BAB 1 PENGERTIAN *DATA MINING* DAN FUNGSI-FUNGSI *DATA MINING*

Pendahuluan

Perkembangan yang cepat dalam teknologi pengumpulan dan penyimpanan data telah memudahkan organisasi untuk mengumpulkan sejumlah data berukuran besar sehingga menghasilkan gunung data. Ekstraksi informasi yang berguna dari gunung data menjadi pekerjaan yang cukup menantang. Seringkali alat dan teknik analisis data tradisional tidak dapat digunakan dalam mengekstrak informasi dari data berukuran besar. *Data mining* adalah teknologi yang merupakan campuran metode-metode analisis data dengan algoritme-algoritme untuk memproses data berukuran besar. *Data mining* telah banyak diaplikasikan dalam berbagai bidang, diantanya dalam bidang bisnis dan kedokteran.

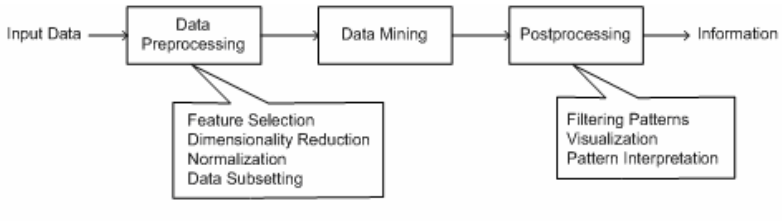
Dalam bidang bisnis, teknik *data mining* digunakan untuk mendukung cakupan yang luas dari aplikasi-aplikasi bisnis inteligen seperti customer profiling, targeted marketing, workflow management, store layout dan fraud detection. Teknik *data mining* dapat digunakan untuk menjawab pertanyaan bisnis yang penting seperti ”Siapakan pelanggan yang akan paling banyak mendatangkan keuntungan ?” dan ” Seperti apa perkiraan pendapatan perusahaan tahun depan ?”.

Dalam bidang kedokteran, peneliti dalam bidang biomolekuler dapat menggunakan teknik *data mining* untuk menganalisis sejumlah besar data genomic yang sekarang ini telah banyak dikumpulkan untuk menjelaskan struktur dan fungsi gen, memprediksi struktur protein, dan lain-lain.

1.1 Pengertian *Data mining*

Data mining adalah sebuah proses percarian secara otomatis informasi yang berguna dalam tempat penyimpanan data berukuran besar. Istilah lain yang sering digunakan diantaranya knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, dan business intelligence. Teknik *data mining* digunakan untuk memeriksa basis data berukuran besar sebagai cara untuk menemukan pola yang baru dan berguna. Tidak semua pekerjaan pencarian informasi dinyatakan sebagai *data mining*. Sebagai contoh, pencarian *record* individual menggunakan *database management system* atau pencarian halaman we tertentu melalui kueri ke semua search engine adalah pekerjaan pencarian informasi yang erat kaitannya dengan *information retrieval*. Teknik-teknik *data mining* dapat digunakan untuk meningkatkan kemampuan sistem-sistem *information retrieval*.

Data mining adalah bagian integral dari *knowledge discovery in databases* (KDD). Keseluruhan proses KDD untuk konversi raw data ke dalam informasi yang berguna ditunjukkan dalam Gambar 1.1.

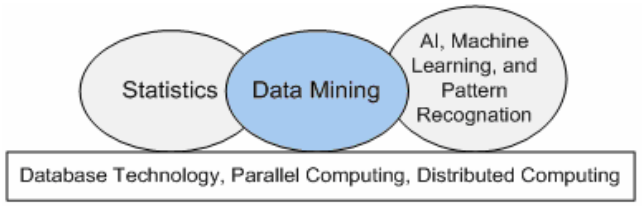


Gambar 1.1 Proses dalam KDD (Tan et al, 2005)

Data input dapat disimpan dalam berbagai format seperti *flat file*, *spreadsheet*, atau tabel-tabel relasional, dan dapat menempati tempat penyimpanan data terpusat atau terdistribusi pada banyak tempat. Tujuan dari preprocessing adalah mentransformasikan data input mentah ke dalam format yang sesuai untuk analisis selanjutnya. Langkah-langkah yang terlibat dalam preprocessing data meliputi mengabungkan data dari berbagai sumber, membersihkan (cleaning) data untuk membuang noise dan observasi duplikat, dan menyeleksi *record* dan fitur yang relevan untuk pekerjaan *data mining*. Karena terdapat banyak cara mengumpulkan dan menyimpan data, tahapan preprocessing data merupakan langkat yang banyak menghabiskan waktu dalam KDD.

