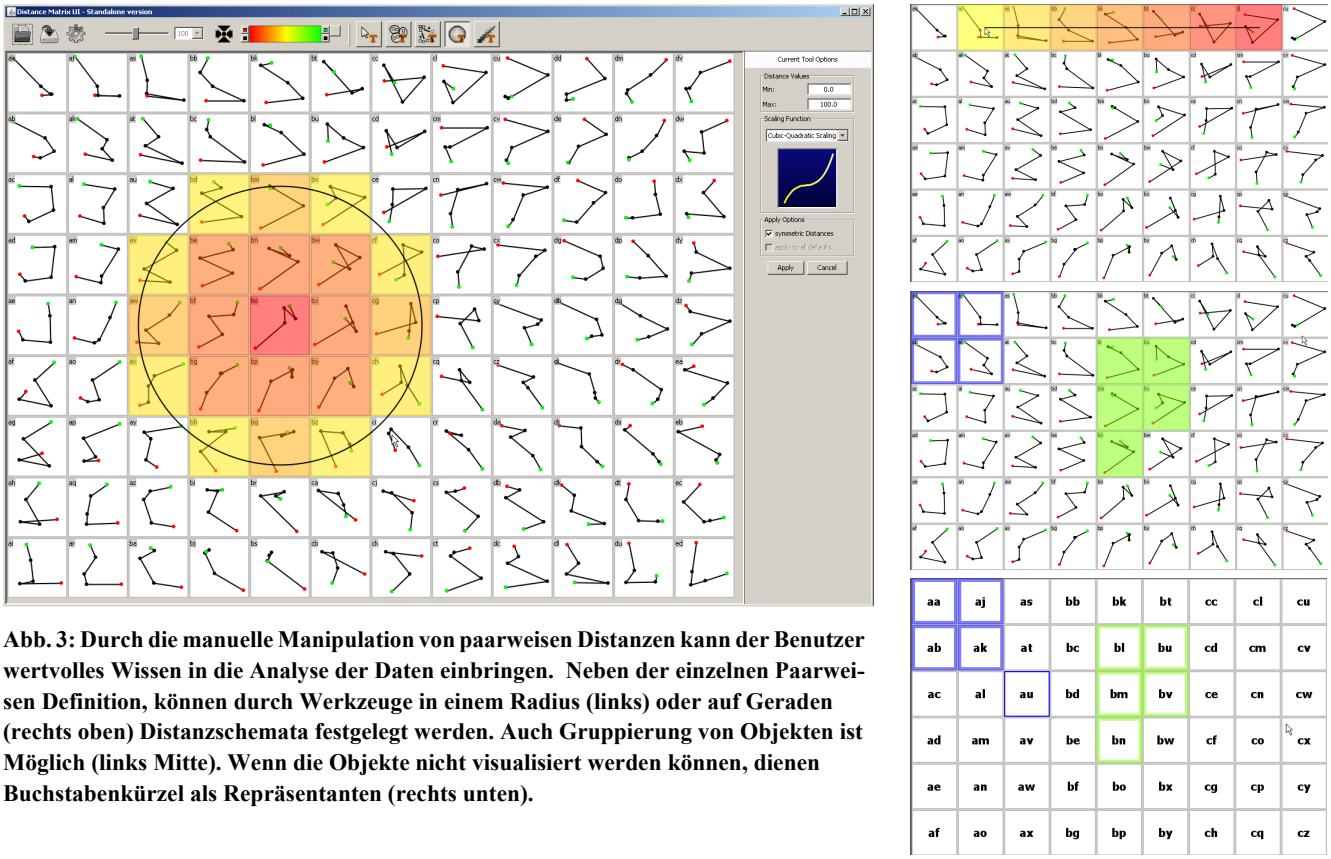

Abb. 3: Durch die manuelle Manipulation von paarweisen Distanzen kann der Benutzer wertvolles Wissen in die Analyse der Daten einbringen. Neben der einzelnen Paarweisen Definition, können durch Werkzeuge in einem Radius (links) oder auf Geraden (rechts oben) Distanzschemata festgelegt werden. Auch Gruppierung von Objekten ist Möglich (links Mitte). Wenn die Objekte nicht visualisiert werden können, dienen Buchstabenkürzel als Repräsentanten (rechts unten).

liche Aspekte der Daten zu untersuchen.

Ein prominentes, einfach zu bedienendes System zur visuellen Analyse von Zeitserien bietet Timesearcher [Hochheiser & Schneiderman 2004]. Es umfasst verschiedene Analysemöglichkeiten, die sich aber fast nur auf eine Linien-Graph basierte Visualisierung stützen.

Nicht unerwähnt sollen verschiedene Ansätze bleiben die zeitabhängige Daten nicht in Form von Sequenzen zeigen sondern zur Darstellung Trajektorien und Animation nutzen. Ein sehr prominentes, ausgereiftes und intuitiv zu bedienendes Beispiel stellt Gapminder (www.gapminder.org) da. Die verschiedenen Datensätze werden dabei in einem 2-dimensionalem Koordinatensystem visualisiert.

Zusammenfassend kann man sagen, dass bereits viele unterschiedliche Ansätze zur Analyse verschiedenster Sequenzen vorgestellt wurden. Meist sind diese aber, unter anderem aufgrund von Spezialisierung, in ihrem Funktionsumfang beschränkt. Unser Ansatz baut darauf auf, viele verschiedene Aspekte effektiv miteinander zu Kombinieren um eine hohe Flexibilität zu ereichen.

