BINUS UNIVERSITY ONLINE LEARNING

People
Innovation
Excellence

LECTURE NOTES

COMP8014

Knowledge Data Discovery

Data Preprocessing

LEARNING OUTCOMES

LO 2: Mahasiswa diharapkan mampu menggunakan teknik eksplorasi data dan preprocessing

OUTLINE MATERI :

1. Mengapa data preprocess?
2. Pembersihan data/Data Cleansing
3. Integrasi data dan transformasi
4. Reduksi data

ISI MATERI

Mengapa Data preprocessing diperlukan?

- Raw Data biasanya berupa data yang ‘dirty’ atau TIDAK bersih
    - Incomplete: lacking attribute
        - e.g., occupation=“ ”
    - Noisy: containing errors or outliers
        - e.g., Salary=“-10”
    - Inconsistent: containing discrepancies in codes or names
        - e.g., Age=“42” Birthday=“03/07/1997”
        - e.g., Was rating “1,2,3”, now rating “A, B, C”
        - e.g., discrepancy between duplicate records

Mengapa raw data adalah data dirty?

- Data tidak lengkap
    - "Tidak berlaku" nilai data saat dikumpulkan
    - Pertimbangan yang berbeda antara waktu ketika data dikumpulkan dan saat dianalisis *)
    - Masalah manusia / perangkat keras / perangkat lunak
- Data bising (nilai salah)
    - Instrumen pengumpulan data yang salah
    - Kesalahan manusia atau komputer pada entri data
    - Kesalahan dalam transmisi data
- Data tidak konsisten
    - Sumber data yang berbeda
    - Pelanggaran ketergantungan fungsional (misalnya, memodifikasi beberapa data yang ditautkan) **)
- Duplikasi data

Mengapa preprocessing data penting?

- Tidak ada data kualitas, tidak ada hasil penambangan yang berkualitas!
    - Keputusan kualitas harus didasarkan pada data kualitas
    - Mis., Data rangkap atau data yang hilang dapat menyebabkan statistik yang salah atau bahkan menyesatkan.
- Data warehouse membutuhkan integrasi data kualitas yang konsisten
    - Ekstraksi data, pembersihan, dan transformasi terdiri dari sebagian besar pekerjaan membangun gudang data (sampai 90%)

Bagaimana pengukuran Kualitas Data Multi Dimensi?

- Accuracy
- Completeness
- Consistency
- Timeliness
- Believability
- Non-redudancy
- Relevance
- Interpretability
- Accessibility

Apa tugas utama dalam Preprocessing Data?

- Data cleaning
    - Isi nilai yang hilang, data berisik yang halus, identifikasi atau hapus outlier, dan atasi inkonsistensi
- Data integration
    - Penggabungan multiple databases, data cubes, atau files
- Data transformation
    - Normalization dan aggregation

- Data reduction
  - Memperoleh penurunan representasi dalam volume tetapi menghasilkan hasil analitis yang sama atau serupa
- Data discretization
  - Bagian dari reduksi data namun sangat penting, terutama untuk data numerik

SIMPULAN

1. Data mentah yang kotor biasanya karena tidak lengkap, berisik dan tidak konsisten.
2. Data Preprocessing penting untuk memastikan kualitas hasil penambangan.
3. Pembersihan data adalah salah satu dari tiga masalah terbesar dalam pergudangan data.
4. Integrasi data menggabungkan data dari berbagai sumber menjadi toko yang koheren.
5. Reduksi/Pengurangan data mengurangi representasi kumpulan data yang jauh lebih kecil dalam volume namun menghasilkan hasil analitis yang sama (atau hampir sama).

BINUS UNIVERSITY ONLINE LEARNING

People Innovation Excellence

DAFTAR PUSTAKA

1. Han, J., Kamber, M., & Pei, Y. (2006). “Data Mining: Concepts and Technique”. Edisi 3. Morgan Kaufman. San Francisco
2. Tan, P.N., Steinbach, M., & Kumar, V. (2006). “Introduction to Data Mining”. Addison-Wesley. Michigan
3. Witten, I. H., & Frank, E. (2005). “Data Mining : Practical Machine Learning Tools and Techniques”. Second edition. Morgan Kaufmann. San Francisco