Hasil dari *data mining* sering kali diintegrasikan dengan decision support system (DSS). Sebagai contoh, dalam aplikasi bisnis informasi yang dihasilkan oleh *data mining* dapat diintegrasikan dengan tool manajemen kampanye produk sehingga promosi pemasaran yang efektif yang dilaksanakan dan dapat diuji. Integrasi demikian memerlukan langkah postprocessing yang menjamin bahwa hanya hasil yang valid dan berguna yang akan digabungkan dengan DSS. Salah satu pekerjaan dan postprocessing adalah visualisasi yang memungkinkan analyst untuk mengeksplor data dan hasil *data mining* dari berbagai sudur pandang. Ukuran-ukuran statistik dan metode pengujian hipotesis dapat digunakan selama postprocessing untuk membuang hasil *data mining* yang palsu.

Secara khusus, *data mining* menggunakan ide-ide seperti (1) pengambilan contoh, estimasi, dan pengujian hipotesis, dari statistika dan (2) algoritme pencarian, teknik pemodelan, dan teori pembelajaran dari kecerdasan buatan, pengenalan pola, dan machine learning. *Data mining* juga telah mengadopsi ide-ide dari area lain meliputi optimisasi, evolutionary computing, teori informasi, pemrosesan sinyal, visualisasi dan information retrieval. Sejumlah area lain juga memberikan peran pendukung dalam *data mining*, seperti sistem basis data yang dibutuhkan untuk menyediakan tempat penyimpanan yang efisien, indexing dan pemrosesan kueri. Gambar 1.2 menunjukkan hubungan *data mining* dengan area-area lain.



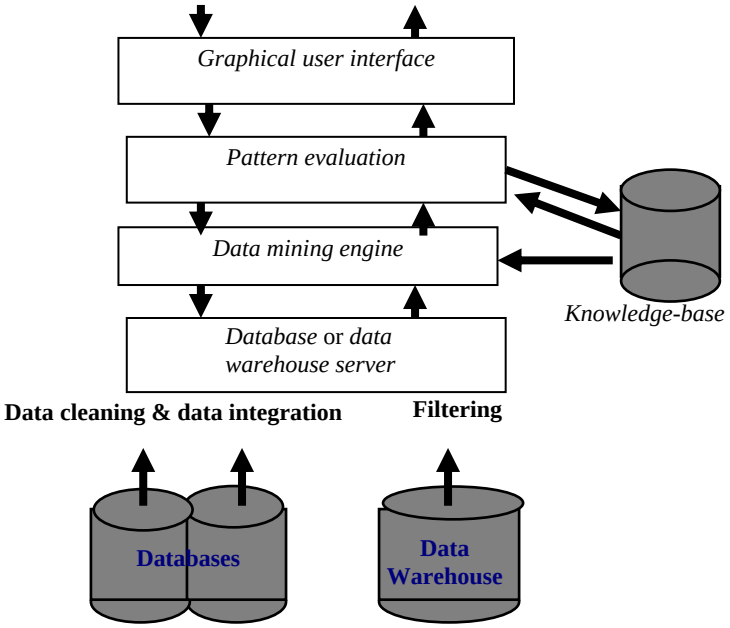
Gambar 1.2 *Data mining* sebagai pertemuan dari banyak disiplin ilmu (Tan et al, 2005)

1.2 Arsitektur Sistem *Data mining*

Data mining merupakan proses pencarian pengetahuan yang menarik dari data berukuran besar yang disimpan dalam basis data, *data warehouse* atau tempat penyimpanan informasi lainnya. Dengan demikian arsitektur sistem *data mining* memiliki komponen-komponen utama yaitu:

- Basis data, *data warehouse* atau tempat penyimpanan informasi lainnya.
- Basis data dan *data warehouse* server. Komponen ini bertanggung jawab dalam pengambilan relevant data, berdasarkan permintaan pengguna.
- Basis pengetahuan. Komponen ini merupakan domain knowledge yang digunakan untuk memandu pencarian atau mengevaluasi pola-pola yang dihasilkan. Pengetahuan tersebut meliputi hirarki konsep yang digunakan untuk mengorganisasikan atribut atau nilai atribut ke dalam level abstraksi yang berbeda. Pengetahuan tersebut juga dapat berupa kepercayaan pengguna (user belief), yang dapat digunakan untuk menentukan kemenarikan pola yang diperoleh. Contoh lain dari domain knowledge adalah threshold dan metadata yang menjelaskan data dari berbagai sumber yang heterogen.
- *Data mining* engine. Bagian ini merupakan komponen penting dalam arsitektur sistem *data mining*. Komponen ini terdiri modul-modul fungsional *data mining* seperti karakterisasi, asosiasi, klasifikasi, dan analisis cluster.
- Modul evaluasi pola. Komponen ini menggunakan ukuran-ukuran kemenarikan dan berinteraksi dengan modul *data mining* dalam pencarian pola-pola menarik. Modul evaluasi pola dapat menggunakan threshold kemenaikan untuk mem-filter pola-pola yang diperoleh.
- Antarmuka pengguna grafis. Modul ini berkomunikasi dengan pengguna dan sistem *data mining*. Melalui modul ini, pengguna berinteraksi dengan sistem mengan menentukan kueri atau task *data mining*. Antarmuka juga menyediakan informasi untuk memfokuskan pencarian dan melakukan eksplorasi *data mining* berdasarkan hasil *data mining* antara. Komponen ini juga memungkinkan pengguna untuk mencari (browse) basis data dan skema *data warehouse* atau struktur data, evaluasi pola yang diperoleh dan visualisasi pola dalam berbagai bentuk.

Arsitektur sebuah sistem *data mining* dapat dilihat dalam Gambar 1.3.



Gambar 1.3 Arsitektur sistem *data mining* (Han dan Kamber, 2001)

Data mining dapat diaplikasikan pada berbagai jenis penyimpanan data seperti basis data relational, *data warehouse*, transactional database, object-oriented and object-relational databases, spatial databases, time-series data and temporal data, text databases and multimedia databases, heterogeneous and legacy databases dan WWW.

a. Basis data Relasional

Basis data relasional merupakan koleksi dari table. Setiap table berisi atribut (field) dan biasanya menyimpan sejumlah besar tuple (*record*). Setiap tuple dalam table relasional merepesentasikan sebuah objek yang diidentifikasikan oleh kunci unik dan dideskripsikan oleh sekumpulan nilai atribut. Data relasional dapat diakses oleh kueri basis data yang ditulis dalam bahasa kueri relasional seperti SQL atau dengan bantuan antarmuka pengguna grafis.

b. *Data warehouse*

Data warehouse merupakan tempat penyimpanan informasi yang dikumpulkan dari berbagai sumber, disimpan dalam skema yang dipersatukan (unified schema) dan biasanya bertempat pada tempat penyimpanan tunggal. *Data warehouse* dikonstruksi melalui sebuah proses data *cleaning*, *data transformation*, *data integration*, *data loading* dan *periodic data refreshing*. Untuk memfasilitasi proses pembuatan keputusan, data dalam *data warehouse* diorganisasikan ke dalam subjek utama seperti customer, item, supplier atau aktivitas. Data disimpan untuk menyediakan informasi dari perspektif sejarah (seperti 5-10 tahun yang lalu) dan biasanya data tersebut diringkas (*summarized*). Sebagai contoh, daripada menyimpan data rinci dari transaksi penjualan, *data warehouse* dapat menyimpan ringkasan dari transaksi per tipe item untuk setiap toko atau diringkas dalam level yang lebih tinggi seperti daerah pemasaran.

Data warehouse biasanya dimodelkan oleh struktur basis data multidimensional, dimana setiap dimensi berkaitan dengan sebuah atribut atau sekumpulan atribut dalam skema, dan setiap sel menyimpan nilai dari ukuran agregasi seperti count dan sales_amount. Struktur fisik dari *data warehouse* dapat berupa penyimpanan basis data relasional atau sebuah kubus data multidimensional.

