Universität Leipzig
Institut für Informatik

Data Warehousing und Data Mining

Problemseminar: Sommersemester 1998

Thema 8: *Data Mining. Überblick*

1. Einleitung

2. Anforderungen ans Data-Mining-System

3. Klassifizierungsmöglichkeiten

4. Klassifikation nach den Endergebnissen

- Gewinnung von Assoziationsregeln
- Datengeneralisierung und Datenaggregation
- Voraussage
- Datenpartitionierung (*Clustering*)
- Mustersuche (*Pattern-based Similarity Search*)
- Zugriffsanalyse in verteilten DBS

5. Anwendungsgebiete für Data-Mining-Verfahren

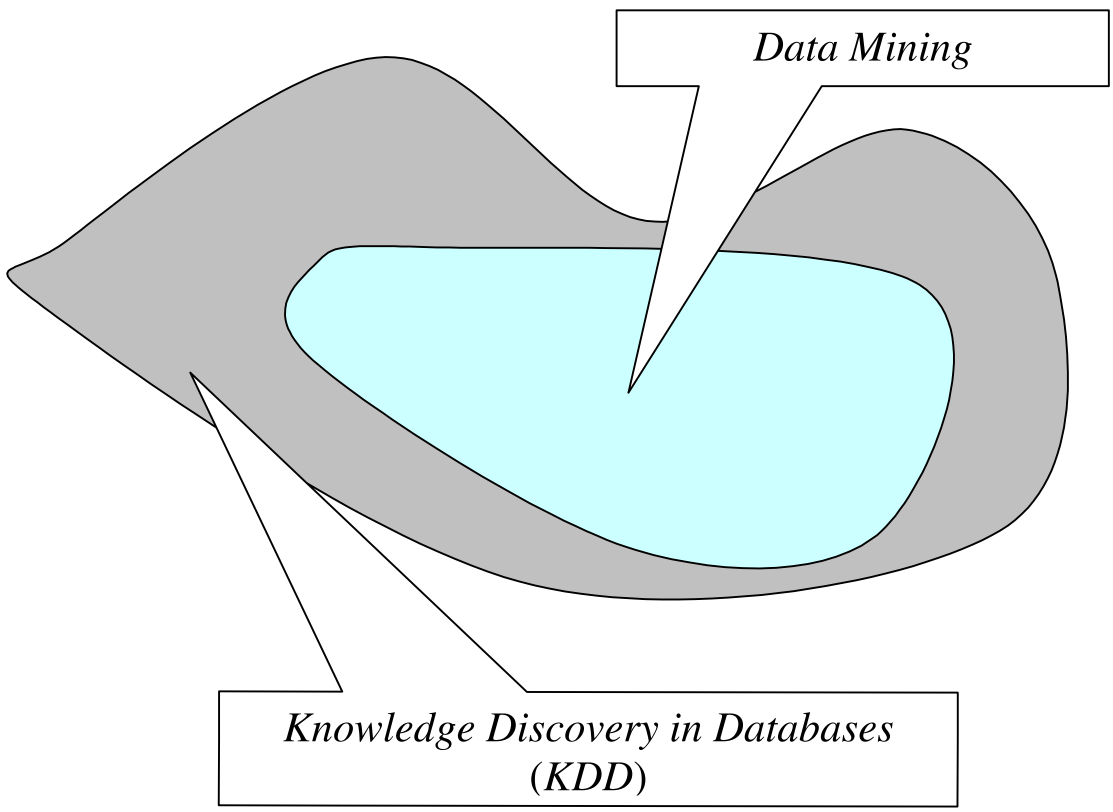
6. Zusammenfassung

Knowledge Discovery in Databases (KDD)

- *Finden der Muster von Daten in Datenbanken*
- *Interpretation und Analyse von Untersuchungsergebnisse*
- *Bereitstellen von Resultaten für andere Systeme*

Data Mining

- *Großer Teil des KDD*
- *Kompletten Satz von Algorithmen, Methoden und Werkzeugen*
- *Automatisierte und intelligente Datenbankanalyse*



2. Anforderungen ans Data-Mining-System

- *Mehrheit und Komplexität von Datentypen*
- *Effizienz und Skalierbarkeit der Data-Mining-Algorithmen*
- *Unterstützte Verteiltheit und Heterogenität*
- *Interaktive Verwaltung und Kontrolle*
- *Datenschutz*
- *Anfrageanalyse*