3 Hintergrund

Die hier exemplarisch verwendeten Rohdaten enthalten Informationen über Risiko und Ertrag gemessen am Wertpapierkurs von 60 Firmen über jeweils mehr als 500 Tage. Das Ziel der Explorativen Analyse dieser Daten ist es, Muster und Besonderheiten zu finden, etwa ein ähnliches Verhalten verschiedener Firmen, einzelne Firmen die sich einem allgemeinen Trend entgegengesetzt verhalten, periodische Auftreten bestimmter Verläufe usw. Der Analyst erhofft sich durch Finden solcher Muster Rückschlüsse auf deren Ursachen ziehen zu können und somit evtl. Vorhersagen über zukünftiges Verhalten treffen zu können.

Ein wichtiger Leitsatz bei der Gestaltung von visuellen Analysesystemen ist das Mantra „Analyse First - Show the Important - Zoom, Filter and Analyse Further - Details on Demand“ [Keim 2008]. Seine Idee findet sich in diesem Systemen wieder.

Ein üblicher Ansatz zur visuellen Analyse 2-dimensionaler Daten ist die Verwendung eines Streudiagramms. Im Fall von zeitabhängigen Daten werden Trajektorien verwendet [Tekusová, Kohlhammer 2007], es wird also für jeden Zeitpunkt ein Marker gezeichnet, die dann chronologisch durch Geraden verbunden werden (Abb. 1 oben). Bei kleinen Datensätzen ist diese Methode durchaus zielführend, sie skaliert aber schlecht und kann so auf große Datensätze nicht angewendet werden, da es zu einer Überfüllung der Zeichenfläche kommt (Abb. 1 unten).

4 Pipeline

Um die beschriebene Problematik des Datenüberflusses zu bewältigen, wird eine Datenreduktion in mehreren Schritten durchgeführt.

Der erste Schritt hierzu ist eine Unterteilung der Sequenz in kürzere Teilsequenzen. Die Länge dieser Sequenzen ist beliebig, wir haben eine Länge von fünf gewählt, da dies einer Handelswoche entspricht und daher inhaltlich sinnvoll erscheint. Unvollständige, etwa durch einen Feiertag unterbrochene Wochen werden verworfen. Im nächsten Schritt haben wir die nun erhaltenen Sequenzen aus Fünf-Tages-Trajektorien normalisiert. Dieser Schritt liegt in der Fragestellung begründet; große Unternehmen reagieren

auf äußere Einflüsse meist weniger als kleine Unternehmen, um aber diese Reaktionen vergleichbar zu machen, werden die Daten normalisiert.

Der nächste Schritt der Datenreduktion ist eine Berechnung von Prototypen. Dazu werden die Daten mit geeigneten Algorithmen gruppiert und ein Referenzprototyp von jeder Gruppe erstellt. Wir benutzen hierfür eine selbstorganisierende Karte, kurz SOM (Self Organizing Map) [Kohonen 2001]. Dieser Algorithmus bietet uns viele Vorteile. Er akzeptiert n-dimensionale Eingabevektoren und liefert eine 2-dimenisonale Karte in Form eines Netzes, das die Nachbarschaft und damit die Ähnlichkeit der Eingaben repräsentiert. Jeder Knoten des Netzes ist dabei mit seinen direkten Nachbarn verbunden. Durch die Dimensionsreduzierung lassen sich die Ergebnisse sehr intuitiv visualisieren.

In unserem Fall besitzt der Eingabevektor zehn Dimensionen, die sich aus Risiko und Ertrag für jeweils fünf Tage ergeben. Jedem Knoten des Netzes ist ein zehndimesnionaler Vektor zugeordnet, der durch den Clusterungsprozess definiert wurde. Dieser Vektor repräsentiert gleichzeitig unseren Prototypen. Der Vorteil dieses Ansatzes besteht darin, dass Eingabevektoren und Prototypen eine intuitive Visualisierung in Form der Trajektorie besitzen (Abb. 1, rechts).

Die Realisierung der Visualisierung der SOM und deren Trainingsprozesses sind in [Schreck et al. 2008] beschrieben. Dem Benutzer werden neben umfangreicher interaktiver Möglichkeiten zur Parametereinstellung auch verschiedene Möglichkeiten zur Qualitätseinschätzung des Ergebnisses geboten. Darüber hinaus wird eine inhaltliche Analyse des Ergebnisses durch verschiedene Werkzeuge unterstützt um letztendlich die gewonnenen Erkenntnisse sowohl zur Verbesserung der SOM-Parameter als auch in Folgeschritten der Sequenzanalyse zu verwenden.

5 Sequenzanalyse

Durch die von der SOM berechneten Prototypen werden nun die Sequenzen dargestellt. Ziel der Sequenzanalyse ist es, Muster und Besonderheiten in Sequenzen, also Folgen von Zeichen zu finden.

Die Eingabe besteht aus beliebig vielen Sequenzen. Jede dieser Sequenzen besitzt eine beliebige Länge. Als Subsequenzen bezeichnet man Zeichenfolgen die in mindestens einer der Sequenzen enthalten sind.

5.1 Definition von Ähnlichkeit

Um zwei (Sub-)Sequenzen miteinander vergleichen zu können, muss zunächst ein Ähnlichkeitsmaß definiert werden. Der erste Schritt ist die Definition einer paarweisen Ähnlichkeit zweier Zeichen. In manchen Fällen wird auch die Distanz definiert, beide Werte können ineinander überführt werden. Besitzen zwei Zeichen die Distanz 0, so besitzen sie die maximal mögliche Ähnlichkeit. Bei einer Ähnlichkeit von 0 trennt sie die maximal mögliche Distanz. Diese paarweisen Distanzen werden üblicherweise durch eine Distanzmatrix repräsentiert. Häufig sind die paarweisen Distanzen symmetrisch.

