Data Mining mit Microsoft SQL-Server 2005/2008

Marcel Winkel

Hochschule für Technik, Wirtschaft und Kultur Leipzig
Fachbereich Informatik, Mathematik und Naturwissenschaften

19. Mai 2010

- relationale Datenbank (Speicherung der Daten)
- Integration Services (Transformation und Zusammenführung von Daten)
- Analysis Services (Auswertung von Daten)
- Reporting Services (Erstellung von Berichten)

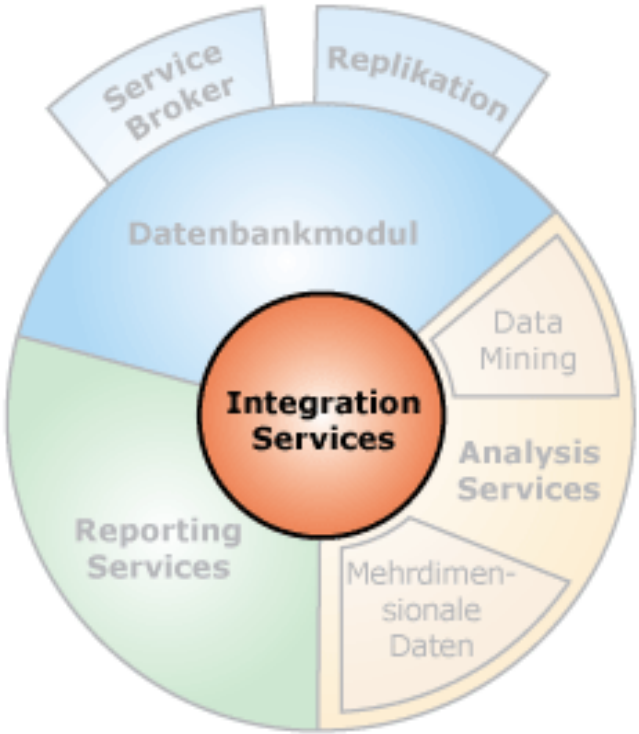


Abbildung 1: Komponenten in SQL Server 2008

Reporting Services

- bietet umfangreiches Sortiment an Tools und Diensten zum Erstellen von Berichten
- interaktive, tabellarische oder grafische Berichte

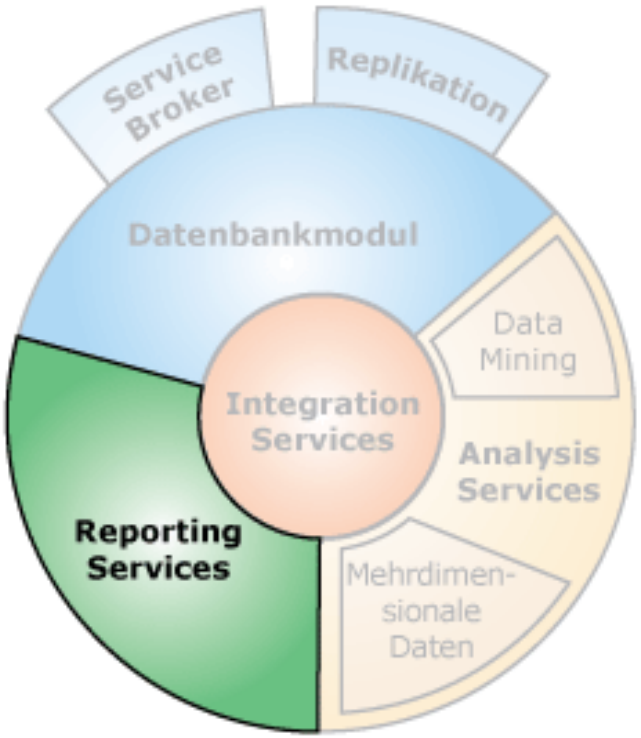


Abbildung 2: Reporting Services

Mehrdiemensionale Daten

- mehrdimensionale Daten in Form von OLAP-Cubes
- Daten aus relationaler DB, XML, ...
- Analyse von großen Datenmengen
- Datenquelle für Miningmodelle

