# Neue Herausforderungen für dichtebasiertes Clustering

Peer Kröger

Institut für Informatik, LFE Datenbanksysteme
Ludwig-Maximilians-Universität München
kroegerp@dbs.ifi.lmu.de

**Abstract:** *Knowledge Discovery in Databases* (KDD) ist der Prozess der semi-automatischen Extraktion von Wissen aus Datenbanken, das gültig, bisher unbekannt und potentiell nützlich für eine gegebene Anwendung ist. Clustering ist ein möglicher Teilschritt in diesem Prozess. Dabei sollen die Objekte einer Datenbank anhand ihrer Ähnlichkeit in Gruppen (Cluster) partitioniert werden. Das dichtebasierte Clustermodell ist unter einer Vielzahl anderer Clustering-Ansätze eine der erfolgreichsten Methoden zum Clustering. Dieser Beitrag befasst sich mit neuartigen Problemfeldern des Clusterings. Anhand aktueller Anwendungen werden neue Herausforderungen an den dichtebasierten Clustering-Ansatz identifiziert und innovative Lösungen dazu erarbeitet. Details zu den vorgestellten Techniken können in [Krö04] gefunden werden.

## 1 Einleitung

Dank stetig wachsenden Datensammlungen haben Techniken zu Knowledge Discovery in Datenbanken (KDD) in den letzten Jahren immer größere Bedeutung erlangt. KDD ist der Prozess der (semi-)automatischen Extraktion von Wissen aus Datenbanken, das gültig, bisher unbekannt und potentiell nützlich für eine gegebene Anwendung ist. Der zentrale Schritt des KDD-Prozesses ist das Data Mining, in dem ein Algorithmus zur Extraktion von Mustern aus der Datenmenge angewandt wird. Eine der wichtigsten Aufgaben des Data Mining ist Clustering. Dabei sollen die Objekte einer Datenbank so in Gruppen (*Cluster*) partitioniert werden, dass Objekte eines Clusters möglichst ähnlich und Objekte verschiedener Cluster möglichst unähnlich zu einander sind. Objekte, die keinem Cluster zugeordnet werden können, werden als *Rauschen* bezeichnet. Clustering spielt in verschiedenen Anwendungsbereichen eine entscheidende Rolle in der Datenanalyse. Im Bereich der Molekularbiologie ist Clustering eine wichtige Technik zur Analyse von Genexpressions-Experimenten und damit für das Verstehen des komplexen Zusammenspiel der Genregulation in Zellen. Im Bereich der Medizin ist die Clusteranalyse dieser Expressionsdaten der zentrale Schritt für die Erforschung, Klassifizierung und Diagnose von genetischen Erkrankungen wie Krebs. Hier ermöglicht die Clusteranalyse vernünftige Diagnoseentscheidungen sowie den rechnergestützten Medikamentenentwurf, und kann dadurch helfen, unnötige Operationen zu vermeiden und die medizinische Forschung nachhaltig zu stimulieren. Im Bereich der Fertigungsindustrie liefert die Clusteranalyse einen entschei-

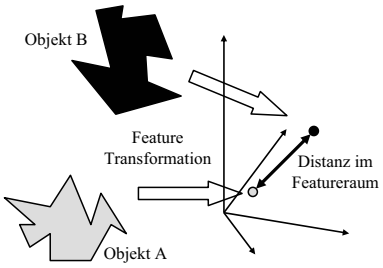


Abbildung 1: Das Prinzip der Feature Transformation.

denden Stellhebel zur Produktivitätssteigerung und Kostensenkung, denn sie erlaubt einen schnellen Überblick über bereits gefertigte und wieder verwendbare Teil. Das dichtebasierte Clustermodell und die darauf aufbauenden Algorithmen DBSCAN [EKSX96] und OPTICS [ABKS99] zählen zu den erfolgreichsten Methoden zum Clustering. Im Rahmen dieses Beitrags werden innovative Erweiterungen des dichtebasierten Clustering-Ansatzes vorgestellt. Dazu werden eine Reihe von neuen Herausforderungen für das dichtebasierte Clustermodell anhand moderner, industrie-naher Anwendungen erarbeitet und innovative Lösungen hierfür vorgeschlagen. Nach einer Einführung in grundlegende Konzepte des dichtebasierten Clusterings in Kapitel 2 wird zunächst die Entwicklung des industriellen Prototyps BOSS (Browsing OPTICS plots for Similarity Search) in Kapitel 3 beschrieben. BOSS erlaubt einen schnellen Überblick über die Objekte in einer Datenbank und ist dadurch unter anderem für die Fertigungsindustrie interessant. Kapitel 4 widmet sich dann den Problemen die hochdimensionale Daten bei der Clusteranalyse mit sich bringen. Die Lösungen, die in diesem Bereich vorgeschlagen werden, finden Einsatz in den oben beschriebenen Anwendungsbereichen der Medizin und Molekular Biologie. Der Artikel schließt mit einer kurzen Zusammenfassung in Kapitel 5.

