Hochschule für Angewandte Wissenschaften Hamburg
*Hamburg University of Applied Sciences*

# Ausarbeitung Seminar

Sven Elvers

Data Mining: Clustering

Betreuender Prüfer: Prof. Dr. Kai v. Luck

*Fakultät Technik und Informatik*
*Studiendepartment Informatik*

**Sven Elvers**

**Thema der Ausarbeitung**
**Seminar**

Data Mining: Clustering

**Stichworte**

Data Mining, Clustering

**Kurzzusammenfassung**

Dieser Artikel stellt kurz Data Mining vor und seine Bedeutung für Wissenschaft und Wirtschaft, um sich dann genauer mit dem Teilbereich Clustering zu beschäftigen. Dazu werden einige aktuelle Arbeiten vorgestellt, um die Aktualität des Themas zu zeigen und die Master Thesis einzuleiten.

# Inhaltsverzeichnis

**1 Einleitung** **1**

**2 Data Mining** **1**

**3 Clustering** **2**
3.1 Hierarchische Clusteringverfahren . . . . . 4
3.2 Partitionierende Clusteringverfahren . . . . . 5

**4 Aktuelle Arbeiten** **6**
4.1 Incremental and Effective Data Summerization for Dynamic Hierarchical Clustering . . . . . 7
4.2 Computing Clusters of Correlation Connected Objects . . . . . 8
4.3 Clustern mit Hintergrundwissen . . . . . 9

**5 Ausblick (Thesis outline)** **10**

# 1 Einleitung

In der Industrie sowie der Wissenschaft werden heutzutage viele Informationen gespeichert, wie zum Beispiel über Kundenverhalten oder Messergebnisse. In diesen Daten können Informationen enthalten sein, die nicht offensichtlich sind, aber zu neuen Erkenntnissen führen oder zur Verbesserung von Strategien genutzt werden können.

Diese Informationen lassen sich mit Techniken und Technologien aufdecken, die unter dem Begriff *Data Mining* zusammengefasst sind. Es werden unter anderem statistische Verfahren angewendet oder intelligente Systeme eingesetzt.

Ein Bereich, in dem das Data Mining seit Jahren angewendet wird, sind Banken und Versicherungen. Diese Unternehmen analysieren ihre Daten für Marketingzwecke. Es wird unter anderem untersucht, welche Kundengruppen mit gleichem oder ähnlichem Verhalten es gibt. So können u.a. Prognosen erstellt und entsprechend gehandelt werden. (vgl. Otte und Nathansen, Hotho (2005))

# 2 Data Mining

Die Aufgabe aus Daten versteckte Informationen, Muster oder Zusammenhänge zu gewinnen, wird als Data Mining bezeichnet. Dabei werden Methoden aus verschiedenen Bereichen verwendet und um die so gewonnenen Informationen zu überprüfen, ob sie sinnvoll sind, werden sie nach folgenden Punkten beurteilt:

- Verständlichkeit
- Gültigkeit (statistischer Rahmen)
- Neuheit
- Nützlichkeit

In Abbildung 1 auf Seite 2 sind die von Knowledge Discovery verwendeten Gebiete dargestellt. Obwohl Data Mining und Knowledge Discovery häufig synonym verwendet werden, interessieren beim Data Mining nur die Statistik, explorative Analyse, das maschinelle Lernen und die Fuzzy-Techniken.

Die Methoden beim Data Mining lassen sich in folgende Gruppen aufteilen, da es eine begrenzte Menge an Problemstellungen gibt:

- Clustering
- Klassifikation
- Regressionsanalyse

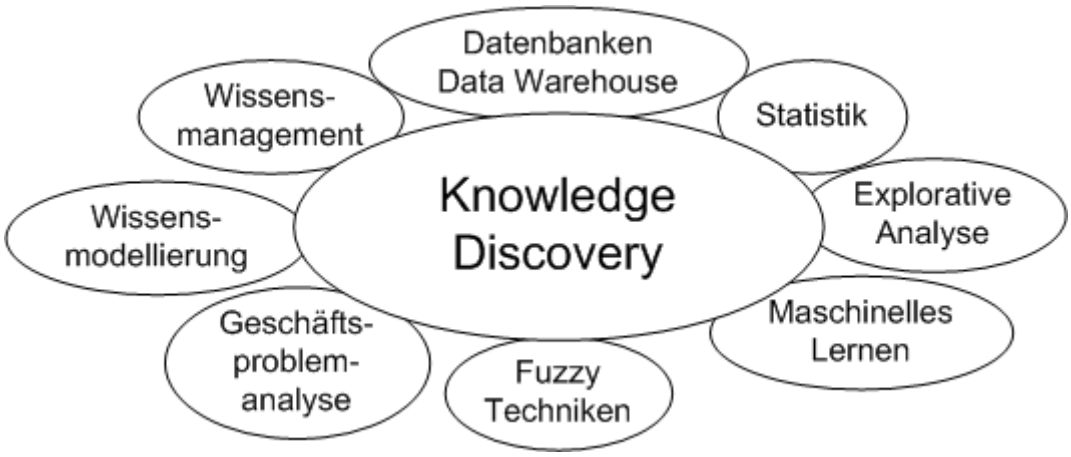



Abbildung 1: Knowledge Discovery (Hotho, 2005)

- Assoziationsanalyse
- Varianzanalyse

Beim Clustering sollen die Daten auf nützliche Gruppen untersucht werden, die gleiche oder ähnliche Merkmale und Beziehungen haben. Auf diese Verfahren wird in Abschnitt 3 genauer eingegangen.

Im Gegensatz zum Clustering sind bei der Klassifikation die Gruppen (Klassen) bereits bekannt und die Daten sind diesen zugeordnet. Wenn jetzt neue Daten hinzukommen, werden diese durch ein Klassifikationsverfahren einem der Klassen zugeordnet.

Bei der Regressionsanalyse wird untersucht, wie sich bei bestimmten Ereignissen die Daten geändert haben, um Voraussagen treffen zu können oder selber Einfluss auf das Verhalten zu nehmen.

Um herauszufinden, welche Merkmale häufig gemeinsam in dem Datenbestand auftreten, werden Assoziationsanalysen verwendet. Ein oft verwendetes Beispiel ist die Warenkorbanalyse, dass wenn jemand Produkt A einkauft, dieser Kunde mit einer bestimmten Wahrscheinlichkeit auch Produkt B kauft.

Um Abweichung zu erkennen, z.B. beim Controlling, werden Varianzanalysen verwendet. (vgl. Bauer (2004), Görz und Rollinger (2000), Vazirgiannis u. a. (2003), Hotho (2005))

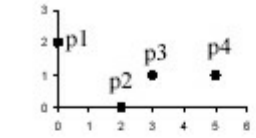
# 3 Clustering

Im Gegensatz zum Klassifizieren, wo die Objekte bekannten Gruppen (Klassen) zugeordnet werden, sind beim Clustering die Gruppen (Cluster) nicht bekannt, sondern es sollen aussagekräftig oder nützliche Cluster ermittelt werden, um Daten zusammenzufassen oder diese

besser verstehen zu können. Die Merkmale eines Clusters wird durch seine Objekte und dem verwendeten Algorithmus bestimmt.

Die Verwendung von Clustering wird in verschiedenen Gebieten aus unterschiedlichen Gründen genutzt. In der Biologie kann mit diesen Verfahren z.B. Gene nach ihren Funktionen gruppiert werden, während in der Bildverarbeitung Clustering genutzt wird, um Muster bzw. Objekte zu erkennen. Wie in der Einleitung bereits erwähnt wird im Marketing Clustering verwendet, um unter anderem Interessensgruppen zu ermitteln und beim Browsen im Web können so Dokumente vorgeschlagen werden, die einen ähnlichen Inhalt haben. (vgl. Hotho (2005))

