DATAMINING

Anief Rufiyanto.ST (Universitas Pandanaran)

ABSTRAK

Data mining adalah salah satu solusi pelayanan proses pengolahan informasi dalam suatu basis data yang berskala besar. Saat sebuah organisasi perusahaan atau institusi yang mempunyai banyak sekali data-data. Tidak menutup kemungkinan banyak sekali informasi yang dapat diperoleh darinya, serta bagaimana solusi data mining bisa diterapkan dengan berbagai teknik diantaranya yaitu classification, Association dan Clustering. Singkatnya, data mining adalah ekstraksi informasi atau pola yang penting atau menarik dari kumpulan data yang ada dalam sebuah database yang besar.

Dengan data mining dimana serangkain prosesnya akan menghasilkan suatu nilai tambah berupa informasi atau pengetahuan baru yang selama ini tidak diketahui secara manual dari sekumpulan data.

Kata kunci: Data , Data Mining, Basis data, informasi, pengetahuan

PENDAHULUAN

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.

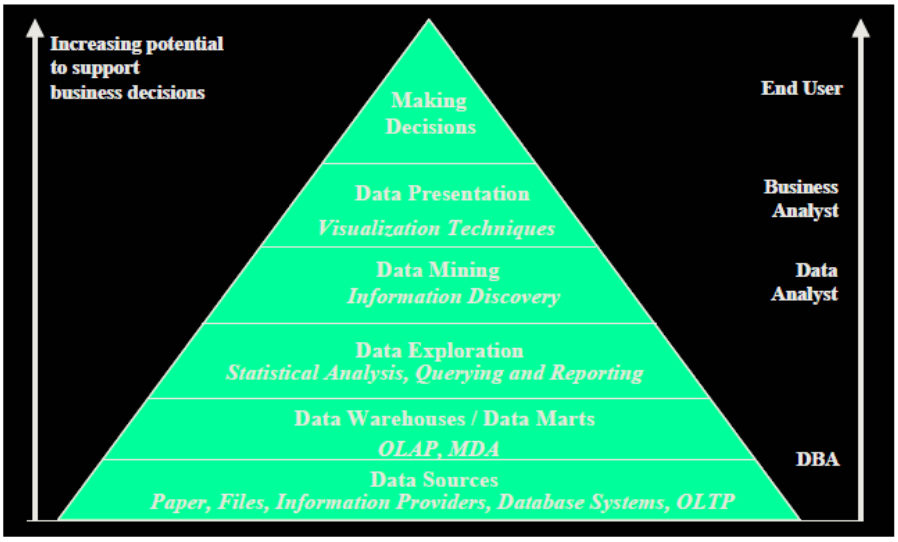
Definisi sederhana dari data mining adalah ekstraksi informasi atau pola yang penting atau menarik dari data yang ada di database yang besar.
Kehadiran data mining dilatar belakangi dengan problema data explosion yang dialami akhir-akhir ini dimana banyak organisasi telah mengumpulkan data sekian tahun lamanya (data pembelian, data penjualan, data nasabah, data transaksi dsb.). Hampir semua data tersebut dimasukkan menggunakan aplikasi komputer yang digunakan untuk menangani transaksi sehari-hari yang kebanyakan adalah OLTP (On Line Transaction Processing).

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.

