Data Integration

Week 3

Nurul Afiqah Burhanuddin
nurul.afiqah@ukm.edu.my
Room 2119

Introduction

- Data mining often requires data integration - merging data from multiple data sources.
- The need for integration may arise due to new applications that span multiple databases, e.g.:
  - an organization may want an application that carries out an enterprise-wide analysis of operations
  - corporate mergers and acquisitions
- However, pre-existing databases in most organizations are defined and populated by different people at different times in response to different organizational or end-user requirements.
- Data integration is necessary to have a complete view of the data and to make sure that the data is accurate and up-to-date.
- Careful integration avoids redundancies and inconsistencies in the resulting data set and thus helps improve the accuracy and speed of the subsequent data mining process.

Issues in Data Integration

1. **Domain mismatch:** The domains of similar attributes are not compatible in semantic or structure.
   - semantic mismatch: the currency attribute in one database being in USD while the corresponding currency attribute in another database is in MYR is a case of semantic mismatch.
   - structural mismatch: when the name attribute in one database has a data type of string while the name attribute in another database is composed of three subattributes of string data type, namely, lastname, firstname, and middlename.
2. **Schema mismatch:** This problem arises when the structures and of two databases are not compatible.
   - the Employee column in one database may correspond to a union of part-time employee and full-time employee while the Employee column in another database only consider full-time employee

Issues in Data Integration

3. **Constraint mismatch:** The constraints specified in the participating databases may be incompatible.
    - the mathematics department database may have the constraint of requiring all students to have a CGPA of greater than 3.5, whereas the computer science department database may have the constraint of requiring all students to have a CGPA of greater than 3.75.
4. **Entity identification:** This is the problem of identifying instances from different database that correspond to the same real-world entity.
    - different employee ID assigned to the same employee in different databases.

Issues in Data Integration

5. **Attribute value conflict:** arises when the attribute values in the two databases, modelling the same property of a real-world entity, do not match.
   - Data scaling conflict occurs when the domains of semantically related attributes use different units of measurement.
   - Inconsistent data occur when semantically equivalent attributes have different values.
   - Missing data refers to the situation when instances modelling the same real world do not have the same set of attributes.

Merging & Binding Data in R

- Merging from y to x:

	dplyr	base
includes all rows in x or y	full_join()	merge(…,all=T)
includes all rows in x and y	inner_join()	merge()
includes all rows in y	left_join()	merge(…, all.x = T)
includes all rows in y	right_join()	merge(…, all.y = T)

- Binding:

	dplyr	base
bind by row	bind_rows()	rbind()
bind by column	bind_cols()	cbind()