Damit das Clustering ein korrektes Ergebnis liefert, müssen die Daten vorbereitet werden. Sie werden bereinigt, indem z.B. fehlerhafte Sätze aus der Datenbasis des Algorithmus (nicht der Stammdaten) entfernt werden. Danach werden die Entfernungen zwischen den Objekten ermittelt und in einer Proximity Matrix (siehe Abb. 2) eingetragen. Bei der Berechnung dieser Werte muss beachtet werden, dass es nicht nur quantitative Skalen gibt, wie Intervall- und Verhältnisskala, die berechenbar sind, sondern auch qualitative Skalen, wie Nominal- und Ordinalskala, bei denen sich nur aussagen lässt, dass sie gleich oder ungleich sind bzw. bei der Ordinalskala, dass ein Element vor dem anderen einzuordnen ist. (vgl. Steinbach u. a. (2003))




a) points

| point | x | y |
|---|---|---|
| p1 | 0 | 2 |
| p2 | 2 | 0 |
| p3 | 3 | 1 |
| p4 | 5 | 1 |

b) data matrix


d) proximity graph

| | p1 | p2 | p3 | p4 |
|---|---|---|---|---|
| p1 | 0.000 | 2.828 | 3.162 | 5.099 |
| p2 | 2.828 | 0.000 | 1.414 | 3.162 |
| p3 | 3.162 | 1.414 | 0.000 | 2.000 |
| p4 | 5.099 | 3.162 | 2.000 | 0.000 |

c) proximity matrix

Abbildung 2: Proximity Matrix (Steinbach u. a., 2003)

Die Anforderungen, die an ein Clusteringverfahren gestellt werden, sind

- Effizienz, dass der Algorithmus in einer ängemessenen"Laufzeit beendet ist,
- Effektivität, dass anwendungsbezogene Cluster erstellt werden,
- Erklärungsfähigkeit, dass es nachvollziehbar ist, warum ein Cluster entstanden ist und

- Benutzerinteraktivität, um das Ergebnis beeinflussen zu können.

Die Algorithmen fürs Clustern haben verschiedene Ziele und liefern deshalb als Ergebnis verschiedene Clustertypen:

- Well-Separated
- Center-based
- Contiguous
- Density-based
- Similarity-based

Bei *Well-Separated* Cluster ist die Entfernung zwischen zwei beliebigen Objekten innerhalb eines Clusters kleiner als zu einem Objekt eines anderen Clusters. Dies sind zum Beispiel hierarchische Verfahren (siehe Abschnitt 3.1). Dies ist bei den *Center-based* Cluster abgeschwächt. Hier muss nur noch die Entfernung zum Zentroid/Median innerhalb des eigenen Clusters kleiner sein als zu dem eines anderen Clusters. Das k-Means Verfahren, dass in Abschnitt 3.2 beschrieben wird, generiert diesen Typ von Cluster. Und bei *Contiguous* Cluster reicht es, wenn die Entfernung zu mindestens einem Objekt innerhalb eines Clusters geringer ist als zu jedem Objekt eines anderen Clusters.

Bei *Density-based* Cluster sind Objekte zusammengefasst, die eine bestimmte Dichte im Raum einnehmen. Dadurch kommen Algorithmen, die diesen Clustertyp erstellen mit Rauschen und Ausreißern klar.

Und ein *Similarity-based* Cluster beschreibt eine Region, in der alle Objekte ein einheitlich lokales Merkmal besitzen. (vgl. Hotho (2005), Steinbach u. a. (2003))

## 3.1 Hierarchische Clusteringverfahren

Hierarchische Clusterverfahren erstellen eine Menge von möglichen Clustern, welche in einem Dendogram ausgegeben werden (siehe Abb. 3 auf Seite 5). In diesem Graphen kann dann abgelesen werden, welche Cluster es bei welcher erlaubten Entfernung zwischen Objekten innerhalb eines Clusters gibt.

