Mining in speziellen Daten und Data Mining Anwendungen

Holger Klus
Betreuer: Boris Stumm

13.01.2004

Inhaltsverzeichnis

1 Einleitung

Unter Multimedia-Mining versteht man im Allgemeinen Data-Mining in unstrukturierten oder semistrukturierten Daten. Die Kernziele unterscheiden sich jedoch meist nicht von denen des Data-Mining in strukturierten Daten wie sie z. B. in relationalen Datenbanken existieren. Folgende Kernziele haben die meisten Mining-Verfahren gemeinsam:

- Klassifizieren von Datensätzen
- Finden von Assoziationsregeln
- Erkennen von Trends
- Clusterbildung

Im Folgenden sollen Verfahren zum Erreichen einiger dieser Kernziele in unterschiedlichen Datentypen erläutert werden. Der erste Abschnitt behandelt das Mining in unstrukturierten Texten. Es wird ein Textmining-Prozess zur Klassifizierung von Textdokumenten vorgestellt und anschließend anhand eines Beispiels erläutert. Der zweite Abschnitt behandelt Image-Mining. Hier wird ein informationsorientierter Image-Mining-Ansatz vorgestellt und im Anschluss durch ein Beispiel vertieft. Im dritten Abschnitt wird ein Prozess zum Mining von Videodaten vorgestellt. Insbesondere wird anhand eines Beispiels beschrieben, wie Objekte in Videos automatisch verfolgt werden können. Im vierten Abschnitt geht es dann um Mining in räumlichen Daten. Abschließend folgt eine Diskussion über die Konsequenzen von Data-Mining für die Gesellschaft.

2 Text-Mining

Unter Textmining, auch *Text-Data-Mining* oder *Knowledge Discovery in Textdatenbanken* versteht man die Extraktion von interessanten, nichttrivialen Mustern oder Wissen aus Textdokumenten. Studien zeigen, dass 80% der Informationen in Unternehmen als einfache Textdateien vorhanden sind [Tan99]. Dies zeigt das Potential und die Wichtigkeit von Text-Mining-Werkzeugen in der Zukunft. Im Folgenden wird zunächst ein Text-Mining-Prozess vorgestellt und anschließend anhand eines Beispiels verdeutlicht.

2.1 Klassifizierung von Textdokumenten

Im Folgenden werden nun die allgemeinen Schritte zur Klassifizierung von Textdokumenten erläutert [AaEi99].

2.1.1 Preprocessing

Der erste Schritt besteht darin, die Dokumente derart zu bearbeiten, dass die effiziente Anwendung von Lern- und Klassifizierungsalgorithmen möglich wird. Hierzu werden meist die folgenden Schritte ausgeführt:

- Entfernen von Formatierungszeichen, HTML-Tags oder Ähnliches
- Entfernen von Stoppworten
- Stammbildung

Stoppwörter sind Wörter wie Artikel oder Präpositionen, die in der Regel keinen großen Informationsgehalt besitzen. Meist stellt man eine Liste von Stoppworten der jeweiligen Sprache zusammen und filtert anhand dessen diese Worte aus.

Unter **Stammbildung** versteht man die Zusammenfassung von Worten, die in unterschiedlichen syntaktischen Formen im Dokument auftreten, aber vom selben Wort abstammen (Suche, suchen, gesucht, suchte...).

2.1.2 Indexierung

Ein häufig verwendetes Modell zur Repräsentation von Textdokumenten ist das Vektorraummodell. Dokumente sind in diesem Modell Vektoren von Wörtern. Eine Sammlung von Dokumenten wird mit Hilfe einer Dokumentmatrix **A** repräsentiert, mit

$$\mathbf{A} = (a_{ik})$$

wobei a_{ik} das Gewicht des Wortes i im Dokument k angibt. Diese Matrix ist im Allgemeinen dünn besetzt, da nicht jedes Wort in jedem Dokument enthalten ist. Zudem ist die Matrix sehr groß, da für jedes verwendete Wort eine Zeile existiert. Die Größe dieser Matrix ist eine besondere Schwierigkeit im Bereich Textmining allgemein und bei der Klassifizierung im Besonderen. Im nächsten Abschnitt werden Möglichkeiten gezeigt, die Dimension der Dokumentmatrix zu reduzieren.

Es gibt mehrere Wege, das Gewicht eines Wortes i im Dokument k zu bestimmen. Die meisten Ansätze berücksichtigen folgende Regeln:

- Je öfter ein Wort in einem Dokument enthalten ist, desto relevanter ist es für das im Dokument behandelte Thema und desto höher sein Gewicht.
- Je öfter das Wort in allen Dokumenten enthalten ist, desto weniger charakteristisch ist dieses Wort für ein Dokument und desto geringer ist sein Gewicht.

Ein weit verbreiteter Ansatz zur Berechnung des Gewichtes eines Wortes i im Dokument k ist folgender.

$$a_{ik} = f_{ik} * \log\left(\frac{N}{n_i}\right)$$

wobei f_{ik} die Häufigkeit ist, mit der das Wort i in Dokument k vorkommt. N gibt die Anzahl aller Dokumente in der Datenbank an und n_i ist die Häufigkeit, mit der das Wort i in allen Dokumenten vorkommt. Der erste Multiplikand berücksichtigt den ersten oben aufgeführten Punkt und der zweite Multiplikand den zweiten Punkt.
Zu diesem Verfahren gibt es noch einige Verfeinerungen, die z. B. die Länge eines Dokumentes bei der Berechnung des Gewichtes berücksichtigen.

2.1.3 Dimensionsreduktion

Wie oben bereits erwähnt, besteht ein Problem bei der Klassifizierung von Dokumenten mit Hilfe der Darstellung im Vektorraummodell in der sehr großen Dokumentmatrix. Ein einfaches, aber effizientes Verfahren ist das *Document Frequency Thresholding*. Die *Document Frequency* eines Wortes ist definiert als die Anzahl der Dokumente, in denen dieses Wort vorkommt. Für jedes Wort wird die *Document Frequency* berechnet. Anschließend wird ein Schwellwert festgelegt. Jedes Wort, das diesen Wert nicht überschreitet, wird dann aus der Matrix entfernt.
Auch hier existieren noch zahlreiche weitere Verfahren.

2.1.4 Anwendung eines Klassifizierungsalgorithmus

Aufgabe der Klassifizierung ist es, eine oder mehrere vorgegebene Kategorien Textdokumenten zuzuordnen. Es existieren eine ganze Reihe von Klassifizierungsalgorithmen. Einige von ihnen seien hier genannt:

- Naive Bayesche Klassifikation
- Algorithmus nach Rocchoi
- kNN-Algorithmus (k-nearest neighbour)

- Entscheidungsbäume
- Support Vector Machines

Die genannten Klassifikationsverfahren sind bereits im Data-Mining in strukturierten Daten bekannt und können nun auf die zuvor generierte Darstellung der Textdokumente in Form von Vektoren angewandt werden. Im Folgenden wird anhand des kNN-Algorithmus die Anwendung eines Klassifikationsverfahrens erläutert und anschließend die Effektivität anhand einiger Messwerte überprüft.
Meist handelt es sich bei Klassifikationsproblemen um binäre Probleme, d. h. jedem Textdokument wird genau eine von zwei exitierenden Klassen zugeordnet. Im Bereich *Information Retrieval* ist eine häufige Anwendung die Klassifikation eines Textdokumentes in „relevant“oder „irrelevant“bezüglich eines Themas.

Der kNN-Algorithmus Der Algorithmus hat als Eingabe einen Dokumentvektor d, der klassifiziert werden soll. Weiter ist eine Menge von Trainingsdokumenten je Klasse vorhanden. Der Algorithmus läuft in folgenden Schritten ab:

1. Bestimmen der k ähnlichsten Nachbarn von d
2. Gewichtung der Klassen, in denen die k ähnlichsten Nachbarn von d enthalten sind.

