KONSEP DATA MINNING

Ari Fadli
fadli.te.unsoed@gmail
http://fadli84.wordpress.com

Pada tulisan kali ini penulis akan sedikit berbagi ilmu mengenai Konsep Dasar dari Data Mining atau biasa diterjemahkan kedalam bahasa Indonesia sebagai Menggali Informasi yang Terpendam

Pendahuluan

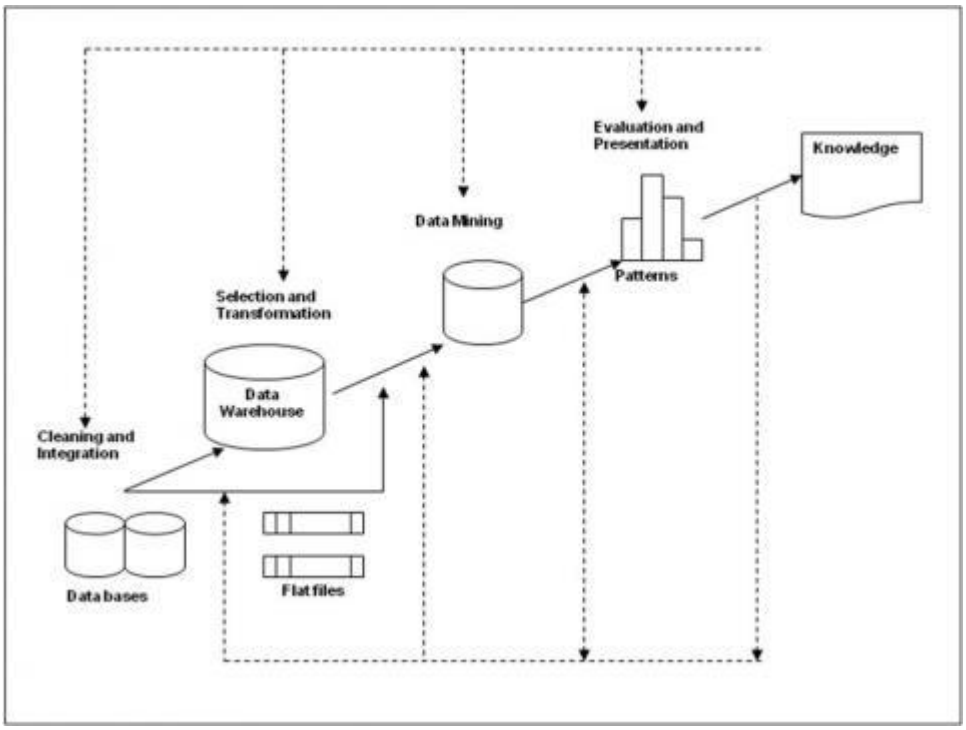
Sebagai cabang ilmu baru di bidang komputer (lihat artikel sebelumnya berjudul ‘Data Mining’) cukup banyak penerapan yang dapat dilakukann oleh Data Mining. Apalagi ditunjang ke-kaya-an dan ke-anekaragam-an berbagai bidang ilmu (artificial intelligence, database, statistik, pemodelan matematika, pengolahan citra dsb.) membuat penerapan data mining menjadi makin luas.

Konsep Data Mining

Apa sebenarnya yang memotivasi datamining dan mengapa data mining begitu penting ?

> *Alasan utama mengapa data mining sangat menarik perhatian industri informasi dalam beberapa tahun belakangan ini adalah karena tersedianya data dalam jumlah yang besar dan semakin besarnya kebutuhan untuk mengubah data tersebut menjadi informasi dan pengetahuan yang berguna.*

Data mining adalah kegiatan mengekstraksi atau menambang pengetahuan dari data yang berukuran/berjumlah besar, informasi inilah yang nantinya sangat berguna untuk pengembangan. Dimana langkah-langkah untuk melakukan data mining adalah sebagai berikut :



gambar 1 langkah-langkah untuk melakukan data mining

1. Data cleaning (untuk menghilangkan noise data yang tidak konsisten) Data integration (di mana sumber data yang terpecah dapat disatukan)
2. Data selection (di mana data yang relevan dengan tugas analisis dikembalikan ke dalam database)
3. Data transformation (di mana data berubah atau bersatu menjadi bentuk yang tepat untuk menambang dengan ringkasan performa atau operasi agresi)
4. Data mining (proses esensial di mana metode yang intelejen digunakan untuk mengekstrak pola data)
5. Pattern evolution (untuk mengidentifikasi pola yang benar-benar menarik yang mewakili pengetahuan berdasarkan atas beberapa tindakan yang menarik)
6. Knowledge presentation (di mana gambaran teknik visualisasi dan pengetahuan digunakan untuk memberikan pengetahuan yang telah ditambang kpada user).