Bei agglomerativen Verfahren wird zuerst angenommen, dass jedes Objekt ein Cluster bildet und mit jedem Schritt werden jeweils zwei Cluster, die zueinander am dichtesten ligen,zu einem Cluster vereinigt. Die umgekehrte Variante, dass alle Objekte zuerst in einem Cluster sind und dann jeweils ein Cluster in zwei neue aufgeteilt wird, wird als diversiv bezeichnet.

Zusätzlich werden hierarchische Verfahren als single, complete oder average linkage bezeichnet. Bei *single linkage* werden die Cluster verbunden, deren Elemente die kleinste Distanz aufweisen. Dadurch sind diese Verfahren aber empfindlich gegenüber Rauschen

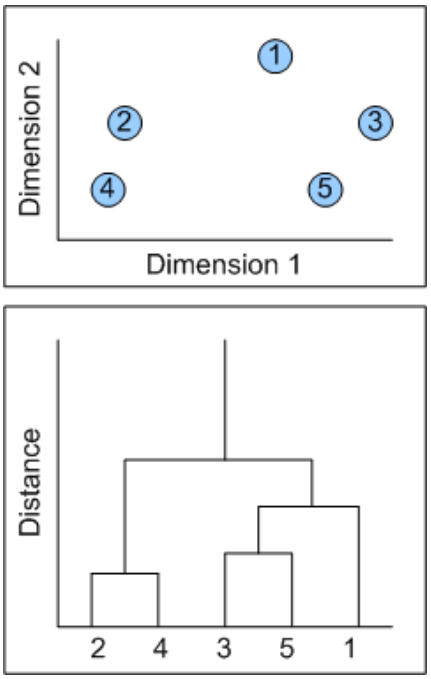



Abbildung 3: Hierarchische Clusteringverfahren

und Ausreißern. Um dieses zu vermeiden, können *complete linkage* Algorithmen verwendet werden. Hier werden die Entfernung zwischen zwei Clustern anhand der größten Distanz zwischen zwei ihrer Elemente bestimmt. Dadurch hat diese Art von Algorithmen Probleme mit convexen Clustern, z.B. wenn ein Clustern ein Kreis ist und ein anderer sich im inneren des Clusters befindet. Bei der dritten Variante, die *average linkage* Algorithmen, ist die Entfernung zweier Cluster die mittlere Distanz zwischen ihren Elementen. (vgl. Görz und Rollinger (2000), Bauer (2004), Steinbach u. a. (2003), Hotho (2005))

## 3.2 Partitionierende Clusteringverfahren

Zu den partitionierenden Clusteringverfahren gehört unter anderem der k-Means Algorithmus. Zu Begin wird vom Anwender festgelegt auf wieviele Cluster die Objekte aufgeteilt werden sollen. Danach werden per Zufall die Mediane der Cluster bestimmt. Im nächsten Schritt werden alle Objekte ihrem jeweiligen dichtesten Median zugeordnet. Danach werden die Mediane neu berechnet. Jetzt wiederholen sich das Zuordnen der Objekte zu den Medianen und die Berechnung neuer Mediane, bis keine Änderung vorhanden ist oder eine bestimmte Anzahl von Schritte durchgeführt wurde. Abbildung 4 zeigt zum Beispiel ein Ergebnis mit drei Medianen.

Da k-Means nach einer bestimmten Zeit ein Ergebnis liefert, ist er bei großen Daten-

