Bab 1

Konsep Data Mining

POKOK BAHASAN:

- ✓ Konsep dasar dan pengertian Data Mining
- ✓ Tahapan dalam Data Mining
- ✓ Model Data Mining
- ✓ Fungsi Data Mining

TUJUAN BELAJAR:

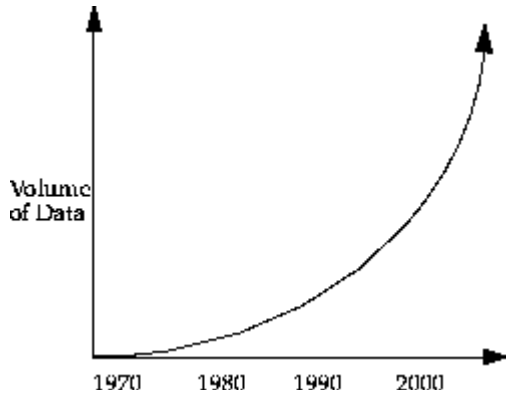
Setelah mempelajari materi dalam bab ini, mahasiswa diharapkan :

- ✓ Memahami konsep dasar dan definisi dari Data Mining
- ✓ Memahami setiap tahapan dalam Data Mining
- ✓ Memahami tentang model dan fungsi dalam Data Mining

1.1. PENDAHULUAN

Selama dua decade terakhir telah terjadi peningkatan yang dramatis terhadap jumlah informasi atau data yang disimpan secara elektronis. Telah diperkirakan sebelumnya bahwa jumlah informasi di dunia akan berlipat ganda setiap 20 bulan dan jumlah ukuran basis data akan bertambah lebih cepat lagi dari itu.

Teknologi database saat ini memungkinkan untuk menyimpan sejumlah data dalam jumlah yang sangat besar dan terakumulasi. Disinilah awal timbulnya persoalan ledakan data (jumlah data yang tiba-tiba begitu sangat besar). Data perlu disimpan, tapi yang lebih penting dari itu adalah proses penemuan pengetahuan (knowledge) dari data yang disimpan ! Oleh karenanya data yang tersimpan dalam sebuah gudang data yang disebut dengan data warehouses perlu dianalisa.



Gambar 1.1 Perkembangan Database

Permasalahannya kemudian adalah apa yang harus dilakukan dengan data-data itu. Sudah diketahui secara umum bahwa informasi merupakan hal yang sangat penting dalam menunjang operasi-operasi bisnis dan membantu para pengambil keputusan untuk mendapatkan gambaran lebih tentang bisnis mereka. Sistem manajemen basis data memberikan akses terhadap data namun hanya sebagian kecil kontribusinya terhadap apa yang seharusnya dapat dihasilkan dari data-data itu. Sistem pemrosesan transaksi online (OLPT) tradisional sangat baik dalam menyimpan data secara cepat, aman, dan efisien ke dalam basis data namun tidak cukup baik dalam hal kemampuan melakukan analisa terhadap data-data yang ada, padahal analisa terhadap data memberikan pengetahuan lebih mendalam tentang bisnis yang dilakukan. Solusi untuk persoalan penemuan pengetahuan dalam database berukuran besar adalah dengan menggunakan Data warehousing dan data mining. Disinilah peran *Data Mining* akan memberikan konstribusi besar bagi setiap perusahaan yang mengimplementasikannya.

1.2. PENGERTIAN DATA MINING

Ada beberapa definisi dari Data mining. Secara umum data mining dapat didefinisikan sebagai berikut :

- Proses penemuan pola yang menarik dari data yang tersimpan dalam jumlah besar. Merupakan evolusi alami dari teknologi database, dan merupakan metode yang paling banyak dibutuhkan, dengan aplikasi yang sangat luas.
- Ekstraksi dari suatu informasi yang berguna atau menarik (non-trivial, implisit, sebelumnya belum diketahui, potensial kegunaannya) pola atau pengetahuan dari data yang disimpan dalam jumlah besar.
- Ekplorasi dari analisa secara otomatis atau semiotomatis terhadap data-data dalam jumlah besar untuk mencari pola dan aturan yang berarti.