Arsitektur data mining

1. Database, data warehouse, atau tempat penyimpanan informasi lainnya.
2. Server database atau data warehouse.
3. Knowledge base
4. Data mining engine.
5. Pattern evolution module.
6. Graphical user interface.

Jenis data dalam data mining

1. Relation Database : Sebuah sistem database, atau disebut juga database management system (DBMS), mengandung sekumpulan data yang saling berhubungan, dikenal sebagai sebuah database, dan satu set program perangkat lunak untuk mengatur dan mengakses data tersebut.
2. Data Warehouse : Sebuah data warehouse merupakan sebuah ruang penyimpaan informasi yang terkumpul dari beraneka macam sumber, disimpan dalam skema yang menyatu, dan biasanya terletak pada sebuah site.

Kemudian pola seperti apa yang dapat ditambang ?

Kegunaan data mining adalah untuk menspesifikasikan pola yang harus ditemukan dalam tugas data mining. Secara umum tugas data mining dapat diklasifikasikan ke dalam dua kategori: deskriptif dan prediktif. Tugas menambang secara deskriptif adalah untuk mengklasifikasikan sifat umum suatu data di dalam database. Tugas data mining secara prediktif adalah untuk mengambil kesimpulan terhadap data terakhir untuk membuat prediksi.

1. **Konsep/Class Description**

   Data dapat diasosiasikan dengan pembagian class atau konsep. Untuk contohnya, ditoko All Electronics, pembagian class untuk barang yang akan dijual termasuk komputer dan printer, dan konsep untuk konsumen adalah big Spenders dan budget Spender. Hal tersebut sangat berguna untuk menggambarkan pembagian class secara individual dan konsep secara ringkas, laporan ringkas, dan juga pengaturan harga. Deskripsi suatu class atau konsep seperti itu disebut class/concept descripition.

2. **Association Analysis**

   Association analysis adalah penemuan association rules yang menunjukkan nilai kondisi suatu attribute yang terjadi bersama-sama secara terus-menerus dalam memmberikan set data. Association analysis secara luas dipakai untuk market basket atau analisa data transaksi.

3. **Klasifikasi dan Predikasi**

   Klasifikasi dan prediksi mungkin perlu diproses oleh analisis relevan, yang berusaha untuk mengidentifikasi atribut-atribut yang tidak ditambahkan pada proses klasifikasi dan prediksi. Atribut-atribut ini kemudian dapat di keluarkan.

4. **Cluster Analysis**

   Tidak seperti klasifikasi dan prediksi, yang menganalisis objek data dengan kelas yang terlabeli, clustering menganalisis objek data tanpa mencari keterangan pada label kelas yang diketahui. Pada umumnya, label kelas tidak ditampilkan di dalam latihan data simply, karena mereka tidak tahu bagaimana memulainya. Clustering dapat digunakan untuk menghasilkan label-label.

5. **Outlier Analysis**

   § Outlier dapat dideteksi menggunakan test yang bersifat statistik yang mengambil sebuah distribusi atau probabilitas model untuk data, atau

menggunakan langkah-langkah jarak jauh di mana objek yang penting jauh dari cluster lainnya dianggap outlier.

§ Sebuah database mungkin mengandung objek data yang tidak mengikuti tingkah laku yang umum atau model dari data. data ini disebut outlier.

6. Evolution Analysis

   Data analisa evolusi menggambarkan ketetapan model atau kecenderungan objek yang memiliki kebiasaan berubah setiap waktu. Meskipun ini mungkin termasuk karakteristik, diskriminasi, asosiasi, klasifikasi, atau clustering data berdasarkan waktu, kelebihan yang jelas seperti analisa termasuk analisa data time-series, urutan atau pencocockkan pola secara berkala, dan kesamaan berdasarkan analisa data.

   Untuk melakukan data mining yang baik ada beberapa persoalan utama yaitu menyangkut metodologi mining dan interaksi user, performance dan perbedaan tipe database. Hal inilah yang sering kali dihadapi disaat kita ingin melakukan data mining.

Data Mining

Data Mining memang salah satu cabang ilmu komputer yang relatif baru. Dan sampai sekarang orang masih memperdebatkan untuk menempatkan *data mining* di bidang ilmu mana, karena *data mining* menyangkut *database*, kecerdasan buatan (*artificial intelligence*), statistik, dsb. Ada pihak yang berpendapat bahwa *data mining* tidak lebih dari *machine learning* atau analisa statistik yang berjalan di atas *database*. Namun pihak lain berpendapat bahwa *database* berperanan penting di *data mining* karena *data mining* mengakses data yang ukurannya besar (bisa sampai terabyte) dan disini terlihat peran penting *database* terutama dalam optimisasi *query*-nya.