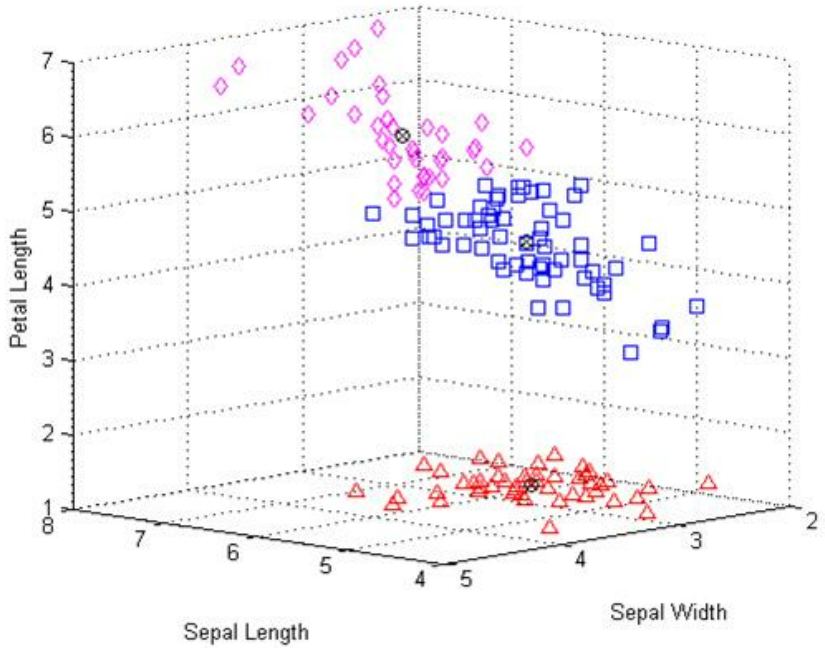



Abbildung 4: k-Means Clusteringverfahren (MathWorks 2005)

mengen den hierarchischen Algorithmen vorzuziehen. Doch ein mehrfaches Ausführen von k-Means kann unterschiedliche Ergebnisse liefern, da diese von dem Startpunkten der Mediane abhängen.

Eine Verbesserung ist der Bi-Sec-kMeans. Dieser fängt mit einem großen Cluster an, halbiert diesen und führt k-Means aus. Danach teilt er den nächsten Cluster und führt wieder k-Means aus. Dieses wird solange wiederholt, bis die gewünschte Anzahl an Clustern vorhanden ist und k-Means durchgelaufen ist. Dieser Algorithmus liefert durch das Zerteilen des z.B. jeweils größten Clusters immer dasselbe Ergebnis. (vgl. Görz und Rollinger (2000), Bauer (2004), Steinbach u. a. (2003), Hotho (2005))

# 4 Aktuelle Arbeiten

In diesem Abschnitt werden drei Arbeiten vorgestellt, die sich mit Clustering beschäftigt haben. Bei der ersten Arbeit geht es darum, wie sich hierarchische Algorithmen auf großen Datenmengen effektiv anwenden lassen. Beim zweiten Bericht wurde untersucht wie sich Cluster finden lassen, deren Objekte sich durch eine Beziehung beschreiben lassen und bei

der letzten Arbeit wurde ein Algorithmus vorgestellt, wie sich die Anzahl der Dimensionen verringern lässt, um unter anderem eine einfache Beschreibung der Cluster zu bekommen.

## 4.1 Incremental and Effective Data Summerization for Dynamic Hierarchical Clustering

In diesem Dokument (Nassar u. a., 2004) wird über Ïncremental Data Bubbles"gesprochen. Ein incremental Data Bubble ist ähnlich wie ein Cluster eine Menge von Objekten, die einen Repräsentant (Seed) haben und sich während der Datenmanipulation aktualisieren. Aber im Gegensatz zu den Clustern, werden hier Objekte einfach dem nahesten Seed zugeordnet.

Ziel dieser Arbeit war es, dass ein hierarchischer Clusteringalgorithmus auf große Datenmengen in ängemessenerSZeit bearbeiten kann, indem dieser auf den Seeds läuft, statt alle Objekte einzubeziehen.

