

LECTURE NOTES

COMP8014

Knowledge Data Discovery

Introduction

LEARNING OUTCOMES

L0 1: Mahasiswa diharapkan mampu menjelaskan konsep dasar Knowledge Data Discovery

OUTLINE MATERI :

1. Motivasi: Mengapa data mining?
2. Apa itu data mining?
3. Data Mining: Pada jenis data apa?
4. Fungsi data mining
5. Klasifikasi sistem data mining
6. Top-10 algoritma data mining yang paling populer
7. Isu utama dalam data mining

ISI MATERI

Motivasi : Mengapa Data Mining?

Pertumbuhan data yang eksplosif dari terabyte hingga petabyte

- Kebutuhan pengumpulan data dan ketersediaan data
 - Alat pengumpulan data otomatis,
 - sistem database, Web, masyarakat terkomputerisas
- Tersedianya sumber utama data berlimpah
 - Bisnis: Web, e-commerce, transaksi, saham, ...
 - Ilmu Pengetahuan: Penginderaan jarak jauh, bioinformatika, simulasi ilmiah,
 - Masyarakat dan semua orang: berita, kamera digital, YouTube

Problem: Kita tenggelam dalam data, tapi kelaparan akan pengetahuan (Knowledge)!

Bagaimana cara mendapatkan pengetahuan tersebut.

"Kebutuhan adalah ibu dari penemuan" -Data mining-Analisis otomatis kumpulan data masif.

- Seperti gambar dibawah, seseorang seperti kebingungan untuk menganalisis data yang bermacam-macam dan besar.
- Diperlukan suatu teknis tertentu untuk menganalisa data tersebut.



Apakah Data Mining itu?

- Data mining (penemuan pengetahuan dari data) adalah Ekstraksi pola atau pengetahuan yang menarik (tidak sepele, implisit, tidak diketahui dan berpotensi berguna) dari sejumlah besar data
- Secara umum, data mining adalah kumpulan metode untuk mengekstrak pengetahuan dari suatu data yang besar.

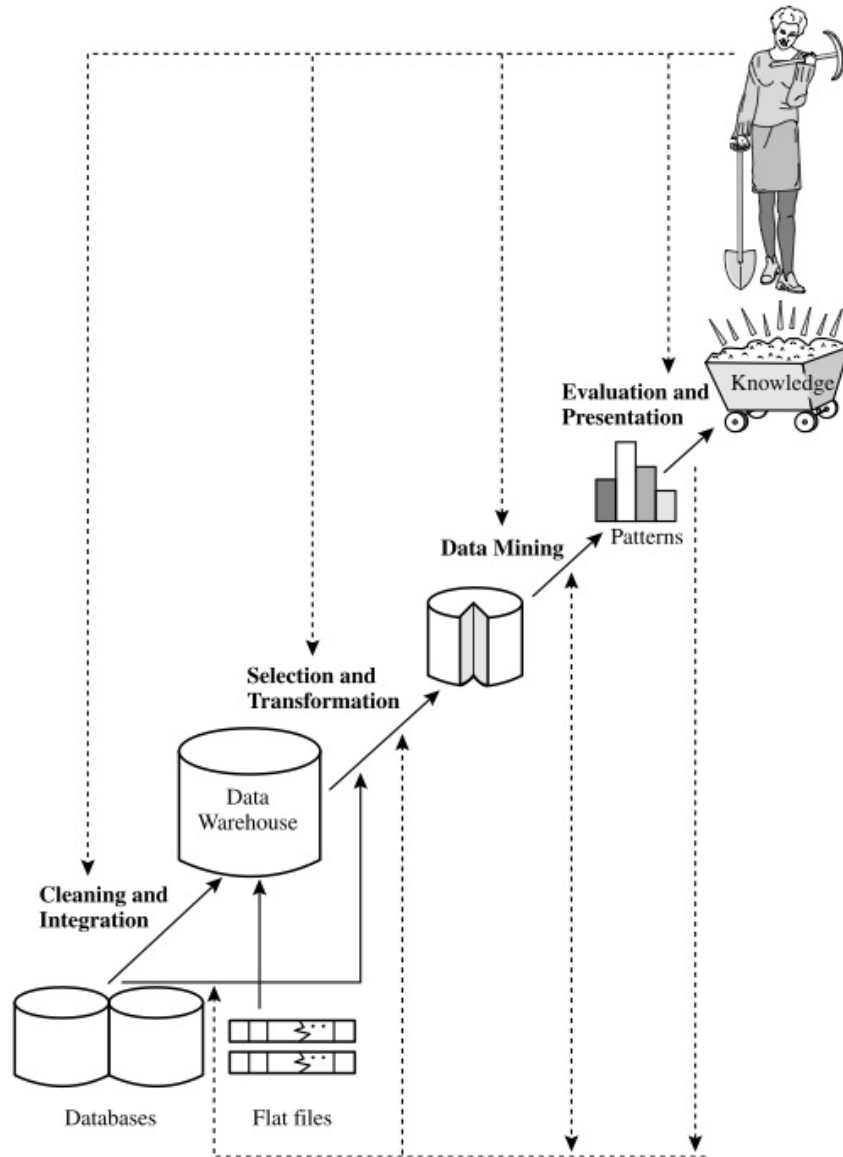
Nama alternatif?