Pada dasarnya data mining berhubungan erat dengan analisa data dan penggunaan perangkat lunak untuk mencari pola dan kesamaan dalam sekumpulan data. Ide dasarnya sangat menggali sumber yang berharga dari tempat yang sama sekali tidak diduga seperti perangkat lunak data mining mengekstrasi pola yang sebelumnya tidak terlihat atau tigak begitu jelas sehingga tidak seorang pun yang memperhatikan sebeumnya.

Analisa data mining berjalan pada data yang cenderung terus membesar dan teknik terbaik yang digunakan kemudian beorientasi kepada data beukuran sangat besar untuk mendapatkn kesimpulan dan keputusan paling layak. Data mining memiliki beberapa sebutan atau nama lain yaitu : Knowledge discovery (mining) in databases (KDD), ekstraksi pengetahuan (knowledge extraction), Analisa data/pola, kecerdasan bisnis (business intelligence), dll.

Meskipun sebagian besar *teknik data mining* seperti yang akan dijelaskan pada bagian- lain laporan tugas akhir ini sudah ada sejak lama, namun hanya pada beberapa tahun terakhir ini *data mining* benar-benar berperan yaitu sejak dilakukan komersialisasi *data mining*.

Beberapa faktor yang mendukung perlunya dilakukan data mining adalah :

1. *Data telah mencapai jumlah dan ukuran yang sangat besar*

   Hasil dan proses *data mining* merupakan suatu informasi yang akan mendasari tindakan tertentu sehingga tingkat kebenaran informasi tersebut menjadi sangat signifikan, dan makin besar serta makin banyak data yang digunakan maka akan semakin valid hasilnya. Perkembangan data dalam hal jumlah dan ukuran telah mencapai kecepatan yang sangat cepat, sehingga ukuran basis data yang dimiliki oleh sebuah perusahaan bisa mencapai kisaran gigabyte atau bahkan terabyte.

2. *Telah dilakukan proses data warehousing*

   Untuk mencapai hasil yang memuaskan, maka sumber data yang digunakan dalam proses *data mining* seringkali merupakan data gabungan dari banyak departemen, daerah operasi bahkan dari sumber-sumber lain seperti data kependudukan. Oleh karena itu maka disarankan perlunya proses data warehousing untuk menjaga konsistensi, memberikan prespektif yang lebih baik terhadap data dan menjaga integritas data.

3. *Kemampuan Komputasi yang semakin terjangkau*

   Pada dasarnya proses data mining melakukan banyak akses terhadap data yang sangat besar. Selain itu juga melakukan proses komputasi yang membutuhkan sumber daya sangat besar. Penurunan harga yang cukup cepat terhadap perangkat keras computer serta semakin tingginya kinerja yang berhasil dicapai oleh perangkat computer maupun teknologi pengolahan data seperti teknologi pararel proses saat ini, menjadikan proses saat ini, menjadikan proses *data mining* suduh cukup layak untuk dilakukan secara komersial.

4. *Persaingan bisnis yang semakin ketat*

   Tekanan persaingan bisnis yang semakin ketat mendorong perusahaan-perusahaan untuk selalu berinovasi agar mampu meningkatkan daya saingnya dipasar global. Beberapa tren yang berkembang saat ini adalah

   a. Setiap bisnis adalah bisnis pelayanan
   b. Adanya fenomena kustomisasi produk oleh masyarakat
   c. Informasi adalah produk

1.3. MODEL DALAM DATA MINING

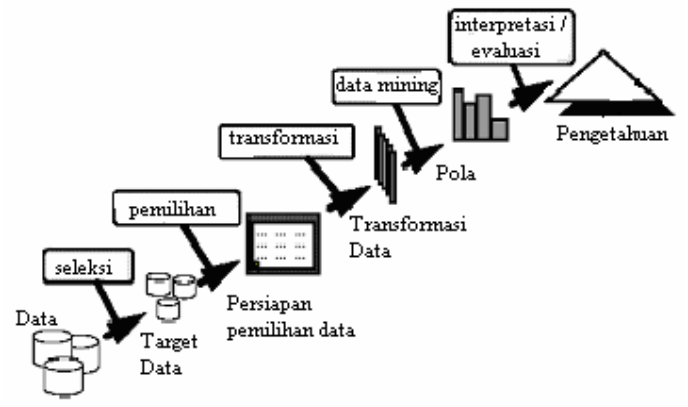
Terdapat dua tipe atau mode operasi yang bisa digunakan untuk mencari informasi yang dibutuhkan user lewat proses data mining, yaitu model verifikasi dan knowledge discovery.