## 2 Dichtebasiertes Clustering

Ein wichtiger Aspekt beim Clustering ist die Definition der Ähnlichkeit zweier Datenobjekte, wie z.B. Autoteile oder Proteinmoleküle. Dieser Beitrag baut auf dem featurebasierten Ansatz zur Modellierung der Ähnlichkeit von Datenobjekten auf. Die Grundidee eines featurebasierten Ähnlichkeitsmodells ist es, für jedes Datenobjekt $d$ numerischen Merkmale (*Features*) zu extrahieren. Durch diese *Feature Transformation* genannte Extraktion werden Objekte auf Featurevektoren (Punkte) im $\mathbb{R}^d$ abgebildet. Die Ähnlichkeit zwischen zwei Objekten wird über deren Nähe im Featureraum definiert. Das Prinzip der Feature Transformation ist in Abbildung 1 illustriert. Die meisten Clustering Algorithmen arbeiten auf Featurevektoren und benutzen typischerweise die Euklidische Distanz als Maß für die Ähnlichkeit, bzw. die Unähnlichkeit, zweier Datenobjekte. Man unterscheidet partitionierende Algorithmen, die eine eindeutige Zuordnung aller Datenobjekte zu genau einem Cluster erzeugen, und hierarchischen Algorithmen, die eine hierarchische Zerlegung der Datenobjekte berechnen, die meist als Baum (*Dendrogramm*) visualisiert

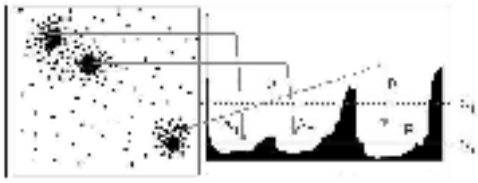

Abbildung 2: Ein Erreichbarkeitsdiagramm (rechts) für einen 2-dimensionalen Datensatz (links).

wird.

In [EKSX96] werden die Grundlagen des dichtebasierten Clustermodells formalisiert und der partitionierende Algorithmus DBSCAN vorgestellt. Das dichtebasierte Clustermodell definiert Cluster als dichte Regionen im Featureraum — Regionen, in denen die Datenobjekte sehr ähnlich sind, d.h. die Distanz zwischen den Objekten (Punkten) ist klein —, abgegrenzt von Regionen geringerer Dichte. Dichte wird dabei relativ zu einem Punkt $p$ über zwei Parameter bestimmt: $\varepsilon$ bestimmt ein Volumen, die so genannte $\varepsilon$-Nachbarschaft von $p$ bei der es sich um eine Hyperkugel mit Radius $\varepsilon$ um $p$ handelt. Der zweite Parameter, $minPts$, bestimmt eine minimale Anzahl von Punkten, die innerhalb dieses Volumens liegen muss. Enthält die $\varepsilon$-Nachbarschaft von $p$ mindestens $minPts$ andere Punkte, so ist $p$ ein *Kernpunkt*. Zusätzlich zur Dichte soll ein Cluster auch "verbunden" sein, d.h. benachbarte Kernpunkte sollen zu einem Cluster zusammengeschlossen werden. Ebenso sollen Punkte, die am Rand eines Clusters liegen und möglicherweise keine Kernpunkte mehr sind, auch dem Cluster zugeordnet werden. Daher wird in [EKSX96] die direkte Dichteerreichbarkeit eingeführt. Ein Punkt $q$ ist von einem Kernpunkt $p$ aus direkt dichteerreichbar, wenn $q$ in der $\varepsilon$-Nachbarschaft von $p$ liegt. Die transitive Hülle der direkten Dichteerreichbarkeit definiert die (allgemeine) Dichteerreichbarkeit. Schließlich wird die Dichteverbundenheit als symmetrische Erweiterung der Dichteerreichbarkeit formalisiert. Basierend auf der Dichteverbundenheit kann formal gezeigt werden, dass man dichtebasierte Cluster finden kann, indem man von einem beliebigen Kernpunkt des Clusters alle dichteerreichbaren Punkte bestimmt. Der Algorithmus DBSCAN arbeitet nach diesem Prinzip und hat mehrere Vorteile gegenüber vergleichbaren partitionierenden Verfahren, insbesondere werden Cluster unterschiedlicher Größe und Form erkannt.