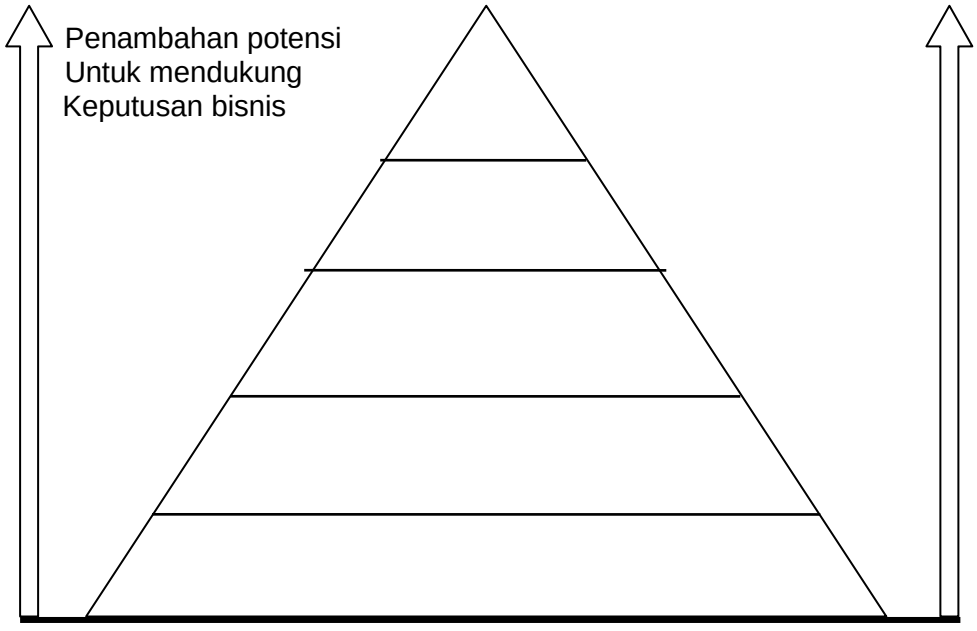
PROSES IDENTIFIKASI DATA

Persoalan di atas merupakan salah satu persoalan yang dapat diatasi oleh data mining dari sekian banyak potensi permasalahan yang ada. Data mining dapat menggali 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



Data Mining merupakan teknologi yang sangat berguna untuk membantu perusahaan perusahaan menemukan informasi yang sangat penting dari gudang data *(Data warehouse)* mereka. Dengan data mining dapat meramalkan tren dan sifat-sifat perilaku bisnis yang sangat berguna untuk mendukung pengambilan keputusan penting. Analisis yang diotomatisasi yang dilakukan oleh data mining melebihi yang dilakukan oleh sistem pendukung keputusan tradisional yang sudah banyak digunakan. Data Mining dapat menjawab pertanyaan-pertanyaan bisnis yang dengan cara tradisional memerlukan banyak waktu dan cost tinggi.
Data Mining mengeksplorasi basis data untuk menemukan pola-pola yang tersembunyi, mencari informasi untuk memprediksi yang mungkin saja Terlupakan oleh para pelaku bisnis karena terletak di luar ekspektasi mereka. Sebagai contoh adalah beberapa solusi yang bisa diselesaikan dengan data mining diantaranya yaitu **menebak target pasar,** yaitu dengan melakukan pengelompokan dari model-model pembeli dan melakukan klasifikasi setiap pembeli dari kebiasaan membeli, dari tingkat penghasilan dan karakteristik 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.