3. Klassifizierungsmöglichkeiten

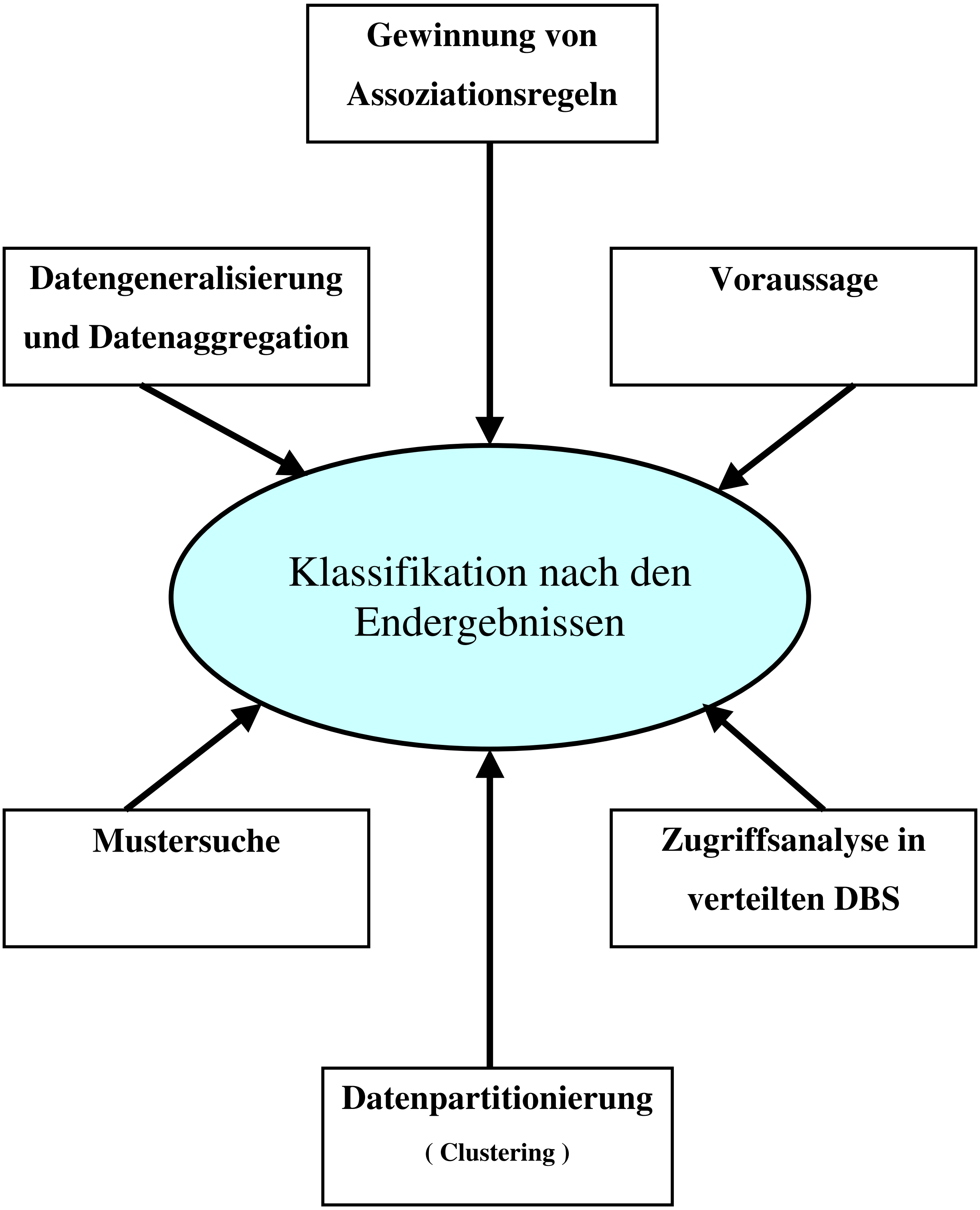
- Nach dem Datenbanktyp

  relationale, objekt-orientierte, deduktive, temporale, heterogene etc.