Der dichtebasierte Clustering-Ansatz wird in [ABKS99] für das hierarchische Clustering erweitert. Basierend auf dem Konzept der *Kerndistanz* wird die *Erreichbarkeitsdistanz* als dichtebasierte Distanzfunktion zwischen Clustern eingeführt. Der resultierende Algorithmus OPTICS erzeugt eine *Clusterordnung*, anhand derer die hierarchische Clusterstruktur sehr übersichtlich dargestellt werden kann. Die Visualisierung erfolgt durch ein *Erreichbarkeitsdiagramm*, das die Clusterordnung der Objekte entlang der x-Achse und deren Erreichbarkeitsdistanz entlang der y-Achse darstellt. Täler in dieser Darstellung repräsentieren dichtere Bereiche (Cluster). Erreichbarkeitsdiagramme ermöglichen — im Unterschied zu Dendrogrammen — eine übersichtliche Repräsentation der hierarchischen Struktur sogar bei sehr großen Datenmengen. Ein Beispiel für ein Erreichbarkeitsdiagramm ist in Abbildung 2 gezeigt.

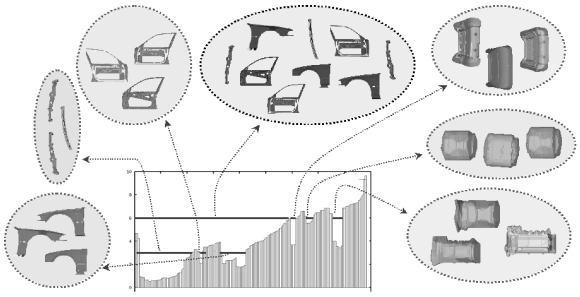

Abbildung 3: Die Idee hinter BOSS.

## 3 Dichtebasiertes Clustering zur Unterstützung von Ähnlichkeitssuche in Datenbanken

Dichtebasiertes hierarchisches Clustering bietet das Potential mehrere wichtige Anwendungsszenarien bei der Ähnlichkeitssuche in Datenbanken zu unterstützen. Zu diesem Zweck haben wir die Entwicklung des Prototyps BOSS (Browsing OPTICS plots for Similarity Search) vorgeschlagen [BKK+04, Krö04]. BOSS ist ein erster Beitrag zu einer umfassenden, skalierbaren und verteilten Softwarelösung, die eine Nutzung der Effizienzvorteile und die analytischen Möglichkeiten des dichtebasierten, hierarchischen Clustering-Algorithmus OPTICS für ein breites Publikum ermöglichen. Basierend auf der von OPTICS erzeugten hierarchischen Clusterordnung wird dabei automatisch eine Clusterhierarchie visualisiert. BOSS ermöglicht dem Anwender, interaktiv durch diese Hierarchie zu navigieren. Dadurch kann der Anwender

- einen schnellen Überblick über alle Objekte in der Datenbank bekommen,
- eine interaktive, visuelle Suche nach Datenbankobjekten starten, ohne das Anfrageobjekt genauer spezifizieren zu müssen,
- eine semi-automatische Clusteranalyse auf einer Datenbank durchführen und
- bequem verschiedene Ähnlichkeitsmodelle evaluieren, d.h. überprüfen, ob das Ähnlichkeitsmodell die intuitive Ähnlichkeitsvorstellung repräsentiert.

Abbildung 3 zeigt schematisch die Idee hinter BOSS: Aus der Clusterordnung werden Clusterhierarchien extrahiert, in der der Anwender interaktiv navigieren kann.

