Knowledge Discovery in Databases

Prof. Gerd Stumme
Dipl.-Math. Robert Jäschke
FG Wissensverarbeitung

http://www.lkw-stefan.de/galerie/tagebau/tagebau.htm

Organisatorisches

Vorlesung

- Beginn: 17. April 2007
- Dienstag, 10.15 – 11.45 Uhr in Raum 0443

Übungen

- Donnerstag, 8.30 h - 10.00 h, in Raum 0443
- Beginn: 26. April 2007
- wird als Präsenzübung abgehalten (s. nächste Folie)
- praktische Übungen mit Weka (http://www.cs.waikato.ac.nz/~ml/weka/)
  (Bonus in der Klausur oder mündlichen Prüfung bei erfolgreicher Teilnahme)

Organisatorisches

Präsenzübung bedeutet

- **selbständiges Bearbeiten** des Übungsblattes in Kleingruppen à 3-4 Personen unter Betreuung des Assistenten
- **kein prinzipielles Wiederholen** des Vorlesungsstoffs
- **kein Vorrechnen** der Musterlösung etc. (Diese wird später zur Verfügung gestellt.)

- **Nötig dafür:**
  - selbständige Vorlesungsnachbereitung **vor** der Übung
  - Mitbringen des Skriptes
  - eigene Aktivität entfalten

Organisatorisches

Warum ein neues Übungskonzept?

- aktives Erarbeiten des Vorlesungsstoffes bringt mehr
- Zusammenhänge im Stoff erkennen
- strukturiertes Denken und selbständiges Arbeiten lernen
- Teamarbeit lernen
- Erklären lernen (als Tutor und als Teilnehmer)
- Klausurtraining ;-)
- *Ihr Studium der ... haben Sie abgeschlossen. Zu Ihren persönlichen Stärken zählen Sie Eigeninitiative, Kommunikations- und Kooperationsbereitschaft, Teamarbeit.* (Typischer Anzeigentext)

Organisatorisches

Sprechstunden nach Absprache:

Prof. Gerd Stumme: stumme@cs.uni-kassel.de, 0561/804-6251
Dipl.-Math. Robert Jäschke: jaeschke@cs.uni-kassel.de, 0561/804-6253

FG Wissensverarbeitung, FB Mathematik/Informatik
Raum 0439, Wilhelmshöher Allee 73

Informationen im Internet: http://www.kde.cs.uni-kassel.de
Hier ist u.a. folgendes zu finden:

- aktuelle Ankündigungen
- Folienkopien
- Übungsblätter
- Literaturempfehlungen
- Termine

1. Einleitung

Inhalt dieses Kapitels

1.1 Grundbegriffe des Knowledge Discovery in Databases
Motivation, Begriffe, Schritte des KDD-Prozesses, Verstehen der Anwendung, Vorverarbeitung, Transformation, Data Mining, Evaluation

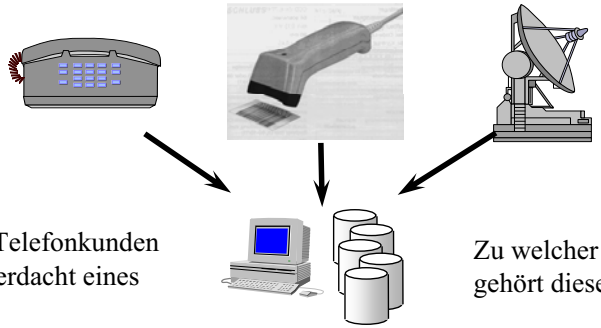
1.2 Typische KDD-Anwendungen
Astronomie, Erdwissenschaften, Marketing, Investment, Betrugserkennung, Individualisierte Werbeanzeigen, Electronic Commerce, Datenschutz

1.3 Inhalt und Aufbau der Vorlesung

Die Folien wurden im wesentlichen aus dem Buch „Knowledge Discovery in Databases“ von M. Ester, J. Sander und den im dazu verfügbaren Folien aus dem Web sowie vom Institut AIFB der Universität Karlsruhe (R. Engels, M. Erdmann, A. Hotho, A. Mädche, S. Staab, R. Studer, G. Stumme) übernommen.

1. 1 Motivation

Riesige Datenmengen werden automatisch gesammelt.