Zur Messung der Ähnlichkeit zweier Dokumentvektoren werden meist der euklidische Abstand oder der Winkel zwischen den Vektoren herangezogen. Der Winkel zwischen zwei Dokumentvektoren a und b berechnet sich wie folgt.

$$\alpha = \arccos \frac{\langle a, b \rangle}{\|a\| \cdot \|b\|}$$

mit

$$\langle a, b \rangle = a_1 \cdot b_1 + ... + a_n \cdot b_n \qquad \text{und} \qquad \|a\| = \sqrt{a_1^2 + ... + a_n^2}$$

Zwei Dokumentvektoren werden dann als ähnlich angenommen, wenn sie einen Grenzwert, d. h. einen bestimmten Winkel unterschreiten, welcher vorher festgelegt werden muss. Für k legt man Erfahrungswerte zugrunde.

2.2 Klassifizierung von Nachrichtenartikeln in der Reuter-Textsammlung

Die Reuter-Datensammlung ist eine Menge von Nachrichten, welche auch häufig zur Effizienzmessung in Text-Retrieval-Anwendungen verwendet wird. Die Nachrichtensammlung enthält 12.902 Zeitungsartikel und 135 mögliche Kategorien. Einigen Artikeln wurden bis zu 14 Kategorien zugeordnet. Die Menge der Dokumente wird nun zunächst aufgeteilt in 9.603 Trainingsdokumente und 3.299 Testdokumente. Hieraus werden unvollständige Berichte entfernt. Hierunter fallen unter anderem Artikel ohne Titel oder Kategorie. Übrig bleiben dann noch 7.063 Trainingsdokumente und 2.658 Testdokumente. Im Folgenden sollen nun den Testdokumenten anhand der Trainingsdaten automatisch Kategorien zugewiesen werden. Dies geschieht nach den im vorherigen Kapitel erläuterten Schritten. Ziel ist es, die Performance des vorgestellten Ansatzes zu verifizieren.

- **Preprocessing**
  Zunächst werden Stoppwörter anhand einer Liste von 525 Stoppwörtern der englischen Sprache entfernt. Anschließend wird ein Algorithmus zur Stammbildung angewendet. Das Resultat sind 15.247 unterschiedliche Wörter.

Thema	# Training	# Test	Recall	Precision
earnings	2.709	1.014	0,95	0,92
aquisitions	1.488	630	1,00	0,91
money-fx	460	133	0,92	0,65
crude	349	160	0,96	0,70
grain	394	130	0,82	0,75
trade	337	106	0,89	0,66
interest	289	95	0,8	0,71
ship	191	82	0,85	0,77
wheat	198	65	0,69	0,73
corn	160	46	0,35	0,76

Tabelle 1: *Precision und Recall der zehn häufigsten Themen bei einem Grenzwert von 11,4°*

- **Indexierung**
  Hier wird nun die im vorherigen Kapitel beschriebene Indexierung durchgeführt. Alle Artikel werden nun durch Vektoren repräsentiert.

- **Dimensionsreduktion**
  Zunächst werden alle Wörter entfernt, die lediglich in einem Dokument vorkommen. Anschließend werden noch zwei weitere Algorithmen zur Dimensionsreduktion angewandt. Es bleiben 200 unterschiedliche Wörter übrig.

- **Anwendung eines Klassifizierungsalgorithmus**
  Hier wird der oben beschriebene kNN-Klassifizierungsalgorithmus angewandt. Der Parameter k wird mit 30 angenommen. Zur Ermittlung der Ähnlichkeit wird der Winkel verwendet. Der Grenzwert, ab dem Ähnlichkeit angenommen wird, ist domänenspezifisch und beruht auf Erfahrungswerten. In diesem Fall wird er auf 11,4° gesetzt.

- **Ergebnisse**
  Zur Effektivitätsmessung des vorgestellten Verfahrens werden die Maßgrößen *Precision* und *Recall* verwendet, die sich wie folgt berechnen.

  $$Recall = \frac{a}{a+c} \qquad Precision = \frac{a}{a+b}$$

  mit

  - a : Anzahl der Dokumente, die der Kategorie korrekt zugewiesen wurden
  - b : Anzahl der Dokumente, die der Kategorie fälschlicherweise zugewiesen wurden
  - c : Anzahl der Dokumente, die der Kategorie fälschlicherweise nicht zugewiesen wurden

  Tabelle 1 zeigt die zehn häufigsten Themen bzw. Kategorien und die jeweiligen Recall-und Precision-Werte bei einer Schwelle von 11,4°. Man kann beobachten, dass sowohl die Recall- als auch Precision-Werte sinken, wenn die Anzahl der Trainings- und Testdokumente sinkt. Doch nicht nur die absoluten Werte von Recall und Precision sind relevant. Da sie sich auch gegenseitig beeinflussen, ist auch ihr Verhältnis zueinander relevant. Hier ist zu beobachten, dass bei einer großen Anzahl von Trainingsdokumenten die Recall- und Precision-Werte nahe beieinander liegen, wobei der Recall größer ist als Precision. Bei sehr wenigen Trainingsdokumenten kehrt sich dies um, und der Precision-Wert steigt im Gegensatz zum Recall wieder an. Das Ziel ist es nun, die Schwelle so zu definieren, dass Recall und Precision möglichst nahe beieinander liegen und einen möglichst hohen Wert erreichen.

2.3 Weitere Text-Mining-Anwendungen

Im Folgenden werden nun noch abschließend zwei weitere Anwendungsdomänen vorgestellt.

- **Erkennen von Trends**
  Diese findet unter anderem Anwendung in der Analyse von Trends in großen Patentsammlungen oder bei der Erforschung von Entwicklungen der Interessen der Öffentlichkeit anhand von Zeitungsartikeln und Kommentaren. Eine Einsatzmöglichkeit ist die automatische Analyse von Konzernberichten und Kommentaren, um z. B. mögliche Veränderungen in den Kernkompetenzen eines Unternehmens zu erkennen.

- **Generieren von Textzusammenfassungen**
  Hierbei wird versucht, aus Texten ganze Sätze zu extrahieren bzw. Sätze zu bilden, die den Inhalt eines Dokumentes möglichst kurz und prägnant wiedergeben.

- **Halbautomatisches Beantworten von Kundenanfragen**
  Hierbei wird versucht, ein System zu realisieren, welches Fragen, die in natürlicher Sprache verfasst sind, automatisch, möglichst exakt und wiederum in natürlicher Sprache beantwortet.

3 Image-Mining

Heutzutage existiert eine sehr große Menge von Bilddaten unter anderem in Form von Satellitenbildern, medizinischen Aufnahmen und digitalen Fotos. Durch die Analyse dieser Bildmenge können interessante bisher versteckte Informationen in Form von Mustern oder Assoziationen extrahiert werden. Unter Image-Mining versteht man die automatische Extraktion von semantisch aussagekräftigen Informationen. Die grundlegende Aufgabe besteht darin, aus Low-Level-Bilddaten wie Pixeln Objekte und deren Beziehungen zueinander zu erkennen. Um dieser Aufgabe gerecht zu werden, wird im Folgenden ein informationsorientierter Image-Mining-Ansatz vorgestellt, der den Image-Mining-Prozess in vier Abstraktionsebenen gliedert. Hierbei sollen die unterschiedlichen Informationsebenen näher untersucht und jeweils Anwendungsbeispiele aufgeführt werden.

3.1 Ein informationsorientierter Image-Mining-Ansatz

Bilddatenbanken enthalten meist Bild-Rohdaten, welche in dieser Form für die Anwendung von Mining-Verfahren ungeeignet sind. Hier soll nun ein Grundgerüst für Image-Mining-Verfahren vorgestellt werden [ZhHL01]. Das Grundgerüst besteht aus vier Informationsebenen. Diese Informationsebenen und deren jeweiliger Informationsgehalt werden nun im Folgenden näher erläutert.

3.1.1 Pixel-Ebene

Die Pixel-Darstellung ist die niedrigste Informationsebene in Image-Mining-Systemen. Die Informationen bestehen ausschließlich aus Pixeln. Auf dieser Ebene werden im Image-Mining-Prozess interessante, d. h. für eine spezifische Mining-Aufgabe relevante Regionen aus dem Bild extrahiert. Daraus resultiert eine Segmentierung des

Bildes. Man unterscheidet zwei Klassen von Segmentierungsalgorithmen. Zum einen Algorithmen, die versuchen, Regionen auf Grund homogener Farbverteilung und homogener Pixelintensität zu finden, und zum anderen Algorithmen, die versuchen, Konturen im Bild zu finden. Zwei Segmentierungsalgorithmen sind das *adaptive local thresholding* und *dynamic water immersion*, deren Funktionsweise im folgenden Kapitel anhand eines Beispiels erläutert wird.