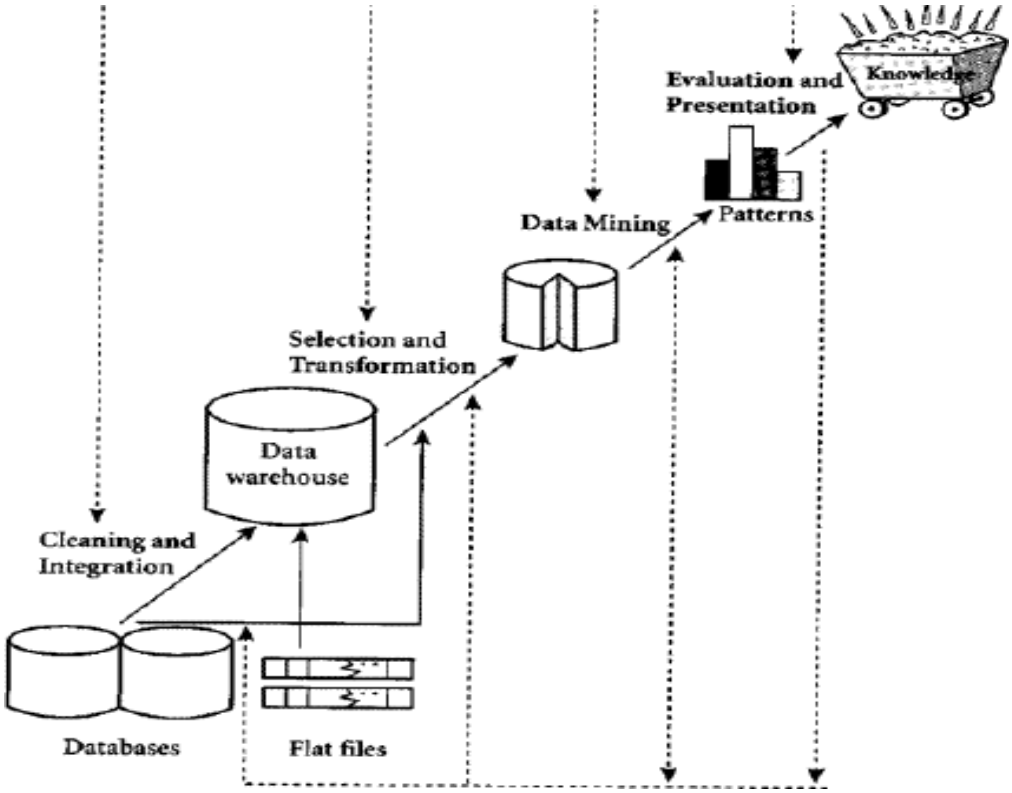
Menentukan profile customer dan **identifikasi kebutuhan customer**, dimana data mining bisa melihat profil customer untuk mengetahui kelompok konsumen tertentu misalnya suka membeli produk apa saja serta mengidentifikasi produk-produk apa saja yang terbaik untuk setiap kelompok konsumen sehingga bisa menyusun faktor-faktor apa saja yang dapat menarik konsumen baru untuk bergabung atau membeli. Serta banyak sekali solusi data mining yang bisa di Implemetasikan sebagai misal di toko buku yang sangat terkenal yaitu www.amazon.com, bidang perbankan , operator GSM dan lain sebagainya. Gambar dibawah adalah hasil dari data mining oleh amazon dalam melihat prilaku konsumen.

Definisi Data Mining
Kemajuan dalam pengumpulan data dan teknologi penyimpanan yang cepat

memungkinkan organisasi menghimpun jumlah data yang sangat luas. Alat dan teknik analisis data yang tradisional tidak dapat digunakan untuk mengektrak informasi dari data yang sangat besar.
Untuk itu diperlukan suatu metoda baru yang dapat menjawab kebutuhan tersebut. *Data mining* merupakan teknologi yang menggabungkan metoda analisis tradisional dengan algoritma yang canggih untuk memproses data dengan volume besar.
Ada beberapa definisi dari data mining yang dikenal diiantaranya adalah :
1. Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual.
2. Data mining adalah analisa otomatis dari data yang berjumlah besar atau kompleks dengan tujuan untuk menemukan pola atau kecenderungan yang penting yang biasanya tidak disadari keberadaannya
3. *Data mining* atau *Knowledge Discovery in Databases* (KDD) adalah pengambilan informasi yang tersembunyi, dimana informasi tersebut sebelumnya tidak dikenal dan berpotensi bermanfaat. Proses ini meliputi sejumlah pendekatan teknis yang berbeda, seperti *clustering*, data *summarization*, *learning classification rules*.
Secara umum, data mining dapat melakukan dua hal yaitu memberikan kesempatan untuk menemukan informasi menarik yang tidak terduga, dan juga bisa menangani data berskala besar.