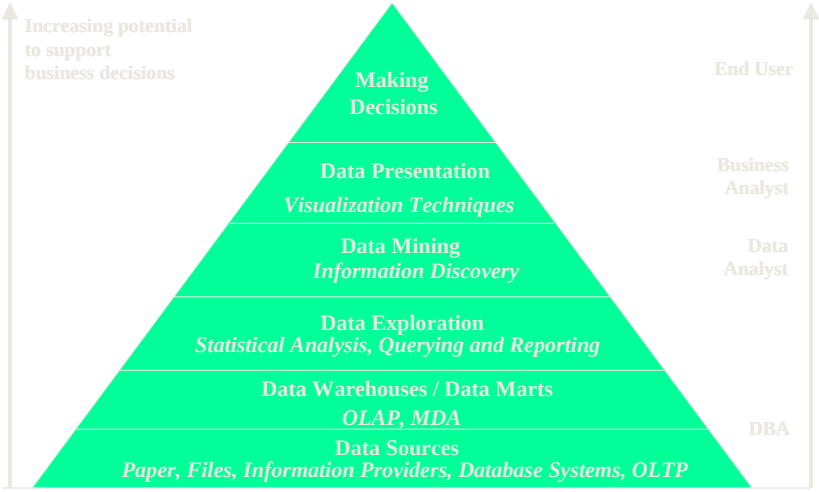
Lalu apakah *data mining* itu? Apakah memang berhubungan erat dengan dunia pertambangan…. tambang emas, tambang timah, dsb. Definisi sederhana dari *data mining* adalah ekstraksi informasi atau pola yang penting atau menarik dari data yang ada di *database* yang besar. Dalam jurnal ilmiah, *data mining* juga dikenal dengan nama *Knowledge Discovery in Databases* (KDD).

Kehadiran *data mining* dilatar belakangi dengan problema *data explosion* yang dialami akhir-akhir ini dimana banyak organisasi telah mengumpulkan data sekian tahun lamanya (data pembelian, data penjualan, data nasabah, data transaksi dsb.). Hampir semua data tersebut dimasukkan dengan menggunakan aplikasi komputer yang digunakan untuk menangani transaksi sehari-hari yang kebanyakan adalah OLTP (*On Line Transaction Processing*). Bayangkan berapa transaksi yang dimasukkan oleh *hypermarket* semacam Carrefour atau transaksi kartu kredit dari sebuah bank dalam seharinya dan bayangkan betapa besarnya ukuran data mereka jika nanti telah berjalan beberapa tahun. Pertanyaannya sekarang, apakah data tersebut akan dibiarkan menggunung, tidak berguna lalu dibuang, ataukah kita dapat me-'nambang'-nya untuk mencari 'emas', 'berlian' yaitu informasi yang berguna untuk organisasi kita. Banyak diantara kita yang kebanjiran data tapi miskin informasi.

Jika Anda mempunyai kartu kredit, sudah pasti Anda bakal sering menerima surat berisi brosur penawaran barang atau jasa. Jika Bank pemberi kartu kredit Anda mempunyai 1.000.000 nasabah, dan mengirimkan sebuah (hanya satu) penawaran dengan biaya pengiriman sebesar Rp. 1.000 per buah maka biaya yang dihabiskan adalah Rp. 1 Milyar!! Jika Bank tersebut mengirimkan penawaran sekali sebulan yang berarti 12x dalam setahun maka anggaran yang dikeluarkan per tahunnya adalah Rp. 12 Milyar!! Dari dana Rp. 12 Milyar yang dikeluarkan, berapa persenkah konsumen yang benar-benar membeli? Mungkin hanya 10 %-nya saja. Secara harfiah, berarti 90% dari dana tersebut terbuang sia-sia.

Persoalan di atas merupakan salah satu persoalan yang dapat diatasi oleh *data mining* dari sekian banyak potensi permasalahan yang ada. *Data mining* dapat menambang data transaksi belanja kartu kredit untuk melihat manakah pembeli-pembeli yang memang potensial untuk membeli produk tertentu. Mungkin tidak sampai presisi 10%, tapi bayangkan jika kita dapat menyaring 20% saja, tentunya 80% dana dapat digunakan untuk hal lainnya.

