Pengantar Data Mining: Menambang Permata Pengetahuan di Gunung Data

Iko Pramudiono
iko@tkl.iis.u-tokyo.ac.jp

Abstrak :
Data Mining (DM) adalah salah satu bidang yang berkembang pesat karena besarnya kebutuhan akan nilai tambah dari database skala besar yang makin banyak terakumulasi sejalan dengan pertumbuhan teknologi informasi. Definisi umum dari DM itu sendiri adalah serangkaian proses untuk menggali nilai tambah berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data. Dalam review ini, penulis mencoba merangkum perkembangan terakhir dari teknik-teknik DM beserta implikasinya di dunia bisnis.

Kata Kunci: data mining, data warehouse, association rule mining, classification, clustering

Pendahuluan
Perkembangan data mining(DM) yang pesat tidak dapat lepas dari perkembangan teknologi informasi yang memungkinkan data dalam jumlah besar terakumulasi. Sebagai contoh, toko swalayan merekam setiap penjualan barang dengan memakai alat POS(point of sales). Database data penjualan tsb. bisa mencapai beberapa GB setiap harinya untuk sebuah jaringan toko swalayan berskala nasional. Perkembangan internet juga punya andil cukup besar dalam akumulasi data.

Tetapi pertumbuhan yang pesat dari akumulasi data itu telah menciptakan kondisi yang sering disebut sebagai "*rich of data but poor of information*" karena data yang terkumpul itu tidak dapat digunakan untuk aplikasi yang berguna. Tidak jarang kumpulan data itu dibiarkan begitu saja seakan-akan "kuburan data" (data tombs).

DM adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. Patut diingat bahwa kata mining sendiri berarti usaha untuk mendapatkan sedikit barang berharga dari sejumlah besar material dasar. Karena itu DM sebenarnya memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (artificial intelligent), machine learning, statistik dan database. Beberapa teknik yang sering disebut-sebut dalam literatur DM antara lain : clustering, classification, association rule mining, neural network, genetic algorithm dan lain-lain.

Yang membedakan persepsi terhadap DM adalah perkembangan teknik-teknik DM untuk aplikasi pada database skala besar. Sebelum populernya DM, teknik-teknik tersebut hanya dapat dipakai untuk data skala kecil saja.

Di sini, penulis mencoba untuk memberi gambaran sekilas atas perkembangan terakhir teknik-teknik DM sambil memberikan juga ilustrasi pemakaian di dunia bisnis. Penulis juga menyajikan pengertian konfigurasi penyimpanan data yang memudahkan pemakai untuk

melakukan DM yang umum disebut dengan *data warehouse*.

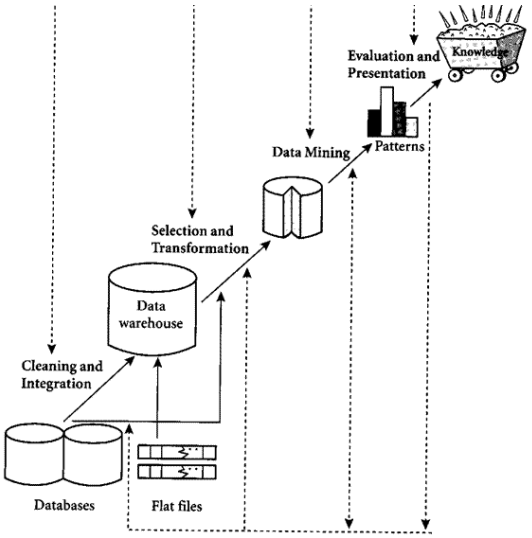
Proses Data Mining

Disini akan diuraikan tahap-tahap DM dan pengertian data warehouse.