Dalam menemukan informasi yang menarik ini, ciri khas data mining adalah kemampuan pencarian secara hampir otomatis, karena dalam banyak teknik data mining ada beberapa parameter yang masih harus ditentukan secara manual atau semi manual. Data mining juga dapat memanfaatkan pengalaman atau bahkan kesalahan di masa lalu untuk meningkatkan kualitas dari model maupun hasil analisanya, salah satunya dengan kemampuan pembelajaran yang dimiliki beberapa teknik data mining seperti klasifikasi.

Tahapan Data Mining
Salah satu tuntutan dari data mining ketika diterapkan pada data berskala besar adalah diperlukan metodologi sistematis tidak hanya ketika melakukan analisa saja tetapi juga ketika mempersiapkan data dan juga melakukan interpretasi dari hasilnya sehingga dapat menjadi aksi ataupun keputusan yang bermanfaat.
Data mining seharusnya dipahami sebagai suatu proses, yang memiliki tahapan-tahapan tertentu dan juga ada umpan balik dari setiap tahapan ke tahapan sebelumnya. Pada umumnya proses data mining berjalan interaktif karena tidak jarang hasil data mining pada awalnya tidak sesuai dengan harapan analisnya sehingga perlu dilakukan desain ulang prosesnya.



Gambar 2. Tahapan Data Mining

Sebagai suatu rangkaian proses, data mining dapat dibagi menjadi beberapa tahap yang diilustrasikan. Tahap-tahap tersebut. bersifat interaktif di mana pemakai terlibat langsung atau dengan perantaraan knowledge base.

1. Pembersihan data

Digunakan untuk membuang data yang tidak konsisten dan noise

2. Integrasi Data

Data yang diperlukan untuk data mining tidak hanya berasal dari satu database tetapi juga berasal dari beberapa database atau file teks. Hasil integrasi data sering diwujudkan dalam sebuah data warehouse karena dengan data warehouse, data dikonsolidasikan dengan struktur khusus yang efisien. Selain itu data warehouse juga memungkinkan tipe analisa seperti OLAP.

3. Transformasi data

Transformasi dan pemilihan data ini untuk menentukan kualitas dari hasil data mining, sehingga data diubah menjadi bentuk sesuai untuk di-Mining.

4. Aplikasi Teknik Data Mining

Aplikasi teknik data mining sendiri hanya merupakan salah satu bagian dari proses data mining. Ada beberapa teknik data mining yang sudah umum dipakai.

5. Evaluasi pola yang ditemukan

Dalam tahap ini hasil dari teknik data mining berupa pola-pola yang khas maupun model prediksi dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai.

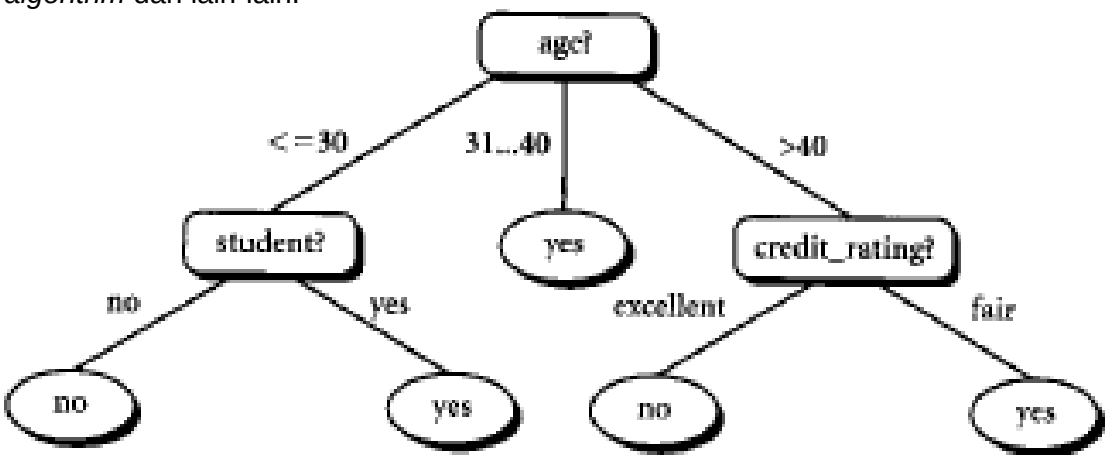
6. Presentasi Pengetahuan

Presentasi pola yang ditemukan untuk menghasilkan aksi tahap terakhir dari proses data mining adalah bagaimana memformulasikan keputusan atau aksi dari hasil analisa yang didapat.