3.1.2 Objekt-Ebene

Die Kernaufgabe auf der Objektebene ist die Identifizierung domänenspezifischer Merkmale wie Objekte und homogene Bildbereiche. Die automatische Erkennung von Objekten in Bildern ist nicht trivial. Meist werden hierzu Objekte in Form von Beispielbildern und deren Bezeichnung gespeichert. Ein neu zu analysierendes Bild wird dann mit den in der Datenbank gespeicherten Beispielobjekten verglichen, um so Objekte zu erkennen und zu bezeichnen. Ein anderer Weg zur Identifizierung von relevanten Objekten ist der, spezifische Eigenschaften und Merkmale der gesuchten Objekte zu beschreiben. Es gibt zahlreiche Algorithmen zur Suche und Erkennung dieser Merkmale in unterschiedlichen Anwendungsdomänen. Oft werden mehrere Merkmale herangezogen, die ein Objekt als relevant identifizieren können, um eine höhere Treffergenauigkeit zu erlangen.

3.1.3 Semantische Ebene

In dieser Ebene wird versucht, Muster in der identifizierten Objektmenge zu extrahieren, Objekte zu klassifizieren und Assoziationsregeln zu finden. Ebenso kann es darum gehen, nicht nur Objekte innerhalb von Bildern Klassen zuzuordnen, sondern auch ganze Bilder zu klassifizieren.

3.1.4 Muster- und Wissensebene

Auf dieser Ebene werden Informationen, die aus dem Bild extrahiert wurden mit weiteren domänenspezifischen, alpha-numerischen Informationen kombiniert.

3.2 Image-Mining bei der Erkennung von Tumoren

In diesem Beispiel wird ein Ansatz vorgestellt, Tumorzellen auf Röntgenbildern zuverlässig zu erkennen und zu zählen. Hierbei wird auf den oben erläuterten informationorientierten Image-Mining-Ansatz zurückgegriffen und der Prozess anhand der vier Ebenen erläutert. Die Aufgabe besteht darin, auf einem gegebenen Röntgenbild Tumorzellen zu erkennen [FaHL02].

3.2.1 Pixel-Ebene

Auf dieser Ebene besteht das Ziel darin, potentiell interessante Regionen auf dem Röntgenbild zu erkennen und vom Hintergrund zu extrahieren. Potentiell interessante Regionen zeichnen sich in diesem Fall durch eine hohe Pixelintensität und helle Grautöne aus. Eine einfache Möglichkeit besteht darin, ein Grautonhistogramm zu erstellen und anschließend einen Grenzwert festzulegen. Abbildung 1 zeigt ein Beispiel eines Grautonhistogramms mit einem eingezeichneten Grenzwert.

Es wird hier jedem im Bild vorkommenden Grauton die Häufigkeit dieses Grautons im Bild zugeordnet. Nun besteht insbesondere bei medizinischen Aufnahmen das Problem, dass es im extremen Fall vorkommen kann, dass einige Tumorzellen in einem Bereich des Bildes dunkler sind als der Hintergrund in einem anderen Teil des Bildes. Dies führt dazu, dass bei der Festlegung eines Grenzwertes für

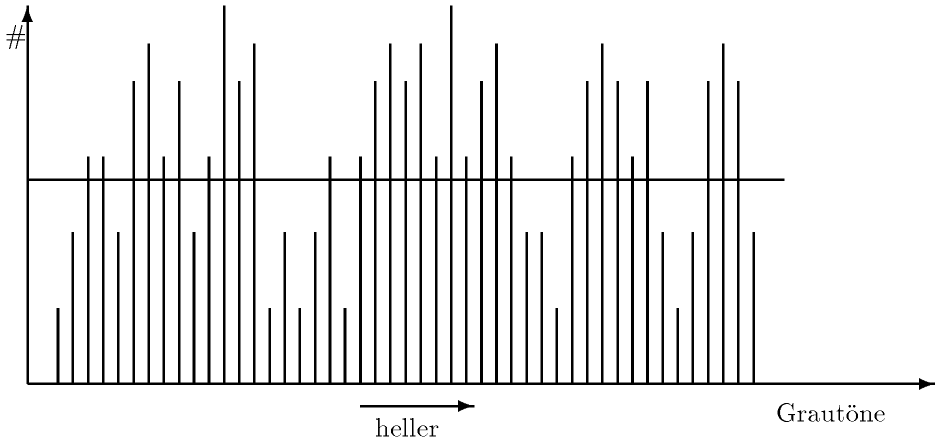

Abbildung 1: *Die Abbildung zeigt ein Grautonhistogramm mit eingezeichnetem Schwellwert. Überschreiten ähnliche Grautöne diesen Schwellwert, so werden diese zu einer Region zusammengefasst*

das gesamte Bild häufig relevante Bereiche nicht erkannt werden bzw. Bildbereiche fälschlicherweise als relevant erkannt werden. Aus diesem Grunde teilt man des Bild in NxN Teilbilder auf, erzeugt für jedes dieser Teilbilder ein Grautonhistogramm und legt einen auf dieses Teilbild beschränkten Grenzwert fest. Abbildung 2 zeigt drei Röntgenaufnahmen. Bei den ersten beiden Aufnahmen wurden jeweils globale Grenzwerte festgelegt, wobei links eine niedrige und rechts eine hohe Schwelle festgelegt wurde. Das rechte Bild zeigt das Ergebnis bei der Anwendung lokaler Grenzwerte.

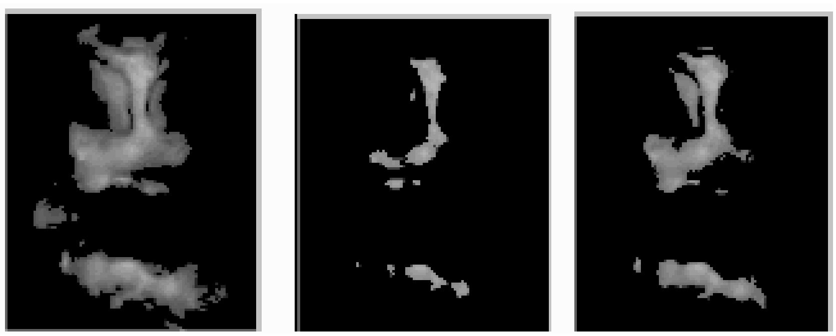

Abbildung 2: *Die Abbildung zeigt unterschiedliche Ergebnisse bei der Erkennung von Regionen. Bei den linken beiden Aufnahmen wurde jeweils ein globaler Grenzwert festgelegt, während bei der rechten Aufnahme lokale Grenzwerte zugrunde gelegt wurden*

Nachdem nun potentiell relevante Regionen extrahiert wurden, besteht die nächste Aufgabe darin, überlappende Zellen und Zellgruppen zu trennen, d. h. Konturen zu extrahieren. Hierzu wird der *water immersion-Algorithmus* angewandt. Hierzu werden zunächst alle lokalen Maxima in dem Bild bestimmt, d. h. Pixel mit der maximalen Intensität bezüglich ihrer näheren Umgebung. Jedes dieser lokalen Maxima ist nun der Repräsentant einer potentiell interessante Zelle. Das Prinzip des *water immersion-Algorithmus* ist es nun, sich ausgehend von einem lokalen Maximum nach bestimmten Regeln auszubreiten, so dass letztlich eine Kontur um dieses lokale Maximum entsteht. Abbildung 3 zeigt die lokalen Maxima und die durch den *water immersion-Algorithmus* erkannten Konturen.

