Ähnlichkeitssuche in der Datenwolke

Thomas Niederberger,
wissenschaftlicher Assistent,
thomas.niederberger@zhaw.ch

Die moderne Informationstechnologie erlaubt es, auf einfache Art und Weise grosse Datenmengen zu sammeln. Da fällt es schwer, die Übersicht nicht zu verlieren. Um einen Überblick über die innere Struktur von Datenmengen zu erhalten, werden häufig Clusteringverfahren eingesetzt. Am Institut für Angewandte Simulation (IAS) werden solche Analysen durchgeführt. Mit Hilfe einer neuen Software, welche am Institut entwickelt wurde, kann der Analyseprozess zusätzlich unterstützt und verbessert werden.

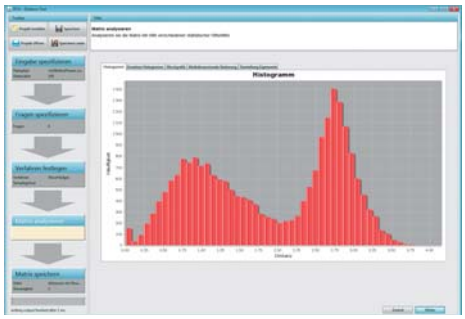
Clusteringverfahren sind ein Mittel, um Strukturen in grossen Datenmengen zu erkennen und daraus Schlüsse zu ziehen. Viele dieser Verfahren setzen voraus, dass für die Daten eine sogenannte Distanzmatrix berechnet wurde. Die Distanz zweier Datenelemente, beispielsweise zweier Fragebögen, gibt an, wie ähnlich sich die Elemente sind: Kleine Distanzen stehen für eine grosse Ähnlichkeit, grosse Distanzen für eine geringe Ähnlichkeit. Zwar gibt es bereits Statistikpakete, welche die Berechnung derartiger Matrizen anbieten, aber ihre Möglichkeiten sind jeweils eingeschränkt. Aus diesem Grund wurde am IAS im Rahmen einer wissenschaftlichen Studie eine Software entwickelt, um die Berechnung von Distanzmatrizen zu vereinfachen und zu unterstützen.

Schrittweise zum Ziel

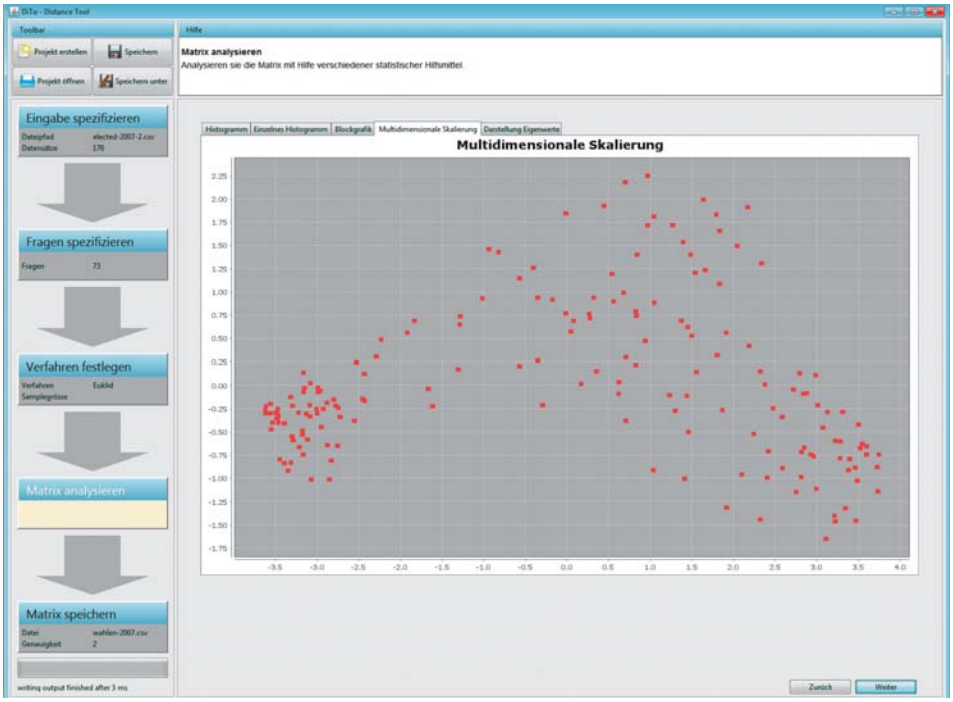
Die neue Software beinhaltet fünf Teilschritte, in denen der Benutzer durch den Berechnungsprozess geführt wird: von der Auswahl der Daten über die Wahl des Berechnungsverfahrens bis zur abschliessenden Speicherung der Daten. Über zahlreiche Parameter kann der Benutzer die Berechnung nach seinen Wünschen gestalten. Dadurch wird es möglich, auf schnelle Art und Weise verschiedene Vergleichsmasse auszuprobieren und auf den Datensatz anzuwenden. Gleich nach Durchführung der Berechnungen werden automatisch erste Visualisierungen erstellt. Diese Visualisierungen zeigen bereits Strukturen in den Daten und unterstützen dadurch den Benutzer beim Entscheid, ob sich die Ergebnisse für eine weitergehende Datenanalyse eignen. Die Software lässt sich für jede Art von Daten nutzen, sofern diese numerisch codiert werden können. Dadurch verfügt sie über ein sehr breites Einsatzspektrum: Geographische Daten, chemische Datensätze und politische Fragebögen wurden bereits erfolgreich analysiert.

Analyse sozialwissenschaftlicher Zusammenhänge

In Zusammenarbeit mit dem Institut für Biomedizinische Ethik der Universität Zürich wird die Software aktuell eingesetzt, um eine Kohärenzanalyse verschiedener Länder durchzuführen. Anhand einer internationalen, sozialwissenschaftlichen Studie werden dabei die Meinungsspektren in der Bevölkerung zu gesellschaftlichen Themen quantifiziert und zwischen den Ländern verglichen.



Ein Histogramm mit mehreren Häufungen deutet darauf hin, dass in den Daten mindestens zwei Cluster vorhanden sind.



Die Analyse politischer Fragebögen führt zu einer Verteilung mit Links-Rechts- und Liberal-Konservativ-Schema.

Forschungsprojekt

Optimale Ähnlichkeitsbestimmung für neuartige Kohärenzanalyse und Clustering

Leitung:	Thomas Ott
Projektdauer:	6 Monate
Partner:	Dr. Markus Christen, Institut für Biomedizinische Ethik, Universität Zürich
Förderung:	Hasler Stiftung, Bern
Projektvolumen:	CHF 44 000