- Nach dem Typ der Algorithmen

  mathematische, heuristische, genetische etc.

- Nach den Endergebnissen

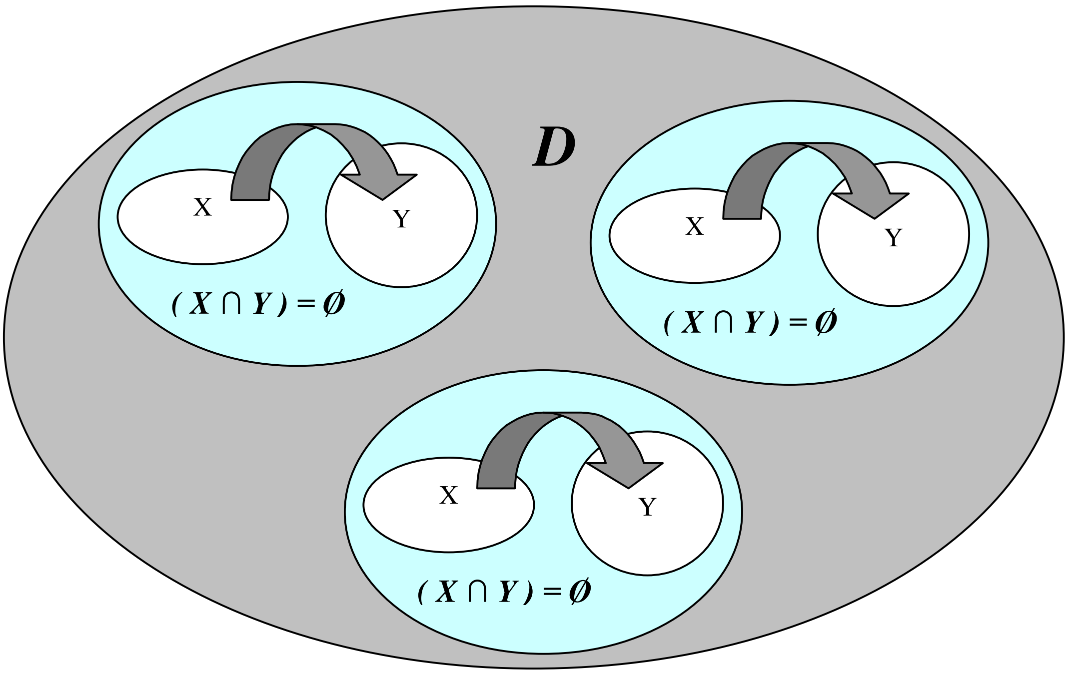

Gewinnung von Assoziationsregeln
Datengeneralisierung und Datenaggregation
Voraussage
Klassifikation nach den Endergebnissen
Mustersuche
Zugriffsanalyse in verteilten DBS
Datenpartitionierung (Clustering)

4.1. Gewinnung von Assoziationsregeln

Die Assoziationsregeln haben die Aufgabe, Sätze von strengen Regeln, die die Gestalt haben:

" A ^ B ^ ... ^ K => L ^ M ^ ... ^ X ",

wobei A ... X Sätze von relevanten Attributwerten in der Datenbank sind, zu gewinnen.



Confidence: **bezeichnet die Implikationsstärke der Regel**

Die Datenbank ***D*** enthält die Regel ***X => Y*** mit *Confidence* ***C*** ∈ [0 , 1] genau dann, wenn ein Anteil von ***C*** aller Tupel in ***D***, die ***X*** enthalten, auch ***Y*** enthält.

$$C = |X \wedge Y| \;/\; |X|$$

Support: **illustriert die Häufigkeit der Regel in der Datenbank**

Die Datenbank ***D*** enthält die Regel ***X => Y*** mit *Support* ***S*** ∈ [0 , 1] genau dann, wenn ein Anteil von ***S*** aller Tupel in ***D*** ***(X*** U ***Y)*** enthält.

$$S = |X \wedge Y| \;/\; |D|$$

Vorgehensweise beim Apriori-Algorithmus

Datenbank D

T_ID	*Waren*
100	a c d
200	b c e
300	a b c e
400	b e

Waren	*Support*
a	0.5
b	0.75
c	0.75
d	0.25
e	0.75

Waren	*Support*
a	0.5
b	0.75
c	0.75
e	0.75

Waren
{ a, b }
{ a, c }
{ a, e }
{ b, c }
{ b, e }
{ c, e }

Scan ***D***

Waren	*Support*
{ a, b }	0.25
{ a, c }	0.5
{ a, e }	0.25
{ b, c }	0.5
{ b, e }	0.75
{ c, e }	0.5

Support ?