Bei der Distanzdefinition gibt es unterschiedliche Ansätze. In unserem Fall, werden die Objekte durch Feature-Vektoren repräsentiert. Hier kann zum Beispiel die Distanz im Feature-Raum oder, nach einer Transformation, die Distanz im Repräsentationsraum als Grundlage für die Definition dienen. In andere Fällen, wie zum Beispiel beim Vergleich von Aminosäuresequenzen wird meist ein anderer Ansatz gewählt, z.B. dienen experimentell bestimmte Mutationswahrscheinlichkeiten von einer zur anderen Aminosäure als Grundlage [Henikoff, Henikoff, 1992]

Wir verfolgen eine Kombination aus beiden Ansätzen, also sowohl eine automatische Berechnung von Ähnlichkeiten auf Grund von Repräsentation der Objekte im Vektorraum, als auch eine Einbindung von zusätzlichem Wissen durch die Möglichkeit der interaktiven Manipulation durch den Benutzer. Die einzelnen Zellen der SOM werden durch den jeweiligen n-dimensionalen Feature-Vektor und ein Zeichen repräsentiert. Zur Definition der Ähnlichkeit zweier Zeichen wählten wir zwei Ansätze. Zum einen dient der schon erwähnte euklidische Abstand im n-dimensionalen Raum als Grundlage. Zum anderen verwenden wir die Kantendistanz im 2-dimensionalen Netz der SOM als Ausgangspunkt.

Um eine einfache Benutzerinteraktion bieten zu können, haben wir eine interaktive Visualisierung der Distanzmatrix realisiert (Abb. 3). Die Distanz eines Objektes zu allen anderen wird durch eine Farbskala mit verschiedenen Normalisierungsoptionen repräsentiert. Der Nutzer hat verschiedene Manipulationsmöglichkeiten. Die einfachste ist die Definition einer einzelnen, paarweisen Distanz. Des weiteren besteht die Möglichkeit einen Pfad zwischen zwei Objekten und deren Distanz zu definieren. Dann werden die Distanzen aller sich auf dem Pfad befindlichen Objekte nach einer frei definierbaren Normalisierung angepasst. Die dritte Möglichkeit ist die Festlegung eines Radius um ein Objekt herum und Anpassung eines Distanzgefälles in Relation zum Radius.

Diese, auf ein Objekt bezogenen Definitionen lassen sich anschließend automatisch auf alle Objekte und deren paarweise Distanzen übertragen.

5.2 Finden ähnlicher Sequenzen

Um die initiale Datenstruktur aufzubauen, suchen wir in allen Sequenzen nach allen möglichen Subsequenzen. Uns interessieren Teilsequenzen deren Häufigkeit mindestens einen gewissen Schwellwert übersteigt.
Der verwendete Algorithmus orientiert sich an der Idee von [Wong et al. 2000]. Dort wird er zur Suche und Speicherung von Sequenzen für die Textanalyse eingesetzt.
Das Resultat ist eine Liste aller Teilsequenzen die häufiger als der gewählte Schwellwert auftreten und wird in Form eines lexikographischen Baumes gespeichert.

6 Visuelle Sequenzanalyse

Das von uns entwickelte System zur visuellen Sequenzanalyse ist in verschiedene Ansichten unterteilt die in Kapitel 6.1 ausführlich erläutert werden. Jede dieser Visualisierungen unterstützt das Finden bestimmter Eigenschaften von Subse-

