TUGAS MTI

“Rangkuman KDD dan Menganalisis kasus Knowledge Discovery from Database tentang Supermarket”

Nama : Dwi Setia Anggraini

NIM : 09031281621052

Knowledge Discovery from Database (KDD) merupakan tahapan atau fase yang menyertai tahapan numeric, seperti Preperasi data, pencarian hipotesis untuk digeneralisasi, formasi pola, evaluasi knowledge, representasi serta manajemen. Proses yang multi fase tersebut merupakan methodology yang sangat penting untuk menemukan pengetahuan dalam data yang sesungguhnya.

Teory Rough set merupakan basic bagi KDD dan digunakan dalam KDD. Istilah knowledge discovery in databases (KDD) dan data mining sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda akan tetapi berkaitan satu sama lain. Dan salah satu tahapan dalam keseluruhan proses KDD adalah data mining. Proses KDD secara garis besar dapat dijelaskan sebagai berikut

**Data Selection**

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.

**Pre-processing/Cleaning**  
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.

**Transformation**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

**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.

**Interpretation/ Evaluation**  
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.

KDD mencakup keseluruhan proses pencarian pola atau informasi dalam basis data, dimulai dari pemilihan dan persiapan data sampai representasi pola yang ditemukan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Data mining merupakan salah satu komponen dalam KDD yang difokuskan pada penggalian pola tersembunyi dalam basis data.  
  
  
**Generalized Distribution Table and Rough Set System**GDT merupakan sebuah sistem induksi yang mudah untuk menemukan aturan-aturan pengklasifikasian dari database dengan data yang tidak lengkap dan tidak pasti (Zhong, dkk., 1998, Dong dkk., 1999a). Sistem tersebut didasarkan atas hibridisasi dari Tabel Distribusi Generalisasi (GDT, Generalization Distribution Table) dan metodologi Rough Set. Dari data pelatihan yang tidak lengkap, system GDT-RS dapat menghasilkan sekumpulan aturan dengan panjang deskripsi yang minimal (semi minimal), yang mempunyai kekuatan yang besar dan mencakup semua contoh-contoh.

Terdapat dua jenis atribut dalam sebuah database yaitu atribut kondisi dan atribut keputusan yang kadang disebut atribut kelas. Atribut kondisi digunakan untuk menggambarkan contoh-contoh yang memungkinkan dalam GDT, sedangkan atribut keputusan berhubungan dengan konsep-konsep (kelas-kelas) yang tergambar dalam sebuah aturan. Biasanya, sebuah atribut keputusan yang tunggal sudah merupakan semua yang dibutuhkan.  
Sebuah GDT terdiri dari tiga komponen, yaitu contoh-contoh yang memungkinkan, generalisasi yang memungkinkan dari contoh-contoh, dan hubungan probabilitas antara contoh-contoh yang memungkinkan dan generalisasi-generalisasi yang memungkinkan.

Contoh-contoh yang memungkinkan, yang digambarkan di puncak baris GDT, ditetapkan oleh semua kombinasi yang memungkinkan dari nilai-nilai atribut dari suatu database. Generalisasi-generalisasi yang memungkinkan dari contoh-contoh, yang digambarkan oleh kolom kiri dari sebuah GDT, semuanya merupakan kasus-kasus yang memungkinkan dari generalisasi untuk semua contoh-contoh yang memungkinkan.

Hubungan-hubungan probabilitas antara contoh-contoh yang memungkinkan dan generalisasi-generalisasi yang memungkinkan , yang digambarkan dengan entri-entri Gij dari sebuah GDT yang diberikan, ditentukan dengan menggunakan distribusi probabilitas yang menggambarkan kekuatan hubungan antara contoh-contoh manapun dan generalisasi yang memungkinkan manapun. Distribusi utama diasumsikan menjadi seragam jika latar belakang pengetahuan tidak tersedia. Karena itu, hal ini dibatasi dengan  
dimana Pij adalah j-th contoh yang memungkinkan, PG adalah i-th generalisasi yang memungkinkan, dan Npg adalah jumlah dari contoh-contoh yang memungkinkan yang memenuhi i-th generalisasi yang memungkinkan tersebut.