Zur Entwicklung von BOSS wurden drei entscheidende Erweiterungen vom dichtebasierten hierarchischen Clustering-Algorithmus OPTICS benötigt:

**Inkrementelles Clustering.** Um die Einsetzbarkeit von BOSS in einem dynamischen Datenbankszenario zu ermöglichen, muss die hierarchische Clusterstruktur, d.h. die Clusterordnung, konsistent gehalten werden. Bei Updateoperationen wie Einfügungen neuer

oder Löschungen bereits vorhandener Datenobjekte muss die Clusterordnung effizient angepasst werden. Da eine komplette Neuberechnung der Clusterordnung nach jeder Updateoperation nicht effizient ist, schlagen wir eine inkrementelle Version von OPTICS [KKG03, Krö04] vor, um nach einer Updateoperation die hierarchische Clustering Struktur effizient zu reorganisieren. Dazu werden die Grundkonzepte des hierarchischen dichtebasierten Clustermodells erweitert und ein Algorithmus zum inkrementellen Aktualisieren der Clusterstruktur nach einer Einfügung bzw. Löschung entwickelt. Grundidee dabei ist, nur die Teile der ursprünglichen Clusterordnung zu reorganisieren, die von der Updateoperation betroffen sind. Die anderen Teile werden unverändert in die neue Clusterordnung übernommen. Diese grundlegenden inkrementellen Verfahren werden zusätzlich erweitert, um auch größere Mengen von Einfügungen oder Löschungen in einem Schritt (sog. Bulk Update Modus) effizient zu verarbeiten. Das Potential der neuen Ansätze wird mit Laufzeitexperimenten auf synthetischen Datensätzen mit 100.000 bis 500.000 2-dimensionalen Punkten und einem realen Datensatz mit ca. 100,000 64-dimensionalen Farbhistogrammen, die TV-Schnappschüsse repräsentieren, demonstriert. Es konnte gezeigt werden, dass die vorgeschlagenen, inkrementellen Algorithmen deutliche Beschleunigungsfaktoren gegenüber dem originalen OPTICS-Algorithmus erzielen.

**Clustererkennung und -repräsentation.** Um dem Anwender eine Navigation durch eine Clusterhierarchie zu ermöglichen, muss diese Hierarchie zunächst automatisch abgeleitet werden. Die dabei entstehenden Cluster sollen dem Anwender durch geeignete Repräsentanten angezeigt werden. Daher schlagen wir einen neuen Algorithmus zur automatischen Clusterextraktion aus hierarchischen Repräsentationen und zwei innovative Methoden zur automatischen Auswahl geeigneter Clusterrepräsentanten vor [BKKP04, Krö04]. Das Verfahren zur Clusterextraktion verallgemeinert bestehenden Verfahren und beseitigt deren Nachteile bezüglich ihrer Verwendbarkeit für BOSS. Mehrere vergleichende Experimente auf zwei realen Datensätzen (bestehend aus ca. 200 CAD Autoteilen, bzw. aus ca. 2.000 Proteinstrukturen) bestätigen den Nutzen des neuen Verfahrens. Die Ansätze zur automatischen Auswahl von Clusterrepräsentanten werden ebenfalls vergleichend auf den CAD und Protein Daten evaluiert. Die Ergebnisse dieser Evaluation zeigen die Einsetzbarkeit der entwickelten Methoden in der Praxis.

**Der Prototyp BOSS** In [BKK+04] werden einige Details zur Systemarchitektur des BOSS Prototypen gezeigt. [Krö04] beschreibt zwei exemplarische Anwendungen von BOSS. Insbesondere wird BOSS dabei zur semi-automatischen Clusteranalyse auf einer Proteinstruktur-Datenbank eingesetzt und zur Evaluation von vier Ähnlichkeitsmodellen [KKM+03, KBK+03] verwendet. Die beschriebenen Anwendungen zeigen den praktischen Nutzen von BOSS auf.

