Knowledge Management
Data Warehouses and Data Mining

Dr. Michael Hahsler <hahsler@ai.wu-wien.ac.at>
Dept. of Information Processing
Vienna Univ. of Economics and BA

11. December 2001
Table of Contents

● Introduction

● Data Warehouses
  - Operational Data <-> Data in a Warehouse
  - Components of a Data Warehouse
  - How does data get into a Warehouse
  - How can we use the data in a Warehouse
  - Examples from the Virtual University

● Data Mining
  - Applications for Data Mining
  - Common Techniques
  - Market Basket Analysis
  - Examples from the Virtual University and Amazon.com
Introduction and Motivation

- Strong competitive pressure in a service based economy (1-to-1 marketing, mass customization)
- IS to support a learning relationships with other entities
  - EDI (Electronic data interchange)
  - CRM (Customer relationship management)
  - SCM (Supply chain management)
- IS for the organizational memory
  - ERP (Enterprise resource planning)
- Mergers and acquisitions cause nonuniform IT infrastructures
- Automation: Transaction records are available and contain valuable information but they are hard to analyze
- Technology for analysis is available and already mainstream
Data Warehouses

- A **central repository** for all or significant parts of the data that an enterprise’s various business systems collect
- **Data** from various online transaction processing (OLTP) applications and other sources **is selectively extracted and organized**
- Provides **access to data** for use by analytical applications and user queries
A Data Warehouse in a Company

- EDI
- ERP
- CRM
- Operational Transaktion Records
- Data Warehouse
- Feedback
- Analytical Application

Supplier
Customer
Supplier
Customer
Supplier
Customer
Architecture of Data

- Business rules
- Metadata
- Database schema
- Summary data
- Operational data

Abstraction level vs. data size
Differences between the Operational System and the Data Warehouse

Data in Operational System

- High volume, detailed
- High update frequency
- Record oriented, optimized for performance
- Current data only
- Internal data of one application

Data in a Data Warehouse

- Medium volume, summarized
- Low update frequency (daily, weekly)
- Optimized for queries, accessible for analysis
- Past and present data
- Used for several application (OLAP, DSS,...)
Physical Structure of Data Warehouses

Central Architecture

Federal Architecture with Data Marts

End Users in different depts

End Users
Components of Data Warehouses

1. Extraction & Integration
   - Operational Data (Source Schema)
   - Business and Technical Metadata (Meta Schema)
   - Derived Data for Business Users (User Schema)
   - Reconciled Data (Data Warehouse Schema)

2. Aggregation & Customization
Data Extraction & Integration

Getting heterogenous data into the Warehouse:
Data from different DBMSs (Data base management system), external information providers, various standard applications,...

Tasks:

- Extraction (accessing different databases)
- Cleaning (resolving inconsistencies)
- Transformation (different formats, languages)
- Replication (importing a whole DB)
- Analyzing (detecting invalid values)
- Checking for data quality (correctness, completeness)
- Update metadata, if necessary
Example: Extraction & Integration from the Virtual University

Link Collection Transaction log

Data Collection Transaction log

Authentication Service log

Personalized Service log

Reconciled Data

vu

ePub

vugate

myvu

web services
Original transaction data vu (raw Web server log)

rumba.wu-wien.ac.at -- [03/Dec/2001:13:53:12 +0100] "GET /dyn/virlib/wu_org/mediate?ID=wu01_4da HTTP/1.0" 302 205 "session=wu01_session230f1-1007383972" 0 "http://vu.wu-wien.ac.at/dyn/virlib?type=doquery &lib=wu_org&from=wu_query&style=wuhome&sortby=score &query=Griller" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 4.0)"

Original transaction data vugate (application level log)

[Wed Dec 5 14:13:40 2001] :cn=myvue4368d5213-1007557998,ou=cookies,o=myvu,state=good, uid=h8951527@powernet,ou=user,o=myvu

Extracted data

[Mon Dec 3 13:53:12 2001] "wu01_4da" "" "session=wu01_session230f1-1007383972" "137.208.3.45"
Data Aggregation & Customization

Getting (multidimensional) data out of the Warehouse as the input for:

- Reporting (summarized by: who, when, where, what)
- Query tools
  - Online analytical processing (OLAP)
  - Geographic information systems (GIS)
- Decision support systems (DSS)
- Executive information systems (EIS)
- Data Mining
Example: Aggregation & Customization from the Virtual University

Simple reporting: Weekly Usage Statistics by int. and ext. Users
OLAP Cube: Usage of the Information Broker of the Virtual University

OLAP: Queries that take long with RDBMS and SQL (multiple joins) are fast and easy with OLAP-cubes (or the denormalized Star schema).