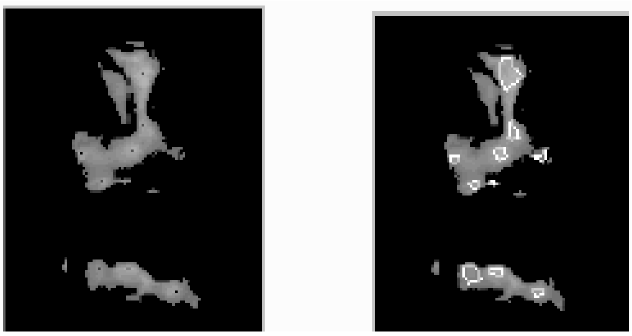

Abbildung 3: *Die Abbildung zeigt links die lokalen Maxima und rechts die durch den water-immersion-Algorithmus erkannten Konturen*

3.2.2 Objekt-Ebene

Nachdem nun im vorangegangenen Schritt potentiell interessante Regionen extrahiert wurden, geht es nun darum, relevante Objekte, d. h. Tumorzellen von nicht relevanten Objekten zu unterscheiden. Hierzu werden auf dieser Abstraktionsebene zunächst folgende drei Merkmale je Objekt extrahiert, die im Folgenden aufgeführt sind.

- Fläche
  Die Anzahl der Pixel innerhalb der Region
- Rundheit

$$Rundheit = \frac{4 \cdot \pi \cdot Fläche}{Umfang^2}$$

- Dehnung/Streckung
  Die Dehnung wird berechnet durch den Quotienten aus der Länge der Hauptachse des Objektes und der Länge der Nebenachse.

$$Dehnung = \frac{L_{Hauptachse}}{L_{Nebenachse}}$$

Im Ergebnis wird auf dieser Abstraktionsebene nun jedes Objekt durch diese drei Merkmale beschrieben. Auf der nächsten Abstraktionsebene geht es nun darum, relevante Objekte, d. h. Tumorzellen von nichtrelevanten Objekten zu unterscheiden.

3.2.3 Semantische Ebene

In dieser Phase werden nun aus dem Data-Mining bekannte Klassifizierungsalgorithmen angewandt, um Tumorzellen von anderen nicht relevanten Objekten zu unterscheiden. Hierzu wird zunächst eine Trainingsmenge angelegt, in der Ausprägungen der drei Merkmale gesammelt werden, die auf einen Tumor schließen lassen. Anschließend kann ein neu zu klassifizierendes Objekt anhand dessen Merkmalsausprägungen klassifiziert werden. Typische Klassifizierungsalgorithmen sind der Bayesche Klassifizierungsalgorithmus und der CBA-Algorithmus [JWB+03].

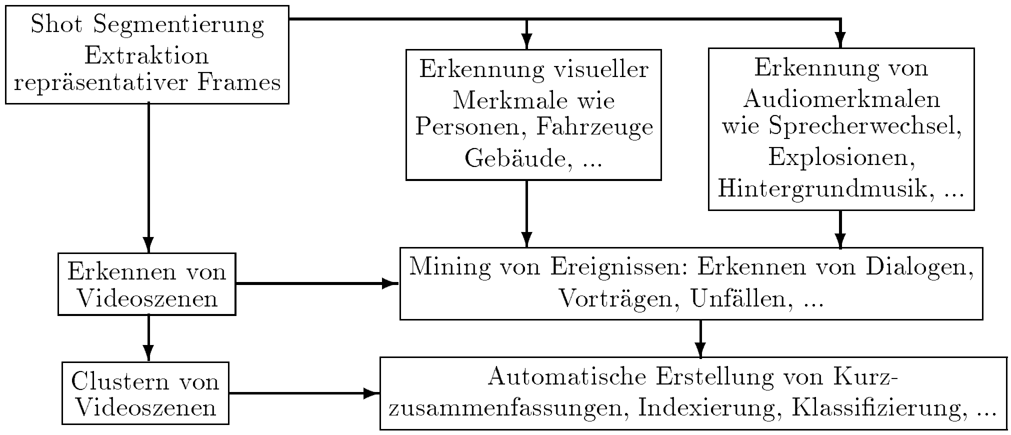

Abbildung 4: *Die Abbildung zeigt die schematische Darstellung des Video-Mining-Prozesses*

4 Video-Mining

Einfach ausgedrückt besteht das Ziel von Video-Mining darin, Videodaten so zu organisieren, dass bisher unbekannte Informationen extrahiert werden können, wobei Informationen in diesem Fall aus Ereignissen, Clustern, Klassifikationen etc. bestehen können. Hier soll dargestellt werden, wie Videodaten so verwaltet und verarbeitet werden können, dass brauchbare Informationen daraus gewonnen werden können. Im folgenden Abschnitt wird zunächst ein Data-Mining-Prozess vorgestellt und anschließend anhand eines Beispiels erläutert. Abschließend werden noch weitere Anwendungsdomänen für Video-Mining vorgestellt.

4.1 Der Video-Mining-Prozess

Ein Ziel des Video-Mining-Prozesses ist die automatische Zusammenfassung gleichartiger Szenen auf unterschiedlichen Abstraktionsniveaus [ZAF+03]. Hierbei wird angestrebt, Frames, welche die kleinste Einheit in Videodaten bilden, zu sogenannten *Shots* zusammenzufassen. Diese sind Filmsequenzen, die ohne Unterbrechung bzw. ohne Schnitt von der Kamera aufgezeichnet wurden. Auf der nächsten Ebene werden die Shots bezüglich ihrer zeitlichen und semantischen Zusammengehörigkeit zu Szenen zusammengefasst. In einigen Fällen wird noch eine Abstraktionsebene zwischen den beiden genannten eingefügt, indem die *Shots* zunächst zu Gruppen zusammengefasst werden, die jeweils zeitlich oder räumlich zusammengehörige *Shots* enthalten. Auf der höchsten Abstraktionsebene werden ähnliche Szenen, welche nicht notwendigerweise konsekutiv angeordnet sind, zu Clustern zusammengefasst. Auf jeder dieser Abstraktionebenen existieren eine Vielzahl von Mining-Anwendungen. Abbildung 4 zeigt den Ablauf eines Video-Mining-Prozesses mit Beispielanwendungen für Mining-Verfahren auf den einzelnen Ebenen.

In den folgenden Abschnitten werden zunächst die drei Schritte (1)video shot detection, (2)scene detection und (3)scene clustering behandelt. Im Anschluss wird eine Mining-Anwendung vorgestellt.

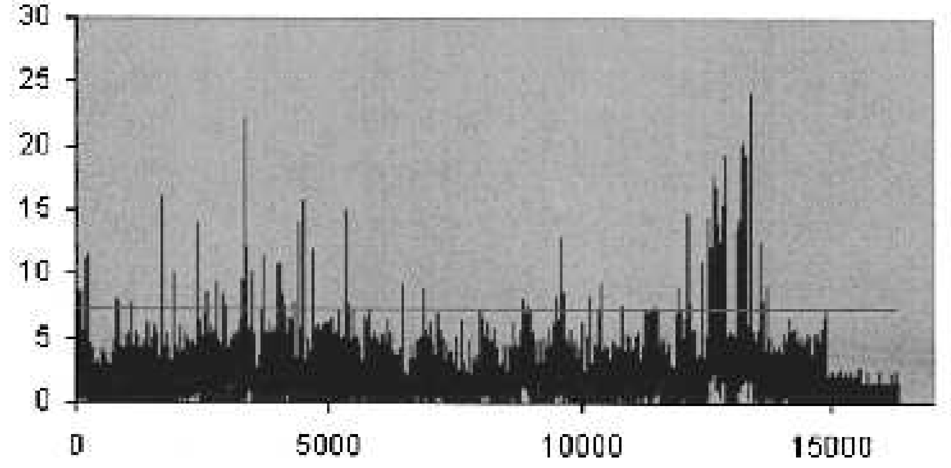

Abbildung 5: *Die Abbildung zeigt die Farb-Differenzen jeweils zweier aufeinanderfolgender Frames*

Testvideos	news1.mpg	news2.mpg	news3.mpg
Anzahl Frames	5177	6288	7024
Break shots	86	98	129
Gradual shots	6	11	7
Missed shots	4	7	15
False alarms	5	6	13

Tabelle 2: *Die Tabelle zeigt in der ersten Zeile die Anzahl der harten Schnitte und in der zweiten Zeile die Anzahl der langsamen Übergänge. In den letzten beiden Zeilen sind die nicht erkannten und die zuviel erkannten Schnitte aufgeführt*