- Beberapa peneliti atau penulis menggunakan nama lain untuk Data Mining, antara lain:
 - Knowledge discovery (mining) in databases (KDD),
 - knowledge extraction,
 - data/pattern analysis,
 - data archeology,
 - data dredging,
 - information harvesting,
 - business intelligence.

Process dari KDD?

Seperti dijelaskan pada gambar di bawah, secara umum proses dari KDD meliputi:

- data cleaning,
- data integration,
- data selection,
- transformation,
- data mining,
- pattern evaluation,
- knowledge presentation



Aplikasi dari Data Mining?

- Data analysis and decision support
 - Market analysis and management
 - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
 - Risk analysis and management
 - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
 - Fraud detection and detection of unusual patterns (outliers)

- Other Applications
 - Text mining (news group, email, documents) and Web mining
 - Stream data mining (cctv, etc.)
 - Bioinformatics and bio-data analysis

Fungsionalitas data mining?

- characterization,
- discrimination,
- association,
- classification,
- clustering,
- outlier
- trend analysis
- etc.

Fungsionalitas data mining ini akan dibahas pada topic-topik selanjutnya.

Apakah semua knowledge dari data mining itu penting?

- Data mining dapat menghasilkan ribuan pola/knowledge:
- Tidak semuanya menarik
 - Pendekatan yang disarankan: Pertambangan yang berpusat pada manusia, berbasis query, terfokus
- Beberapa hal yang perlu diperhatikan dalam penentuan pola/knowledge yang penting;
- Interestingness measures
 - Sebuah pola menarik jika mudah dipahami oleh manusia, berlaku pada data baru atau uji coba dengan beberapa tingkat kepastian, berpotensi berguna, baru, atau memvalidasi beberapa hipotesis bahwa pengguna berusaha untuk mengkonfirmasi

- Objective vs. subjective interestingness measures
 - berdasarkan statistik dan struktur pola, misalnya, dukungan, kepercayaan, dll.
 - Subjektif: berdasarkan kepercayaan pengguna terhadap data, mis., Tak terduga, kebaruan, tindakan tindakan, dll.

Bagaimana menemukan semua pola dan pola yang penting saja?

- Temukan semua pola yang menarik: Completeness
 - Bisakah sistem data mining menemukan semua pola yang menarik?
 - Apakah kita perlu menemukan semua pola yang menarik?
 - Heuristik vs pencarian yang melelahkan
 - Asosiasi vs klasifikasi vs pengelompokan
- Cari hanya pola menarik: Masalah optimasi
 - Bisakah sistem data mining hanya menemukan pola yang menarik?
 - Pendekatan
 - Pertama umum semua pola dan kemudian saring yang tidak menarik
 - Buatlah hanya pola menarik-query mining optimization

Isu di Data Mining?

- Metodologi Data Mining
 - Pertambahan berbagai jenis pengetahuan dari beragam jenis data, mis., Bio, streaming, Web
 - Kinerja: efisiensi, efektivitas, dan skalabilitas
 - Evaluasi pola: masalah menarik
 - Penggabungan pengetahuan latar belakang
 - Menangani kebisingan dan data tidak lengkap

- Metode penambangan paralel, terdistribusi dan inkremental **)
- Integrasi pengetahuan yang ditemukan dengan yang sudah ada: fusi pengetahuan
- Interaksi pengguna
 - Data Mining interaktif
 - Ekspresi dan visualisasi hasil data mining
- Masyarakat
 - Dampak sosial data mining
 - Privasi melestarikan data mining
 - Data mining yang invisible

SIMPULAN

1. Data mining: Menemukan pola menarik dari sejumlah besar data
2. Sebuah evolusi alami teknologi database, sangat diminati, dengan aplikasi yang luas
3. Proses KDD meliputi pembersihan data, integrasi data, pemilihan data, transformasi, data mining, evaluasi pola, dan presentasi pengetahuan
4. Data mining bisa dilakukan di berbagai repositori informasi. Fungsi penambangan data: karakterisasi, diskriminasi, asosiasi, klasifikasi, clustering, analisis outlier dan trend, dll.
5. Telah dibahas sistem dan arsitektur data mining
6. Telah dijelaskan isu utama dalam data mining

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