# 4 Dichtebasiertes Clustering in hochdimensionalen Daten

Eine weitere Herausforderung für Clustering-Verfahren stellen hochdimensionale Featureräume dar. Reale Datensätze beinhalten, dank moderner Verfahren zur Datenerhebung, häufig sehr viele Merkmale. Teile dieser Merkmale unterliegen oft Rauschen oder Abhängigkeiten, d.h. sie sind irrelevant für das Clustering da sie redundant und/oder korreliert sind. Meist können diese irrelevanten Features nicht im Vorfeld identifiziert werden, da diese Effekte jeweils in Teilen der Datenbank unterschiedlich ausgeprägt sind. Für verschiedene Cluster können unterschiedliche Teilmengen von Merkmalen relevant bzw. irrelevant sein. Ein globales Verfahren zur Featureselektion oder Korrelationsanalyse kann in diesem Fall nichts ausrichten. Daher muss die Wahl der Features in das entsprechende Data Mining Verfahren integriert werden.

## 4.1 Anwendungen

Im folgenden werden zwei exemplarische Anwendungen aus der Biologisch-medizinischen Forschung vorgestellt und verschiedene Probleme diskutiert, die beim Clustering hochdimensionaler Daten auftreten können.

**Genexpressionsanalyse.** DNA Microarray Technologie ermöglichen die zeitgleiche Erfassung des Expressionlevels tausender von Genen in organischen Zellen. Der Expressionslevel eines Gens erlaubt Aussagen über die Menge an Genprodukten, die aus diesem Gen entstehen und gibt daher einen Überblick über den Zustand einer Zelle. Expressionsdaten erfassen für eine Menge von Genen deren Expressionslevel bezüglich einer Menge von Proben (z.B. Zeitpunkten, Patienten, Organen). Die Clusteranalyse dieser Daten erlaubt Einblicke in die funktionellen Zusammenhänge von Genen und spielt sowohl in der biologischen Grundlagenforschung (z.B. Erforschung der Genregulation) als auch der medizinisch-pharmazeutischen Industrie (z.B. individuelle Arzneimittelverträglichkeit, Tumorforschung) eine entscheidende Rolle. Die Clusteranalyse wird allerdings durch die hohe Dimensionalität und des daraus resultierenden hohen Rauschanteils der Rohdaten erschwert. Hinzu kommt, dass verschiedene Teilmengen von Genen für unterschiedliche Phänotypen, d.h. Erscheinungsformen, verantwortlich sind. Ein Teil der Gene ist z.B. verantwortlich für das Geschlecht der Testpersonen, in einem anderen Teil spiegelt sich die Haarfarbe oder das Alter wieder, an wiederum anderen Genen kann man die Personen nach Tumorart unterscheiden. Mit anderen Worten können die Patienten ganz unterschiedlich in Cluster gruppiert werden, wenn man unterschiedliche Genmengen (Merkmale) zugrunde legt. Meist ist man an möglichst allen unterschiedlichen Gruppierungsmöglichkeiten, die die Rohdaten bieten, interessiert.

**Metabolomanalyse.** Mit modernen Screening-Verfahren der Medizin werden genetische und metabolische Krankheiten erforscht. Patienten wird dabei Blut abgenommen, und die Konzentration verschiedener Metabolite (Stoffwechselprodukte) werden gemes-

sen. Die Analyse dieser Daten erfordert ein Clustering der Patienten anhand der Metabolitkonzentrationen in homogene Krankheitsgruppen. Verschiede Krankheiten äußern sich wiederum dadurch, dass unterschiedliche Teilmengen der Metabolite (linear) korreliert sind. Mit anderen Worten sind auch in dieser Anwendung wieder verschiedene Merkmalsmengen für die einzelnen Cluster relevant.

In [Krö04] werden innovative Erweiterungen des dichtebasierten Clustermodells für hochdimensionale Daten vorgestellt, die die Featureauswahl eng mit dem Data Mining verknüpfen und insbesondere Techniken darstellen, die auf die oben beschriebenen Applikationen anwendbar sind.

### 4.2 Dichtebasiertes Subspace Clustering

Die erste Erweiterung des dichtebasierten Clustering-Ansatzes widmet sich dem so genannten Subspace Clustering. Diese Erweiterung ist unter anderem besonders für die Analyse von Genexpressionsdaten geeignet. Beim Subspace Clustering besteht das Ziel der Clusteranalyse darin, alle Cluster in allen Teilräumen des ursprünglichen Featureraumes zu berechnen. Damit wird speziell auch die Problematik berücksichtigt, dass Datenobjekte in verschiedenen Unterräumen des Featureraumes unterschiedlich gruppiert werden können. Z. B. kann ein Patient anhand einer Genmenge entsprechend seines Alters oder anhand einer anderen Genmenge entsprechend seines Tumortyps eingeteilt werden.

