Wissensentdeckung in Datenbanken / Data Mining

Prof. Dr. Katharina Morik

Datenflut

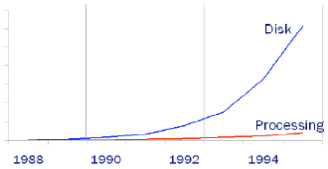
- Immer mehr Daten werden generiert:
  - Banken, Telecom, Geschäftstransaktionen ...
  - Wissenschaftliche Daten: Astronomie, Biologie, ...
  - Web, Text, e-commerce
- Immer mehr Daten werden erfasst:
  - Speichertechnik schneller and billiger
  - DBMS handhaben größere DB

Beispiele

- Europas Very Long Baseline Interferometry (VLBI) hat 16 Teleskope, jedes produziert astronomische Daten von **1 Gigabit/Sekunde** in einer 25 Tage dauernden Beobachtung
- Walmart berichtet, dass sie 24 Tera-byte DB benutzen
- AT&T bearbeitet Milliarden Anrufe pro Tag
  - Daten können nicht gespeichert werden -- Analysis “on the fly”

Wachstumsraten

- Moore’s Gesetz
  - Rechengeschwindigkeit verdoppelt sich alle 18 Monate
- Speichergesetz
  - Speicherkapazität verdoppelt sich alle 9 Monate
- Konsequenz
  - Nur sehr wenige Daten werden je von einem Menschen angesehen!



Wissensentdeckung ist **notwendig**, um sinnvollen Gebrauch von den Daten zu machen.

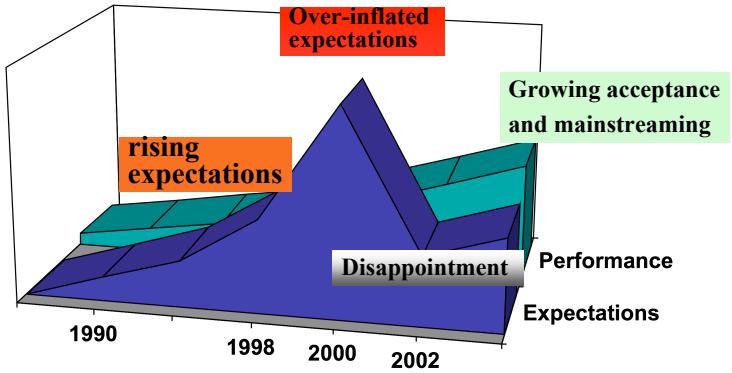
Wissensentdeckung ist...

- ... der nichttriviale Prozess der Identifikation
  - gültiger,
  - neuer,
  - potenziell nützlicher und
  - schlussendlich verständlicher Muster in (sehr großen) Datenbeständen.

Wissensentdeckung

- Wissensentdeckung benötigt Verfahren, die sehr große Datenmengen verarbeiten können.
  - Die gegebenen Daten sind üblicherweise viele Tabellen mit vielen Attributen, z.B. 40 Tabellen mit jeweils 40 Attributen.
  - Die Anzahl der Reihen in jeder Tabelle ist in der Größenordnung 107
  - Datenströme!
- Wissensentdeckung behandelt gegebene Datensätze, die meist nicht für die Analyse erstellt wurden.
  - Ein Prozess muss die gegebenen Daten in einen Datensatz überführen, der für das Data Mining geeignet ist.

Hype-Kurve Wissensentdeckung



Anwendungen

- Entdecken von Missbrauchmustern in Kreditkarten-Transaktionen
- Bestimmung der Personen, die eine bestimmte Werbung per Post bekommen sollen
- Entdecken von latent unzufriedenen Kunden
- Heraussuchen „interessanter“ oder zu einem Thema gehörender Web-Seiten
- Ausfiltern von unerwünschter electronic mail
- Vorhersage des Abverkaufs von Artikeln (Lagerhaltung)
- Finden von Assoziationen zwischen Waren oder zwischen Kunden und Waren
- Analyse von RNA-Aktivität (Expressivität) als Hinweis auf mikrobiologische Prozesse
  ACHTUNG: mehrere Tausend Attribute bei einigen Zehn Proben!

KDnuggets : Polls :
Industry of your Data Mining Clients (Mar 2008)
In what industries/sectors were your data mining clients in 2007-2008? [100 voters]

Banking (36)	36.0%
Financial (21)	21.0%
Telecom and wireless (20)	20.0%
Retail (18)	18.0%
Insurance (16)	16.0%
e-Commerce (15)	15.0%
Utilities (gas (13)	13.0%
Government (10)	10.0%
Pharma (9)	9.0%
Manufacturing (9)	9.0%
Health care/ HR (9)	9.0%
Biotech/Genomics (9)	9.0%
Travel/Hospitality (8)	8.0%
No clients (8)	8.0%
Investment / Stocks (8)	8.0%
Software (6)	6.0%
Non-profit org (6)	6.0%
Security (5)	5.0%
Entertainment/ Music (5)	5.0%
Military (4)	4.0%
Mortgage/Lending (3)	3.0%
Law (2)	2.0%
Other	6.0%