Bei welchen Telefonkunden besteht der Verdacht eines Betrugs?

Zu welcher Klasse gehört dieser Stern?

Welche Assoziationen bestehen zwischen den in einem Supermarkt gekauften Waren?

➡ Solche Analysen lassen sich nicht mehr manuell durchführen.

1.1 Definition KDD

[Fayyad, Piatetsky-Shapiro & Smyth 96]

Knowledge Discovery in Databases (*KDD*) ist der Prozeß der (semi-) automatischen Extraktion von Wissen aus Datenbanken, das

- *gültig*
- *bisher unbekannt*
- und *potentiell nützlich* ist.

Bemerkungen:

- *(semi)-automatisch*: im Unterschied zu manueller Analyse. Häufig ist trotzdem Interaktion mit dem Benutzer nötig.
- *gültig*: im statistischen Sinn.
- *bisher unbekannt*: bisher nicht explizit, kein „Allgemeinwissen“.
- *potentiell nützlich*: für eine gegebene Anwendung.

1.1 Problemstellung

Data Mining

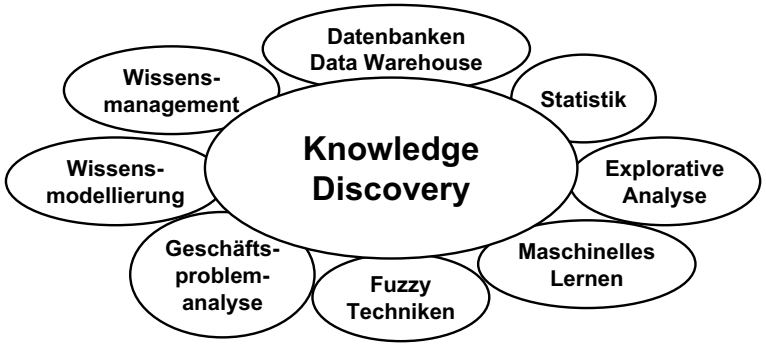
- **zwei alternative Bedeutungen**
- **Bedeutung (1):**
  - Synonym für KDD: beinhaltet alle Aspekte des Prozesses der Wissensgewinnung
  - Diese Bedeutung ist insbesondere in der Praxis verbreitet.
- **Bedeutung (2):**
  - Teil des KDD-Prozesses: Mustergewinnung / Modellierung, Interpretation
  - Anwendung von Algorithmen, die unter gewissen Ressourcenbeschränkungen Muster / Modelle E bei gegebener Faktenmenge F erzeugen

"Data Archeology"

(Brachman)

1.1 Abgrenzung KDD

KDD nutzt und integriert eine Vielzahl von Methoden und Techniken aus verschiedenen Gebieten:



1.1 Abgrenzung KDD

- Statistik
  - Teilgebiet der angewandten Mathematik, fußt auf der statistischen Theorie bzw. Wahrscheinlichkeitstheorie
  - Analyse und Beschreibung von empirischen Daten
  - statistische Methoden werden heute häufig im KDD eingesetzt [Maitra 2002]
- Maschinelles Lernen
  - Teilgebiet der Künstlichen Intelligenz
  - Entwicklung von Verfahren, welche es Computern erlaubt, aus Daten „zu lernen"
  - Schwerpunkt auf symbolischen Daten [Mitchell 1997]
- Datenbanksysteme
  - Notwendig für den Umgang mit sehr großen Datenmengen
  - Skalierbarkeit, konsistenter Zugriff und Speicherung
  - neue Datentypen (z.B. Webdaten)
  - Integration mit kommerziellen Datenbanksystemen [Chen, Han & Yu 1996]

1.1 Der KDD Prozess

Der KDD Prozess

(CRISP: **http://www.crisp-dm.org/CRISPWP-0800.pdf**)
(Fayyad et al. 1996, chapter 2)
(Engels 1999, Han)