Contoh kasus nya pada supermarket

Pertumbuhan data yang ekplosif  disimpan dalam database-database sementara yang bersifat operasional. Dengan berjalannya waktu, database tersebut menjadi gudang data atau lebih dikenal dengan data warehouse. Di dalam data warehouse ini tersimpan banyak sekali data yang telah direkap setiap harinya. Melalui data-data ini lah terdapat informasi yang harus digali untuk menunjang proses bisnis yang ada dalam menjalankan visi misi perusahaan.

Dibutuhkan suatu teknik untuk menggali informasi yang terdapat pada data warehouse tersebut. Lahirlah data mining, yaitu suatu teknik untuk meng-ekstrasi suatu pola dari data. Data mining diterapkan dengan paradigma untuk melihat informasi yang tersembunyi. Proses pencarian ini dilakukan secara otomatis terhadap pola dalam data dengan jumlah besar dengan menggunakan teknik-teknik seperti klasifikasi, clustering, dll. Data mining adalah suatu multidisiplin ilmu yang kompleks dan beririsan dengan ilmu lain seperti statistika, database, information retrieval, machine learning, pattern recognition, algoritma dan ilmu lainnya.

Data mining yang diartikan sebagai penggalian data merupakan suatu proses pencarian otomatis terhadap pola dalam data yang melibatkan data dalam jumlah besar dengan menggunakan perangkat seperti clustering, pengelompokan, dll. Data mining diterapkan dengan tujuan untuk menemukan informasi yang “tersembunyi” didalam data, oleh sebab itu data mining juga dikenal dengan nama Knowledge Discovery in Database (KDD)

.

Sebagai contoh kasus data mining yang terjadi yaitu data transaksi disebuah supermarket dapat berjumlah ribuan dalam periode singkat dan terus bertambah. Seringkali perusahaan hanya membiarkan data tersebut menggunung untuk laporan transaksi, tetapi dengan data mining kita dapat menggali  informasi yang sangat berharga dari data tersebut, misalnya asosiasi antar produk yang dapat digunakan untuk menentukan tata letak produk, dapat juga digunakan untuk menentukan daftar barang yang telah kadaluarsa.

Sebagai contoh lain kasus data mining yaitu   misalnya dalam bidang pemasaran, sebelum sebuah perusahaan mengeluarkan suatu produk baru kepasaran, perusahaan tersebut harus memiliki informasi tentang kecenderungan pelanggan untuk membeli produk yang akan di keluarkan. Perkiraan (hypothesis) dapat disusun untuk mengidentifikasikan pelanggan yang potensial dan karakteristik dari pelanggan yang ada. Data-data tentang pembelian pelanggan sebelumnya dan data tentang keadaan pelanggan, dapat digunakan untuk melakukan perbandingan antara pembelian dan karakteristik pelanggan untuk menetapkan dan menguji target yang telah diperkirakan sebelumnya. Dari keseluruhan operasi yang ada selanjutnya dapat dilakukan penyaringan dengan cermat sehingga jumlah perkiraan (hypothesys) yang sebelumnya banyak akan menjadi semakin berkurang sesuai dengan keadaan yang sebenarnya.Permasalahan utama dengan model ini adalah tidak ada informasi baru yang dapat dibuat, melainkan hanya pembuktian atau melemahkan perkiraan (hypothesys) dengan data-data yang ada sebelumnya. Datadata yang ada pada model ini hanya digunakan untuk membuktikan mendukung perkiraan (hypothesis) yang telah diambil sebelumnya. Jadi model ini sepenuhnya tergantung pada kemampuan user untuk melakukan analisa terhadap permasalahan yang ingin digali dan diperoleh informasinya.