

COMP8400: Algorithms and Techniques for Data Mining

Data issues in data mining

Data size and complexity

- *We are drowning in data but starving of knowledge*
(Jiawei Han)
- Automated data collection and mature database technology
 - Allows data to be stored efficiently, cheap, persistent
 - Using databases, data warehouses and other repositories
 - Data is increasingly stored distributed (storage area networks, grids, etc.)
- Large and massive data collections
 - Millions to billions of records
 - Tens to thousands of attributes (sometimes also called *variables*)
 - Data is rarely collected for data mining (rather for online transaction processing - OLTP)
- A lot of data is *write only* (or *read once only*)

Lecture outline

- Data size and complexity
- Data sources
- Data types and measurements
- Data formats
- Data warehousing
- Meta-data (describing data)
- *Real world data is dirty*

Data sources

- Relational databases
 - Transactional data, mostly normalised into many tables, with keys between them, continuous and frequent updates on (single) records
- Data warehouses
 - Decision support data, processed and cleaned, historical data, aggregated, updated at certain intervals (*more later*)
- Internet
 - Click-stream data, log files, HTML, XML, blogs, e-mails, etc.
- Files
 - Portable text (like comma separated, tabulator, fixed column) or non-portable proprietary binary files
- Scientific instruments, experiments and simulations
 - Astronomy, genomics, seismology, physics, chemistry, etc.
- Sensors (often data streams)

Types and measurements of data

- Numerical data
 - Integer, floating-point, binary, interval, ratio
 - Non-scalar (like velocity: speed and direction)
- Non-numerical data
 - Nominal data (just naming things, for example personal names)
 - Categorical data (grouping things, like postcodes, university course codes)
 - Ordinal data (ordering things, for example wine tasting, movie ratings)
- Series data
 - Ordering is an important feature (otherwise not series data)
 - One attribute must always be monotonic (increasing or decreasing)
 - Most common are *time series*

Formats of data

- Structured data
 - Relational database tables, integrated data warehouses
 - Images, video, audio (can be compressed)
- Semi-structured data
 - XML, HTML, e-mails, SMS, log files
- Free-format data
 - Mainly free-format text - ASCII or Unicode

Types and measurements of data (2)

- Multimedia data
 - Images, video, audio
 - Many standard formats used, binary, often compressed
- Different mappings and conversions between data types are possible and often needed
 - Some conversions are loss-less, others are lossy
- Different data mining techniques can handle different types of data
 - Some are restricted to certain types of data, for example only numerical data

Data warehousing (1)

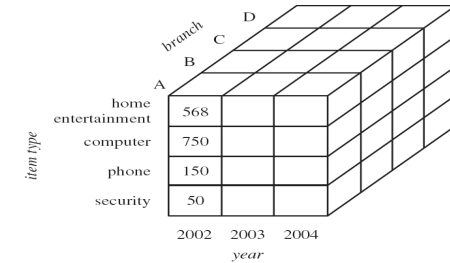
- A data warehouse is a decision support database that is maintained separately from an organisation's operational databases(s)
- Provides a solid platform of consolidated, historical data for analysis and mining
- Organised around major subjects, like customers, products, or sales
 - Provides a simple and concise view around these entities
- Often constructed by cleaning, standardising and integrating multiple heterogeneous data sources
 - To ensure consistency in coding, naming, measurements, etc.
 - *Topic of next lecture and next week's lectures*

Data warehousing (2)

- Longer time horizon than operational systems (that are used for transaction processing)
 - Historical data is important for analysis and mining
 - Separate data warehouse due to performance, data representation, consistency, integration, and data quality
 - Databases: OLTP (On-Line Transaction Processing)
 - Data warehouses: OLAP (On-Line Analytic Processing)
- Contains a time element
 - New data is, for example, loaded into a data warehouse every week or every month
- Only two basic operations: *Initial loading* and *querying* of data (read)
 - While transaction processing systems have *reads*, *writes* and *updates*

Data warehousing (3)

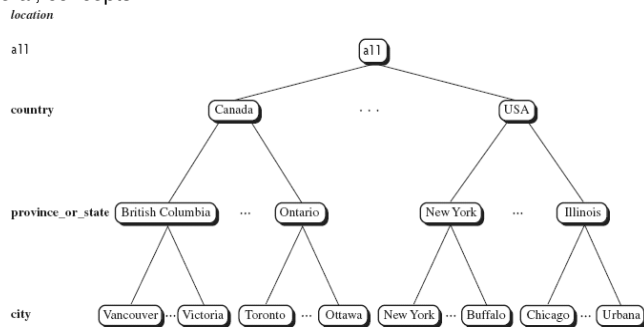
- Data warehouse architecture
 - Data cubes (multi-dimensional aggregated data views)
 - Dimension tables (details of the dimensions) and fact tables (values and names of the facts, e.g. *items_sold*, as well as keys into dimension tables)
 - Data is stored at different levels of details (e.g. *country / state / city*, or *item / item_group / item_category*)



Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

Data warehousing (4)

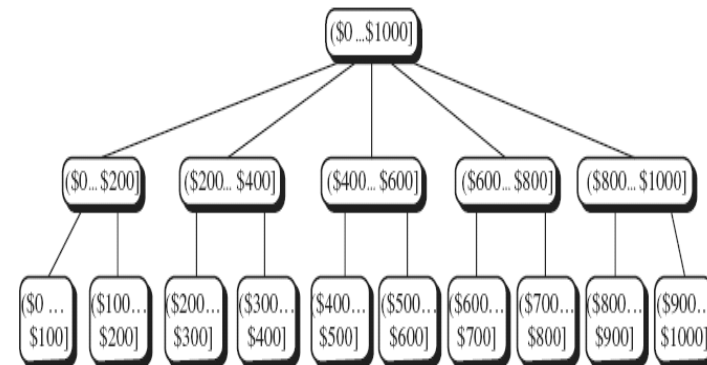
- Concept hierarchies
 - Defines a sequence of mappings from a set of low-level concepts to higher-level, more general, concepts



Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

Data warehousing (5)

- Concept hierarchies can be created by discretising or grouping numerical values

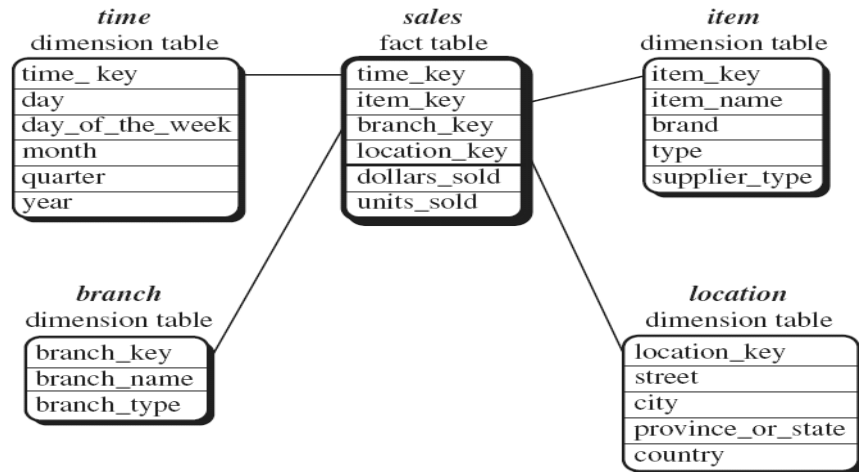


Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

Data warehousing (6)

- For data warehouses, a *multi-dimensional data model* is most popular
 - Compared to *entity-relationship model* for relational databases
- Implemented as:
 - Star schema (a large central *fact* table containing bulk of the data, and a set of smaller *dimension* tables)
 - Snowflake schema (variant of star schema with normalised dimension tables)
 - Fact constellation schema (multiple fact tables who share dimension tables), can be viewed as a collection of star schemas

Star schema



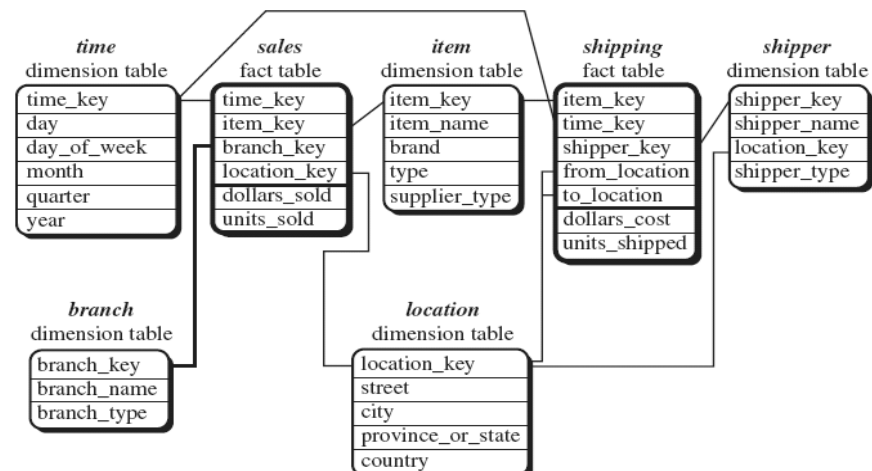
Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

Snowflake schema



Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

Fact constellation schema



Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

Data warehousing (7)