4.1.1 Shot Segmentierung

Ein *video shot* ist ein Filmausschnitt ohne Schnitt, auch Take genannt. Aufgabe der shot cut detection-Einheit ist es, in einem Film die Schnitte zu erkennen, um ihn in *shots* zu gliedern. Für diese Aufgabe stehen zahlreiche effiziente Algorithmen zur Verfügung. Man unterscheidet hierbei zwischen Algorithmen, die auf komprimierten Videodateien wie MPEG basieren und jenen, welche auf unkomprimierten Videodaten angewendet werden können. Die meisten Verfahren basieren im Grunde darauf, Farbhistogramme aufeinanderfolgender Frames zu vergleichen [FAE+01]. Übersteigt die Differenz zweier Farbhistogramme einen Grenzwert, so wird ein Cut bzw. ein Shot erkannt. Abbildung 5 zeigt ein Diagramm, welches die Unterschiede zweier aufeinanderfolgender Farbhistogramme der jeweiligen Frames und einen Grenzwert, bei dem ein Cut erkannt wird. Tabelle 2 zeigt die Präzision aktueller Algorithmen anhand dreier MPEG-komprimierter Videosequenzen. Man kann erkennen, dass die Trefferrate bei der Erkennung eines Schnittes relativ hoch ist. Allerdings ist die Anzahl der weichen Schnitte relativ gering, was eine gute Performance begünstigt.

Nachdem die Videosequenz in shots partitioniert wurde, werden für jeden shot ein oder mehrere repräsentative Frames extrahiert. In den meisten Fällen wird der erste oder der zehnte Frame eines shots gewählt.

4.1.2 Erkennen von Szenen

In diesem Schritt werden shots zu Szenen zusammengefasst. Diese zeichnen sich meist dadurch aus, dass sie sequentiell angeordnet sind und z. B. einen ähnlichen Hintergrund haben. Zum Beispiel werden alle shots, die zu einem Gespräch gehören, zu einer Szene zusammengefasst. Auch die Analyse der Audiosignale kann in diesem Schritt hinzugezogen werden. Im Anschluss wird ein repräsentativer shot in Form des repräsentativen Frames des shots gewählt.

4.1.3 Clustern von Szenen

Im letzten Schritt werden ähnliche Gruppen zu Clustern zusammengefasst. Zum Beispiel können alle Gespräche in einem Cluster zusammengefasst werden.

4.2 Video-Mining in der Verkehrsüberwachung

Im Bereich der Analyse von Verkehrüberwachungsvideos gibt es zahlreiche Anwendungsmöglichkeiten. Mining-Methoden können eingesetzt werden, um Staus oder Unfälle zu erkennen, Fahrzeuge auf videoüberwachten Straßen automatisch zu verfolgen und den Verkehr zu steuern.
Überwachungsvideos haben einige ganz spezielle Eigenschaften. In den meisten Fällen handelt es sich um stationäre Kameras. Dadurch ergibt sich meist ein ähnlicher Hintergrund. Dies unterscheidet Überwachungsvideos von Kinofilmen, in denen Kamerafahrten, wechselnde Hintergründe, sehr viele unterschiedliche Objekte, viele Schnitte und Zwischensequenzen vorkommen. Aus diesem Grunde reicht in diesem Fall die niedrigste Abstraktionsebene aus. Aufgaben im Mining-Prozess beschränken sich in diesem Fall auf die Erkennung visueller Merkmale in Form von Fahrzeugen und in einigen Fällen zusätzlich auf die Erkennung von Audiomerkmalen.
In diesem Abschnitt soll nun eine Beispielanwendung vorgestellt werden, in der eine Kreuzung videoüberwacht wird [CSZ+01]. Hierbei wird erläutert, wie Fahrzeuge erkannt und vom Eingang in die Kreuzung bis zur Ausfahrt verfolgt werden können. Der Mining-Prozess gliedert sich in drei Teilschritte, welche im Folgenden beschrieben werden.

4.2.1 Entfernen des Hintergrundes

Im ersten Schritt wird der Hintergrund in den einzelnen Frames entfernt, d. h. alle sich nicht bewegenden Objekte werden extrahiert. Eine weit verbreitete Möglichkeit besteht darin, eine Referenzaufnahme der leeren Kreuzung zu erzeugen. Mit Hilfe dieses Referenzframes können dann später Fahrzeuge vom Hintergrund unterschieden und damit erkannt werden. Der Einsatz von Referenzframes birgt jedoch einige Nachteile in sich. Zum Beispiel ergeben sich Probleme bei wechselnden Lichtverhältnissen und beim Hinzufügen neuer fester Gegenstände. Hier existieren einige weitaus bessere Methoden. Abbildung 6 zeigt den Ablauf bei der Extraktion des Hintergrundes.

Abbildung 6: *Die Abbildung zeigt den Ablauf bei der Extraktion des Hintergrundes*

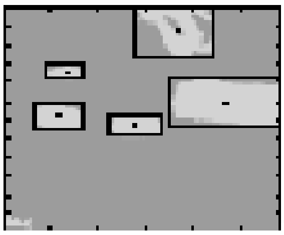

Abbildung 7: *Die Abbildung zeigt eine Verkehrssituation als Blockbild. Erkannte Fahrzeuge sind von einem Rechteck umgeben und besitzt einen repräsentativen Punkt*

4.2.2 Fahrzeugerkennung

Mit Hilfe des SPCPE-Algorithmus (Simultaneous Parition and Class Parameter Estimation) [CSS+00] wird jeder Frame in eine von zwei Klassen zugeteilt, dem Hintergrund oder der Klasse der Fahrzeuge. In jedem Frame werden dann erkannte Fahrzeuge durch Rechtecke umrandet und jeweils dessen Mittelpunkt berechnet.

4.2.3 Verfolgen von Fahrzeugen

Die Aufgabe besteht nun darin, zusammengehörige Objekte zweier aufeinanderfolgender Frames zu verbinden. Dies geschieht durch die Berechnung der euklidischen Abstände zweier Objekte. Jeweils zwei Objekte mit dem geringsten Abstand werden verbunden. Bei der Suche zusammengehöriger Objekte kann ebenfalls die Größe des Rechteckes hinzugezogen werden.
Zusätzlich wird nun jeder Frame in neun Segmente aufgeteilt. Abbildung 7 zeigt die erkannten Objekte in einem Frame und Abbildung 8 die Segmentierung mit den Segmentbeschriftungen.

1	2	3
4	5	6
7	8	9

Abbildung 8: *Die Abbildung zeigt ein Raster, welches über einen Video-Frame gelegt wird. Die Numerierung ermöglicht die Erfassung der Fahrzeugpositionen*

Für jeden Frame läßt sich nun für jedes Objekt das Segment bestimmen, in dem sich dieses Objekt befindet. Jeder Frame bzw. jede Verkehrssituation kann so in eine strukturierte textuelle Form gebracht werden, in sogenannte *Multimedia Input Strings*.

Beispiel
Gegeben sei folgender String: (a1 & b5 & c6)(a2 & b6)
Jeder geklammerte Ausdruck stellt die Verkehrssituation eines Frames dar. Im Beispiel-String sind nun zwei aufeinanderfolgende Frames dargestellt. Aus diesen lässt sich nun u. a. erschließen, dass das Objekt c die Kreuzung verlassen hat, da dieses Objekt im zweiten Klammerausdruck nicht mehr vorkommt.Das Objekt a hat sich nach rechts fortbewegt, von a1 nach a2. Anhand solcher Strings kann nun z. B. das Verkehrsaufkommen an dieser Kreuzung einfach bestimmt werden, indem die Länge der Klammerausdrücken untersucht wird.

4.2.4 Ergebnisse

Mit Hilfe des Multimedia-Inputstrings können nun unter anderem folgende Fragen beantwort werden.

- Wieviele Fahrzeuge sind von links in die Kreuzung eingefahren und haben sie nach oben wieder verlassen ?
- Wie war das Verkehrsaufkommen zwischen 08:00 und 08:30 ?

Weitere Verfeinerungen der Ergebnisse sind u. a. dadurch zu erreichen, Objekte genauer zu klassifizieren, um z. B. Busse und LKW von PKW zu unterscheiden.