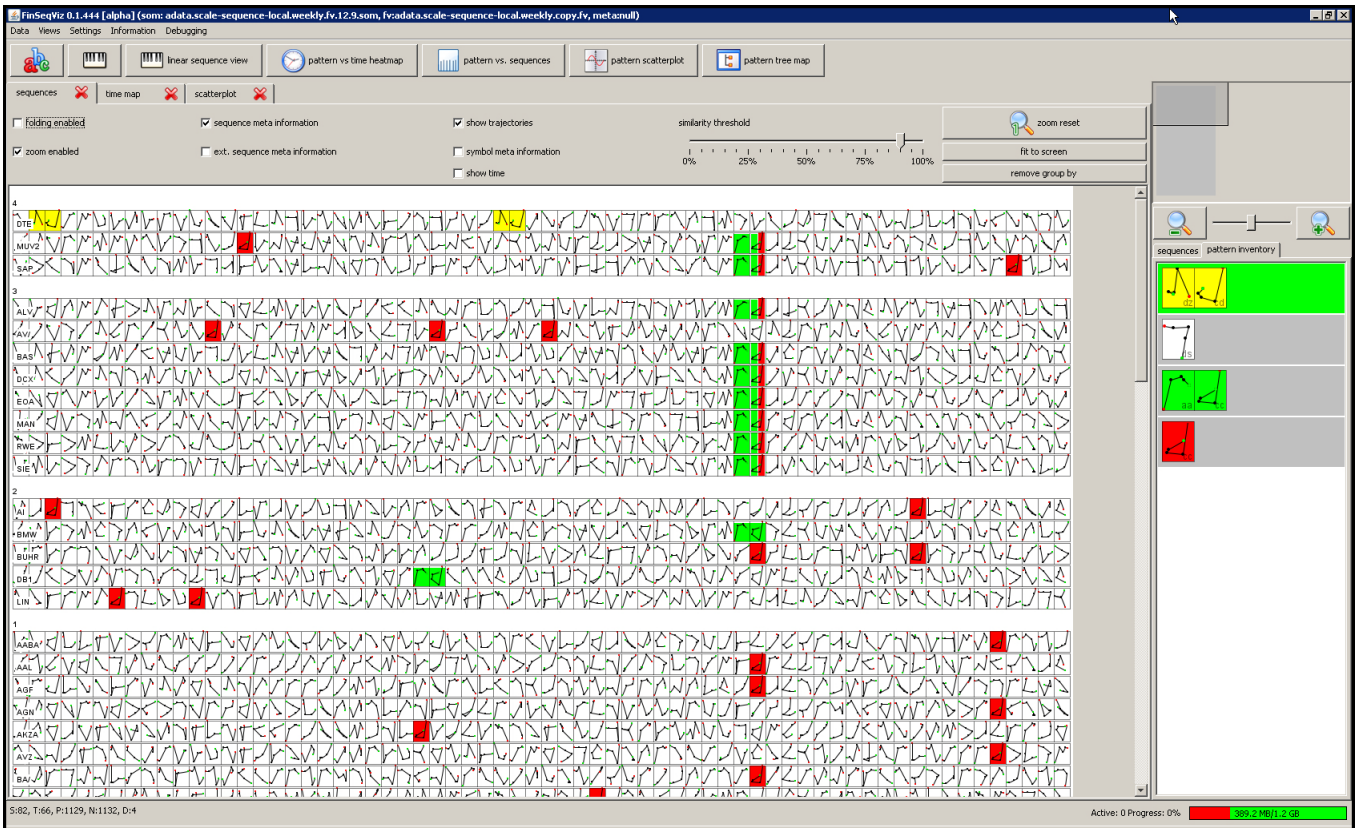

Abb. 4: Auf dieser Abbildung ist die Sequenzansicht und die dazu gehörende Bedienoberfläche zu sehen. Die visualisierten Sequenzen können wahlweise als Folge von Buchstaben oder als Folge visualisierter Objekte, in diesem Fall Trajektorien dargestellt werden. Am rechten Bildrand findet man vom Benutzer markierte Teilsequenzen in verschiedenen Farben. Dieses Element gilt global für alle integrierten Visualisierungen. Markierte Teilsequenzen sind auch in der Sequenzansicht hervorgehoben. Geteilte Zellen zeigen, dass dieses Objekt in verschiedenen Teilsequenzen auftritt. Nur teilweise eingefärbe Zellen zeigen, dass das Muster ähnlich der markierten Teilsequenz ist. Bis zu welchem Grad Ähnlichkeiten betrachtet werden sollen, lässt sich, wie auch verschiedene andere Visualisierungsparameter in der Kopfleiste einstellen.