Model verifikasi menggunakan pendekatan top down dengan mengambil hipotesa dari user dan memeriksa validitasnya dengan data sehingga bisa dibuktikan kebenaran hipotesa tersebut.

Model Knowledge Discovery menggunakan pendekatan bottom up untuk mendapatkan informasi yang sebelumnya tidak diketahui. Model ini terbagi menjadi dua *directed knowledge discovery dan undirected knowledge discovery*. Pada *directed knowledge discovery* data mining akan mencoba mencari penjelasan nilai target field tertentu (seperti penghasilan, respons, usia, dan lain-lain) terhadap field-field yang lain. Sedangkan pada *undirected knowledge discovery* tidak ada target field karena komputer akan mencari pola yang ada pada data. Jadi *undirected knowledge discovery* digunakan untuk mengenali hubungan / relasi yang ada pada data sedangkan *directed knowledge discovery* akan menjelaskan hubungan / relasi tersebut.

1.3. TAHAPAN PROSES DALAM DATA MINING

Ada beberapa tahapan proses dalam data mining. Diagram dibawah menggambarkan beberapa tahap / proses yang berlangsung dalam data mining. Fase awal dimulai dari data sumber dan berakhir dengan adanya informasi yang dihasilkan dari bebrapa tahapan, yaitu :



Gambar 1.2 Fase-fase Dalam Data Mining

Tahapan proses dalam Data Mining dapat dijelaskan sebagai berikut :

1. ***Seleksi Data***

   Pemilihan (seleksi) data daru sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses *data mining*, disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. ***Pre-processing/ Cleaning (pemilihan data)***

   Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD.

   Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi).

   Juga dilakukan proses *enrichment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

3. ***Transformasi***

   Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses *coding* dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data

4. ***Data mining***

   Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. ***Interpretasi / Evaluasi***

   Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut dengan *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

Proses KDD secara garis besar memang terdiri dari 5 tahap seperti yang telah dijelaskan sebelumnya. Akan tetapi, dalam proses KDD yang sesungguhnya, dapat saja terjadi iterasi atau pengulangan pada tahap tahap tertentu. Pada setiap tahap dalam proses KDD, seorang analis dapat saja kembali ke tahap sebelumnya. Sebagai contoh, pada saat *coding* atau *data mining*, analis menyadari proses *cleaning* belum dilakukan dengan sempurna, atau mungkin saja analis menemukan data atau informasi baru untuk "memperkaya" data yang sudah ada.

1.4. METODE DATA MINING

Data mining model dibuat berdasarkan salah satu dari dua jenis pembelajaran *supervised* dan *unsupervised*. Fungsi pembelajaran Supervised digunakan untuk memprediksi suatu nilai. Fungsi pembelajaran Unsupervised digunakan untuk mencari struktur intrinsik, relasi dalam suatu data yang tidak memerlukan class atau label sebelum dilakukan proses pembelajaran. Contoh dari algoritma pembelajaran unsupervised, diantaranya *k*-means clustering dan Apriori association rules. Contoh dari algoritma pembelajaran supervised yaitu NaiveBayes untuk klasifikasi.