4.3 Weitere Video-Mining-Anwendungen

- Fahrzeugidentifizierung
- Stauerkennung
- Verkehrssteuerung
- Unfallerkennung
- Bewertung der Kreuzungsauslastung
- Identifizieren von Strukturen in Videos
- Erkennen und Klassifizieren von Ereignissen

5 Mining in räumlichen Daten

Unter Mining in räumlichen Daten versteht man die Extraktion von Wissen, das Entdecken von Beziehungen zwischen Objekten in einer räumlichen Datenbank und anderen interessanten Mustern, die nicht explizit gespeichert sind. Es geht hierbei insbesondere auch darum, Beziehungen zwischen räumlichen und nicht-räumlichen Daten zu entdecken.
Die Anwendungen von Data-Mining in räumlichen Daten sind weit gefächert und reichen von Anwendungen in geographischen Informationssystemen, Marketing, Sensordatenverarbeitung und Verkehrsplanungssystemen bis hin zu Anwendungen bei der Erstellung von Umweltstudien.
Eine wichtige Herausforderung hierbei ist es zunächst, räumliche Daten, die meist

aus heterogenen Quellen und Systemen stammen, zu integrieren und einen effizienten Zugriff für die Anwendung von Data-Mining-Methoden zu ermöglichen. Insbesondere müssen Operationen wie die Verbundoperation, das finden der nächsten Nachbarn (Nearest Neighbours) und andere effizient ausgeführt werden können, da ein Großteil der Data-Mining-Algorithmen, die in räumlichen Daten angewendet werden, hiervon Gebrauch machen. Einen hohen Stellenwert haben hierbei die Nachbarschaftsrelationen. Denn im Gegensatz zu Daten in nicht-räumlichen Daten, sind Datensätze bzw. Objekte nicht unabhängig verteilt. In den meisten Fällen werden Objekteigenschaften von Objekten in der direkten Nachbarschaft beeinflusst. Ein Beispiel wird in Kapitel 5.4 näher erläutert. Aus diesem Grunde ist auch die Anwendung von statistischen Methoden auf räumlichen Daten problematisch, da hier die Datensätze im Gegensatz zu nicht-räumlichen Daten nicht unabhängig voneinander sind und dies eine wichtige Voraussetzung für die Anwendung der meisten statistischen Methoden ist.
Auf Grund der besonderen Bedeutung von Nachbarschaftsrelationen wird nun zunächst eine Datenstruktur Namens *Nachbarschaftsgraph* bzw. *Neighborhood Graph* vorgestellt und anschließend Operationen auf diese Datenstruktur definiert, die in Data-Mining-Verfahren Anwendung finden. Anschließend werden Verfahren zum Clustern von räumlichen Daten (Spatial Clustering), zur Charakterisierung (Spatial Characterization) und zur Erkennung von Trends in räumlichen Daten (Spatial Trend Detection) vorgestellt und jeweils Anwendungsdomänen erläutert [EsKS99].

5.1 Neighborhood Graphs und Neighborhood Paths

Wie oben bereits erläutert, spielt die Nachbarschaftsrelation im Data-Mining in räumlichen Daten eine entscheidende Rolle. Aus diesem Grund werden nun die verschiedenen Nachbarschaftsrelationen im Folgenden definiert. Anschließend werden *Neighborhood Graphs* und *Neighborhood Paths* vorgestellt und Funktionen hierauf definiert.

Nachbarschaftsrelationen Man unterscheidet im Allgemeinen folgende drei Nachbarschaftsrelationen.

- Topologie
  Unter Topologie-Relationen zwischen zwei Objekten A und B versteht man Relationen wie A schneidet B, A verdeckt B, A enthält B, A liegt in B, A berührt B und A ist gleich B.

- Distanz
  Distanz-Relationen vergleichen die Entfernung zweier Objekte mit einer gegebenen Konstante mit Hilfe eines arithmetischen Operators.
  Gegeben sei eine Distanzfunktion $dist$. Sei σ eins der drei arithmetischen Prädikate $<, >$ oder $=$ und c eine reelle Zahl. A und B seien räumliche Objekte. Dann gilt folgende Relation.

  $$A \ dist_{\sigma c} \ B \Leftrightarrow dist(A,B)_{\sigma c}$$

- Richtung
  Es wird nun eine 2-dimensionale Richtungsrelation definiert, indem zunächst aus dem Bezugsobjekt A ein repräsentativer Punkt $rep(A)$ bestimmt wird. Dieser wird gleichzeitig als Ursprung eines virtuellen Koordinatensystems verwendet. Die zu diesem Objekt A in Relation stehenden Objekte befinden sich nun in einem oder mehreren Quadranten des Koordinatensystems. Auf diese Weise kann nun die Richtung bestimmt werden. Es werden nun folgende drei Richtungsrelationen definiert.

- B nordöstlich von $A \Leftrightarrow \forall b \in B : b_x \geq rep(A)_x \wedge b_y \geq rep(A)_y$
  südöstlich, südwestlich und nordwestlich werden analog definiert
- B nördlich von $A \Leftrightarrow \forall b \in B : b_y \geq rep(A)_y$
  südlich, westlich und östlich werden analog definiert

Ebenso lassen sich Richtungsrelationen im 3-dimensionalen Raum definieren.

Neighborhood Graphs und Neighborhood Paths Basierend auf dem Konzept der Nachbarschaftsrelationen werden nun die Konzepte der Neighborhood Graphs und der Neighborhood Paths eingeführt.

- Neighborhood Graph
  Ein $Neighborhood\ Graph\ G^{DB}_{neighbor} = (N, E)$ ist ein Graph mit einer Knotenmenge N die den Objekten in einer räumlichen Datenbank entsprechen. Die Menge der Kanten $E \subseteq N \times N$ enthält alle Knotenpaare (n_1, n_2) für die eine der Nachbarschaftsrelationen hält.

- Neighborhood Path
  Ein $NeighborhoodPath$ ist eine Sequenz von Knoten $[n_1, n_2, ..., n_k]$ wobei $neighbor(n_i, n_i + 1)$ hält für alle $n_i \in N, 1 \leq i \leq k$. Die Anzahl der Knoten k bezeichnet die Länge des Neighborhood Paths.

Funktionen auf Neighborhood Graphs Im folgenden sollen nun die zwei wichtigsten Operationen auf Neighborhood Graphs vorgestellt werden.

- $neighbors : NGraphs \times Objects \times Predicates \rightarrow 2^{Objects}$
  Der Aufruf der Funktion $neighbors(graph, object, pred)$ liefert die Menge aller Objekte, die mit $object$ über eine Kante aus $graph$ verbunden sind und die angegebenen Prädikate erfüllen. Die Prädikate dienen dazu, die Anwendung von Mining-Algorithmen auf bestimmte, für eine spezielle Anwendungsdomäne relevente Knoten zu beschränken.

- $extensions : NGraphs \times NPaths \times Integer \times Predicates \rightarrow 2^{NPaths}$
  Der Aufruf der Funktion $extentions(graph, paths, max, pred)$ liefert die Menge aller Pfade, die einen Pfad aus $paths$ um maximal max Knoten erweitern. Alle hinzugenommenen Pfade müssen das Prädikat $pred$ erfüllen. Mit Hilfe dieser Prädikate können für eine spezifische Mining-Aufgabe irrelevante Pfade aussortiert werden.

Filter-Prädikate In den meisten Fällen enthalten Neighborhood Graphs sehr viele Pfade, die für eine spezielle Mining-Aufgabe irrelevant sind und eventuell zu irreführenden Ergebnissen führen. Außerdem wäre die Komplexität von Mining-Algorithmen in räumlichen Daten zu groß. Aus diesem Grunde wird die Kantenmenge beschränkt. Ein häufig verwendeter Filter ist der sogenannte Starlike-Filter. Bei der Erkennung von Trends in räumlichen Daten geht man meist von einem Startobjekt aus und betrachtet dann Veränderungen von räumlichen und nicht-räumlichen Attributen, wenn man sich vom Startobjekt entfernt. Der Starlike-Filter sortiert nun alle Pfade aus, die nicht vom Startobjekt wegführen.
In den folgenden Abschnitten werden nun einige Anwendungen von Data-Mining in räumlichen Daten vorgestellt.