Operations: Roll-up, drill-down, slice, dice, pivot
Source data (from the data warehouse)

[Mon Dec 3 13:53:12 2001] "wu01_4da" ""
"session=wu01_session230f1-1007383972" "137.208.3.45"

Aggregated data by session

ID := {wu01_session2246d-1006862104}
date := {Tue Nov 27 12:55:46 2001}
mediation := {wu01_290a;wu01_30e3;wu01_35af;wu01_4d1;
wu01_4bf;wu01_4c1;wu01_a57;wu01_26b9;wu01_c69;wu01_419;
wu01_11a8;wu01_114;wu01_364d;wu01_3396;wu01_2e1a}
user := {myvu4e1245bef9-1006862431}
Implementation of a Data Warehouse

Several providers (IBM, Oracle, ...) offer Data Warehouse Systems.

But:

- Warehouses are not sold as of-the-shelf products
- Available products often only support part of the functionality of a warehouse (middleware for information transport, database)
- Implementation of a valuable warehouse is a major project with major risk factors
- Data Warehouses need constant maintenance to stay usable
Summary: Data Warehouse

A data warehouse is

- a central repository for
- all or significant parts of the data that an enterprise’s various business systems collect.

It enables the management to

- access the available data in an efficient way,
- learn about trends and
- make informed decisions.
Data Mining

From Michael J.A. Berry and Gordon Linoff, Data Mining Techniques:

● Data mining provides the enterprise with Intelligence.

● Data mining is the exploration and analysis, by automatic or semiautomatic means, of large quantities of data to discover meaningful patterns and rules.

Reasons for Data Mining:

● Data is being produced
● Data is being warehoused
● Computing power is affordable
● Competitive pressure is strong
● Commercial Data Mining software packages are available
The Virtuous Cycle of Data Mining

Transform data into actionable information

Identify business problems where analyzing data can provide value

Act on the information

Measure the results

From Berry and Linoff
Some Applications for Data Mining

- Market segmentation
- Identifying ‘good’ and ‘bad’ customers
- Fraud detection
- Detecting cross selling potential
- Basis for marketing decisions (shelfing, sales promotions)
- Mass customization / recommender systems
Common Techniques for Data Mining

Data mining uses mostly techniques from artificial intelligence (AI) research. Examples are:

- Memory-based reasoning
- Automatic cluster detection
- Decession trees
- Neural networks
- Genetic algorithms
- Market basket analysis (MBA)
Market Basket Analysis (MBA)

MBA helps to understand what items are likely to be purchased together (association rules) with the aim to identify cross-selling opportunities.

Example: Supermarket

- Shopping cart (= a market basket), point-of-sale scanner produces transaction data.
  With this information alone, the supermarket can already improve shelving.

- If the customer is member of the supermarket’s ’Value Club’ (using e.g. the ATM Card), the supermarket also has demographic information for data mining.

Famous Example: Young fathers buy diapers and six packs of beer Thursdays nights.
Example: Association Rules used in the Virtual University

Simple Association Rule Generator
Reading Sessions from 2001Oct

minsupport=0.001
minconfidence=0.05
number of transactions=13364
number of unique items=1587

wu01_28e3 -> wu01_22b0  s=0.00172  c=0.2948
wu01_3b   -> wu01_3c    s=0.00310  c=0.6774
wu01_22b0 -> wu01_28e3  s=0.00194  c=0.1780
wu01_34cf -> wu01_4a    s=0.00179  c=0.1318

...

Personen, die diesen Eintrag verwendet haben, verwendeten gemeinsam auch die folgenden Einträge:

1. **Free Programming Source Code**
   - Programmieren/Sprachen/C/C++
   - Programmieren/Sprachen/JAVA
   - Programmieren/Sprachen/PERL
   - Programmieren/Sprachen/MS Visual Basic

   Welcome to the FreeCode software archive! All of the programs found here are: Free for personal and commercial use. Compatible with Unix and/or Windows. Written in C/C++, Java, Perl, or Visual Basic.

2. **A UML tutorial**
   - Software Engineering/Object Oriented Unified Modeling Language

   Recommender Value: 13/83.16% - Region.
The Recommender System

Hahsler Michael: Einführung in das Objektorientierte Programmieren mit C++ (Introduction C++)

People, who used this site, also used the following sites:

1. Free Programming Source Code
2. A UML tutorial
3. UML Quick Reference
4. Overview of UML diagrams (Rational Software)
5. Nicolai Josuttis Die C++-Standardbibliothek
6. Vinny Carpenter Learn C/C++ today
Example: Amazon.com
Summary

- Information technology constantly changes the relationship between customers and a company.
- Convenience and better service for customers are key factors for success.
- Intelligent gathering, integration and usage of information about the customer is vital in order to survive competition.
- Data Warehouses and Data Mining provide the components for mass customization.
Readings


3. Rhonda Delmater and Monte Hancock, Data Mining Explained, Butterworth-Heinemann, Woburn, MA, 2001

These slides are available at:

http://wwwai.wu-wien.ac.at/~hahsler/research/datawarehouse_webster2001/talk/

(without the line break!)