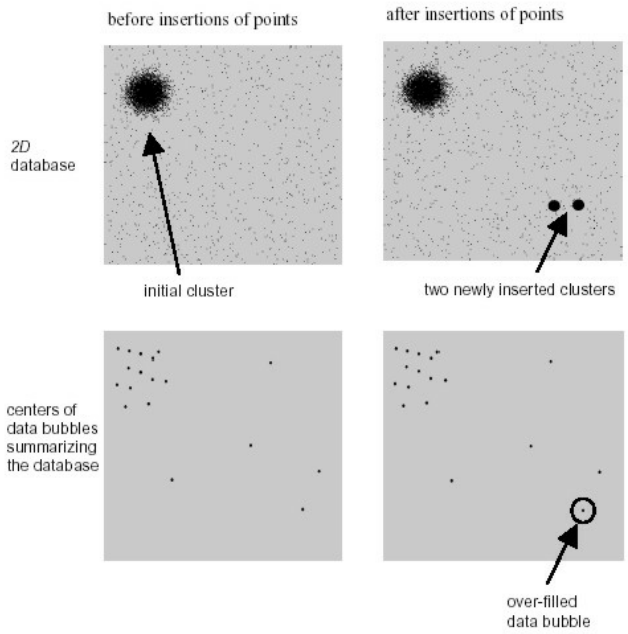



Abbildung 5: Data Bubbles Beispiel: overfilled (Nassar u. a., 2004)

Damit anhand der Seeds die Verteilung aller Objekte sichtbar ist und nicht wie in Abb. 5, wurde ein Qualitätskriterium für eingeführt. Ein Data Bubble muss aus einer bestimmten Anzahl Elementen bestehen $\beta \in [\mu_\beta - k\sigma_\beta, \mu_\beta + k\sigma_\beta]$ . Zusätzlich wurden die Begriffe Good, Under-filled und Over-filled geprägt, um den Zustand der Data Bubbles beschreiben zu können, ob sie die richtige Anzahl, zu wenige oder zu viele Elemente besitzen.

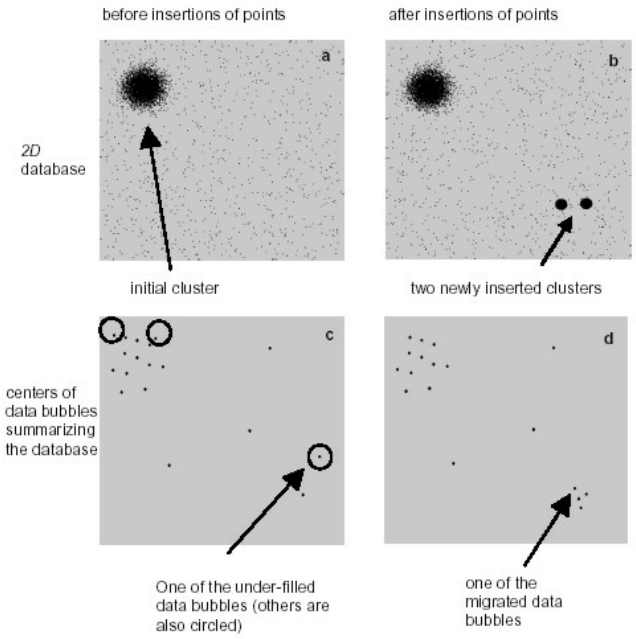



Abbildung 6: Data Bubbles Beispiel: underfilled (Nassar u. a., 2004)

Wenn jetzt ein Data Bubble over-filled ist, wird ein oder mehrere neue Seeds bestimmt und die überschüssigen Elemente auf diese verteilt und falls ein Data Bubble under-filled ist, wird dieser aufgelöst und seine Elemente werden auf die umliegenden Data Bubbles verteilt. Dieses Verhalten wird in Abb. 6 dargestellt.

## 4.2 Computing Clusters of Correlation Connected Objects

In dieser Arbeit von Böhm u. a. (2004) wird der Algorithmus "Computing Correlation Connected Clusters"(4C) vorgestellt. Dieser kombiniert das Density-based Clustering mit der Assoziationsanalyse, um Beziehung zwischen Objekten erkennen zu können, die nicht global sichtbar sind (siehe Abb. 7 auf Seite 9).