5.2 Clustering von räumlichen Daten

Unter Spatial Clustering versteht man den Prozess zur Gruppierung einer Menge von räumlichen Objekten in der Art, dass Objekte innerhalb eines Clusters eine

hohe Ähnlichkeit bezüglich ausgewählter Attribute aufweisen und möglichst verschieden sind von den Objekten, die nicht dem Cluster angehören.
Spatial Clustering findet unter anderem Anwendung in der Verbrechensanalyse in Großstädten. Mit Hilfe von entsprechenden Mining-Algorithmen werden Verbrechensschwerpunkte gefunden. Dies ermöglicht es, Präventivmaßnahmen einzuleiten.
Im folgenden soll anhand des sogenannten *Single-Scan-Algorithmus* gezeigt werden, wie die oben definierten Nachbarschaftsgraphen beim Clustern von räumlichen Objekten verwendet werden.

```
SingleScanClustering(Nachbarschaftsgraph g)
    Setzte AktuelleObjekte = leer;
    Für jeden Knoten k in g
        Wenn k noch nicht in einem Cluster, dann
            Erzeuge neuen Cluster C;
            Füge k zu AktuelleObjekte hinzu;
            Solange AktuelleObjekte nicht leer
                Entferne ein Element aus AktuelleObjekte als o;
                Setzte Nachbarn = neighbors(g, o, true);
                Wenn Nachbarn die Cluster-Bedingung erfüllt
                    Füge o zu Cluster C hinzu;
                    Füge Nachbarn zu AktuelleObjekte hinzu;
Ende SingleScanClustering;
```

Entscheidend für des Ergebnis dieses Algorithmus ist die Definition der Cluster-Bedingung. Weit verbreitet sind sogenannte dichte-basierte Cluster. Die Grundidee hierbei ist, dass jeder Punkt innerhalb eines Clusters in seiner Nachbarschaft eine minimale Anzahl von Knoten haben muss. Mit Hilfe einer solchen Definition der Cluster-Bedingung können z. B. Städte und Siedlungen in räumlichen Datenbanken gefunden werden.

5.3 Charakterisieren von räumlichen Daten

Die Aufgabe der Charakterisierung besteht darin, zu untersuchen, ob bestimmte Objekte immer, häufig oder nie in der Umgebung eines anderen Objektes vorkommen. Eine durch einen Data-Mining-Algorithmus generierte Regel könnte z. B. lauten:

In einer Gemeinde leben viele Rentner $\Rightarrow$
  Anzahl der Wohnungen pro Gebäude = sehr niedrig $\wedge$
  Mittlere Größe von Unternehmen = sehr niedrig $\wedge$
  ...
  besondere Objekte = Gebirge

Auch in Bereichen wie etwa bei der Vorhersage von Naturkatastrophen wird das Charakterisieren von räumlichen Daten eingesetzt, um z. B. bestimmte Eigenschaften und Voraussetzungen für das Eintreffen von bestimmten Naturereignissen zu bestimmen. Weitere Beispiele für Mining-Ergebnisse können z. B. auch Regeln sein wie:

Campingplätze liegen meist „in der Nähe“eines „großen“Gewässers.
Schulen liegen meist„in der Nähe“eines Parks.

5.4 Erkennen von Trends in räumlichen Daten

Unter Spatial Trend Detection versteht man die Suche nach einer regelmäßigen Veränderung nicht-räumlicher Attribute mit wachsender Entfernung von einem Referenzobjekt. Als Beispiel wird im Folgenden kurz eine Anwendung in einem geographischen Informationssystem vorgestellt.

Ein geographisches Informationssystem dient u. a. der Datenrepräsentation der Erdoberfläche zusammen mit der Erfassung relevanter Einrichtungen wie Straßen oder Gebäuden.
Eine Aufgabe von Data-Mining ist es nun, geographische Eigenschaften zu finden, die die Wirtschaftskraft in einer Region positiv wie negativ beeinflussen. Hierzu wird zunächst ein Indikator zur Messung der wirtschaftlichen Leistungsfähigkeit einer Region bestimmt. In diesem Fall werden die Arbeitslosenzahlen in den Gemeinden als Indikator herangezogen. Der Mining-Prozess durchläuft nun die folgenden vier Schritte.

1. Entdecken von Zentren
   Zentren sind allgemein Objekte, die eine lokal extreme Ausprägung von meist nicht-räumlichen Attributen aufweisen. Als Zentren werden in diesem Fall Regionen mit niedrigen Arbeitslosenzahlen gewählt.

2. Festlegen eines theoretisch berechneten Trends eines nicht-räumlichen Attributs, der zu beobachten ist, wenn man sich vom vorher definierten Zentrum entfernt
   Die theoretische Verteilung der Arbeitslosenzahlen wird nach folgendem Muster berechnet: Die Arbeitslosenzahlen steigen, je weiter man sich von einem Zentrum entfernt.

3. Suchen nach Abweichungen von diesem theoretischen Trend
   Hier werden z. B. Abweichungen wie die folgende gesucht: Je weiter man sich von München in Nord-Östlicher Richtung entfernt, desto geringer wird die Arbeitslosenquote. Solche Abweichungen weisen meist auf Besonderheiten bei räumlichen Daten bzw. Objekten hin.

4. Begründung der Abweichungen
   Eine durch einen Data-Mining-Prozess gelieferte Begründung könnte z. B. wie folgt lauten: Wenn man sich von München in Richtung Flughafen wegbewegt, sinkt die Arbeitslosenquote.

Weitere Anwendungsmöglichkeiten dieser Methode finden sich im Bereich Verbrechensbekämpfung, in dem man Verbrechensschwerpunkte als Zentren definiert und annimmt, dass die Verbrechensrate sinkt, desto weiter man sich vom Zentrum entfernt. Mit Hilfe von Data-Mining können nun Abweichungen von dieser Annahme gefunden werden um so Möglichkeiten der Verbrechensprävention zu finden.

5.5 Spatial Association Rules

Ein weiterer typischer Ansatz zur Wissensgenerierung in geographischen Datenbanken ist es, Korrelationen zwischen verschiedenen Charakteristiken in bestimmten Regionen zu finden. Beispielsweise könnte ein Ergebnis sein, zu beobachten, dass Regionen mit einer hohen Anzahl von Rentnern meist in der Nähe von Gebirgen und Flüssen liegen.

6 Soziale Auswirkungen der Datenanalyse und Trends

Im Folgenden soll es nun beispielhaft um die möglichen Auswirkungen von Datensammlung und Datenauswertung gehen. Insbesondere soll hier auf Probleme im Umgang mit Unternehmensdaten und Kundendaten eingegangen werden [Luep03].

6.1 Soziale Auswirkungen der Datenanalyse

Am Beispiel des Customer Relationship Management eines Unternehmens sollen im folgenden mögliche positive und negative Auswirkungen des Data-Mining für die Gesellschaft verdeutlicht werden. Das Customer Relationship Management eines Unternehmens hat die Verbesserung und Stärkung der Kundenbindung an das Unternehmen zum Ziel. Aus diesem Grunde versuchen Unternehmen möglichst viele Daten über Konsumenten zu sammeln. Eine immer weiter verbreitete Möglichkeit an Kundendaten zu kommen ist die Einführung von Rabatt- oder Kundenkarten. Hierbei legt der Kunde bei jedem Einkauf seine Karte vor und sorgt somit aktiv für Daten über sein Konsumverhalten. Dies kann sich für den Kunden zunächst positiv auswirken. Zum einen werden ihm meist Rabatte zugeschrieben, und zum anderen bekommt der Kunde weniger Werbung, und nur die, die für ihn potentiell interessant ist. Folgendes Beispiel zeigt jedoch, wie sich die über ihn gesammelten Daten auch negativ auswirken können.