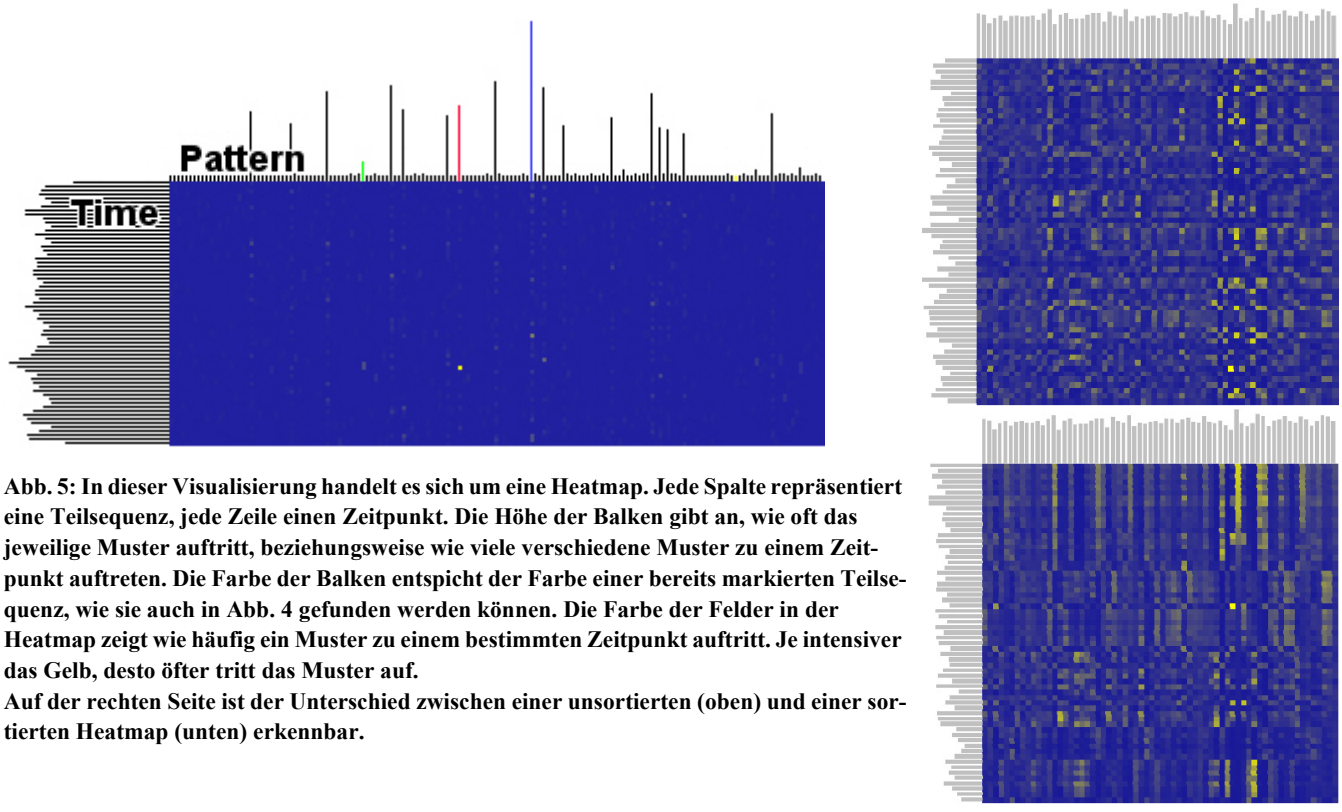

Abb. 5: In dieser Visualisierung handelt es sich um eine Heatmap. Jede Spalte repräsentiert eine Teilsequenz, jede Zeile einen Zeitpunkt. Die Höhe der Balken gibt an, wie oft das jeweilige Muster auftritt, beziehungsweise wie viele verschiedene Muster zu einem Zeitpunkt auftreten. Die Farbe der Balken entspicht der Farbe einer bereits markierten Teilsequenz, wie sie auch in Abb. 4 gefunden werden können. Die Farbe der Felder in der Heatmap zeigt wie häufig ein Muster zu einem bestimmten Zeitpunkt auftritt. Je intensiver das Gelb, desto öfter tritt das Muster auf.
Auf der rechten Seite ist der Unterschied zwischen einer unsortierten (oben) und einer sortierten Heatmap (unten) erkennbar.

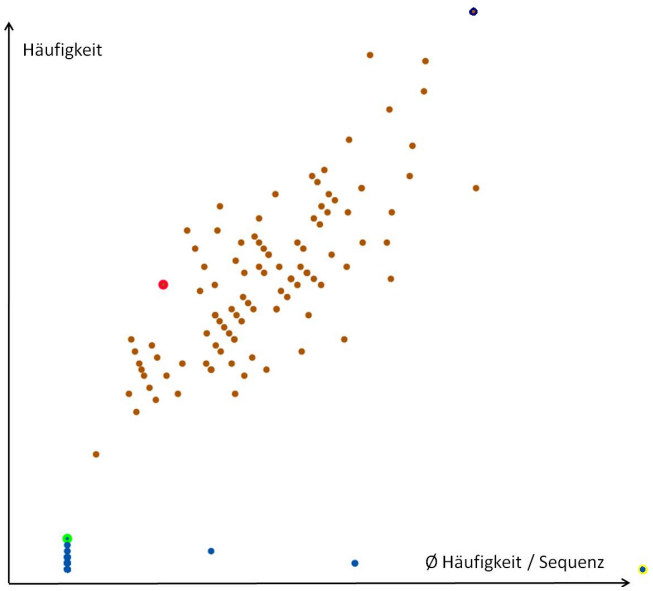

Abb. 6: Die Visualisierung zeigt ein Streudiagramm. Die X Achse repräsentiert das durchschnittliche Auftreten einer Teilsequenz pro Sequenz, die Y Achse repräsentiert die absolute Häufigkeit der Subsequenz. Die Farbe wird durch die Länge der Teilsequnz bestimmt.

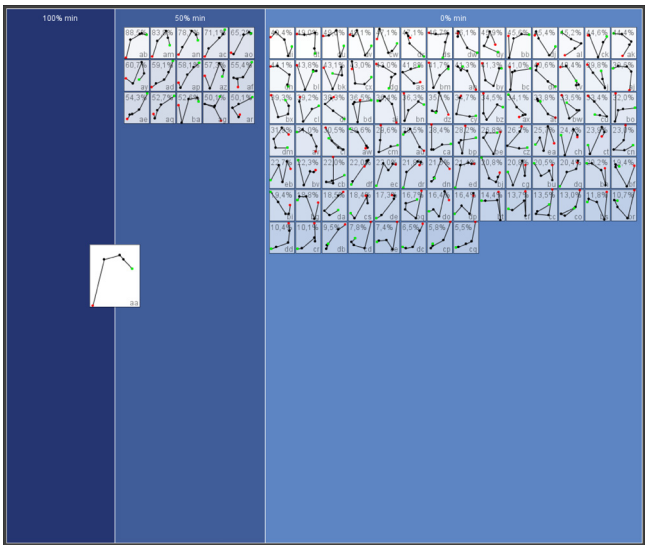

Abb. 7: Das Bild zeigt die in das System integrierte Ansicht zur Manipulation paarweiser Ähnlichkeiten. Größer hervorgehoben ist das ausgewählte Zeichen. Alle anderen Symbole sind zur besseren Übersichtlichkeit in farblich getrennte Gruppen gemäß Ähnlichkeit eingeteilt.
Durch Bewegen der einzelnen Elemente in andere Gruppen können die Ähnlichkeiten bearbeitet werden.

quenzen.

Die *Sequenzansicht* bietet eine Übersicht über die Zeichenfolgen und ermöglicht das Sortieren nach Metadaten (Abb. 4 & 8).
Die *Heatmap* verschafft durch eine Dimensionsreduzierung einen kompakten Überblick (Abb. 5).
Durch ein 2-dimensionales *Streudiagramm* lassen sich die Daten schnell auf Zusammenhänge verschiedener Merkmale untersuchen (Abb. 6).
Die einzelnen Bestandteile sind eng miteinander verknüpft, so sind alle Einstellungen zu Filtern und Markierungen global. Dadurch ist es dem Benutzer möglich, die in einer Visualisierung als besonderes Muster gefundenen Subsequenzen zu Markieren und deren Verhalten bezüglich anderer Aspekte zu untersuchen.

Die einzelnen Visualisierungen werden im Folgenden nacheinander im Detail vorgestellt und anschließend wird ihr Zusammenspiel anhand eines Beispiels verdeutlicht.

6.1 Visualisierungen im Detail

Die *Sequenzansicht* ist die Standardansicht, in der die Sequenzen als Folge von Buchstaben gezeigt werden. Alternativ können auch andere Objekte, wie in diesem Beispiel Trajektorien gerendert und zur Darstellung verwendet werden (Abb. 4).

