Data Linkage for Bioinformatics

Peter Christen

Data Mining Group
Department of Computer Science
Australian National University

Contact: peter.christen@anu.edu.au

Data Linkage is the task of linking together records representing the same entity (patient, customer, protein, genome) from one or more data sources.

Real world data is often dirty, therefore data cleaning and data standardisation are important first steps for successful data linkage.

Three records, which represent the same person?

1. Dr Smith, Peter; 42 Miller Street 2602 O'Connor
2. Pete Smith; 42 Miller St 2600 Canberra A.C.T.
3. P. Smithers, 24 Mill Street 2600 Canberra ACT

Applications and Usage

Applications of data linkage
- Remove duplicate records in a data set (internal linkage)
- Merge new records into a larger master data set
- Create patient or customer oriented statistics
- Compile data for longitudinal (over time) studies
- Clean data sets for further data analysis or data mining

Usage of data linkage
- Business mailing lists and customer statistics
- Health and biomedical research (epidemiology)
- Fraud and crime detection
- Bioinformatics (?)

Data Linkage Techniques

Deterministic or exact linkage
- A unique identifier is needed, which is of high quality (precise, robust, stable over time, highly available)
- Examples: Medicare number, Tax file number, genome identifier (are they really unique, stable, trustworthy?)

Probabilistic linkage
- Apply linkage using information available in records (can be missing, wrong, coded differently, outdated, etc.)
- Examples: name, address, date of birth, etc.

Other techniques
Rule-based, fuzzy approach, information retrieval, etc.
Data Issues in Bioinformatics

"In a dynamic heterogeneous environment such as bioinformatics, many different databases and software systems are used." (Limsoon Wong)

- Databases grow from small to large, from simple to complex, with often non-standard query software
- Researchers demand flexible access and queries in ad-hoc combinations (data exploration)
- Databases are distributed world-wide

The challenge: How to manipulate and integrate data retrieved from various databases so it can be used to investigate a specific biomedical problem?

Issues for Data Linkage in Bioinformatics

- Meta data issues (attributes, formats, coding)
  → Standards are needed to describe biological information
- Distributed databases (remote access, availability, local caches, updates)
- Specific operations for bioinformatics
  → Data cleaning and standardisation is different from name and address cleaning
  → Sequence and protein comparisons (algorithms like FASTA, BLAST, Smith/Waterman, etc.)
- No gold standard (not possible to get true results, accuracy not known)

Data Linkage in Bioinformatics?

- Example: Assume database $A$ has attributes $v$, $w$ and $y$; and database $B$ has attributes $v$, $w$ and $z$
  → A researcher needs both attributes $A.y$ and $B.z$
  → Attribute $v$ is used to link records from $A$ and $B$
    (link two records if $A.v = B.v$)
  → What happens if $v$ is recorded wrongly for some records?
  → For linked records, what to do if $A.w$ differs from $B.w$?

Many experiments result in fuzzy data

- How are errors and missing data represented?
- Amount of data forces automatic analysis methods

ANU Data Linkage Project

- Project with NSW Department of Health
  → ANU Industry-Collaboration Scheme / APAC 2002 - 03
  → ARC Linkage Grant 2004 - 06 (PhD Scholarship on offer)
- Commercial software for data linkage is often cumbersome to use and expensive
- Project aims
  → Allow linkage of larger data sets (high-performance and parallel computing techniques)
  → Reduce the amount of human resources needed
    (improve linkage quality by using machine learning)
  → Reduce costs (free open source software)
ANU Data Linkage Project (cont.)

-Prototype open source software "Febrl"
  -Freely extensible biomedical record linkage
  -Probabilistic data cleaning and standardisation (based on hidden Markov models)
  -Probabilistic data linkage (based on Fellegi & Sunter model)
  -Parallelism transparent to the user
  -Based on scripting language Python [www.python.org](http://www.python.org)


Outlook

-Data quality is an important issue in many application areas, including bioinformatics
-Data linkage helps to enrich data and enables research with more details and of better quality
-Many open issues regarding data quality and data linkage in bioinformatics
-ANU Data Mining group has one ARC PhD scholarship (APAI) on offer starting in early 2004