Teknik Data Mining

Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. Perlu diingat bahwa kata *mining* sendiri berarti usaha untuk mendapatkan sedikit data berharga dari sejumlah besar data dasar. Karena itu *data mining* sebenarnya memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (*artificial intelligent*), *machine learning*, statistik dan basisdata.

Beberapa teknik yang sering disebut-sebut dalam literatur *data mining* antara lain yaitu *association rule mining*, *clustering*, *klasifikasi*, *neural network*, *genetic algorithm* dan lain-lain.



Gambar 3 : Contoh decision tree

A. Classification

Suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan.Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dan dengan menggunakan hasilnya untuk memberikan sejumlah aturan. Aturan-aturan tersebut digunakan pada data-data baru untuk diklasifikasi. Teknik ini menggunkan *supervised induction*, yang memanfaatkan kumpulan pengujian dari record yang terklasifikasi untuk menentukan kelas-kelas tambahan.

Salah satu contoh yang mudah dan popular adalah dengan Decision tree yaitu salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi. Decision tree adalah model prediksi menggunakan struktur pohon atau struktur berhirarki.

Decision tree adalah struktur *flowchart* yang menyerupai *tree* (pohon), dimana setiap simpul internal menandakan suatu tes pada atribut, setiap cabang merepresentasikan hasil tes, dan simpul daun merepresentasikan kelas atau distribusi kelas. Alur pada *decision tree* di telusuri dari simpul akar ke simpul daun yang memegang prediksi kelas untuk contoh tersebut. *Decision tree* mudah untuk dikonversi ke aturan klasifikasi (*classification rules*).

