DATA INTEGRATION

STQD6414 PERLOMBONGAN DATA

Assoc. Prof. Dr. Nurulkamal Masseran
Department of Mathematical Sciences
Universiti Kebangsaan Malaysia

INTRODUCTION:

- Data which obtained from various sources need to be integrated into a common form before the data mining analysis can be carried out.
- However, data from different sources often have different data structures and format.
- Data integration needs to be done to avoid inconsistent data problems.
- In addition, data integration can also overcome the problem of information overlap in the data.
- Through a uniform set of data, it will facilitate the process of data mining analysis.
- This process is known as an entity identification problem.

ENTITY IDENTIFICATION PROBLEM:

- Entity identification problems refer to techniques for managing a variety of different structures and forms of data entry between multiple data sources.

Example:

- How can data scientists be sure that "customer id" in one data file and "cust number" in another data files are actually refer to the same attribute?.

- How to combine "customer id" information with "cust number" ?.

IMPORT DATA SETS FROM DIFFERENT SOURCES/FILES:

- Data from different sources/files can be imported into R.

- **Example:**

i) R file.

ii) Excell file.

iii) Text file.

iv) SPSS and SAS file.

v) Unstructured Data (Text data).

vi) Web data/Database.

vii) Media social data (facebook, twitter, and etc.).

INTEGRATION OF DATA FROM DIFFERENT SOURCES:

- Among the important things to consider when integrating data from different sources are the following:

i) Integration of data with different attributes.

ii) Data integration based on inconsistent attribute names with some mismatched attribute values

iii) Rename an attribute in a data set.

iv) Customize data with inconsistent attribute values.

data integration from different databases/clouds will be discussed in the data management course

CUSTOMIZE DATA ATTRIBUTES:

- Among the techniques of modifying data attributes:

i) Extract specific attributes in the data
ii) Adds new observations in the data
iii) Adds new attributes to the data.
iv) Remove certain attributes from the data.
v) Data Subset
vi) Sorting

EKSPORT DATA FROM R:

- Data from R can be exported out to various types of storage files. Among them are:

i) Text File.

ii) CSV File.

iii) R File.

iv) And etc.

ASSIGNMENTS 2:

1. Combine the data from the custdata2i and custdata3i files through the identifying entity to the same "customer id" attribute. Ignore observations that do not contain complete attribute information.

2. Create a new data set for male customers with a salary greater thab 7000 dollars and also contain information for the following attributes:

- state.of.res , custid, marital.stat, health.ins, housing.type , num.vehicles , sex, income

3. Show the data for each customer in the form of an ascending salary order.

4. Suppose new information is known as follows:

- state.of.res: alabama, Louisiana, new York
- ID customer: 567891, 33421, 21134
- marital.stat: Married, Never Married, bercerai
- Ins.health: TRUE, FALSE, TRUE
- Home Status: Sewa, Not Available, loan
- num.vehicles: 2, 1, 2
- sex: M, Male, lelaki
- is.employed: TRUE, FALSE, TRUE
- income: 99200, Not Available, 150341

5. Add the new observation information in your data set.

6. Suppose you know new attribute information (personal loan) for each customer (newinfo file), combine the new attribute information with your data set.

REFERENCES:

- Aggarwal, C.C. (2015). *Data Mining. The Textbook*. Springer, New York.
- García, S., Luengo, J., Herrera, F. (2015). *Data Preprocessing in Data Mining*. Springer, New York.
- Jafari, R. (2022). *Hands-On Data Preprocessing in Python: Learn how to effectively prepare data for successful data analytics 1st Edition*. Packt Publishing
- Kuhn, M., Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Chapman and Hall.
- Luengo, J., García-Gil, D., Ramírez-Gallego, S., García, S., Herrera, F. (2020). *Big Data Preprocessing*. Springer, Switzerland.

NEXT TOPIC:

Data Cleaning