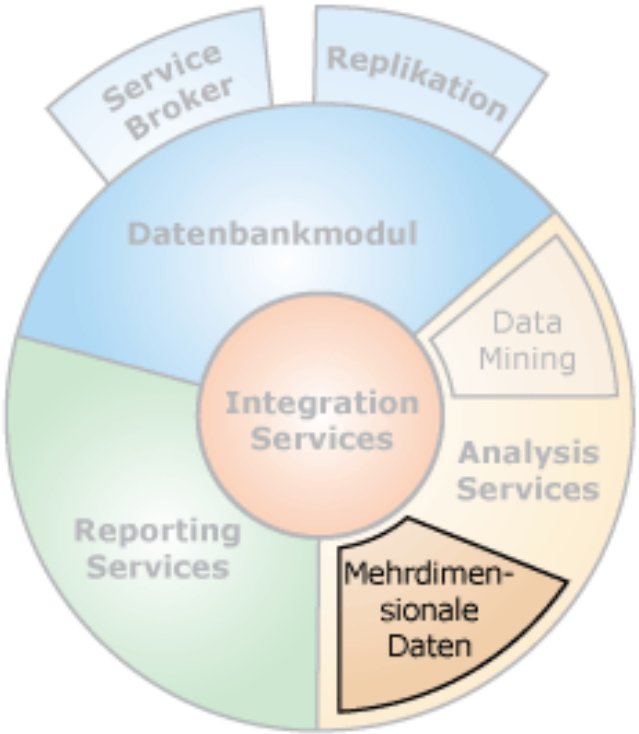


Abbildung 3: SQL Server Analysis Services (OLAP)

Data Mining

- stellt mehrere Algorithmen für unterschiedliche Analysemethoden bereit
- zahlreiche Tool, Abfragesprachen sowie Objektbibliotheken für Programmierung stehen zur Verfügung
- DMX (Data Mining eXtensions), Erstellen und Verwalten von Data Mining-Modellen

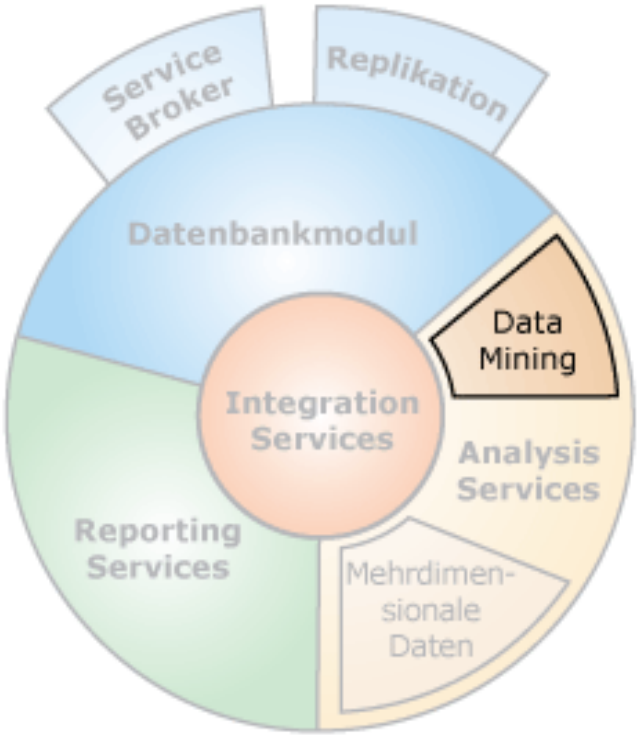


Abbildung 4: SQL Server Analysis Services (Data Mining)

Miningstruktur

- tabellenähnliche logische Datenstruktur
- beinhaltet die zu analysierenden Daten
- optionale Partitionierung in Testsatz und Trainingssatz
- kann von beliebig viele Miningmodellen verwendet werden

Miningmodell

- erhält Daten aus Miningstruktur
- analysiert Daten mittels Data Mining-Algorithmus
- Struktur und Modell sind separate Objekte
- speichert Informationen aus der Verarbeitung der Daten (gefundene Muster)
- enthält Metadaten (Namen des Modells, Liste der Spalten aus der Miningstruktur)
- leer bis Daten verarbeitet wurden
- Miningmodell wird mittels DMX oder Data Mining Assistenten erstellt