Tahap-Tahap Data Mining

Karena DM adalah suatu rangkaian proses, DM dapat dibagi menjadi beberapa tahap yang diilustrasikan di Gambar 1[4]:

1. Pembersihan data (untuk membuang data yang tidak konsisten dan noise)
2. Integrasi data (penggabungan data dari beberapa sumber)
3. Transformasi data (data diubah menjadi bentuk yang sesuai untuk di-mining)
4. Aplikasi teknik DM
5. Evaluasi pola yang ditemukan (untuk menemukan yang menarik/bernilai)
6. Presentasi pengetahuan (dengan teknik visualisasi)



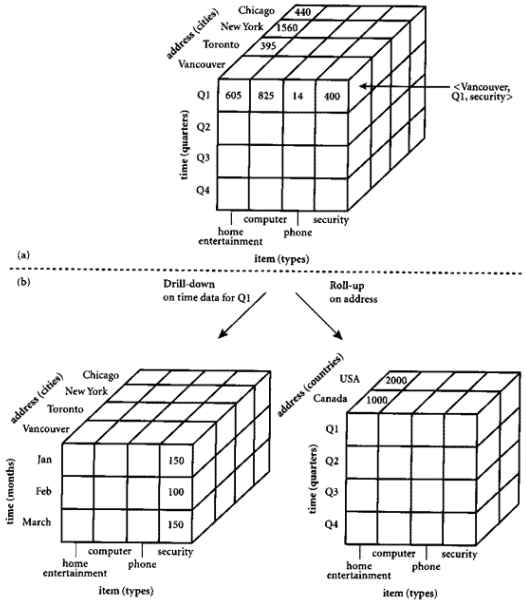
Gambar 1 : Tahap-Tahap Data Mining

Tahap-tahap tsb. bersifat interaktif di mana pemakai terlibat langsung atau dengan perantaraan knowledge base.

Data Warehouse

Biasanya perusahaan-perusahaan memakai database dalam operasi sehari-harinya seperti pencatatan transaksi jual-beli, administrasi pengiriman barang, inventori, penggajian dsb yang lazim disebut dengan OLTP (online transaction processing). Dengan makin besarnya kebutuhan akan analisa data untuk mempertahankan keunggulan dalam kompetisi, banyak perusahaan yang juga membangun database tersendiri yang khusus digunakan untuk menunjang proses pengambilan keputusan (decision making) atau lazim juga disebut dengan OLAP (online analytical processing).

Berbeda dengan OLTP yang hanya memakai operasi query yang sederhana dan berulang-ulang, query untuk OLAP biasanya lebih rumit , bersifat adhoc, dan tidak melibatkan operasi data update. OLAP juga tidak memakai data operasi sehari-hari begitu saja, tetapi memakai data yang sudah terangkum dengan model data yang disebut *data cube*. Data cube adalah presentasi data multidimensi seperti jenis barang, waktu, lokasi dsb. Ilustrasi dari data cube ditunjukkan di Gambar 2.



Gambar 2: Data Cube Pada Data Warehouse

Dimensi pada data cube dapat dibuat bertingkat, contohnya dimensi lokasi dapat dibagi menjadi kota, propinsi dan negara. Sedangkan dimensi waktu mencakup jam, hari, minggu, bulan, tahun dsb. Dengan ini pemakai dapat dengan mudah mendapat rangkuman informasi dari tingkatan dimensi yang lebih luas/umum seperti negara atau tahun dengan operasi yang disebut *roll-up* seperti ditunjukkan di Gambar 2. Sebaliknya dengan operasi *drill-down*, pemakai dapat menggali informasi dari tingkatan dimensi yang lebih detil seperti data harian atau data di lokasi yang spesifik.

Data cube yang tersedia pada data warehouse memungkinkan pemakai untuk menganalisa data operasi sehari-hari dengan berbagai sudut

pandang, dan sangat berguna untuk mengevaluasi suatu asumsi bisnis. Akan tetapi untuk mendapatkan informasi yang tidak diketahui secara eksplisit diperlukan satu tahap lagi yaitu aplikasi teknik DM. Disini data warehouse merupakan data mentah untuk DM. Data warehouse sendiri secara periodik diisi data dari OLTP setelah menjalani pembersihan dan integrasi data. Karena itu ada pula anggapan bahwa DM adalah tahap lanjut dari OLAP.

Teknik-Teknik Data Mining

Dengan definisi DM yang luas, ada banyak jenis teknik analisa yang dapat digolongkan dalam DM. Karena keterbatasan tempat, disini penulis akan memberikan sedikit gambaran tentang tiga teknik DM yang paling populer.