In [Krö04] wird daher zunächst der dichtebasierten Subspace Clustering Algorithmus SUBCLU [KKK04] vorgestellt, der auf einer innovative Erweiterung von DBSCAN beruht. Um eine effiziente Suchstrategie durch den Suchraum aller möglichen Teilräume des Datenraumes zu ermöglichen (bei $d$ Dimensionen sind dies $O(2^d)$), werden Monotonie-Eigenschaften für die zentralen dichtebasierten Konzepte — Kernpunkt, (direkte) Dichteerreichbarkeit, Dichteverbundenheit — hergeleitet. Es wird insbesondere formal gezeigt, dass wenn eine Menge $C$ von Datenobjekten dichteverbunden in einem Unterraum $S$ ist (d.h. Teil eines Clusters in Unterraum $S$ ist) so ist die Menge $C$ auch dichteverbunden in allen Projektionen von $S$. Die umgekehrte Folgerung, dass eine Menge, die in einem Teilraum $S$ nicht dichteverbunden ist, in allen Oberräumen von $S$ ebenfalls nicht dichteverbunden sein kann, ermöglicht den frühen Ausschluss von Unterräumen, die keine Cluster mehr enthalten können. SUBCLU erzeugt daher sehr effizient, beginnend mit den 1-dimensionalen Teilräumen alle Cluster in allen Teilräumen die DBSCAN gefunden hätte, wenn man diesen Algorithmus auf alle $2^d$ Teilräume angewendet hätte. Dabei werden nur relevante Teilräume untersucht, und Teilräume die anhand der Monotonie keine Cluster enthalten können, frühzeitig aussortiert. Experimente mit synthetischen und realen Genexpressionsdaten zeigen den Qualitätsgewinn, den SUBCLU gegenüber dem bekanntesten bereits etablierten Subspace Clustering Verfahren CLIQUE [AGGR98] erzielt.

Einen Hauptnachteil der bisher bestehenden Subspace Clustering Verfahren ist die Verwendung eines globalen Dichtegrenzwertes. Daher wird in [KKKW03, Krö04] als Erweiterung von SUBCLU das Verfahren RIS präsentiert, das kein Subspace Clustering Verfahren im eigentlichen Sinne ist. Vielmehr erzeugt RIS eine Liste von Teilräumen, nach

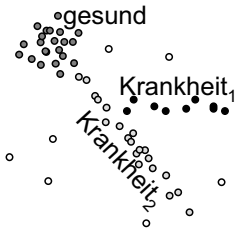


Abbildung 4: Beispielhafte Anwendung von Correlation Clustering.

Qualität der enthaltenen Clustering Struktur sortiert, statt die Subspace Cluster direkt zu berechnen. Die Suchstrategie von RIS basiert auf einem Unterraumausschlussverfahren das ähnlich zu SUBCLU ist. Der Vorteil von RIS ist, dass in den gefundenen Teilräumen ein hierarchisches Clustering ermittelt werden kann, d.h. es können Cluster unabhängig von einem globalen Dichtegrenzwert gefunden werden. Dieser Vorteil gegenüber traditionellen Subspace Clustering Verfahren wird in einer Reihe von Experimenten gezeigt, in denen RIS mit OPTICS kombiniert wird, um eine hierarchische Clustering Struktur in Teilräumen zu erzeugen.

### 4.3 Dichtebasiertes Correlation Clustering

Correlation Clustering ist eine Technik, bei der man Datenobjekte in Cluster gruppieren will, die nicht nur ähnlich zueinander sind, sondern auch eine einheitliche Korrelation aufweisen. Korrelation ist in diesem Zusammenhang eine lineare Abhängigkeit von beliebig vielen Attributen. Abbildung 4 zeigt exemplarisch die Anwendung von Correlation Clustering bei der Metabolomanalyse. Während die gesunden Patienten keine Korrelation aufweisen, gibt es zwei Cluster, die eine unterschiedliche Korrelation aufweisen. Jede dieser Korrelationen besteht dabei aus Patienten, die an der gleichen Krankheit leiden. Ziel des Correlation Clusterings ist es, die beiden Mengen der kranken Patienten zu finden, nicht jedoch die unkorrelierten gesunden Patienten.