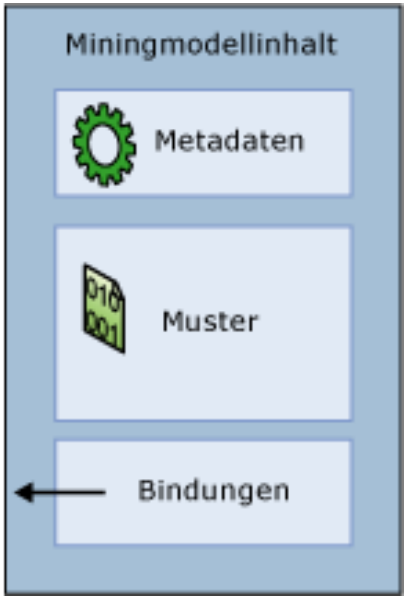


Abbildung 5: Miningmodell

ASSL

- Kommunikation zwischen Client und Server über SOAP-Nachrichten
- ASSL ist XML-Dialekt für diese Nachrichten
- besteht aus:
  - Datendefinitionssprache (Data Definition Language, DDL)
  - Befehlssprache: XMLA (XML for Analysis)

XMLA - XML for Analysis

- von Hyperion und Microsoft entwickelt, 2001 veröffentlicht
- XMLA ist ein Industriestandard für Zugriff auf Daten in multidimensionalen Datenbanken
- basiert auf HTTP, XML und SOAP
- native Protokoll, welches für sämtliche Interaktionen zwischen Client und Server verwendet wird
- auch Business Intelligence Development Studio und SQL Server Managment Studio kommunizieren über XMLA
- Entwicklung und Verwendung von Miningmodellen ist mit Hilfe von XMLA möglich, aber nicht üblich
- durch XML basierende Kommunikation, ist Integration in andere Umgebungen möglich

DMX - Data Mining eXtensions

- Erweiterung des SQL-Standards um die Fähigkeit, mit Data Mining Modellen zu arbeiten
- Syntax von DMX ähnelt der von SQL
- besteht aus DDL (Data Definition Language), DML (Data Manipulation Language) sowie Funktionen und Operatoren
- Data Mining-Modelle erstellen und verwenden
- Data Mining-Struktur erstellen
- diese Modelle trainieren sowie durchsuchen, verwalten und Vorhersagen treffen

Sprachelemente von DMX

- Operatoren
  - Arithmetische $(+, -, *, /)$
  - Logische (AND, NOT, OR)
  - Vergleich $(<, >, =, ...)$
  - Unäre $(-, +)$
- Datendefinitionsanweisungen
  - **CREATE**/**DROP** MINING MODEL/STRUCTURE
  - **ALTER** MINING STRUCTURE
  - **IMPORT** / **EXPORT**
- Datenbearbeitungsanweisungen
  - **DELETE**
  - **INSERT INTO**
  - **UPDATE**
  - **SELECT**, kann mit *TOP, ORDER BY, WHERE* beschränkt werden

DMX - Mining-Struktur erstellen

```
CREATE MINING STRUCTURE [Miningstruktur]
(
   [Schluessel] LONG KEY,
   [Jahreseinkommen] DOUBLE CONTINUOUS,
   [Autos] LONG DISCRETE,
   [Alter] LONG DISCRETIZED,
   [Fahrradkaeufer] LONG DISCRETE
)
```

DMX - Mining-Modell erstellen

```
ALTER MINING STRUCTURE [Miningstruktur]
ADD MINING MODEL [MiningmodelI_DT]
(
   [Schluessel],
   [Jahreseinkommen],
   [Autos],
   [Alter],
   [Fahrradkaeufer] Predict
)USING Microsoft_Decision_Trees
```

DMX - Mining-Modell trainieren

```
INSERT INTO MINING MODEL [Miningmodell_DT]
(
   [Schluessel],
   [Jahreseinkommen],
   [Autos],
   [Alter],
   [Fahrradkaeufer]
)OPENQUERY(
   [Adventure Works DW],
   'SELECT
   [CustomerKey] As [Schluessel],
      [YearlyIncome] As [Jahreseinkommen],
      [NumberCarsOwned] As [Autos],
      [Age] As [Alter],
      [BikeBuyer] As [Fahrradkaeufer]
  FROM dbo.vTargetMail'
)
```

Entwicklungsumgebungen

SQL Server Management Studio

- Konfiguration und Verwaltung von Miningmodellen und anderen Objekten
- es können DMX-Abfragen an Analysis Services gesendet werden