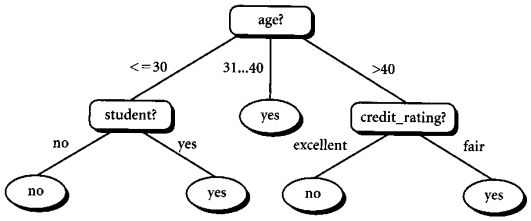
Association Rule Mining

Association rule mining adalah teknik mining untuk menemukan aturan assosiatif antara suatu kombinasi item. Contoh dari aturan assosiatif dari analisa pembelian di suatu pasar swalayan adalah bisa diketahui berapa besar kemungkinan seorang pelanggan membeli roti bersamaan dengan susu. Dengan pengetahuan tsb. pemilik pasar swalayan dapat mengatur penempatan barangnya atau merancang kampanye pemasaran dengan memakai kupon diskon untuk kombinasi barang tertentu. Penting tidaknya suatu aturan assosiatif dapat diketahui dengan dua parameter, *support* yaitu persentase kombinasi item tsb. dalam database dan *confidence* yaitu kuatnya hubungan antar item dalam aturan assosiatif.

Algoritma yang paling populer dikenal sebagai Apriori dengan paradigma generate and test, yaitu pembuatan kandidat kombinasi item yang mungkin berdasar aturan tertentu lalu diuji apakah kombinasi item tsb memenuhi syarat support minimum. Kombinasi item yang memenuhi syarat tsb. disebut *frequent itemset*, yang nantinya dipakai untuk membuat aturan-aturan yang memenuhi syarat confidence minimum[1]. Algoritma baru yang lebih efisien bernama FP-Tree[5].

Classification

Classification adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Model itu sendiri bisa berupa aturan “jika-maka”, berupa *decision tree*, formula matematis atau *neural network*.



Gambar 3: Decision Tree

Decision tree adalah salah satu metode classification yang paling populer karena mudah untuk diinterpretasi oleh manusia. Contoh dari decision tree dapat dilihat di Gambar 3. Disini setiap percabangan menyatakan kondisi yang harus dipenuhi dan tiap ujung pohon menyatakan kelas data. Contoh di Gambar 3 adalah identifikasi pembeli komputer, dari decision tree tsb. diketahui bahwa salah satu kelompok yang potensial membeli komputer adalah orang yang berusia di bawah 30 tahun dan juga pelajar.

Algoritma decision tree yang paling terkenal adalah C4.5[7], tetapi akhir-akhir ini telah dikembangkan algoritma yang mampu menangani data skala besar yang tidak dapat ditampung di main memory seperti RainForest[3]. Metode-metode classification yang lain adalah Bayesian, neural network, genetic algorithm, fuzzy, case-based reasoning, dan k-nearest neighbor.

Proses classification biasanya dibagi menjadi dua fase : *learning* dan *test*. Pada fase learning, sebagian data yang telah diketahui kelas datanya diumpankan untuk membentuk model perkiraan. Kemudian pada fase test model yang sudah terbentuk diuji dengan sebagian data lainnya untuk mengetahui akurasi dari model tsb. Bila akurasinya mencukupi model ini dapat dipakai untuk prediksi kelas data yang belum diketahui.

Clustering

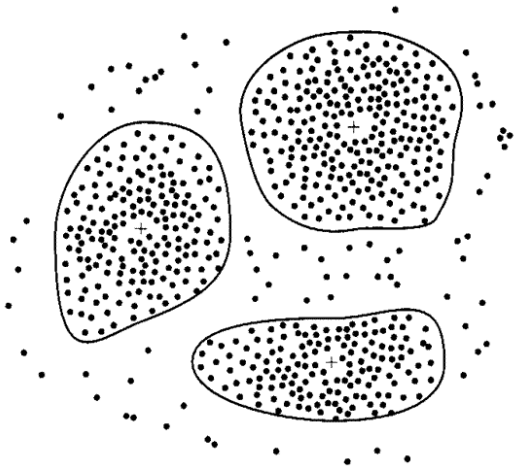
Berbeda dengan association rule mining dan classification dimana kelas data telah ditentukan sebelumnya, clustering melakukan penge-lompokan data tanpa berdasarkan kelas data tertentu. Bahkan clustering dapat dipakai untuk memberikan label pada kelas data yang belum diketahui itu. Karena itu clustering sering digolongkan sebagai metode *unsupervised learning*.