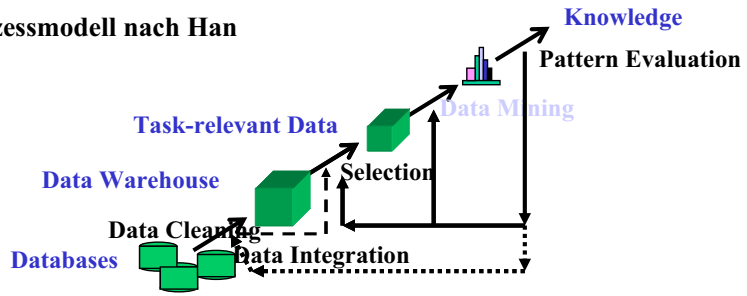
> **"Knowledge discovery is a knowledge-intensive task consisting of complex interactions, protacted over time, between a human and a (large) database, possibly supported by a heterogeneous suite of tools"**
>
> (Brachman/Anand 1996)

1.1 Der KDD Prozess

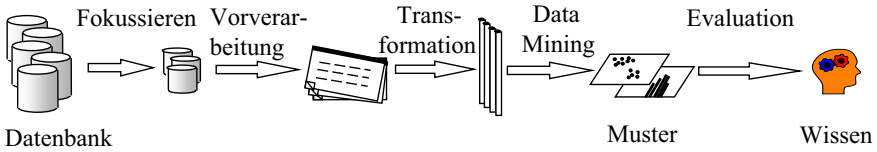
- **Der KDD-Prozess muss in Beziehung zur Anwendungsaufgabe und dem Anwender (Prozessentwickler) stehen.**
- **Der Entwicklung benötigt einiges Wissen über Datenbanken, Datenanalysemethoden und das Anwendungsgebiet.**
- **Der KDD-Prozess besteht aus eine Menge verschiedener Schritte.**
- **Der KDD-Prozess ist interaktiv und iterativ.**
  - Anwender muss entscheiden
  - einige Schritte müssen mehrfach wiederholt werden

1.1 KDD-Prozess

Prozessmodell nach Han



Prozessmodell nach Fayyad, Piatetsky-Shapiro & Smyth



1.1 Der KDD-Prozess

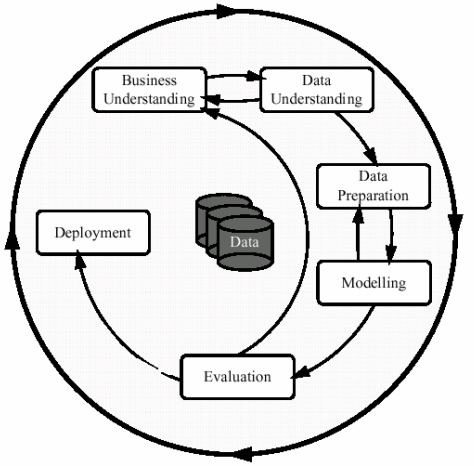


Figure 2: Phases of the CRISP-DM Reference Model

1.1 Der KDD-Prozess – Crisp-DM Modell

Hierarchisches Prozess Model mit vier Stufen der Abstraktion: phase, generic task, specialized task and process instance

- **Phasen: Top-Level-Prozesse**
  - business understanding
  - data understanding
  - data preparation
  - modelling
  - evaluation
  - deployment
- **Allgemeine Aufgabe (generic task): Jede Phase wird in einzelne allgemeine Aufgaben zerlegt**
  - Abdeckung des gesamten Prozesses (Vollständigkeit)
  - Abdeckung aller möglichen Anwendungen (dauerhaft)