Waren	*Support*
{ a, b }	0.5
{ b, c }	0.5
{ b, e }	0.75
{ c, e }	0.5

Waren
{ b, c, e }

Scan ***D***

Waren	*Support*
{ b, c, e }	0.5

Support ?

Waren	*Support*
{ b, c, e }	0.5

Entdeckte Regeln: e ^ c => b, ***C = 1 ;***
b => c ^ e, ***C = 0.66 ;***
c => b ^ e, ***C = 0.66 ;***
usw.

4.2. Datengeneralisierung und Datenaggregation

- *Überflüssige, zu detaillierte Information*
- *Materialisierte Sichten*
- *Data-Warehousing-Anfrageverarbeitungstechnik*

Data-Mining-Algorithmen gibt Hinweise für die Erzeugung materialisierter Sichten:

- Data Cube Approach

  Rechnet am häufigsten gestellten Anfragen im voraus und merkt sich als materialisierte Sichten

- Attribute-Oriented Induction Approach

  erzeugt neue bzw. erweitert alte materialisierte Sichten im Laufe der Anfrageverarbeitung

Bei der Anfrageverarbeitung nach der *Attribute-Oriented-Induction-Approach-*Methode kann die Testsuche nach Sichtdefinition uneffizient werden.

4.3. Voraussage

Es wird versucht, aufgrund der vorhandenen Attributwerte das weitere Verhalten der einzelnen Objekte in der Datenbank zu erraten.

Datenklassifikation:

Antworten der Art : "Ja" oder "Nein"

Als Verallgemeinerung: " gehört zu" oder " gehört nicht zu"

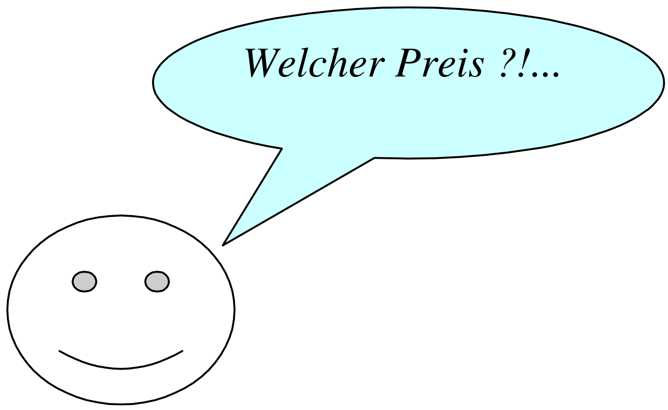
Korrektheit der Antwort: **accuracy = 1 - errors / cases**

Autohändlerbeisbiel:

Datenregression:

Fragen der Art : "Wie groß ?", "Wie klein ?"

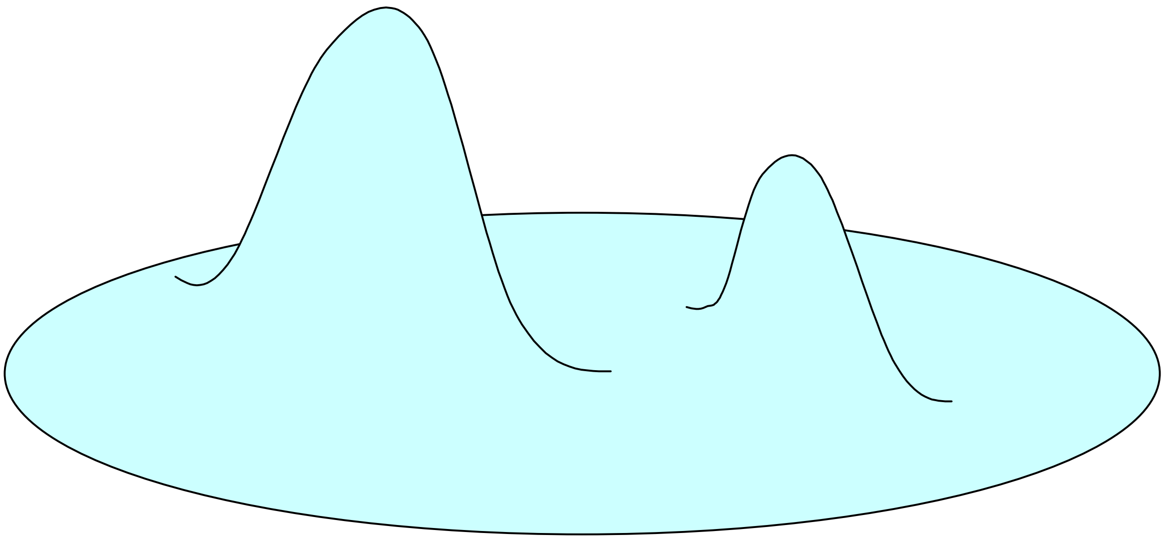
Autohändlerbeisbiel:



Maschinelles Lernen *als Musterbeispiel für das interaktive System für das Erarbeiten und Prüfen von Hypothesen*

4.4. Datenpartitionierung (*Clustering*)

- *Unterteilung von großen Datenbanken in kleinere*
- *Entdeckung der Regionen gemäß der Dichte an bedeutenden Werten*
- *Unterscheidung von Klassen, deren innere Ähnlichkeit und äußere Differenz am größten sind*



Regionen von bedeutenden Werten in einer DB als Erkennungsziel für Clustering-Algorithmen

Simulated Annealing   *als*   *Beispielswerkzeug*

4.5. Mustersuche

- Temporale Datenbanken als Objekte für die Mustersuche-Algorithmen

  medizinische, wirtschaftliche etc.

- Antworten auf die Fragen nach dem Risiko, der Kausalität, den Entwicklungstrends als Ziele der Mustersuche

  Identifiziere die Firmen mit ähnlichem Verhalten des Wachstums ...

  Identifiziere die Waren mit ähnlichem Verkaufsmuster ...

- Zwei Varianten der Anfragestellung

  1) Vergleich aller Objekte in der Datenbank mit dem Musterobjekt
  2) Vergleich aller Objekte in der Datenbank paarweise zwischen einander

- Klassifizierungsvariante für die Vergleichsalgorithmen

  1) ... anhand des Ähnlichkeitsmaßes
  2) ... beim Unterscheiden zwischen Domänentypen
  3) ... bei der Existenz der Möglichkeiten für das Umrechnen des Maßstabs, das Verwenden von Übersetzungsmechanismen und das Vergleichen von Unterketten variabler Längen

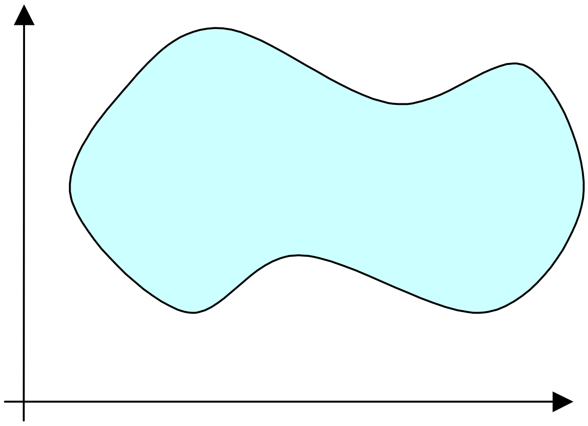


Bild (T_1)