Dem Benutzer werden unterschiedliche Interaktionsmöglichkeiten geboten. So können zum Beispiel Teilsequenzen markiert oder zusätzliche Metainformationen eingeblendet werden. Bei den in diesem Beispiel vorgestellten Daten kann das die Branche oder die Herkunft des Unternehmens sein. Aufgrund von Markierungen, Metadaten und anderen Informationen wie z.B. Ähnlichkeit kann die Liste der Sequenzen sortiert und gruppiert werden. Dabei können lange Sequenzen auf Wunsch umgebrochen werden, um eine bessere Übersichtlichkeit zu gewährleisten.

Die Visualisierung einer 2-dimensionalen *Heatmap* ist eine weitere Möglichkeit Informationen über die Daten zu erhalten. Hier werden Heatmaps verwendet, um die Häufigkeit von Teilsequenzen in Abhängigkeit von der Zeit (Timemap) oder der zugrundeliegenden Sequenz (Sequencemap) darzustellen. Manche Sequenzen wie in der Biologie und der Textanalyse, sind nicht Zeit- sondern Positionsabhängig. Diese dient dann als Achsenmerkmal. Abbildung 5, links zeigt die Häufigkeit von Teilsequenzen über die Zeit. Das Histogramm zur Linken zeigt die Anzahl der unterschiedlichen Teilsequenzen zum jeweiligen Zeitpunkt, das Histogramm am oberen Rand die Summe der Häufigkeit der jeweiligen Teilsequenz und der ihr ähnlichen Teilsequenzen. Bis zu welchem Grad Ähnlichkeiten dabei betrachtet werden, richtet sich nach dem zughörigen globalen Filter.

Grundsätzlich können alle hier erkennbaren Muster auch in der Sequenzansicht gefunden werden. Durch Weglassen einer Dimension, im Falle von Abb. 5 der Sequenzdimension, wird die Übersichtlichkeit erhöht. Während ein Muster in der Sequenzansicht sich auf z.B. drei Sequenzen verteilt und damit im schlechtesten Fall räumlich weit getrennt wird, erscheint es in der Heatmap in einer einzigen Zeile. Die Ansicht ist also im Vergleich zur Sequenzansicht komprimiert, um den Preis der fehlenden Information über die Verteilung des Musters auf verschiedene Sequenzen. Gleich verhält es sich bei der Heatmap, welche die Teilsequenzen in Relation zu den Sequenzen setzt. Hier wird auf die Zeitdimension verzichtet.

In der Timemap können die Teilse-

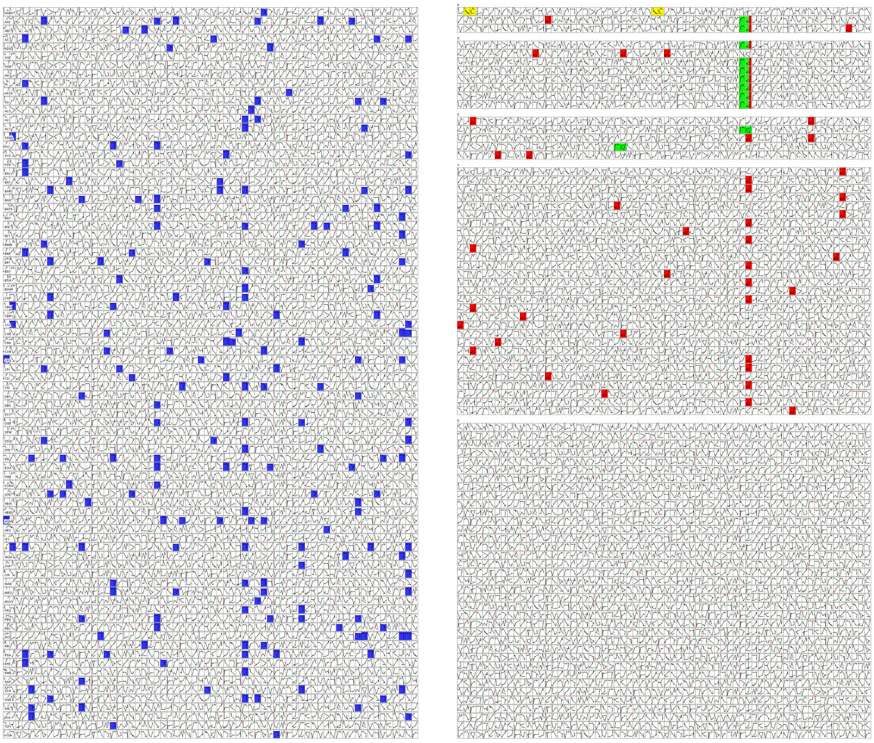

Abb. 8: Zu sehen sind Visualisierungen der analysierten Sequenzen. Auf der rechten Seite sind, wie auch schon in Abb. 4 die rot, gelb und grün markierten Teilsequenzen hervorgehoben. Sie waren entweder im Streudiagramm (grün, gelb) oder in der Heatmap (rot) aufgefallen. Die Zeilen sind nach Anzahl der enthaltenen Markierungen sortiert. Eine häufiges auftreten der markierten Sequenzen zum gleichen Zeitpunkt ist deutlich erkennbar. Auf der linken Abbildung sind die gleichen Sequenzen zu sehen. Blau hervorgehoben ist die, im Streudiagramm gefundene, am häufigsten auftretende Subsequenz. Da sie im Streudiagramm aber kein besonderer Ausreißer war und auch in der Heatmap nicht auffiel, ist es nicht verwunderlich, dass auch hier kein eindeutiges Muster zu erkennen ist.

quenzen nach einzelnen Zeitpunkten sortiert werden, so dass die Muster in der Reihenfolge ihres Vorkommens zum gegebenen Zeitpunkt sortiert werden. Abb. 5 rechts oben zeigt eine ungeordnete Heatmap und Abb. 5 rechts unten eine nach den Werten zum Zeitpunkt 48 sortierte Heatmap. Zeitliche Zusammenhänge von Teilsequenzen können so deutlich hervorgehoben werden. Identisch können in der Sequencemap die Teilsequenzen nach einzelnen Sequenzen sortiert werden.