In [BKKZ04, Krö04] wird eine neuartige Lösung für dieses Problem vorgeschlagen. Die dichtebasierten Konzepte von DBSCAN werden dabei mit einem Korrelationsprimitiv zur lokalen Korrelationsanalyse verbunden, um Correlation Cluster zu berechnen. Als Korrelationsprimitiv wird dabei die Hauptachsenzerlegung (Principal Component Analysis, PCA) verwendet. Jedem Punkt $p$ wird ein Distanzmaß zugeordnet, das die Korrelation in seiner $\varepsilon$-Nachbarschaft widerspiegelt. Dabei wird eine quadratische Form-Distanz verwendet, d.h. die Distanz von $p$ relativ zu einem anderen Objekt $q$ ist wie folgt definiert:

$$dist_p(p,q) = (p-q)\mathbf{M}_p(p-q)^{\mathbf{T}},$$

wobei $\mathbf{M}_p$ eine lokale Ähnlichkeitsmatrix darstellt, die aus der Kovarianzmatrix der $\varepsilon$-Nachbarschaft von $p$ entsteht, wenn man die Eigenwerte geeignet invertiert. Die geometrische Interpretation dieser Distanzfunktion ist ein Ellipsoid, das sich optimal an die lokale Korrelation anpasst. Die dichtebasierten Konzepte werden anhand dieses lokalen Distanz-

maßes verändert, sodass letztlich ein ähnliches algorithmisches Schema wie bei DBSCAN anwendbar ist, um Correlation Cluster zu bestimmen. Hauptvorteil des vorgestellten Verfahrens 4C (Computing Correlation Connected Clusters) gegenüber etablierten Verfahren wie ORCLUS [AY00] ist dabei die Determiniertheit der Resultate und die Robustheit gegenüber Rauschen. Ausführliche Experimente auf synthetischen und realen Metabolom-Daten zeigen den praktischen Nutzen von 4C im Vergleich zu mehreren Konkurrenzverfahren.

## 5 Zusammenfassung

In diesem Beitrag wurde der Stand der Technik im Bereich KDD und speziell dichtebasiertes Clustering vorangetrieben. Zunächst wurde der Prototyp BOSS zur Unterstützung industrienaher Anwendungen in der Ähnlichkeitssuche vorgestellt. Beispielsweise erleichtert BOSS die semi-automatische Clusteranalyse und erlaubt die interaktive, visuelle Suche nach ähnlichen Objekten in Multimedia- oder CAD-Datenbanken. Dazu wurden die Grundlagen des dichtebasierten, hierarchischen Clustering-Algorithmus OPTICS um wichtige Konzepte erweitert. Ein zweiter Bereich, der von dieser Arbeit abgedeckt wird, ist die Clusteranalyse hochdimensionaler Daten. Anhand von zwei wichtigen Anwendungen aus der biologisch-medizinischen Forschung und Industrie wurden die Probleme von hochdimensionalen Daten aufgezeigt und innovative Lösungen hierfür vorgeschlagen. Insbesondere wurden die Subspace Clustering Verfahren SUBCLU und RIS sowie der Correlation Clustering Algorithmus 4C vorgestellt. Details zu den einzelnen Techniken und umfassende experimentelle Evaluationen der Verfahren können in [Krö04] gefunden werden.

## Literatur

[ABKS99] M. Ankerst, M. M. Breunig, H.-P. Kriegel und J. Sander. ÖPTICS: Ordering Points to Identify the Clustering Structure". In *Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'99), Philadelphia, PA*, Seiten 49–60, 1999.

[AGGR98] R. Agrawal, J. Gehrke, D. Gunopulos und P. Raghavan. Äutomatic Subspace Clustering of High Dimensional Data for Data Mining Applications". In *Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'98), Seattle, WA*, 1998.

[AY00] C. C. Aggarwal und P. S. Yu. "Finding Generalized Projected Clusters in High Dimensional Space". In *Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'00), Dallas, TX*, 2000.

[BKK+04] S. Brecheisen, H.-P. Kriegel, P. Kröger, M. Pfeifle, M. Pötke und M. Viermetz. "BOSS: Browsing OPTICS-Plots for Similarity Search". In *Proc. 19th Int. Conf. on Data Engineering (ICDE'04), Boston, MA*, 2004.