1.1 Der KDD-Prozess

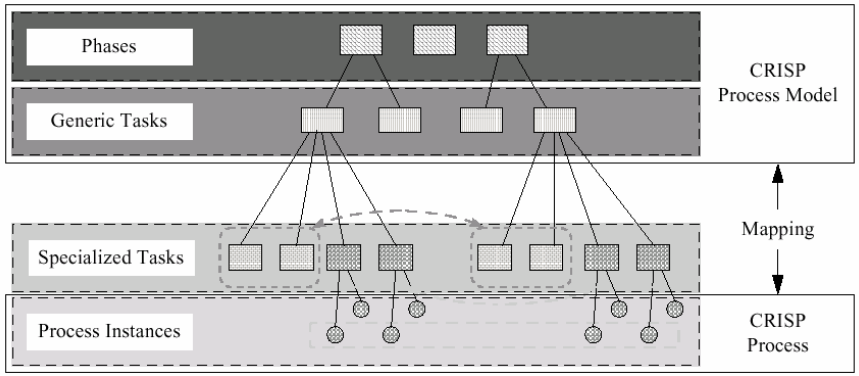


Figure 1: Four Level Breakdown of the CRISP-DM Methodology

1.1 Der KDD Prozess

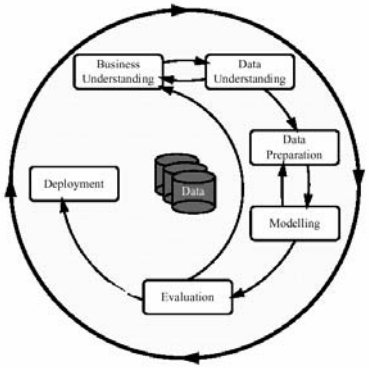
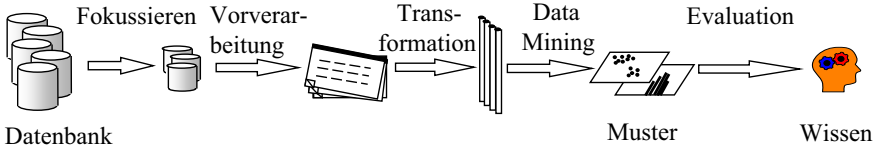


Figure 2: Phases of the CRISP-DM Reference Model



1.1 Fokussieren (Business Understanding, Data Understanding)

- Verständnis der gegebenen Anwendung

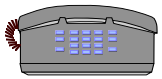
  z.B. Tarifgestaltung in der Telekommmunikations-Branche
- Definition des Ziels des KDD

  z.B. Segmentation der Kunden
- Beschaffung der Daten

  z.B. aus operationaler DB zur Abrechnung
- Klärung der Verwaltung der Daten

  File System oder DBS?
- Selektion der relevanten Daten

  z.B. 100 000 ausgewählte Kunden mit allen Anrufen in 1999



Bsp.-Anwendung

1.1 Fokussieren (Business Understanding, Data Understanding)

- "File Mining"
  - Daten meist in Datenbanksystem (DBS)
  - Data Mining bisher auf speziell vorbereiteten Files
- Integration des Data Mining mit DBS
  - Vermeidung von Redundanzen und Inkonsistenzen
  - Nutzung der DBS-Funktionalität (z.B. Indexstrukturen)
- Basisoperationen zum Data Mining
  - Standard-Operationen für eine Klasse von KDD Algorithmen
  - effiziente DBS-Unterstützung
    - schnellere Entwicklung neuer KDD Algorithmen
    - bessere Portabilität der Algorithmen

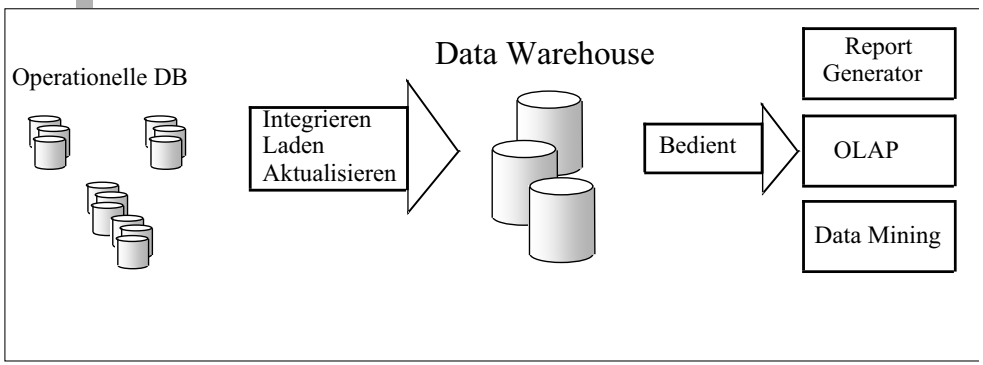
1.1 Vorverarbeitung (Preprocessing)

- Integration von Daten aus unterschiedlichen Quellen
  - einfache Übersetzungen von Attributnamen (z.B. KNr --> KundenSchl)
  - Nutzen von Anwendungswissen um ähnliche Daten zusammenzufassen (z.B. regionale Zuordnung von Postleitzahlen)
- Konsistenzprüfung
  - Test anwendungsspezifischer Konsistenzbedingungen
  - Bereinigung von Inkonsistenzen
- Vervollständigung
  - Ersetzen von unbekannten Attributwerten durch Defaults
  - Verteilung der Attributwerte soll i.A. erhalten bleiben!