4C untersucht jedes Objekt auf Beziehungen zu anderen. Wenn der Algorithmus eine Beziehung gefunden hat, wird das Objekt zu dem entsprechenden Cluster hinzugefügt oder wenn der Cluster noch nicht existiert, wird dieser neu angelegt. Objekte, die nicht zu einem Cluster zugeordnet werden können, werden als Rauschen markiert.

In der Abbildung 8 auf Seite 10 wird der Unterschied zwischen dem 4C Algorithmus und dem Density-based Clusteringverfahren DBSCAN gezeigt.

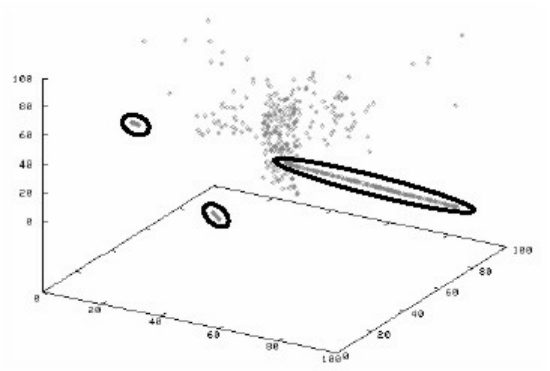


Abbildung 7: C4 Beispiel (Böhm u. a., 2004)

## 4.3 Clustern mit Hintergrundwissen

Der Algorithmus "Concept Selection and Aggregation"(COSA) wird in der Dissertation (Hotho, 2005) vorgestellt. Es handelt sich hierbei um einen ontologie-basierten Vorverarbeitungsschritt, bevor das eigentliche Clustering von Dokumenten durchgeführt wird.

Ziel ist es die subjektiven Informationen des Anwenders zu verwenden, um für den jeweiligen Anwender bzw. Anwendungsfall angepasste Cluster zu bekommen. Des weiteren wird die Anzahl der Dimensionen (Merkmalen) reduziert, so dass die Cluster für den Anwender verständlich beschrieben werden können und der Clusteringalgorithmus eine beschränkte Auswahl aller Dimensionen betrachten braucht.

COSA analysiert zuerst die Dokumente. Der Algorithmus sammelt wie häufig Wörter in den Dokumenten vorkommen, dazu werden alle Wörter auf ihre Stammwörter abgebildet und sogenannte Stoppwörter entfernt. Diese sind Wörter die keine alleine keine Aussagekraft haben, wie z.B. und, oder, der, die, das. Dieser Bearbeitungsschritt wird als SStemming"bezeichnet. Danach wird die Ontologie vom übergebenen Startkonzept aus durchgearbeitet. Die Ontologie ist eine Struktur aus Knoten und Kanten. Dabei ist ein Knotenpunkt ein Begriff, welcher als Konzept bezeichnet wird und die Kanten zwischen den Konzepten beschreiben deren Beziehung.

Der Algorithmus baut eine Liste auf, bei der an forderster Stelle das am stärksten in den Dokumenten vorkommende Konzept (Merkmal) steht, an zweiter Stelle das zweitstärkste usw. Mit jedem Schritt wird anhand der Ontologie das stärkste Merkmal verfeinert. Wenn die Liste die vorgegebene Länge erreicht bzw. überschreitet, werden die überschüssigen Merkmale verworfen. Danach wird die Liste der Ergebnismenge hinzugefügt und die Ontologie weiter durchgearbeitet.

Durch dieses Verfahren stehen dem Anwender nur die Kombinationen von Merkmale fürs

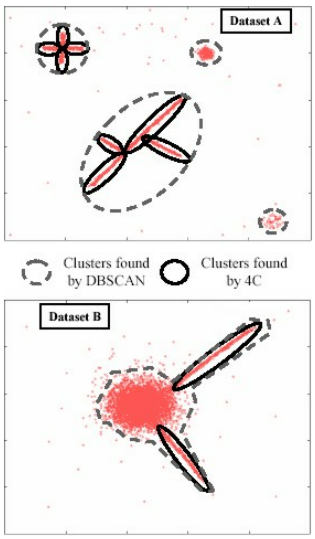



Abbildung 8: Vergleich zwischen DBSCAN und C4 (Böhm u. a., 2004)

Clustering zur Verfügung, die den Anforderungen entsprechen und Merkmale enthalten, die nicht zu häufig sowie zu selten in den Dokumenten vorkommen.