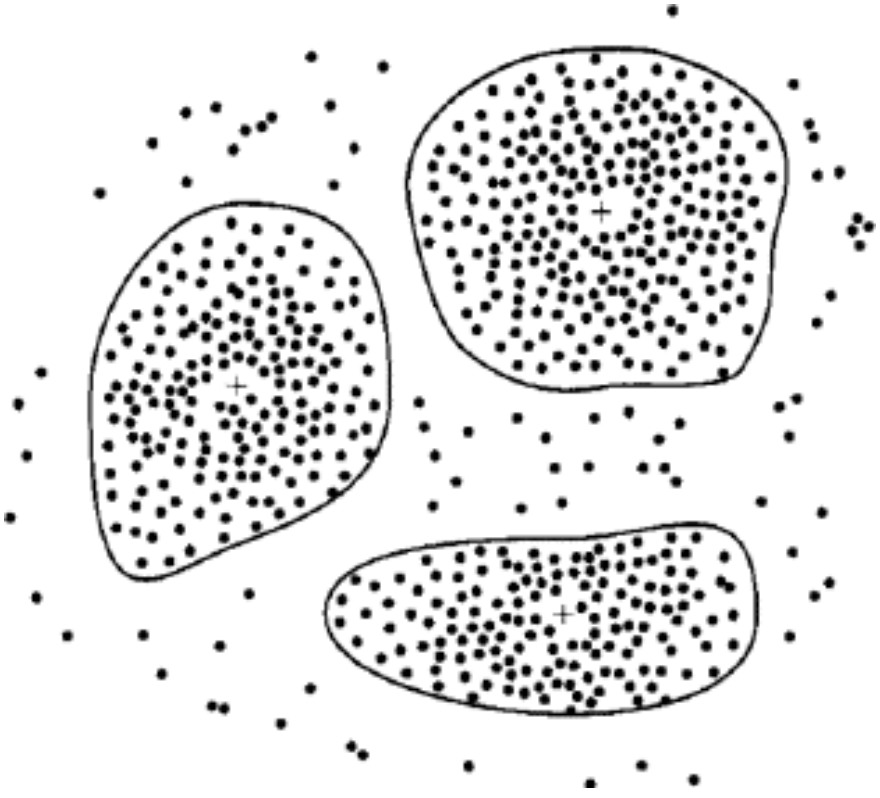
B. Association

Digunakan untuk mengenali kelakuan dari kejadian-kejadian khusus atau proses dimana link asosiasi muncul pada setiap kejadian. Contoh dari aturan assosiatif dari analisa pembelian disuatu pasar swalayan adalah bisa diketahui berapa besar kemungkinan seorang pelanggan membeli roti bersamaan dengan susu.

Dengan pengetahuan tersebut 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 prosentasi kombinasi atribut tersebut dalam basisdata dan *confidence* yaitu kuatnya hubungan antar atribut dalam aturan asosiatif.Motivasi awal pencarian association rule berasal dari keinginan untuk menganalisa data transaksi supermarket, ditinjau dari perilaku customer dalam membeli produk. Association rule ini menjelaskan seberapa sering suatu produk dibeli secara bersamaan. Sebagai contoh, association rule **"*beer* => *diaper* (80%)"** menunjukkan bahwa empat dari lima customer yang membeli *beer* juga membeli *diaper*. Dalam suatu association rule *X* => *Y*, *X* disebut dengan *antecedent* dan *Y* disebut dengan *consequent.*Rule.

C. Clustering

Digunakan untuk menganalisis pengelompokkan berbeda terhadap data, mirip dengan klasifikasi, namun pengelompokkan belum didefinisikan sebelum dijalankannya tool data mining. Biasanya menggunkan metode *neural network* atau statistik. Clustering membagi item menjadi kelompok-kelompok berdasarkan yang ditemukan tool data mining. Prinsip dari *clustering* adalah memaksimalkan kesamaan antar anggota satu kelas dan meminimumkan kesamaan antar *cluster*. *Clustering* dapat dilakukan pada data yang 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 took 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 metoda untuk normalisasi bermacam atribut yang dimiliki data.



Gambar 4 : Contoh klasterisasi

Dalam pembahasan ini akan di ambil salah satu bentuk solusi dari data mining untuk menganalisis Perilaku Pengunjung pada suatu toko X. Beberapa metoda atau teknik yang dikenal di dalam data mining salah satunya adalah *association rule* (aturan asosiasi) yang berusaha menemukan aturan-aturan tertentu yang mengasosiasikan data yang satu dengan data yang lain, dapat digunakan untuk kasus di atas.

Untuk mencari association rule dari data tersebut pertama-tama kita harus mencari lebih dulu yang disebut "frequent item" (item sering), yaitu barang yang sering dibeli oleh seorang pengunjung. Misalnya kita dapatkan data barang yang beberapa barang dibeli oleh seorang pengunjung di bulan X seperti berikut:

1. beras, **minyak goreng**, daging sapi
2. **gula pasir**, **minyak goreng**, **telur ayam**
3. beras, **gula pasir**, **minyak goreng**, **telur ayam**
4. **gula pasir**, **telur ayam**

Jika kita menetapkan bahwa yang dikatakan "sering" adalah pembelian sebanyak 2 kali atau lebih, maka kita dapatkan frequent item dari seorang pengunjung adalah:

Gula pasir 3 kali
Minyak Goreng 3 kali
Telur ayam 3 kali

Sementara daging sapi bukan frequent item karena hanya 1 kali dibeli oleh pengunjung. Frequent item ini mencakup juga untuk pembelian secara bersama-sama, sehingga kita dapatkan juga frequent item seperti berikut:

Gula pasir dan telur ayam 3 kali
Gula pasir dan minyak goreng 2 kali
Minyak goreng dan telur ayam 2 kali
Minyak goreng dan beras 2 kali
goreng dan gula pasir dan telur ayam 2 kali

Jumlah pembelian barang di sini dalam bahasa data mining dikenal dengan istilah "support"(dukungan) dan batas minimal "2" disebut *minimum support* (dukungan minimal).