Selain *data warehouse*, terdapat istilah penyimpanan data yang lain yaitu *data mart*. Sebuah *data warehouse* mengumpulkan informasi mengenai subjek-subjek yang menjangkau seluruh organisasi, dengan demikian cakupannya *enterprise-wide*. Sedangkan *data mart* merupakan sub bagian dari *data warehouse*. Fokus *data mart* adalah pada subjek yang dipilih dan dengan demikian cakupannya adalah *department-wide*.

c. Basis data Transaksional

Secara umum, basis data transaksional terdiri dari sebuah file dimana setiap *record* merepresentasikan transaksi. Sebuah transaksi biasanya meliputi bilangan identitas transaksi yang unik (trans_id), dan sebuah daftar dari item yang membuat transaksi (seperti item yang dibeli dalam sebuah took). Basis data transaksi dapat memiliki tabel tambahan, yang mengandung informasi lain berkaitan dengan penjualan seperti tanggal transaksi, customer ID number, ID number dari sales person dan dari kantor cabang (branch) dimana penjualan terjadi.

1.3 Tugas-tugas dalam *Data mining*

Tugas-tugas dalam *data mining* secara umum dibagi ke dalam dua kategori utama:

- Prediktif. Tujuan dari tugas prediktif adalah untuk memprediksi nilai dari atribut tertentu berdasarkan pada nilai dari atribut-atribut lain. Atribut yang diprediksi umumnya dikenal sebagai target atau variabel tak bebas, sedangkan atribut-atribut yang digunakan untuk membuat prediksi dikenal sebagai *explanatory* atau variabel bebas.
- Deskriptif. Tujuan dari tugas deskriptif adalah untuk menurunkan pola-pola (korelasi, *trend*, *cluster*, trayektori, dan anomali) yang meringkas hubungan yang pokok dalam data. Tugas *data mining* deskriptif sering merupakan penyelidikan dan seringkali memerlukan teknik postprocessing untuk validasi dan penjelasan hasil.

Berikut adalah tugas-tugas dalam *data mining*:

- Analisis Asosiasi (Korelasi dan kausalitas)

  Analisis asosiasi adalah pencarian aturan-aturan asosiasi yang menunjukkan kondisi-kondisi nilai atribut yang sering terjadi bersama-sama dalam sekumpulan data. Analisis asosiasi sering digunakan untuk menganalisa *market basket* dan data transaksi.

  Aturan-aturan asosiasi memiliki bentuk $X \Rightarrow Y$, bahwa $A_1 \wedge A_2 \wedge \ldots \wedge A_m \rightarrow B_1 \wedge B_2 \wedge \ldots \wedge B_n$, dimana A_i (untuk i = 1, 2, …, m) dan B_j (untuk j = 1, 2, …,

n) adalah pasangan-pasangan nilai atribut. Aturan asosiasi X $\Rightarrow$ Y diinterpretasikan sebagai tuple-tuple basis data yang memenuhi kondisi-kondisi dalam X juga mungkin memenuhi kondisi dalam Y.

Contoh dari aturan asosiasi adalah

- age(X, "20..29") ^ income(X, "20..29K") $\Rightarrow$ buys(X, "PC") [support = 2%, confidence = 60%]
- contains(T, "computer") $\Rightarrow$ contains(x, "software") [1%, 75%]

- Klasifikasi dan Prediksi

  Klasifikasi adalah proses menemukan model (fungsi) yang menjelaskan dan membedakan kelas-kelas atau konsep, dengan tujuan agar model yang diperoleh dapat digunakan untuk memprediksikan kelas atau objek yang memiliki label kelas tidak diketahui. Model yang turunkan didasarkan pada analisis dari training data (yaitu objek data yang memiliki label kelas yang diketahui). Model yang diturunkan dapat direpresentasikan dalam berbagai bentuk seperti aturan IF-THEN klasifikasi, pohon keputusan, formula matematika atau jaringan syarf tiruan.

  Dalam banyak kasus, pengguna ingin memprediksikan nilai-nilai data yang tidak tersedia atau hilang (bukan label dari kelas). Dalam kasus ini biasanya nilai data yang akan diprediksi merupakan data numeric. Kasus ini seringkali dirujuk sebagai prediksi. Di samping itu, prediksi lebih menekankan pada identifikasi trend dari distribusi berdasarkan pada data yang tersedia.