Prinsip dari clustering adalah memaksimalkan kesamaan antar anggota satu kelas dan meminimumkan kesamaan antar kelas/cluster. Clustering dapat dilakukan pada data yan memiliki beberapa atribut yang dipetakan sebagai

ruang multidimensi. Ilustrasi dari clustering dapat dilihat di Gambar 4 dimana lokasi, dinyatakan dengan bidang dua dimensi, dari pelanggan suatu toko dapat dikelompokkan menjadi beberapa cluster dengan pusat cluster ditunjukkan oleh tanda positif (+).

Banyak algoritma clustering memerlukan fungsi jarak untuk mengukur kemiripan antar data, diperlukan juga metode untuk normalisasi bermacam atribut yang dimiliki data.

Beberapa kategori algoritma clustering yang banyak dikenal adalah metode partisi dimana pemakai harus menentukan jumlah *k* partisi yang diinginkan lalu setiap data dites untuk dimasukkan pada salah satu partisi, metode lain yang telah lama dikenal adalah metode hierarki yang terbagi dua lagi : bottom-up yang menggabungkan cluster kecil menjadi cluster lebih besar dan top-down yang memecah cluster besar menjadi cluster yang lebih kecil. Kelemahan metode ini adalah bila bila salah satu penggabungan/pemecahan dilakukan pada tempat yang salah, tidak dapat didapatkan cluster yang optimal. Pendekatan yang banyak diambil adalah menggabungkan metode hierarki dengan metode clustering lainnya seperti yang dilakukan oleh Chameleon[6].



Gambar 4: Clustering

Akhir-akhir ini dikembangkan juga metode berdasar kepadatan data, yaitu jumlah data yang ada di sekitar suatu data yang sudah teridentifikasi dalam suatu cluster. Bila jumlah data dalam jangkauan tertentu lebih besar dari nilai ambang batas, data-data tsb dimasukkan dalam cluster. Kelebihan metode ini adalah bentuk cluster yang lebih fleksibel. Algoritma yang terkenal adalah DBSCAN[2].

Penutup

Ada bermacam-macam teknik DM termasuk yang tidak diulas disini. Untuk mendapatkan hasil DM yang optimal tidak hanya diperlukan pemahaman akan teknik tsb. tapi juga model permasalahan yang dihadapi.

Referensi

[1] Agrawal and R. Srikant, “Fast algorithms for mining association rules in large databases”, In Proc. 1994 Int. Conf. Very Large Data Bases(VLDB), 1994.

[2] M. Ester et. al. “A density based algorithm for discovering clusters in large spatial databases” In Proc. 1996 Int. Conf. Knowledge Discovery and Data Mining (KDD’96), 1996.

[3] J. Gehrke, R. Ramakrishnan and V. Ganti. “Rainforest: A framework for fast decision tree construction of large datasets”. In Proc. 1998 Int. Conf Very Large Data Bases (VLDB), 1998.

[4] J. Han and M. Kamber. Data Mining : Concepts and Techniques. Morgan Kaufmann, 2001.

[5] J. Han, J. Pei and Y. Yin. “Mining frequent patterns without candidate generation”, In Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD’00), 2000.

[6] G. Karypis, E.-H. Han and V. Kumar. “CHAMELEON:A hierarchical clustering algorithm using dynamic modeling” COMPUTER, 32:68-75, 1997.

[7] J.R. Quinlan. “C4.5:Programs for Machine Learning”, Morgan Kaufmann, 1993.

Biografi Penulis

Iko Pramudiono, Menyelesaikan program M.Eng. dalam Electro Communications Engineering pada tahun 1999, dari The University of Tokyo. Saat ini kandidat Doktor di tempat yang sama dengan bidang riset data mining, khususnya untuk aplikasi di data dari web. Iko Pramudiono adalah anggota society ilmiah IECI.