Metode data mining dapat diklasifikasikan berdasarkan fungsi yang dilakukan atau berdasarkan jenis aplikasi yang menggunakannya:

- Klasifikasi (supervised)
- Clustering (unsupervised)
- Association Rules (unsupervised)
- Attribute Importance (supervised)

1.4.1. KLASIFIKASI (SUPERVISED)

Pada persoalan klasifikasi, kita memiliki sejumlah kasus (sampel data) dan ingin memprediksi beberapa class yang ada pada sampel data tersebut. Tiap instan data berisi banyak atribut, dimana masing-masing atribut memiliki satu dari beberapa kemungkinan nilai. Hanya satu atribut diantara banyak atribut tersebut yang disebut dengan atribut target, sedangkan atribut yang lain disebut sebagai atribut prediktor. Tiap kemungkinan nilai yang dimiliki oleh atribut target menunjukkan class yang diprediksi berdasarkan nilai-nilai dari atribut prediktor.

Klasifikasi digunakan untuk segmentasi customer, pemodelan bisnis, analisa kartu kredit, dan banyak aplikasi yang lain. Sebagai contoh, perusahaan kartu kredit ingin memprediksi customer berdasarkan tipe pembayaran.

1.4.2. CLUSTERING (UNSUPERVISED)

Clustering adalah teknik yang berguna untuk mengeksplorasi data. Digunakan pada saat banyak kasus dan tidak memiliki pengelompokan secara alami. Dalam hal ini algoritma data mining dapat digunakan untuk mencari pengelompokan yang ada pada data..

Analisa Clustering mengidentifikasi cluster yang ada pada data. Cluster adalah kumpulan obyek data yang mirip satu sama lain. Metode clustering yang bagus menghasilkan cluster yang berkualitas untuk memastikan kesamaan pada datadata yang ada dalam satu cluster. Clustering model berbeda dari model prediktif dikarenakan pada clustering tidak perlu ada atribut target. Clustering yang diorganisasi ke dalam struktur hirarkikal akan mendefinisikan taksonomi dari data.

Dalam ODM, suatu cluster dikarakterisasi oleh *centroid*, attribute histograms, dan clustering model hierarchical tree. ODM membentuk hierarchical clustering dengan menggunakan versi perbaikan dari algoritma *k*-means dan O-Cluster.

1.4.3. ASSOCIATION RULES (UNSUPERVISED)

Fungsi Association Rules seringkali disebut dengan "market basket analysis", yang digunakan untuk menemukan relasi atau korelasi diantara himpunan item2. Fungsi ini paling banyak digunakan untuk menganalisa data dalam rangka keperluan strategi pemasaran, desain katalog, dan proses pembuatan keputusan bisnis. Tipe association rule bisa dinyatakan sebagai misal : "70% dari orangorang yang membeli mie, juice dan saus akan membeli juga roti tawar".

Aturan asosiasi mengcapture item atau kejadian dalam data berukuran besar yang berisi data transaksi. Dengan kemajuan teknologi, data penjualan dapat disimpan dalam jumlah besar yang disebut dengan "basket data." Aturan asosiasi yang

didefinisikan pada basket data, digunakan untuk keperluan promosi, desain katalog, segmentasi customer dan target pemasaran.

1.4.4. ATTRIBUTE IMPORTANCE (SUPERVISED)

Attribute Importance, disebut juga dengan *feature selection*, menyediakan solusi otomatis untuk meningkatkan kecepatan dan akurasi dari model klasifikasi yang dibangun pada table data yang memiliki jumlah atribut yang sangat banyak.. Attribute Importance meranking atribut prediktif dengan melakukan eleminasi nilai yang redundant, tidak relevant atau tidak informative dan mengidentifikasi atribut predictor yang banyak paling berpengaruh dalam pengambilan keputusan.

Dengan menggunakan atribut yang lebih sedikit akan mereduksi waktu untuk membangun suatu model, juga dapat meningkatkan akurasi dari kemampuan prediksi. Jika terlalu banyak atribut yang dilibatkan maka akan banyak pula noise yang terlibat yang akan berpengaruh terhadap model karena dapat menurunkan performansi dan akurasi.

LATIHAN SOAL :

1) Mengapa perlu dilakukan Data Mining ?
2) Jelaskan definisi dari Data Mining !
3) Jelaskan setiap tahapan yang ada dalam Data Mining
4) Jelaskan tipe mode operasi yang digunakan oleh user untuk mencari informasi dalam Data Mining !
5) Pada dasarnya Data Mining menggunakan dua jenis pembelajaran, sebutkan !