INTEGRASI DATA

STQD6414 PERLOMBONGAN DATA

Prof. Madya Dr. Nurulkamal Masseran
Jabatan Sains Matematik
Universiti Kebangsaan Malaysia

PENGENALAN:

- Data yang diperolehi dari pelbagai sumber perlu diintegrasikan kepada satu bentuk yang sama sebelum di analisis.
- Namun, data dari sumber yang berbeza seringkali mempunyai struktur data dan format yang berbeza.
- Integrasi data perlu dibuat untuk mengelakkan masalah data tidak konsisten dan juga masalah data bertindan maklumat.

- Menerusi satu bentuk data yang seragam, ianya akan memudahkan proses analisis perlombongan data.
- Ini dikenali sebagai masalah pengecaman entiti (*entity identification problem*).

MASALAH PENGECAMAN ENTITI:

- **Masalah pengecaman entiti** merujuk kepada struktur dan bentuk pemasukan data yang berbeza antara beberapa sumber data.

Contoh:

- Bagaimana Saintis data pasti bahawa *"customer id"* dalam satu file data dan "*cust number*" dalam file data lain merujuk kepada atribut yang sama ?.

- Bagaimana untuk menggabungkan maklumat *"customer id"* dengan *"cust number"*?.

IMPORT SET DATA DARIPADA SUMBER/FILE YANG BERBEZA:

- Data daripada sumber/file yang berbeza boleh diimport ke dalam R.

- **Contoh Data:**

i) R file.

ii) Excell file.

iii) Text file.

iv) Data Tak Berstruktur (data teks).

v) Web data/Pangkalan data.

vi) **SPSS, SAS dan lain-lain.** (library(foreign); dataset = read.spss(file.choose(), to.data.frame=TRUE))

vii) Data dari laman sosial(facebook, twitter, dll).

INTEGRASI DATA DARI SUMBER BERBEZA:

- Antara perkara yang menjadi isu penting apabila mengintegrasi data dari sumber berbeza:

i) Integrasi data yang berlainan atribut.

ii) Integrasi data berdasarkan nama atribut yang tidak konsisten, saiz sampel yang sama, namun mengandungi sebahagian nilai atribut yang tidak sepadan.

iii) Menamakan semula atribut dalam set data.

iv) Ubahsuai data dengan nilai-nilai atribut tidak konsisten.

###integrasi data dari pangkalan data yang berbeza akan dibincangkan dalam kursus pengurusan data###

UBAHSUAI ATRIBUT DATA:

- Antara kaedah pengubahsuaian atribut data:

i) Ambil atribut tertentu dalam data.
ii) Menambah cerapan baru dalam data.
iii) Menambah atribut baru dalam data.
iv) Buang atribut tertentu dalam data.
v) Data Subset (Memilih cerapan tertentu).
vi) Pengisihan (*Sorting*).

EKSPORT DATA DARIPADA R:

- Data dari R boleh dieksport keluar kepada pelbagai jenis fail simpanan. Antaranya:

i) Text File.

ii) CSV File.

iii) R File.

iv) Dan lain-lain.

TUGASAN:

1. Gabungkan data dari file custdata2i dan custdata3i menerusi entiti pengecaman terhadap atribut "customer id" yang sama. Abaikan cerapan yang yang tidak mengandungi maklumat atribut yang lengkap.

2. Bentukkan data set baru bagi pelanggan lelaki yang mempunyai gaji melebihi 7000 dollar dan juga mengandungi maklumat bagi atribut-atribut berikut:

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

3. Tunjukkan data bagi setiap pelanggan dalam bentuk susunan gaji yang semakin tinggi.

4. Misalkan diketahui maklumat baru seperti berikut:

- 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. Tambahkan maklumat cerapan baru tersebut dalam data set anda.

6. Misalkan diketahui maklumat atribut baru (personal loan) untuk setiap pelanggan (file newinfo), gabungkan maklumat atribut baru tersebut dengan data set anda.

RUJUKAN:

- 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.

TOPIK SETERUSNYA:

Pembersihan Data