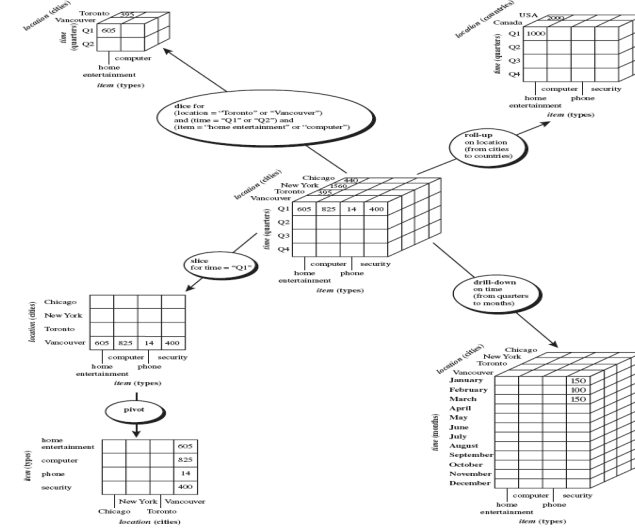
• Data warehouse operations

- Roll-up (summarise data)
- Drill-down or roll-down (get detailed view)
- Slice and dice (project and select)
- Pivot (rotate), re-orient the cube, 2D to 2D visualisation

• Applications of data warehousing:

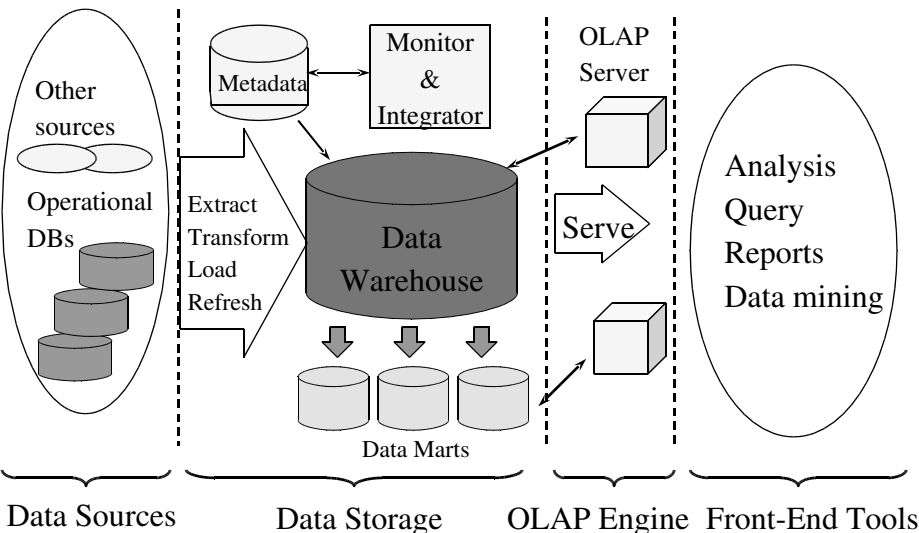
- Information processing (basic statistics, reporting, tables, charts, graphs, Web-based reporting, etc.)
- Analytic processing (further drill down, multi-dimensional analysis, on both summarised and detailed data)
- Data mining: Use a clean, stable, high-quality source for data mining algorithms

Data warehouse operations



Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

Data warehouse architecture



Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

Meta-data / describing data

• Meta-data: Data about data

- Structure of the data (types, names, format, etc.)
- Summary statistics (for example minimum / maximum values, histograms)
- Quality of the data (percentage of missing values, etc.)
- Information about pre-processing done
- Data source and owner, business information, charging policies
- Information about data access, retrieval, updates

• Sometimes called *data dictionary*

- Within a large organisation, or between organisations, e.g. health sector

• Stored in a database, as XML schema, etc.

• Meta-meta-data: Data about meta-data...?

Real world data is dirty (1)

- Various sources of errors
 - Misinterpretation of the data
 - Errors during data entry
 - Missing data
 - Out-of date data
 - Data from different sources
- Personal information (like names and addresses) are especially prone to data entry errors
- A great effort is often needed to *clean* and *standardise* raw data (data pre-processing)

Real world data is dirty (2)

- What does *dirty data* mean?
 - Incomplete data (missing attributes, missing attribute values, only aggregated data, etc.)
 - Inconsistent data (different coding, impossible values or out-of-range values)
 - Noisy data (data containing errors, outliers, not accurate values)
- For quality mining results, quality data is needed
 - *Garbage-in garbage-out* principle
- Transactional database systems should be designed with data quality and data mining in mind
- Pre-processing is an important step for successful data mining and data analysis
 - *More on this in the next lecture*