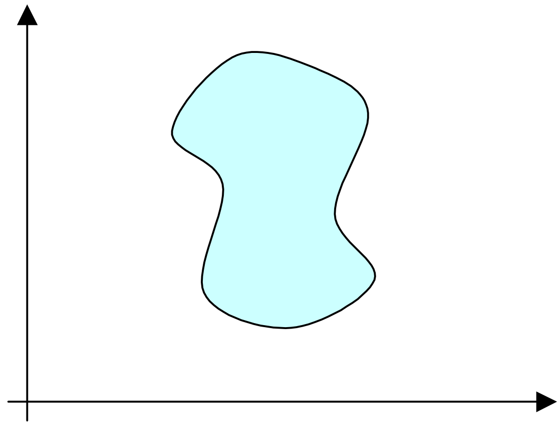
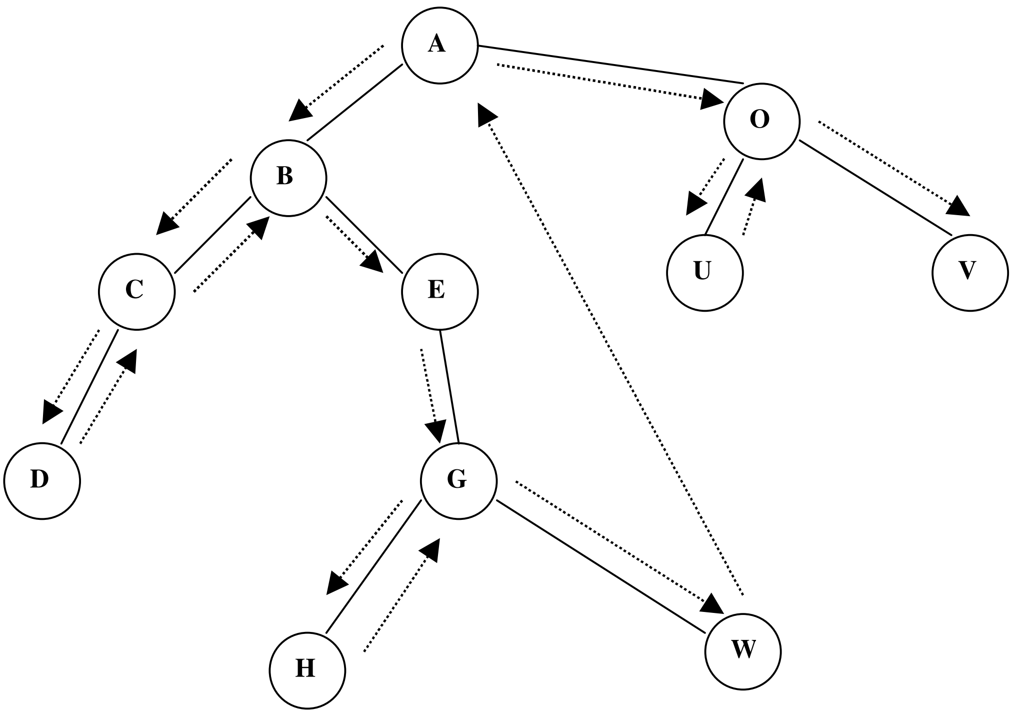


Bild (T_2)

4.6. Zugriffsanalyse in verteilten Systemen

- *Zusammenhalten von Objekten in Verteilten Systemen*
- *Verständnis der Struktur der Zugriffspfade*
- *Erforschung der Struktur auf die entstandenen Interessen*



Illustrierendes Beispiel für eine Zugriffsfolge
(Traversal Patterns)

Beispielserklärung:

Schritt 1:

Baum {A,B,C,D,C,B,E,G,H,G, W,A,O,U,O,V} als ein Referenzenpfad

Schritt 2:

Folgen: {A,B,C,D}, {A,B,E,G,H}, {A,B,E,G,W}, {A,O,U}, {A,O,V}

Schritt 3:

Unterfolgen, die wegen ihrer Häufigkeit herauskristallisiert wurden:

{ABE, CGH, AD, BG}

5. Anwendungsgebiete

- **Chemie- und Pharmakologie-Industrie**

Ziele:

- Klassifikation von Molekülen nach Ähnlichkeit
- Aussuchen aller Moleküle, die einem bestimmten Muster entsprechen
- Aussuchen aller Moleküle, die andere Zielmoleküle binden können
- Voraussage des Resultats eines Tests aufgrund der gespeicherten Daten

- **Einzelhandel**

Ziele:

- Verkaufsvorhersagen
- Optimierung von Marketingstrategien

- **Automatische Verarbeitung von Fernerkundungsdaten**

Ziele:

- Militärische und Wissenschaftliche Überwachungsanalyse

- **Finanzielle Anwendung**

Ziele:

- Untersuchung der Tätigkeit der Banken als finanzielle Vermittlungsstellen
- Finanzielle Voraussage