

COMP8400: Algorithms and Techniques for Data Mining

More on data linkage

Lecture outline

- Probabilistic data or record linkage
 - Record pair comparison
 - Record pair classification
 - Linkage example: Month-of-birth weight calculation
 - Value specific matching weights
- Why blocking / indexing?
- Improved blocking
- Improved record pair classification
- Deduplication and geocoding
- Data linkage research at the ANU

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 and Kennedy* (1962)
- Theoretical foundation by *Fellegi and Sunter* (1969)
 - Compare common attributes from record pairs and calculate similarity values (also called *matching weights*)
 - Using matching weights based on frequency ratios
 - Summation of matching weights is used to classify a pair of records as *match*, *possible match* or *non-match*
 - Manual clerical review required to determine the match status of the possible matches

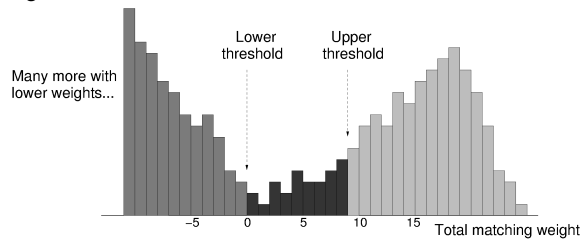
Record pair comparison example

- Attributes are compared using various comparison functions (like exact or approximate string, numeric, date, age, etc.)

Record A:	[Dr,	Peter,	Paul,	Miller]	
Record B:	[Mr,	John,	,	Miller]	
	[0.2,	-3.2,	0.0,	2.4]	-0.6

Record pair classification

- The final *matching weight* is the sum of the attribute comparison weights (similarity values)
 - Record pairs with a weight above an upper threshold are classified as a *match*
 - Record pairs with a weight below a lower threshold are classified as a *non-match*
 - Record pairs with a weight between the thresholds are classified as a *possible match*



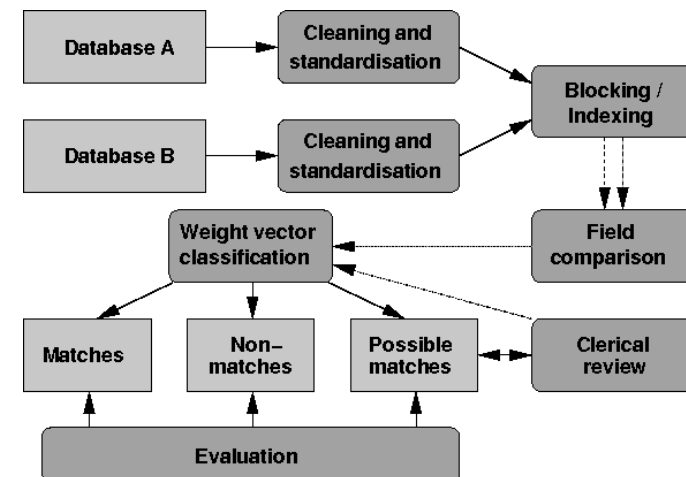
Value specific frequencies

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

Linkage example: Month-of-birth weight calc

- Assume two databases that have a 3 % error in the month-of-birth attribute
 - Probability that two matched records (that represent the same person) have the same month is 97 % (*M agreement*)
 - Probability that two matched records (that represent the same person) do not have the same month is 3 % (*M disagreement*)
 - Probability that two un-matched records (randomly picked) have the same month is $1/12 = 8.3\%$ (*U agreement*)
 - Probability that two un-matched records (randomly picked) 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_{disag} / U_{disag}): $\log_2(0.03 / 0.917) = -4.92$

Data linkage process



Why blocking / indexing?