Das *Streudiagramm* ist eine einfache Möglichkeit die Objekte bezüglich ihrer Abhängigkeit von verschiedenen Werten zu untersuchen. Dabei richtet sich die Anzahl der zu vergleichenden Parameter nach der Anzahl der Achsen. Zwar ist es möglich noch mehr Dimensionen, etwa wie in unserem Fall durch Einfärben der Objekte hinzuzufügen. Die Auffassungsgabe des Menschen ist allerdings begrenzt, weswegen für Visualisierungen auch meist nicht mehr als zwei Achsen gewählt werden.

Das Streudiagramm dient in dieser Arbeit dem Vergleich verschiedener Teilsequenzen, um Unterschiede und Gemeinsamkeiten zu beobachten. Dazu stehen eine Reihe an Bewertungsmöglichkeiten für die Teilsequenzen zur Verfügung.

Dazu gehören:

- Datenspezifische Werte wie in diesem Fall z.B. maximales Risiko oder durchschnittlicher Ertrag. Bei biologischen Sequenzen wären z.B. Eigenschaften der Aminosäuren wie Löslichkeit oder Ladung relevant.
- Häufigkeit der Teilsequenz
- Länge der Teilsequenzen
- Innere Ähnlichkeit: Ein Wert der die Ähnlichkeit der Symbole innerhalb einer Teilsequenz untereinander angibt.
  Bei Teilsequenzen der Länge 1 ist er nicht aussagekräftig und ergibt immer 1. Gleiches gilt für Teilsequenzen, deren Symbole als identisch definiert sind.
- Zeitbezogene Werte wie zum Beispiel Varianz, arithmetisches Mittel oder Anzahl der Zeitpunkte zu denen die Teilsequenz auftritt.
- Sequenzbezogene Werte wie die Anzahl der Sequenzen in denen die Teilsequenz auftritt oder deren durchschnittliches Auftreten in einer Sequenz.

Die Visualisierung stellt drei Dimensionen zur Verfügung, die vertikale und die horizontale Achse, sowie die Farbe der Markierung des Punktes. Jede Dimension ist in der Lage, alle Bewertungsmöglichkeiten darzustellen und kann vom Benutzer frei belegt werden.

Zusätzlich steht auch eine, in das System integrierte Visualisierung der paarweisen Ähnlichkeit zur Verfügung. Mit Hilfe dieser können schnell einzelne Distanzen manipuliert werden (Abb. 7). Zur besser Übersichtlichkeit sind die Objekte dabei in beliebig definierbare Ähnlichkeitskategorien unterteilt.

7 Anwendung

Im Folgenden werden wir anhand der vorgestellten Daten beispielhaft einige mögliche Analyseschritte erläutert. Dabei wird Risiko und Ertrag Der Wertpapiere von 60 Firmen über einen Zeitraum von 500 Tagen untersucht. Ziel der Sequenzanalyse ist es, interessante Teilsequenzen zu finden.

Zu Beginn der Analyse ist es vorteilhaft, sich in abstrakteren Visualisierungen einen Überblick über den Datensatz zu verschaffen. In diesem Fall bieten sich das Streudiagramm und die Heatmaps an, um erste Auffälligkeiten zu finden und die entsprechenden Teilsequenzen zu markieren. Diese können in späteren Analyseschritten, zum Beispiel in der Sequenzansicht genauer auf ihr Verhalten im Gesamtzusammenhang untersucht werden.

In diesem Beispiel wurden im Streudiagramm (Abb. 6) die Häufigkeit des

Auftretens als Y-Achse und das durchschnittliche Auftreten pro Sequenz als X-Achse gewählt. Die Farbe der Punkte repräsentiert die Länge der Teilsequenz. Auf den ersten Blick fällt auf, dass die Merkmale der Achsen stark korrelieren und kurze häufiger als längere Sequenzen auftreten. Daneben Fallen einige andere Aufälligkeiten ins Auge.

Als erstes wurde der Ausreißer rechts unten (gelb) markiert. Hinzu kommen die Subsequenz mit dem absolut häufigsten Auftreten (blau) und die Subsequenz mit dem häufigsten Auftreten aus dem Pool derer, die nur in wenigen Sequenzen auftreten (grün).

In der Visualisierung von Zeit vs. Teilsequenz (Abb. 5) fällt ein besonders hell leuchtender Punkt auf. Hier tritt eine Teilsequenz zu einem Zeitpunkt sehr häufig auf. Auch sie wird zu dem Pool der markierten Objekte hinzugefügt (rot).

In der Sequenzansicht (Abb. 4 & Abb. 8 rechts) wird deutlich, dass viele Firmen das gleiche oder ein ähnliches Verhalten zum gleichen Zeitpunkt, in Woche 46 & 47 aufweisen. Die rot markierte Teilsequenz ist offensichtlich ein Teil der grün markierten und tritt zu diesem Zeitpunkt aber häufiger als diese auf.

Die gelbe Teilsequenz tritt exklusiv in einer Zeile, also einer Firma, in diesem Fall ein deutscher Telefonanbieter auf.

Nächster Schritt des Finanzanalysten wäre es folglich zum einen den Telefonanbieter und seine Besonderheiten zu untersuchen und zum anderen die Ursache für das gemeinsame Verhalten der Unternehmen in Woche 46 & 47 zu prüfen.