Miles & More Ein Beispiel für solche Auswirkungen soll anhand der Lufthansa Karte Miles & More gegeben werden. Hierbei erhalten die 1000 besten Vielflieger sogenannte VIP-Karten, mit denen sie auf jedem Flug einen Platz erhalten, auch wenn der Flug bereits ausgebucht ist. Im Zweifel wird hierfür ein normaler Fluggast wieder ausgebucht.
Negative Konsequenzen können sich allerdings dann ergeben, wenn die Fluggesellschaft nicht nur die geflogenen Flugmeilen erfasst, sondern zusätzlich auch das Flugziel, wer bezahlt hat und wer Begleitperson war. Hieraus können Rückschlüsse auf den Zweck der Reise geschlossen werden und temporäre Veränderungen erfasst werden. Wenn ein Kunde z. B. früher häufig geschäftlich im Ausland war und mit der Familie teuere Urlaubsreisen gemacht hat und nun nur noch Inlandsflüge einer niedrigen Klasse bucht, könnte die Fluggesellschaft daraus schließen, dass es seinem Arbeitgeber schlechter geht, oder der Kunde einen Abstieg innerhalb des Unternehmens erlitten hat. Wenn nun die Fluggesellschaft mit anderen Unternehmen kooperiert, können solche Vermutungen für die Betroffenen ungeahnte Folgen haben, z. B. beim Abschluss von Kreditverträgen.

Doch nicht nur der Kunde kann ein Opfer von gesammelten und elektronisch verfügbaren und auswertbaren Informationen über ihn sein. Im Unternehmensbereich gibt es ähnliche Szenarien. Hier lassen sich unter anderem aus Bestellmengen, aus Unternehmenspräsentationen oder auf Grund von rechtlichen Regelungen wie z. B. Veröffentlichungspflichten bei börsennotierten Unternehmen Informationen gewinnen, die für Konkurenzunternehmen Rückschlüsse auf Unternehmensgeheimnisse zulassen.
Und waren die Analysemethoden vor nicht allzu langer Zeit noch auf strukturierte Daten in meist relationalen Datenbanken beschränkt, so vervielfältigen sich die Anwendungsmöglichkeiten durch Mining-Verfahren für unstrukturierte Daten enorm.
Im folgenden Abschnitt werden nun mögliche künftige Entwicklungen im Data-Mining-Sektor beleuchtet.

6.2 Trends im Data-Mining

Die Vielfältigkeit der vorhandenen Daten und Data-Mining-Aufgaben stellen auch weiterhin große Herausforderungen an die Forschung. Einige dieser möglichen Herausforderungen, der sich die Wissenschaft in Zukunft zu stellen hat, sollen nun im Folgenden kurz erläutert werden.

- Vervielfältigung der Anwendungsmöglichkeiten
  In der Anfangszeit wurde Data-Mining nahezu ausschließlich angewandt,

um in Unternehmen wettbewerbsfähig zu bleiben, indem unter anderem Kauf-Transaktionsdaten ausgewertet wurden. Mittlerweile finden Mining-Algorithmen unter anderem Anwendung in der Medizin, Finanzanalyse und der Telekommunikation, und mit der wachsenden Verbreitung von e-Commerce und e-Business kommen weitere Anwendungsfelder hinzu. Insgesamt kann man einen Trend hin zu sehr anwendungsspezifischen Mining-Algorithmen beobachten, während generische Data-Mining-Verfahren seltener werden.

- Skalierbare Data-Mining-Methoden
  Im Gegensatz zu traditionellen Datenanalyse-Verfahren müssen aktuelle Data-Mining-Algorithmen mit einer großen Menge von oft heterogenen Daten umgehen können. In den letzten Kapiteln wurden einige Datenquellen vorgestellt, auf die Data-Mining-Methoden angewandt werden. Insbesondere beim Mining in räumlichen Daten sind die Datenmengen extrem groß und oft in unterschiedlichen Formaten abgelegt. Hier ist ein Trend zu erkennen hin zu Constraint-basiertem Mining, um zum einen die Laufzeit der Analysealgorithmen zu beschränken und zum anderen irreführende Daten mit Hilfe von domänenspezifischem Wissen zu entfernen.

- Neue Methoden zum Mining von komplexen Datentypen
  Hier besteht immer noch eine große Lücke zwischen aktuellen Bedürfnissen und der verfügbaren Technologie. Insbesondere bei der Integration von Data-Mining-Methoden und existierenden Datenanalyseverfahren für diese komplexen Datentypen gibt es noch große Herausforderungen.

In allen in dieser Arbeit vorgestellten Data-Mining-Domänen steht die Wissenschaft noch vor großen Herausforderungen. Insbesondere die durch das Internet hervorgerufene Informationsflut macht die Entwicklung weiterer effizienter Mining-Algorithmen unabdingbar, um diese Informationen in Wissen umzuwandeln und sinnvoll zu nutzen.

Literatur

[AaEi99] Aas, K.; Eikvil, L.: Text Categorisation - A survey. Technical Report No.941, Juni, 1999.
(Elektronisch verfügbar unter: www.nr.no/documents/samba/research_areas/BAMG/Publications/tm_survey.ps)

[CSS+00] Chen, S.C.; Sista, S.; Shyu, M.L.; Kashyap, R.L.: An indexing and searching structure for Multimedia Database Systems.In:*Storage and Retrieval for Media Databases, San Jose, USA, Januar 2000*, S.262-270
(Elektronisch verfügbar unter: www.eng.miami.edu/˜shyu/Paper/2000/SPIE00.ps)

[CSZ+01] Chen, S.C.; Shyu, M.L.; Zhang, C.; Strickrott, J.: Multimedia Data Mining for Traffic Video Sequences.In:*Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), San Francisco, USA, 26.August 2001*, S.78-85
(Elektronisch verfügbar unter: http://citeseer.nj.nec.com/chen01multimedia.html)

[EsKS99] Ester, M.;Kriegel, H.-P.;Sander, J.: Knowledge Discovery in Spatial Databases.In:*Proc. 23rd Annual German Conf. on Artificial Intelligence (KI'99), Springer Verlag, Berlin, 1999*, S.61-74.
(Elektronisch verfügbar unter: www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/SpatialDM.invited.pdf)

[FAE+01] Fan, J.; Aref, W.G.; Elmagarmid, A.K.; Hacid, M.S.; Marzouk, M.S.; Zhu, X.: MultiView: Multilevel video content representation and retrieval. In:*Journal of Electronic Imaging, Oktober 2001*, S.895-908
(Elektronisch verfügbar unter: www.cs.uvm.edu/˜xqzhu/publications/jei005.pdf)

[FaHL02] Fang, B.; Hsu, W.; Lee, M.L.: Tumor Cell Identification using Features Rules.In: *8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , Edmonton, Canada, Juli 2002*, S.495-500
(Elektronisch verfügbar unter: www.comp.nus.edu.sg/˜leeml/papers/kdd02.pdf)

[JWB+03] Janssens, D.; Wets, G.; Brijs, T.; Vanhoof, K.; Chen, G.: Adapting the CBA Algorithm by means of Intensity of Implication.In:*Proceedings of the First International Conference on Fuzzy Information Processing Theories and Applications, Beijing, China, März 2003*, S.397-403

[Luep03] Lüpkes, C.: Soziale und rechtliche Aspekte der Datenanalyse, 2003.
(Elektronisch verfügbar unter: www.diko-project.de/dokumente/ausarbeitungen/luepkes.pdf)

[Tan99] Tan, A.H.: Text Mining: The state of the art and the challenges.In:*Workshop an Knowledge Discovery from Advanced Databases, Beijing, China, April 1999*, S.65-70
(Elektronisch verfügbar unter: www.ewastrategist.com/papers/text_mining_kdad99.pdf)

[ZAF+03] Zhu, X.;Aref, W.G.; Fan, J.; Catlin, A.C.; Elmagarmid, A.K.: Medical Video Mining for Efficient Database Indexing, Management and Access.In:*International Conference on Data Engineering, 2003*, S.569-580
(Elektronisch verfügbar unter: www.cs.purdue.edu/vdbms/papers/icde03.pdf)

[ZhHL01] Zhang, J.; Hsu, W.; Lee, M.L.: Image Mining: Issues, Frameworks and Techniques.In: *Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), San Francisco, USA, August, 2001*, S.13-20
(Elektronisch verfügbar unter: www.comp.nus.edu.sg/˜whsu/publication/2001/MDM_paper5.pdf)