- Analisis *Cluster*

  Tidak seperti klasifikasi dan prediksi, yang menganalisis objek data yang diberi label kelas, *clustering* menganalisis objek data dimana label kelas tidak diketahui. *Clustering* dapat digunakan untuk menentukan label kelas tidak diketahui dengan cara mengelompokkan data untuk membentuk kelas baru. Sebabai contoh *clustering* rumah untuk menemukan pola distribusinya. Prinsip dalam *clustering* adalah memaksimumkan kemiripan *intra-class* dan meminimumkan kemiripan *interclass.*

- Analisis *Outlier*

  Outlier merupakan objek data yang tidak mengikuti perilaku umum dari data. Outlier dapat dianggap sebagai noise atau pengecualian. Analisis data outlier dinamakan *outlier mining*. Teknik ini berguna dalam *fraud detection* dan *rare events analysis*.

- Analisis Trend dan Evolusi

  Analisis evolusi data menjelaskan dan memodelkan trend dari objek yang memiliki perilaku yang berubah setiap waktu. Teknik ini dapat meliputi karakterisasi, diskriminasi, asosiasi, klasifikasi, atau *clustering* dari data yang berkaitan dengan waktu.

Data mining merupakan bidang interdisplin. Disiplin ilmu ini banyak dipengaruhi oleh disiplin sistem basis data, statistika, ilmu informasi, mesin

pembelajaran, dan visualisasi. Sistem *data mining* dapat diklasifsikasikan berdasarkan beberapa kategori, yaitu

- Klasifikasi berdasarkan data yang akan di-*mine* seperti *relational*, *transactional*, *object-oriented*, *object-relational*, *spatial*, *time-series*, *text*, multi-media dan www.
- Klasifikasi berdasarkan pengetahuan yang akan di-*mine*, yaitu berdasarkan fungsionalitas *data mining* seperti karakterisasi, diskriminasi, asosiasi, klasifikasi, *clustering*, analisis *outlier* dan analisis evolusi. Sistem *data mining* yang komprehensif biasanya menyediakan beberapa fungsi-fungsi *data mining*.
- Klasifikasi berdasarkan teknik yang akan digunakan seperti *database-oriented*, *data warehouse* (OLAP), *machine learning*, *Statistics*, *Visualization* dan neural network.
- Klasifikasi berdasarkan aplikasi yang diadaptasi, sebagai contoh system *data mining* untuk keuangan, telekomunikasi, DNA, dan e-mail.

Penutup – Soal Latihan

Tugas Individu

Jawablah pertanyaan berikut secara singkat dan jelas. Carilah literatur pendukung untuk memperkaya jawaban anda.

1. Apakah data mining itu?
2. Sebutkan dan jelaskan secara singkat area-area yang berhubungan dengan data mining
3. Apa yang dimaksud dengan *descriptive data mining* dan *predictive data mining*?
4. Jelaskan secara singkat apa itu teknik asosiasi, klasifikasi, prediksi dan *clustering*. Berikan contoh pengunaan teknik-teknik tersebut menggunakan basis data yang telah anda kenal sehari-hari.

Tugas Kelompok

Diskusikan dengan kelompok anda jawaban untuk pertanyaan-pernyataan berikut. Carilah literatur pendukung untuk memperkaya jawaban anda.

1. Berikan contoh aplikasi dari *data mining* dalam berbagai bidang.
2. Tentukan apakah aktivitas-aktivitas berikut adalah tugas dalam *data mining*?
    a. Membagi pelanggan sebuah perusahaan berdasarkan jenis kelamin
    b. Membagi pelanggan sebuah perusahaan berdasarkan profitabilitas-nya
    c. Menghitung total penjualan dari sebuah perusahaan
    d. Mengurutkan basis data mahasiswa berdasarkan NRP mahasiswa
    e. Memprediksi keluaran dari hasil pelemparan sepasang dadu

f. Memprediksi harga *stock* mendatang dari sebuah perusahaan berdasarkan *record* historis

g. Memonitor kecepatan jantung dari seorang pasien

h. Memonior gelombang yang berkaitan dengan gempa bumi untuk aktivitas gempa bumi

i. Mengekstrak frekuensi dari gelombang suara.