- The number of record pair comparisons equals the product of the sizes of the two data sets
(for example, linking two data sets with 1 and 5 million records will result in $1,000,000 \times 5,000,000 = 5 \times 10^{12}$ record pairs)
- Performance bottleneck in a data linkage system is usually the (expensive) comparison of field values between record pairs (similarity measures / field comparison functions)
- Blocking / indexing / filtering techniques are used to reduce the large amount of comparisons
- Aim of blocking: Cheaply remove candidate record pairs which are obviously not matches

Improved blocking

- Recent research methods
 - Sorted neighbourhood approach (sliding window over sorted blocking variable)
 - Fuzzy blocking using n-grams (for example: *bigrams*: 'peter' -> ['pe', 'et', 'te', 'er'], 'pete' -> ['pe', 'et', 'te'])
 - Overlapping canopy clustering (where records are inserted into several clusters)
- Post-blocking filtering (like length differences or n-grams count differences)
- US Census Bureau: *BigMatch*
(pre-process 'smaller' data set so its values can be directly accessed in main memory; with all blocking passes in one go)

Traditional blocking

- Traditional blocking works by only comparing record pairs that have the same value for a blocking variable
(for example, only compare records which have the same postcode value)
- Problems with traditional blocking
 - An erroneous value in a blocking variable results in a record being inserted into the wrong block (several passes with different blocking variables can solve this)
 - Values of blocking variable should be uniformly distributed (as the most frequent values determine the size of the largest blocks)
 - Example: Frequency of 'Smith' in NSW: 25,425

Improved record pair classification

- Summing of weights results in loss of information
(like same name but different address, or different address but same name)
- View record pair classification as a multidimensional binary classification problem
(use weight vector to classify record pairs as matches or non-matches, but no possible matches)
- Many machine learning techniques can be used
 - Supervised: Decision trees, neural networks, learnable string comparisons, active learning, etc.
 - Un-supervised: Various clustering algorithms
- Major issue: Lack of training data

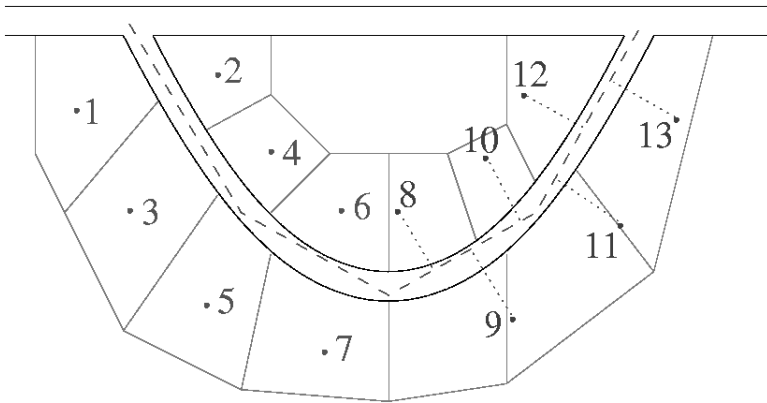
Classification challenges

- In many cases there is no training data available
 - Possible to use results of earlier linkage projects?
 - Or from clerical review process?
- How confident can we be about correct manual classification of possible links?
- Often there is no *gold standard* available (no data sets with true known linkage status)
- No large test data set collection available (like in information retrieval or machine learning)

Deduplication and geocoding

- Deduplication
 - Find duplicate records within a database
 - Important for longitudinal (over time) studies, business mailing lists, etc.
- Geocoding
 - Match addresses against *geocoded* reference data (addresses and their geographic locations: latitudes and longitudes)
 - 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 address data)
 - Australia has a *Geocoded National Address File (G-NAF)* since early 2004 (all Australian property addresses and their locations)
 - Commercial geocoding systems mainly work on *street centreline* data

Geocoding example (1)



Geocoding example (2)



Data linkage research at the ANU

- We have been working in data linkage since 2002
(see: <http://datamining.anu.edu.au/linkage.html>)
- Research project in collaboration with the New South Wales Department of Health, Sydney
- We have developed an open source data linkage system called *Febrl* (Freely Extensible Biomedical Record Linkage)
- Only data cleaning, deduplication and record linkage system in the world that is free and has a graphical user interface
- Interested in this area? We have projects at various levels (implementation, honours, PhD/MPhil)

What now.. things to do

- Read through assignment 1 draft specifications and **start working on task 1!**
(ask questions about assignment on Forum or e-mail me)
- Download papers for tutorial 1 (available tomorrow) and question sheet – read papers before tutorial 1 (on Thursday 19 March)
- Read through student presentations specifications
- Read chapters 1, 2 and 3 in text book (if not yet done)