KDnuggets : Polls :

Analysierte Datentypen (Juli 2007)

Types of Data Analyzed/Mined in the past 12 months: [120 voters total]

table data, fixed # of columns (103)	85.8%
time series (46)	38.3%
text, free-form (41)	34.2%
itemsets/transactions (26)	21.7%
anonymized data (22)	18.3%
web content (17)	14.2%
spatial data (2-D, 3-D) (16)	13.3%
web clickstream (14)	11.7%
email (14)	11.7%
XML data (13)	10.8%
other (12)	10.0%
links or networks (11)	9.2%
images/video (9)	7.5%
music / audio (5)	4.2%

Wissensentdeckung in Datenbanken/Data Mining

- Interdisziplinär: Statistik, maschinelles Lernen, Datenbanken (und ein Anwendungsbereich)
- Verschiedene Theorien
- Verschiedene Werkzeuge
- Verschiedene Datenformate
- Verschiedene Geschäftsprozesse, Organisationsformen

Fragen bei der Wissensentdeckung

- Welche Attribute sind relevant? (Merkmalsauswahl)
- Welche Repräsentation ist für das Data Mining geeignet?
- Müssen neue Attribute aus vorhandenen erstellt werden? (Merkmalsgenerierung, Merkmalsextraktion)
  - Diskretisierung
  - Funktionen/Aggregationen (Durchschnitt, Häufigkeit, Varianz)
  - Transformationen (Fourier, Datumsberechnung)
- Welches Verfahren ist für das Data Mining geeignet?
- Wie viele Beobachtungen sollen ausgewählt werden und wie?
  - Stichproben

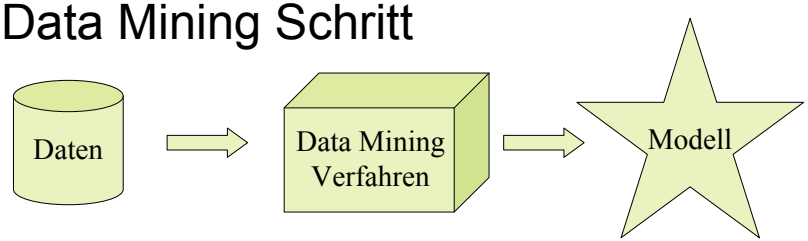
Zeitaufwand für die Datenvorverarbeitung KDnuggets Oktober 2003

What % of time in your data mining project(s) is spent on data cleaning and preparation [187 votes total]	
over 80% (46)	25%
61 to 80% (73)	39%
41 to 60% (46)	25%
21 to 40% (7)	4%
20% or less (15)	8%

Data Mining

- Ein Schritt innerhalb der vielen, die zur Entdeckung von Mustern führen.
- Maschinelles Lernen oder statistische Verfahren werden in der Wissensentdeckung insbesondere als Data Mining step verwendet.
- Neu: auch in der Datenvorverarbeitung werden maschinelle Lernverfahren und statistische Verfahren eingesetzt.
- Maschinelles Lernen besteht aus Hypothesengenerierung und Hypothesentest.

Data Mining Schritt



- Aus einem (aufbereiteten, s.o.) Datensatz bzw. Stichprobe wird ein Verfahren des Data Minings angewendet
- Dieser liefert ein Modell der Daten, welches **beschreibend** oder auch **vorhersagend** sein kann
- Beschreibung der Eingaberepräsentation: **LE**
  Beschreibung der Modellrepräsentation: **LH**

Lernen ist …

- ... jeder Vorgang, der ein System in die Lage versetzt, bei der zukünftigen Bearbeitung derselben oder einer ähnlichen Aufgabe diese besser zu erledigen. (Simon 1983)
  - Was heißt „besser“?
  - Was für den einen besser ist, ist für den anderen schlechter!
- ... das Konstruieren oder Verändern von Repräsentationen von Erfahrungen (Michalski 1986)
- Lernen ohne Ziel

Lernen ist …

- Wissenserwerb (Begriffe, Theorie, Sprache)
- Definition eines Begriffs aus seinen Beispielen, zusammenhängende Definitionen ergeben eine Theorie
- Funktionslernen (Klassifikation, Regression)

$$geg.: \ \vec{X}, Y \qquad ges.: \ f \ mit \ f(\vec{X}) \cong Y, \quad Y = R \vee Y = \{0,1\}$$

- Suche im „Hypothesenraum“ („Modellklasse“)
- Mögliche Hypothesen („statistische Modelle“) werden geordnet aufgezählt und untersucht, bei der richtigen wird angehalten
- Der induktive Schluss
  - Uta ist ein Mensch, Uta ist sterblich, so auch Uli, Vroni, Sokrates...,
  - also: Alle Menschen sind sterblich.