# 5 Ausblick (Thesis outline)

Der Autor beabsichtigt, in seiner Masterarbeit die im Masterprojekt gesammelten Erfahrungen mit Analyseverfahren im Bereich Business Intelligence zu vertiefen. Zu diesem Zweck soll untersucht werden, ob es weitere sinnvolle Kombinationen von Clustering mit anderen Data Mining Verfahren gibt.

Im Projekt wurde ein Pentaho System aufgesetzt (<http://www.pentaho.org/> (30.Dezember 2005)). Dieses kann in der Masterarbeit als Testumgebung genutzt werden.

Die Risiken bei dieser Aufgabe sind der Aufwand und die Datenbasis. Da die vorgestellten Algorithmen von mehreren Personen erarbeitet wurden, muss die Abgrenzung der Aufgabe mit besondere Sorgfalt gewählt werden. Genauso ist die Datenbasis wichtig, da vorher schon bekannt sein muss, welches Resultat nach dem Clustern zu erwarten ist. Zum Beispiel musste die Datenbasis für den 4C-Algorithmus aus Abschnitt 4.2 so gewählt werden, dass mit einer Assoziationsanalyse oder einem Density-based Clusteringverfahren die Cluster nicht erkannt werden konnten, deren Objekte eine Beziehung zueinander hatten.

# Literatur

[Bauer 2004] BAUER, Andreas ; GÜNZEL, Holger (Hrsg.): *Data Warehouse Systeme, 2.Aufl.* dpunkt, 2004. – ISBN 3-89864-251-8

[Böhm u. a. 2004] BÖHM, Christian ; KAILING, Karin ; KRÖGER, Peer ; ZIMEK, Arthur: *Computing Clusters of Correlation Connected Objects*. 2004. – URL http://www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/SIGMOD_2004.pdf. – (9. November 2005)

[Görz und Rollinger 2000] GÖRZ, G. ; ROLLINGER, C.-R. ; SCHNEEBERGER, J. (Hrsg.): *Handbuch der Künstlichen Intelligenz, 3.Aufl.* Oldenbourg, 2000. – ISBN 3-486-25049-3

[Hotho 2005] HOTHO, Andreas: *Clustern mit Hintergrundwissen, 2.Aufl.* Aka GmbH, 2005. – ISBN 3-89838-286-9

[MathWorks 2005 ] *The MathWorks, Statistics Toolbox 5.0.2, Demos*. – URL http://www.mathworks.com/products/statistics/demos.jsp. – (17.Mai 2005)

[Nassar u. a. 2004] NASSAR, Samer ; SANDER, Jörg ; CHENG, Corrine: *Incremental and Effective Data Summarization for Dynamic Hierarchical Clustering*. 2004. – URL http://www.sigmod.org/sigmod/sigmod04/eproceedings/content/bytrack.html#research. – (10. November 2005)

[Otte und Nathansen ] OTTE, Ralf ; NATHANSEN, Martin: *Data Mining Applications for Industry Germany*. – URL http://www.datmin.de/index.html. – (28. Dezember 2005)

[Steinbach u. a. 2003] STEINBACH, Michael ; ERTÖZ, Levent ; KUMAR, Vipin: *The Challenges of Clustering High Dimensional Data*. 2003. – URL http://www-users.cs.umn.edu/~ertoz/papers/clustering_chapter.pdf. – (10. November 2005)

[Vazirgiannis u. a. 2003] VAZIRGIANNIS, Michalis ; HALKIDI, Maria ; GUNOPULOS, Dimitrios: *Uncertainty Handling and Quality Assessment in Data Mining*. Springer, 2003. – ISBN 1-85233-655-2