

## Lecture 4: Data Integration and Data Linkage

- Why data integration?
- Schema integration
- Handling redundant data
- Data linkage / matching
  - Deterministic linkage
  - Probabilistic linkage
- Deduplication / Geocoding
- Summary module 1
  - Data mining system architectures

## Why data integration and data linkage?

- Combine data from multiple sources into a coherent form
- Increasingly data mining projects require data from more than one data source
  - Data distributed (different databases or data warehouses) (for example an epidemiological study that needs information about hospital admissions and car accidents)
  - Geographic distribution or historical data (e.g. integrate historical data into a new data warehouse)
  - Enrich data with additional (external) data (to improve data mining accuracy)

## Schema integration

- Imagine two database tables

PID	Name	DOB
1234	Christine	01-01-1975
4791	Robert	21-10-1969

PID	GivenName	Age
1234	Christina	29
4791	Bob	35

- Integration issues
  - The same attribute may have different names
  - An attribute may be derived from another
  - Attributes might be redundant
  - Values in attributes might be different
  - Duplicate records (under different keys)
- Conflicts have to be detected and resolved
- Integration is made easier if unique entity keys are available in all the data sets (or tables) to be linked

## Handling redundant data

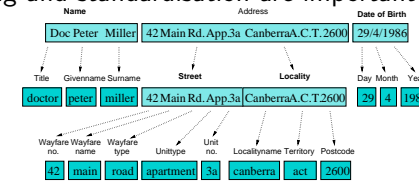
- Use correlational analysis
- Different coding / representation has to be considered (e.g. *metric / imperial* measures)
- Careful (manual) integration of the data can reduce or prevent redundancies (and inconsistencies)
- Deduplication (also called *internal data linkage*)
  - If no unique entity keys are available
  - Analysis of values in attributes to find duplicates
- Process redundant and inconsistent data (easy if values are the same)
  - Delete one of the values
  - Average values (only for numerical attributes)
  - Take majority values (if more than 2 duplicates and some values are the same)

## Data linkage / matching

- Task of linking together information from one or more data sources that represent the same entity
- If no unique entity keys in data, the available attributes have to be used (often personal information like names, addresses, dates of birth, etc.)
- Application areas
  - Health (epidemiology)
  - Census, taxation
  - Business mailing lists
  - Crime, fraud and terror detection
- Different parts of the linked records are of interest
  - Non-personal information (epidemiology, census, data mining)
  - Personal information (crime, fraud and terror detection, mailing lists)

## Data linkage / matching techniques

- Cleaning and standardisation are important first steps



- Deterministic or exact linkage
  - If a unique/stable/accurate key is available → *SQL join* (for example: *Medicare, Tax file number...* really suitable?)
  - Rule-based if no key is available (complicated to set-up and maintain, changes needed for new data sets)
- Probabilistic linkage (use available attributes)

## Probabilistic data or record linkage

- Computer assisted data linkage goes back as far as the 1950s (based on ad-hoc heuristic methods)
- Basic ideas of probabilistic linkage were introduced by *Newcombe & Kennedy* (1962)
- Theoretical foundation by *Fellegi & Sunter* (1969)
  - Using matching weights based on frequency ratios (global or value specific ratios)
  - Compute matching weights for all attributes used in linkage
  - Summation of matching weights is used to designate a pair of records as *link*, *possible-link* or *non-link*

## Linkage example: Month of birth

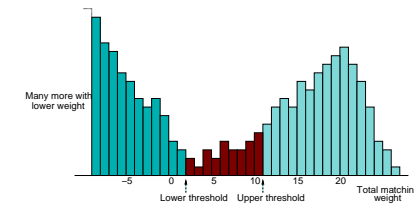
- Assume two data sets with a 3% error in field *month of birth*
  - Probability that two linked records (that represent the same person) have the same month is 97% (*M agreement*)
  - Probability that two linked records do not have the same month is 3% (*M disagreement*)
  - Probability that two (randomly picked) unlinked records have the same month is  $1/12 = 8.3\%$  (*U agreement*)
  - Probability that two unlinked records do not have the same month is  $11/12 = 91.7\%$  (*U disagreement*)
  - Agreement weight ( $M_{ag}/U_{ag}$ ):  $\log_2(0.97/0.083) = 3.54$
  - Disagreement weight ( $M_{di}/U_{di}$ ):  $\log_2(0.03/0.917) = -4.92$

## Value specific frequencies

- Example: Surnames
  - Assume the frequency of *Smith* is higher than *Dijkstra* (*NSW Whitepages*: 25,425 *Smith*, only 3 *Dijkstra*)
  - Two records with surname *Dijkstra* are more likely to be the same person than with surname *Smith*
- The matching weights need to be adjusted
  - Difficulty: How to get value specific frequencies that are characteristic for a given data set
  - Earlier linkages done on same or similar data
  - Information from external data sets (e.g. *Australian Whitepages*)

## Final linkage decision

- The final weight is the sum of weights of all attributes
  - Record pairs with a weight above an *upper threshold* are designated as a *link*
  - Record pairs with a weight below a *lower threshold* are designated as a *non-link*
  - Record pairs with a weight between the thresholds are *possible link*



## Data linkage: Blocking and classification

- Blocking
  - Potentially each record in one data set needs to be compared with all records in a second data set  $\rightarrow O(n^2)$
  - Blocking techniques reduce number of comparisons
  - Use one or more attributes as blocking variables, and only compare records that have the same value of such a *blocking variable*
- Classification
  - *Fellegi & Sunter* approach simply adds matching weights into one number
  - More advanced classifiers are possible (e.g. names and addresses are independent)
- ANU DM group research project (open-source linkage software *Febri* – Freely extensible biomedical record linkage)

## Deduplication and geocoding

- Deduplication
  - Find duplicate records within a data set
  - Important for longitudinal studies, business mailing lists, etc.
- Geocoding
  - Match addresses against *geocoded reference data*
  - Useful for spatial data analysis / mining and for loading data into geographical information systems
  - Matching accuracy is critical for good geocoding (as is accurate geocoded data)
  - Australia has a *Geocoded National Address File* (G-NAF) since early 2004

## Data mining system architectures

- **No coupling**  
(data mining system doesn't use any function of a database or data warehouse, basically file-input file-output)
- **Loose coupling**  
(data mining system is using some database or data warehouse functionality, e.g. querying and storing results back into database)
- **Semi-tight coupling**  
(efficient implementations of essential data mining primitives are integrated into database or data warehouse, e.g. sorting, indexing, aggregation)
- **Tight coupling**  
(data mining system smoothly integrated into database or data warehouse, allows query optimisation, but might lose flexibility)

## Summary module 1

- Data mining is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in large and complex data collections.
- (Many different) data issues are important for data mining
- A large proportion of time and effort in a data mining project is spent on data preprocessing
- Issues not covered: Ethical, privacy, social implications, etc.  
(especially important with techniques involving personal or confidential data, like data linkage)