Business Intelligence Development Studio

- umfangreiche grafische Schnittstelle, zahlreiche Assistenten
- Erstellung von Miningmodellen auf zwei Arten:
  - *Projektmodus:* Änderungen erst mit Übertragen des Projektes wirksam
  - *Onlinemodus:* direkte Verbindung zum Server, Änderungen sofort wirksam
- weitere Tools um Vorhersagen zu erstellen, Miningmodelle zu analysieren, auf Genauigkeit überprüfen

Versionen von SQL Server

- MSDE, Express, Web und Workgroup Editionen bieten keine Unterstützung für Data Mining
- SQL Server Standard Edition umfangreiche Data Mining Funktionalität
- SQL Server 2008 Enterprise Edition bietet umfangreichste Unterstützung für Data Mining
  - erweiterte Konfiguration und Optimierung von Algorithmen
  - Plug-In-Api für Algorithmen
  - unbegrenzt gleichzeitige Data Mining-Abfragen

Klassifikationsalgorithmen

- Naive Bayes
- Entscheidungsbäume
- neuronale Netzwerke

Regressionssalgorithmen

- Lineare Regression
  - z.B. Marktsegmentierung (Alter - Einkommen)
  - Ziel ist eine Formel, welche die Beziehung zwischen den Attributen beschreibt

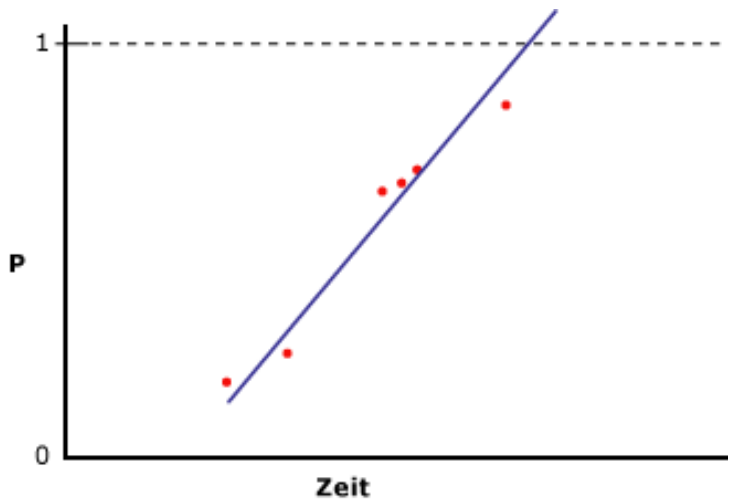


Abbildung 6: Lineare Regression, Alter - Einkommen

Regressionssalgorithmen

- Logistische Regression
  - Anstatt einer Linie → Kurve in "S"-Form
  - berücksichtig obere und untere Schranke

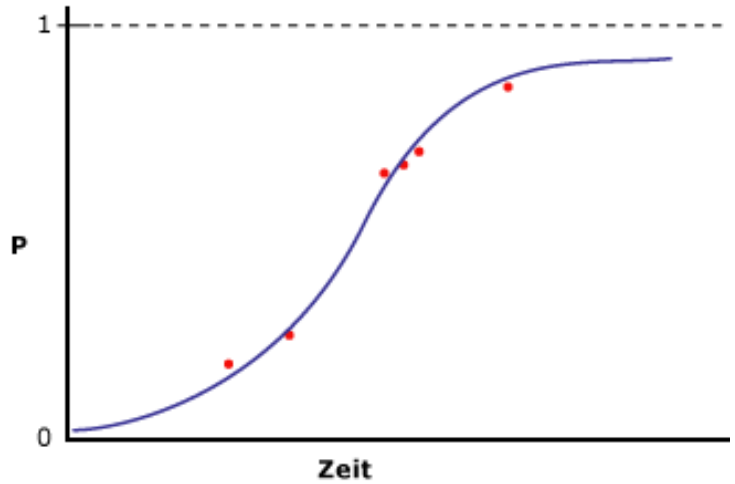


Abbildung 7: Logistische Regression, Alter - Einkommen

Regressionssalgorithmen

- Microsoft Time Series-Algorithmus
  - Vorhersage von Kennzahlen anhand Trendanalyse
  - linke Seite enthält Daten mit denen das Modell trainiert wurde, rechte Seite ist die Vorhersage