Vorverarbeitung ist häufig einer der aufwendigsten KDD-Schritte

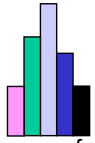
1.1 Vorverarbeitung (Preprocessing)

- Data Warehouse [Chaudhuri & Dayal 1997]
  - dauerhafte
  - integrierte Sammlung von Daten
  - aus unterschiedlichen Quellen
  - zum Zweck der Analyse bzw. Entscheidungsunterstützung



1.1 Transformation (Preprocessing)

- Diskretisierung numerischer Attribute
  - unabhängig von der Data-Mining-Aufgabe
    z.B. Aufteilung des Wertebereichs in Intervalle gleicher Länge
  - abhängig von der Data-Mining-Aufgabe
    z.B. Aufteilung in Intervalle so, daß der Informationsgewinn in Bezug auf die Klassenzugehörigkeit maximiert wird



- Erzeugen abgeleiteter Attribute
  - durch Aggregation über Mengen von Datensätzen
    z.B. von einzelnen Anrufen zu „Gesprächsminuten tagsüber, Wochentag, Stadtgespräch“
  - durch Verknüpfung mehrerer Attribute
    z.B. Umsatzänderung = Umsatz 2000 - Umsatz 1999

1.1 Transformation (Preprocessing)

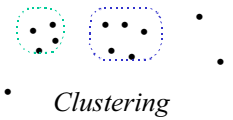
•Attribut-Selektion
  - *manuell*
    wenn Anwendungswissen über die Bedeutung der Attribute und über die gegebene Data-Mining-Aufgabe bekannt ist
  - *automatisch*
    Bottom-Up (ausgehend von der leeren Menge jeweils ein Attribut hinzufügen)
    Top-Down (ausgehend von der Gesamtmenge der Attribute jeweils ein Attribut entfernen)
    z.B. so, daß die Diskriminierung der Klassen optimiert wird

⇒ zuviele Attribute führen zu Ineffizienz und evtl. Ineffektivität des Data Mining

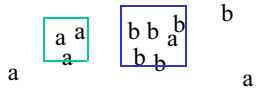
⇒ manche Transformationen können durch OLAP-Systeme realisiert werden

1.1 Data Mining (Modelling)

- Definition [Fayyad, Piatetsky-Shapiro, Smyth 96]
  - *Data Mining* ist die Anwendung effizienter Algorithmen, die die in einer Datenbank enthaltenen Muster liefern.
- Data-Mining-Aufgaben



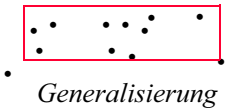
Clustering



Klassifikation

A und B --> C

Assoziationsregeln



Generalisierung

andere Aufgaben: Regression, Entdecken von Ausreißern . . .

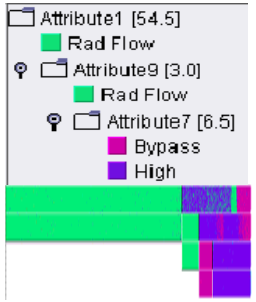
1.1 Data Mining (Modelling)

Anwendungen

- Clustering
  Customer Segmentation, Strukturierung von Mengen von Webdokumenten
- Klassifikation
  - Bewertung der Kreditwürdigkeit, Automatische Analyse astronomischer Beobachtungen
- Assoziationsregeln
  Reorganisation eines Supermarkts, Cross-Selling der Produkte einer Firma
- Generalisierung
  Beschreibung von Clustern, Kundengruppenanalyse

1.1 Evaluation

Ablauf



- Präsentation der gefundenen Muster
  häufig durch entsprechende Visualisierungen
- Bewertung der Muster
  durch den Benutzer
- falls schlechte Bewertung:
  erneutes Data Mining mit
  - anderen Parametern
  - anderem Verfahren
  - anderen Daten
- falls gute Bewertung:
  Integration des gefundenen Wissens in die Wissensbasis
  Nutzung des neuen Wissens für zukünftige KDD-Prozesse

1.1 Evaluation

Bewertung der gefundenen Muster

Vorhersagekraft der Muster
- Verwendete Daten sind Stichprobe aus der Grundgesamtheit aller Daten.
- Wie gut lassen sich die in diesen „Trainingsdaten“ gefundenen Muster auf zukünftige Daten verallgemeinern?
- Vorhersagekraft wächst mit Größe und Repräsentativität der Stichprobe.