8 Zusammenfassung und Ausblick

Die Analyse großer Mengen von Sequenzdaten ist ein aktuelles Problem. Es gibt bereits viele Ansätze und Systeme zur Bearbeitung spezieller Teilprobleme und genau definierter Sequenzarten.

Unser System erreicht durch die Kombination verschiedener Techniken und die Einbettung in eine mächtige Visual Analytics Pipeline eine große Flexibilität. Die unterschiedlichen Visualisierungen für unterschiedliche Aspekte des Datensatzes sind eng miteinander verknüpft und können sich so ideal ergänzen.

Dabei wird der Benutzer sowohl in der Datenaufbereitung als auch in allen Analyseschritten so stark wie möglich eingebunden. Durch diese effektive Kombination der Fähigkeiten von Mensch und Maschine kann ein sehr überzeugendes Ergebnis erzielt werden. Dabei werden dem Benutzer sehr viele Freiheiten bei Werkzeug- und Parameterwahl gelassen. Großer Handlungspielraum birgt immer die Gefahr, dass der Benutzter sich in den Möglichkeiten verliert und ein anfangs zielgerichtetes Arbeiten mit den Daten zu einer randomisierten Suche wird. Diesem Problem begegnen wir durch die Visualisierung jedes Schrittes. Dadurch werden Informationen und Rückmeldungen zu den Aktionen des Analysten geliefert.

Bei der hier vorgestellten Arbeit sehen wir einige nötige und vielversprechende Weiterentwicklungsmöglichkeiten. Auf Architekturseite ist dies die Möglichkeit auch im späteren Verlauf der Analyse noch einzelne Sequenzen hinzufügen zu können, um sie mit dem vorhandenen Sequenzpool zu vergleichen.

Zur Verbesserung der automatischen Analyse sind vor allem bessere Methoden zur Bewertung von Teilsequenzen zu integrieren. Dadurch können dem Benutzer wertvolle Hinweise auf potentiell interessante Merkmale gegeben werden. Entsprechende Vorarbeiten [Ding et al. 2008] sollten in unserem System berücksichtigt werden. Des weiteren Könnten Benutzerstudien weitere Erkenntnisse zur Verbesserung des Systems liefern.

9 Danksagung

Diese Arbeit wurde im Rahmen des Teilprojekts „Visual Feature Space Analysis“ des Schwerpunktprogramms 1335: „Scalable Visual Analytics“ durch die Deutsche Forschungsgemeinschaft (DFG) unterstützt.

10 Literatur

[Chang et al. 2007] *Chang, R.; Ghoniem, M.; Kosara, R.; Ribarsky, W.; Jing Yang; Suma, E.; Ziemkiewicz, C.; Kern, D.; Sudjianto, A.:* WireVis: Visualization of Categorical, Time-Varying Data From Financial Transactions. Visual Analytics Science and Technology (VAST) 2007, S. 155-162.

[Ding et al. 2008] *Ding, H.; Trajcevski, P. Scheuermann, P.; wang, X.; Keaogh, E.:* Querying and mining of time series data: experimental comparison of representations and distance measures. In: Proc. VLDB Endow. 2008; S. 1542-1552.

[Henikoff, Henikoff 1992] *Henikoff, S.; Henikoff, JG.:* Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences USA, 1992, 89(22):10915-9.

[Hochheiser & Shneiderman 2008] *Hochheiser, H.; Shneiderman, B,:* Dynamic Query Tools for Time Series Data Sets, Timebox Widgets for Interactive Exploration. Information Visualization 3, 2004; S. 1-18.

[Keim et al. 2006] *Keim, D.; Oelke, D.; Truman, R.; Neuhaus, K.:* Finding Correlations in Funktionally Equivalent Proteins by Integrationg Automated and Visual Data Exploration. In: Proceedings of the Sixth IEEE Symposium on Bioinformatics and Bioengineering 2006. S. 183-192

[Keim et al. 2008] *Keim, D.; Andrienko, G.; Gorg, C.; Kohlhammer, J.; Melancon, G.:* Visual Analytics: Definition, Process, and Challenges. In Information Visualization 2008; S. 154-175

[Kohonen 2001] *Kohonen. T.:* Self-Organizing Maps 3. Edition. Springer-Verlag, Berling, Deutschland, 2001.

[Korf et al. 2003] *Korf I.; Yandell M.; Bedel* J.: Blast. O'Reilly, Sebastopol, CA, USA, 2003.

[Larin et al. 2007] *Heuer, A.; Saake, G.:* Clustal W and Clustal X version 2.0. In: Bioinformatics, 2007, Band 23, S. 2947-2948.

[Lin et al. 2005] *Lin, J.; Keogh, E.; Lonardi, S.:* Visualizing and discovering non-trivial patterns in large time series databases. In: Information Visualization 4, 2005; S. 61-82.

[Pósfei et al. 1994] *Pósfei, J.; Száraz, Z.; Roberts R.J.:* VISA: Visual Sequence Analysis for the comparison of multiple amino acid sequences. In: Bioinformatics, 1994. S. 537-544.

[Schreck et al. 2008] *Schreck, T.; Bernhard, J.; Tekusová, T.; Kohlhammer, J.:* Visual cluster analysis of trajectory data with interactive Kohonen Maps. In: Visual Analytics Science and Technology (VAST) 2008. S. 3-10.

[Tekusová, Kohlhammer 2007] Tekusová, *T.; Kohlhammer, J.:* Applying animation to the visual analysis of nancial time-dependent data. In: Proc. Int.Conf. on Information Visualization (IV), 2007. S. 101-108.

[Wong et al. 2000] *Wong, P. C.; Cowley W.; Foote H.; Jurrus E.; Thomas J.:* Visualizing sequential patterns for text mining. In: IEEE Symposium on Information Visualization 2000. S. 105-111