Fragen beim Data Mining

- Wie viele Beispiele muss ich mindestens haben, bis ich ein ausreichend korrektes und vollständiges Lernergebnis erzielen kann (Stichprobengröße)?
- Wie mächtig muss mein Repräsentationsformalismus (Modellklasse) sein, damit ich ein annähernd korrektes und vollständiges Lernergebnis ausdrücken kann?
- Unter welchen Umständen wird der Lernalgorithmus zu einem Ergebnis kommen und anhalten? Wie schnell ist er?
- Welche Zusicherungen kann der Algorithmus über sein Ergebnis garantieren?
- z.B.: Diese Regel hat eine Fehlerwahrscheinlichkeit von 10%.
- z.B.: Dieses sind alle Regeln, die in den Daten gültig sind – wenn eine fehlt, fehlen auch die entsprechenden Daten!

Vorlesung

- Kurze Einführung in Datenbanken
- Verfahren für den Data Mining step aus allen drei Bereichen
  - Datenbanken,
  - Maschinelles Lernen,
  - Statistik
- Vorverarbeitung
- Fallstudien
- Übungen mit RapidMiner und R

Verfahren

- Datenbanken:
  - Data Cube
  - Assoziationsregeln, häufige Mengen
- Klassifikation:
  - Diskriminanzanalyse,
  - Entscheidungsbäume,
  - Stützvektormethode
- Clustering
- Vorverarbeitung
  - Hauptkomponentenanalyse
  - Merkmalsgenerierung und -selektion
- Zeitreihen und Sequenzen

Software

KDnuggets : Polls : Data Mining / Analytic Software Tools (May 2007)
Data Mining (Analytic) tools you used in 2007: [534 voters]
The first (narrow) bar corresponds to the number of votes where the tool was selected alone, and the second (wide) bar to the number votes where the tool was select as one among several;

SPSS Clementine	116, 73 alone or with SPSS
Salford CART/MARS/TreeNet/RF	106, 54 alone
Yale (now: RapidMiner)	103, 70 alone
Excel	94, 2 alone
SPSS	91, 49 alone or with Clementine
SAS	80, 8 alone or with SAS E-Miner
Angoss	78, 50 alone
KXEN	70, 51 alone
Weka	48, 3 alone
R	42, 0 alone
SQL Server	38, 2 alone
MATLAB	30, 1 alone
SAS E-Miner	25, 8 alone or with SAS
your own code	61, 7 alone

Lernziele

- Verstehen der verschiedenen Fachsprachen
- Kenntnis der wichtigsten Verfahren für den Data Mining Schritt
- Überblick über den gesamten Prozess der Wissensentdeckung
- Handhabung verschiedener Werkzeuge

Übungsschein

- Zu jeder Vorlesungsstunde kommen und zuhören!
- Nachbereiten, indem Materialien gelesen und Fragen in der Gruppe diskutiert werden.
  Dies erfordert mindestens 2 Stunden pro Woche.
- Alle Übungszettel bis auf 3 müssen bearbeitet werden.
  Dazu sind mindestens noch einmal 4 Stunden pro Woche nötig.

Arbeitsmaterial

- Texte:
  - Proceedings: PKDD, KDD, ICDM
  - Zeitschriften:
    - Knowledge and Information Systems,
    - Data Mining and Knowledge Discovery
- Bücher:
  - Mitchell, T. (1997): Machine Learning, McGraw-Hill
  - Witten, I.H., Frank, E. (2001): Data Mining – Praktische Werkzeuge und Techniken für das maschinelle Lernen
- Folien sind teilweise auch gleich Skript.
- Software:
  - RapidMiner und R

Konferenzen

KDnuggets : Polls : Conferences papers were submitted to (Feb 2008)
To which conferences did you submit a paper in the last 2 years: [109 voters total]

KDD (38)	34.9%
ECML/PKDD (31)	28.4%
None (27)	24.8%
IEEE ICDM (27)	24.8%
SDM (19)	17.4%
ICML (16)	14.7%
PAKDD (15)	13.8%
Other conference (14)	12.8%
Other AI/ML related conference (13)	11.9%
Other KDD-related conference (9)	8.3%
AAAI/IJCAI (9)	8.3%
SIGMOD-PODS (8)	7.3%
ICDE (8)	7.3%
Other DB-related conference (7)	6.4%
VLDB (6)	5.5%

ECML/PKDD 2008

Antwerpen
15.-19. September 2008

European Conference on Machine Learning and
Principles and Practice of Knowledge Discovery in Databases

- Program Chairs:
  Walter Daelemans
  Bart Goethals
  Katharina Morik
- 30 Area Chairs
- >300 program committee members
- >600 eingereichte Papiere erwartet
- 60 Vorträge

Sie erhalten

- Wissen aus erster Hand
  - Sekundärliteratur birgt ”stille Post Effekt”!
- International orientierte Auswahl der Grundlagen
  - Sie sollen auch mit Stanford-Studenten mithalten können!
- Sorgfältig ausgewählte Grundlagen, damit Sie auf den Stand der Kunst kommen können
  - Vielleicht schreiben Sie Ihre Diplomarbeit im Bereich Wissensentdeckung, arbeiten später im Bereich
- Anleitung zur Arbeit mit dem weltweit meist benutzten Open Source Werkzeug