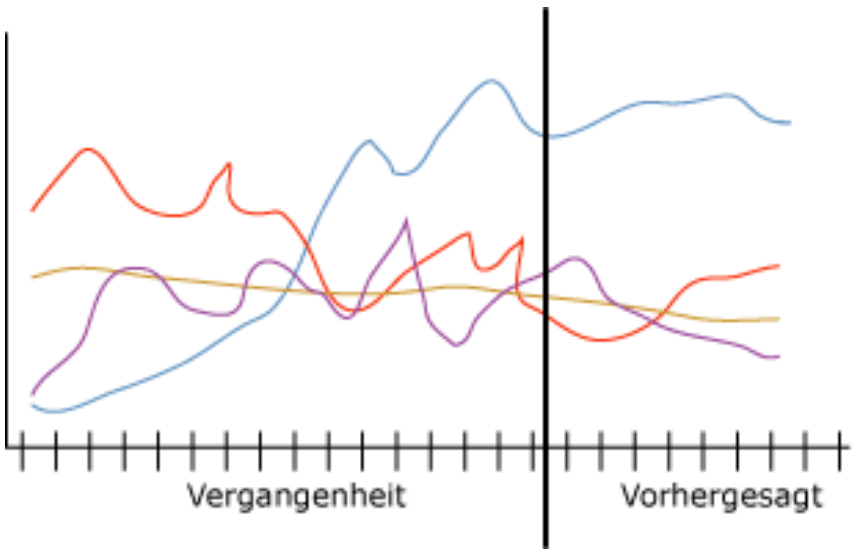


Abbildung 8: Microsoft Time Series-Algorithmus

Zuordnungsalgorithmen

- Microsoft Association-Algorithmus
  - findet häufig in Kombination auftretende Elemente und leitet Regeln ab
  - basiert auf Apriori
  - z.B. Warenkorbanalyse

Regel
Road Bottle Cage = Existing, Cycling Cap = Existing -> Water Bottle = Existing
Mountain-200 = Existing, Mountain Tire Tube = Existing -> HL Mountain Tire = Existing
Mountain-200 = Existing, Water Bottle = Existing -> Mountain Bottle Cage = Existing
Touring-1000 = Existing, Water Bottle = Existing -> Road Bottle Cage = Existing
Road-750 = Existing, Water Bottle = Existing -> Road Bottle Cage = Existing
Touring Tire = Existing, Sport-100 = Existing -> Touring Tire Tube = Existing

Abbildung 9: Microsoft Association-Algorithmus

Segmentierungsalgorithmen

- Microsoft Clustering-Algorithmus
  - gruppiert Daten mit ähnlichen Eigenschaften → Cluster
  - basiert auf k-Means

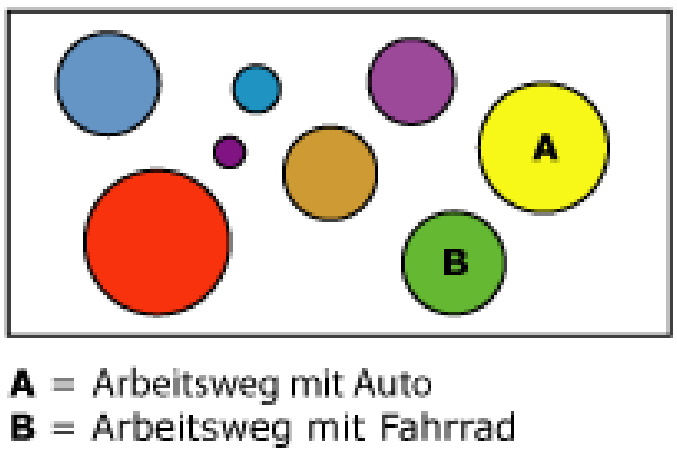


Abbildung 10: Microsoft Clustering-Algorithmus

Sequenzanalysealgorithmen

- Microsoft Sequence Clustering-Algorithmus
  - berücksichtig in den zu analysierenden Daten gegebene Reihenfolge, zeitliche Abfolge von Ereignissen
  - ähnelt dem MS-Clustering-Algorithmus, anstatt nach ähnlichen Attributen zu suchen, sucht er ähnliche Pfade in einer Sequenz
  - z.B. Navigationswege von Benutzern auf einer Webseite
  - Ergebnisse können verwendet werden um Struktur der Webseite zu verbessern
  - Platzierung von Werbebannern