Lalu apa beda *data mining* dengan *data warehouse* dan OLAP (*On-line Analytical Processing*)? Secara singkat bisa dijawab bahwa teknologi yang ada di *data warehouse* dan OLAP dimanfaatkan penuh untuk melakukan *data mining*. Gambar di bawah menunjukkan posisi masing-masing teknologi:



Gambar 1: *Data mining* dan teknologi *database* lainnya

Dari gambar di atas terlihat bahwa teknologi *data warehouse* digunakan untuk melakukan OLAP, sedangkan *data mining* digunakan untuk melakukan *information discovery* yang informasinya lebih ditujukan untuk seorang *Data Analyst* dan *Business Analyst* (dengan ditambah visualisasi tentunya). Dalam prakteknya, *data mining* juga mengambil data dari *data warehouse*. Hanya saja aplikasi dari *data mining* lebih khusus dan lebih spesifik dibandingkan OLAP mengingat *database* bukan satu-satunya bidang ilmu yang mempengaruhi *data mining*, banyak lagi bidang ilmu yang turut memperkaya *data mining* seperti: *information science* (ilmu informasi), *high performance computing*, visualisasi, *machine learning*, statistik, *neural networks* (jaringan syaraf tiruan), pemodelan matematika, *information retrieval* dan *information extraction* serta pengenalan pola. Bahkan pengolahan citra (*image processing*) juga digunakan dalam rangka melakukan *data mining* terhadap data *image*/*spatial*.

Dengan memadukan teknologi OLAP dengan *data mining* diharapkan pengguna dapat melakukan hal-hal yang biasa dilakukan di OLAP seperti *drilling/rolling* untuk melihat data lebih dalam atau lebih umum, *pivoting*, *slicing* dan *dicing*. Semua hal tersebut diharapkan nantinya dapat dilakukan secara interaktif dan dilengkapi dengan visualisasi.

Data mining tidak hanya melakukan *mining* terhadap data transaksi saja. Penelitian di bidang *data mining* saat ini sudah merambah ke sistem *database* lanjut seperti *object oriented database*, *image/spatial database*, *time-series*

data/temporal database, teks (dikenal dengan nama *text mining*), web (dikenal dengan nama *web mining*) dan *multimedia database*.

Meskipun gaungnya mungkin tidak seramai seperti ketika *Client/Server Database* muncul, tetapi industri-industri seperti IBM, Microsoft, SAS, SGI, dan SPSS terus gencar melakukan penelitian-penelitian di bidang *data mining* dan telah menghasilkan berbagai *software* untuk melakukan *data mining*:

- Intelligent Miner dari IBM. Berjalan di atas sistem operasi AIX, OS/390, OS/400, Solaris dan Windows NT. Dijual dengan harga sekitar US$60.000. Selain untuk data IBM juga mengeluarkan produk Intelligent Miner untuk teks. *Web site*:
  www.software.ibm.com/data/iminer/fortext

  www-4.inm.com/software/data/iminer/fordata/index.html

- Microsoft juga telah menambahkan fasilitas *data mining* di Microsoft SQL Server 2000. *Web site*: www.microsoft.com/sql/productinfo/feaover.htm
- Enterprise Miner dari SAS. Berjalan di atas sistem operasi AIX/6000, CMS, Compaq Tru64 UNIX, HP-UX, IRIX, Intel ABI, MVS, OS/2, Open VMS Alpha, Open VMS Vax, Solaris, dan Windows. *Web site*: www.sas.com
- MineSet dari Silicon Graphics. Berjalan di atas sistem operasi Windows 9x/NT dan IRIX. Dijual per *seat* seharga US$995, server (Windows NT) seharga US$35.000 dan untuk IRIX dijual US$50.000. *Web site*: www.sgi.com/software/mineset
- Clementine dari SPSS. Berjalan di atas sistem operasi UNIX dan Windows NT. Web site: www.spss.com/software/clementine

Beberapa penelitian sekarang ini sedang dilakukan untuk memajukan data mining diantaranya adalah peningkatan kinerja jika berurusan dengan data berukuran *terabyte*, visualisasi yang lebih menarik untuk *user*, pengembangan bahasa *query* untuk *data mining* yang sedapat mungkin mirip dengan SQL. Tujuannya tidak lain adalah agar *end-user* dapat melakukan *data mining* dengan mudah dan cepat serta mendapatkan hasil yang akurat.

Biografi Penulis

	**Ari Fadli**, Lahir di Cirebon, 31 Juli 1984. Menamatkan SMU di SMU Negeri 4 Cirebon. Menyelesaikan program S1 dari Jurusan Teknik Program Studi Teknik Elektro, Universitas Jenderal Soedirman Puwokerto pada tahun 2007. Saat ini menjadi dosen di Jurusan Teknik Program Studi Teknik Elektro, Universitas Jenderal Soedirman Puwokerto dan sedang menyelesaikan studi di pascasarjana universitas Gadjah Mada Jurusan Teknik Elektro dengan Spesifikasi Sistem Komputer dan Informasi . Kompetensi awalnya adalah bidang basis data, Sistem informasi, sistem pakar dan saat ini sedang bergerak ke arah open source