Interessantheit der Muster
- Muster schon bekannt?
- Muster überraschend?
- Muster für viele Fälle anwendbar?

1.1 Anwendung (Deployment)

Erstellung einer Anwendung im Unternehmen

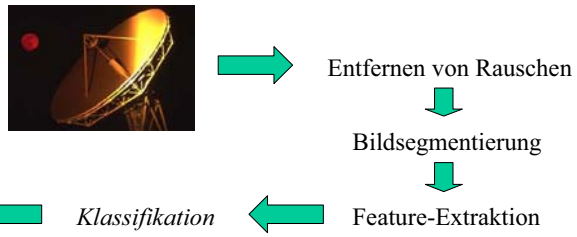
- **Planung des Einsatzes der KDD-Anwendung**
  - Erstellung eines Planes zu Einführung der Anwendung
- **Planung des Monitorens und der Wartung**
  - Wann sollten Modelle nicht mehr verwendet werden?
  - Ändern sich die Geschäftsziele mit der Zeit?
- **Erstellung der endgültigen Berichtes**
  - Wer ist die Zielgruppe für die Präsentation?
- **Review des Projektes**
  - Zusammenfassung der wichtigsten Erkenntnisse und Erfahrungen
  - Integration der Projektergebnisse in die Strategie des gesamten Unternehmens.

1.2 Typische KDD-Anwendungen

- *Astronomie*

SKICAT-System [Fayyad, Haussler & Stolorz 1996]

- Architektur



- Technik der Klassifikation: Entscheidungsbaum-Klassifikator
- Evaluation
  - wesentlich schneller als manuelle Klassifikation
  - Klassifikation auch von sehr entfernten (lichtschwachen) Objekten

1.2 Typische KDD-Anwendungen

- Email-Spam-Filterung
  - auf Text-Ebene
    - Einfache Bayes-Klassifikatoren
      - Techniken sind effektiv
      - Einsatz erfolgt in Tools wie SpamAssassin
      - Problem: werden von Spammern mittlerweile umgangen
    - Good Word Attacks on Statistical Spam Filters. Daniel Lowd and Christopher Meek. Second Conference on Email and Anti-Spam (CEAS) (2005)
      - Finden von Worten, die von Spammern aktiv genutzt werden, um den Spamschutz zu unterlaufen
  - Graph-basiert
    - Boykin, P., & Roychowdhury, V. (2004). Personal email networks: an effective anti-spam tool. Preprint, arXiv id 0402143.
      - Analysiert das Netzwerk aus Adressen der eigenen Emails, um Teilnetze aus Freunden und Spammern zu identifizieren

1.2 Typische KDD-Anwendungen

Marketing

Kundensegmentierung [Piatetsky-Shapiro, Gallant & Pyle 2000]

- Ziel: Aufteilung der Kunden in Segmente mit ähnlichem Kaufverhalten
- Nutzen
  - Ideen für Produkt-Pakete (Product Bundling)
  - Entwickeln einer neuen Preispolitik (Pricing)

Projektablauf

- Entwicklung verschiedener automatischer Modelle (Bayesian Clustering)
  zu komplex, keine Berücksichtigung von Anwendungswissen
- manuelle Entwicklung einer Entscheidungsliste
  aufgrund der gewonnenen Erkenntnisse
- Umsetzung der Erkenntnisse im Marketing der Firma
- Integration der Entscheidungsliste in Software-Umgebung

1.2 Typische KDD-Anwendungen

Electronic Commerce

Elektronische Vergabe von Kreditkarten [Himelstein, Hof & Kunii 1999]

- Bisher
  - manuelle Analyse der Kreditwürdigkeit
  - erfordert Zugriffe auf verschiedene Datenbanken
  - dauert u.U. Wochen
- Mit Data-Mining
  - Analyse des Kunden in wenigen Sekunden
  - wesentlich schnellerer Service
  - erlaubt es einer Kreditkartenfirma, zahlreiche neue Kunden zu gewinnen
- Technik: Entscheidungsbaum-Klassifikator