[BKKP04] S. Brecheisen, H.-P. Kriegel, P. Kröger und M. Pfeifle. "Visually Mining Through Cluster Hierarchies". In *Proc. SIAM Int. Conf. on Data Mining (SDM'04), Lake Buena Vista, FL*, Seiten 400–412, 2004.

[BKKZ04] C. Böhm, K. Kailing, P. Kröger und A. Zimek. ”Computing Clusters of Correlation Connected Objects”. In *Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'04), Paris, France)*, 2004.

[EKSX96] M. Ester, H.-P. Kriegel, J. Sander und X. Xu. Ä Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD'96), Portland, OR*, Seiten 291–316, 1996.

[KBK+03] H.-P. Kriegel, S. Brecheisen, P. Kröger, M. Pfeifle und M. Schubert. “Using Sets of Feature Vectors for Similarity Search on Voxelized CAD Objects”. In *Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'03), San Diego, CA)*, 2003.

[KKG03] H.-P. Kriegel, P. Kröger und I. Gotlibovich. ”Incremental OPTICS: Efficient Computation of Updates in a Hierarchical Cluster Ordering”. In *5th Int. Conf. on Data Warehousing and Knowledge Discovery (DaWaK'03), Prague, Czech Republic*, Jgg. 2737 of *Lecture Notes in Computer Science (LNCS)*, Seiten 224–233. Springer, 2003.

[KKK04] K. Kailing, H.-P. Kriegel und P. Kröger. ”Density-Connected Subspace Clustering for High-Dimensional Data”. In *Proc. SIAM Int. Conf. on Data Mining (SDM'04), Lake Buena Vista, FL*, 2004.

[KKKW03] K. Kailing, H.-P. Kriegel, P. Kröger und S. Wanka. ”Ranking Interesting Subspaces for Clustering High Dimensional Data”. In *Proc. 7th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'03), Cavtat-Dubrovnic, Croatia*, Jgg. 2838 of *Lecture Notes in Artificial Intelligence (LNAI)*, Seiten 241–252. Springer-Verlag, 2003.

[KKM+03] H.-P. Kriegel, P. Kröger, Z. Mashael, M. Pfeifle, M. Pötke und T. Seidl. “Effective Similarity Search on Voxelized CAD Objects”. In *Proc. 8th Int. Conf. on Database Systems for Advanced Applications (DASFAA'03), Kyoto, Japan*, 2003.

[Krö04] P. Kröger. *Coping with new challenges for density-based clustering*. Dissertation, Ludwig-Maximilians-University Munich, 2004.

**Peer Kröger** wurde am 21. Januar 1975 in Kösching geboren. Er besuchte die Grundschule Zorneding von 1981 bis 1985, und das Gymnasium Vatterstetten von 1985 bis 1994. Von August 1994 bis October 1995, leistete er Zivildienst bei der Nachbarschaftshilfe Vaterstetten, Zorneding, and Grasbrunn. Im November 1995 begann er ein Studium der Informatik mit Nebenfach Physiologische Chemie an der Ludwig-Maximilians-Universität München (LMU). Seine Diplomarbeit trägt den Titel “Molecular Biology Data: Database Overview, Modelling Issues, and Perspectives” und wurde von Prof. François Bry und Prof. Rolf Backofen betreut. Seit Oktober 2001 arbeitet Peer Kröger an der LMU zunächst als wissenschaftlicher Mitarbeiter und seit November 2004 als wissenschaftlicher Assistent am Department “Institut für Informatik”, Lehrstuhl für Datenbanksysteme unter Prof. Hans-Peter Kriegel. In dieser Zeit publizierte er 18 Beiträge auf wissenschaftlichen Tagungen sowie in wissenschaftlichen Zeitschriften. Den Grad des Dr. rer. nat. erhielt er im Juli 2004 (Rigorosum am 8. Juli 2004), seine Dissertation trägt den Titel “Coping with new challenges for density-based clustering” und wurde im Mai 2004 eingereicht. Sein Forschungsschwerpunkt ist Knowledge Discovery in großen Datenbanken räumlicher und Multimediaobjekte, Data Mining für zur Analyse biologischer Daten und Ähnlichkeitssuche in räumlichen Datenbanken.