Dari data pengunjung di atas kita bisa membuat association rule seperti misalnya:

Pengunjung yang membeli beras akan membeli juga minyak goreng. Dalam bahasa data mining association rule ini kita tulis seperti berikut:

beras ***=>*** *minyak goreng, atau "****if*** *beras* ***then*** *minyak goreng".*

Tentu saja aturan ini tidak bersifat pasti tetapi hanya kemungkinan-kemungkinan berdasarkan kebiasaan pembelian seorang konsumen di bulan X ini. Kemungkinan dari association rule yang kita buat dapat kita hitung seperti berikut ini : ***Support(minyak goreng & beras)/support(beras)= 2/2 = 1***

Jika melihat data seorang pengunjung saja seperti contoh data diatas, ternyata kemungkinan *association rule* ini adalah 100%, atau sering disebut "*confidence*" (kepercayaan).

Sehingga sebagian *association rule* yang dapat kita buat dari data seorang pengunjung di atasKetika perhitungan data seperti di atas kita lakukan terhadap data seluruh pengunjung, maka kita akan mendapatkan *association rule* yang valid yang benar-benar mencerminkan kecenderungan pola pembelian dari pengunjung toko tersebut.

Pengetahuan yang diperoleh dari hasil analisis diatas dapat dipercaya tentang kecenderungan dan perilaku pengunjung sebuah toko dengan akan menghasilkan anfaat bagi pemilik toko untuk mengambil keputusan-keputusan

strategis tentang tokonya tersebut. Seperti yang telah disinggung di muka, antara lain dalam mengatur tata letak barang, penyiapan stok barang, pemberian rekomendasi-rekomendasi tertentu kepada pengunjung dan lain-lain.

KESIMPULAN

Teknologi basisdata modern telah mengaplikasikan gudangdata dengan basis kemampuan yang sangat besar dalam menyimpan dan mentransmisikan data. Datamining punya potensi pengembangan untuk mendapat informasi dan pengetahuan yang lebih valid dan efisien dengan berbagai metode analisis. Sehingga kita dapat menganalisis, memahami, bahkan memvisualisasikannya, melalui pencarian pengetahuan dalam basisdata dalam proses identifikasi pola-pola yang valid, berpotensi manfaat, dan dapat dipahami secara mudah. Data mining sebagai solusi pengambilan keputusan yang memungkinkan menghasilkan "pengetahuan baru".

DAFTAR PUSTAKA

Fayyad, U., Piatetsky-Shapiro, G. dan Smyth, P. (1996), *From Data Mining to Knowledge Discovery in Databases*, AAAI and The MIT Pres, 37-53.

Han, Jiawei and Micheline Kamber, "*Data Mining Concepts and Techniques"*, Morgan Kaufmann, California, 2001.

José Hernández-Orallo, I*ntroduction to Data Mining (Presentation)*, Dpto. de Sistemas Informáticos y Computación Universidad Politécnica de Valencia, Spain Horsens, Denmark, 26th September 2005

Marsela ,Veronica S. Moertini Analsis Keranjang Pasar Dengan Algortima Hash-Based Pada Data Transaksi Penjualan Apotek, INTEGRAL, Vol. 9 No. 3, November 2004

Yudi Agusta, PhD, *Data Warehouse and Data Mining*, Lecture 9 , 2006, http://kuliah.stikombali.net

----, Data Mining: Penerapan yang Penting, namun Tidak Disadari, di akses 7 Februari 2007.
http://www.pcmedia.co.id/detail.asp?id=1330&Cid=22&cp=1&Eid=27