1.2 Typische KDD-Anwendungen

Purchase Circles [Beck 1999]

– Idee

– Methode
  - Gruppieren der Käufe nach PLZen und Domains
  - Aggregation dieser Daten
  - Konstruktion von Bestseller-Listen, die in dieser Kundengruppe populärer sind als in der Gesamtheit aller Kunden

1.2 Typische KDD-Anwendungen

Datenschutz

- Große Gefahren des Missbrauchs der Data-Mining-Techniken,
- insbesondere dann, wenn persönliche Daten ohne Kenntnis der betreffenden Person gesammelt und analysiert werden.
- Datenschutz (*privacy*) muß im Kontext des KDD neu diskutiert werden!

Beispiel Amazon.com
[http://www.amazon.com/exec/obidos/subst/misc/policy/privacy.html]

- sammelt persönliche Daten.

  „*... when you search, buy, bid, post, participate in a contest or questionnaire, or communicate with customer service.*“
- erlaubt dem Kunden, die persönlichen Daten zu überprüfen und zu korrigieren.
- bildet z.B. Purchase Circles nur für mindestens 200 Kunden.

1.3 Inhalt und Aufbau der Vorlesung

Lernziele

- Überblick über KDD
- Kenntnis der wichtigsten Aufgaben und Verfahren mit Vor- und Nachteilen
- Auswahl und Einsatz eines Verfahrens für eine gegebene Anwendung
- Entwicklung eigener Verfahren für eine gegebene Anwendung
  - kurze Einführung der Grundlagen aus den Bereichen Statistik und der Datenbanksysteme
  - Vorverarbeitung unterschiedlicher Daten
  - Schwerpunkt auf dem zentralen KDD-Schritt des Data Mining
- Fragestellungen aus verschiedenen Sichten wie Datenbanken, Web- und Textanwendungen, Information Retrieval oder dem Semantic Web:
  - Skalierbarkeit für große Datenmengen
  - unterschiedliche Datentypen (z.B. Webdaten, Worte, Relationen)

1.3 Inhalt und Aufbau der Vorlesung

Aufbau der Vorlesung (1)

1. Einleitung
2. Grundlagen des KDD
   Statistik, Datenbanksysteme, OLAP, Preprocessing

Unüberwachte Verfahren

3. Clustering
   partitionierende und hierarchische Verfahren, Verfahren aus DBS-Sicht, neue Anforderungen und Techniken des Clustering
4. Assoziationsregeln
   einfache Assoziationsregeln, Algorithmus Apriori, Einbeziehung von Taxonomien, numerische Attribute

1.3 Inhalt und Aufbau der Vorlesung

Aufbau der Vorlesung (2)

Überwachte Verfahren

5. Klassifikation
   Bayes-, nächste-Nachbarn- und Entscheidungsbaum-Klassifikatoren, SVM
6. Besondere Datentypen und -anwendungen
   Temporal Data Mining, Spatial Data Mining, Web Mining
7. Text Mining
   Linguistische Vorverarbeitung, Text Clustering
8. Andere Paradigmen
   induktive Logik-Programmierung, genetische Algorithmen, neuronale Netze, Visualisierung großer Datenmengen

1.3 Inhalt und Aufbau der Vorlesung

Literatur

Textbuch der Vorlesung

- Ester M., Sander J., „*Knowledge Discovery in Databases: Techniken und Anwendungen*“, Springer Verlag, September 2000.

Weitere Bücher

- Berthold M., Hand D. J. (eds.), „*Intelligent Data Analysis: An Introduction*“, Springer Verlag, Heidelberg, 1999.
- T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.
- Han J., Kamber M., „*Data Mining: Concepts and Techniques*“, Morgan Kaufmann Publishers, August 2000.
- Mitchell T. M., *„Machine Learning“*, McGraw-Hill, 1997.
- Witten I. H., Frank E., „*Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*“, Morgan Kaufmann Publishers, 2000.

1.3 Inhalt und Aufbau der Vorlesung

- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurasamy. **Advances in Knowledge Discovery and Data Mining**. Cambridge, London. MIT press, 1996.
- P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth: **CRoss Industry Standard Process for Data Mining**, 1999, http://www.crisp-dm.org/
- Weitere Literatur findet sich auf der Homepage der Vorlesung und unter http://www.bibsonomy.org/kdd.