Plug-In-Algorithmen

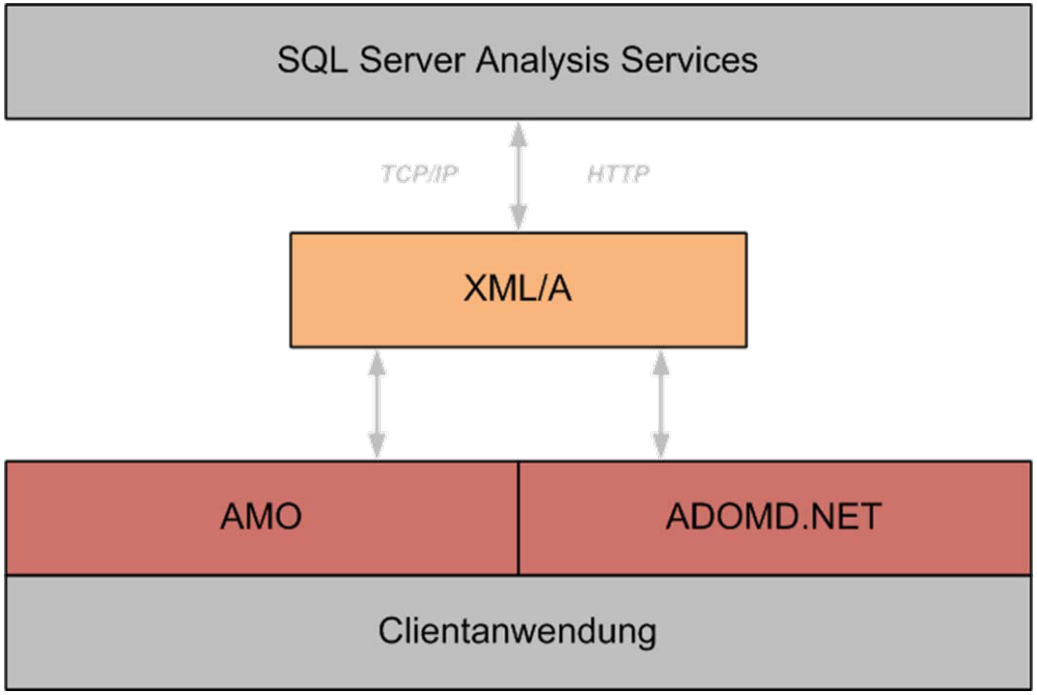
- benutzerdefinierte Algorithmen
- vordefienierte Schnittstellen
- .NET Wrapper erlaubt das Entwickeln mittels managed Code

Office 2007

- Data Mining-Add-Ins bieten umfangreiches Toolset
- Data Mining-Aufgaben direkt aus Excel / Visio heraus
- Vorraussetzungen:
  - .NET Framework 2.0
  - Office 2007 mit .NET-Programmierunterstützung
  - SQL Server 2005/2008 Analysis Services
- Tabellenanalysetools für Excel 2007
- Data Mining-Client für Excel 2007
- Data Mining Vorlagen für Visio 2007

Programmierung mit .NET

- SSAS ermöglicht die Integration von Data Mining-Funktionalität in eigene Anwendungen
- SSAS bietet Objektmodelle und Controls für die Entwicklung
  - AMO (Analysis Management Objects)
  - ADOMD.NET (ActiveX Data Objects MultiDimensional.NET)
  - Data Mining Viewer Control



AMO und ADOMD.NET

- auf .NET basierende Klassenbibliotheken
- mit **AMO** können alle Verwaltungsaufgaben integriert werden
- Verwaltung von Miningstrukturen und -modellen sowie OLAP-Cubes
- mit AMO kein Datenabfrage möglich
- Abfrage von Cubes und Mining Modellen mittels **ADOMD.NET**
- Verbindung aufbauen, Daten abrufen, Befehle ausführen, ...

Beispiele

Laptop

[Mic] Microsoft. *SQL Server Analysis Services - Data Mining*. URL: `http://msdn.microsoft.com/en-us/library/bb510517(v=SQL.100).aspx`.

[TM05] ZhaoHui Tang und Jamie MacLennan. *Data Mining with SQL Server 2005*. Wiley Puplishing, 2005.

[Tit] Jan Tittel. *Data Mining - Mustererkennung in Daten*. URL: `http://www.jan-tittel.de/downloads/DataMining-MustererkennunginDaten.pdf`.

[TS09] Jan Tittel und Manfred Steyer. *Data Mining mit Microsoft SQL Server*. Microsoft Press, 2009.

[Xml] *XML for analysis*. URL: `http://www.xmla.org/`.

Fragen?