**TUGAS**

**DM DAN KDD**

***Data Mining Dan Knowledge Discovery in Database***



**NAMA : GHIYATS SYAFIQ ABRIYANSYAH**

**NRP : 210210189**

**MAGISTER KOMPUTER**

**INSTITUT SAINS TERAPAN DAN TEKNOLOGI SURABAYA**

**2010**

**DATA MNING (DM)**

Data mining adalah salah satu solusi untuk menjelaskan proses pengalian informasi dalam suatu basis data yang berskala besar.(Kusnawi, *Seminar Nasional Teknologi 2007 (SNT 2007)*).

Data Mining adalah kegiatan untuk menemukan informasi atau pengetahuan yang berguna secara otomatis dari data yang jumlahnya besar. Data Mining merupakan salah satu proses dari keseluruhan proses yang ada pada Knowledge Discovery in Databases (KDD). (Data Mining, IT Telkom, <http://www.ittelkom.ac.id/library/>).

Saat suatu oraganisasi baik itu perusahaan atau suatu institusi yang mempunyai banyak sekali data-data, tidak menutup kemungkinan banyak sekali informasi yang dapat diperoleh, serta bagaimana solusi data mining bisa diterapkan dengan berbagai teknik data mining diantaranya yaitu classification, Association dan Clustering. Dengan data mining dimana serangkain prosesnya akan menghasilkan suatu nilai tambah berupa pengetahuan baru yang selama ini tidak diketahui secara manual dari sekumpulan data.

Data Mining merupakan teknologi yang sangat berguna untuk membantu perusahaan-perusahaan dalam 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.

**KNOWLEDGE DISCOVERY IN DATABASE (KDD)**

Knowledge Discovery In Database(KDD) sendiri merupakan sekumpulan proses untuk menemukan pengetahuan yang bermanfaat dari data. (Data Mining, IT Telkom, <http://www.ittelkom.ac.id/library/>).

KDD terdiri dari serangkaian langkah perubahan, termasuk data *preprocessing* dan juga *post processing*. Data *preprocessing* merupakan langkah untuk mengubah data mentah menjadi format yang sesuai untuk tahap analisis berikutnya. Selain itu data *preprocessing* juga digunakan untuk membantu dalam pengenalan atribut dan data segmen yang relevan dengan task *data mining*.

Data *preprocessing* kemungkinan akan membutuhkan waktu yang sangat lama, hal ini dikarenakan data yang mentah kemungkinan disimpan dengan format dan *database* yang berbeda.

*Post processing* meliputi semua operasi yang harus dilakukan agar hasil dari *Data Mining* dapat diakses dan lebih mudah untuk diinterpretasikan oleh para analis. Teknik visualisasi juga dapat digunakan untuk mempermudah para analis untuk menggali dan memahami kegunaan dari *data mining*. Kumpulan proses dalam KDD meliputi : pembersihan data (*data cleaning*), integrasi data (*data integration*), pemilihan data (*data selection*), transformasi data (*data transformation*), penambangan data (*data mining*), evaluasi pola (*pattern evaluation*), dan presentasi pengetahuan (*knowledge presentation*). Berdasarkan definisi ini terlihat bahwa *data mining* hanya merupakan salah satu proses dari keseluruhan proses yang ada pada KDD, tetapi merupakan proses yang sangat penting dalam usaha menemukan pola-pola yang berguna dari sejumlah data yang besar (data tersebut bisa disimpan dalam basisdata, *data warehouse*, atau media penyimpanan informasi lainnya).

Business Objective Determination atau dalam hal ini adalah penentuan sasaran bisnis. Di dalam tahapan ini yang perlku dilakukan adalah mendefinisikan permasalahan atau tantangan bisnis dengan jelas. Dalam hal ini adalah merupakan aspek yang sangat esensial dalam setiap proyek data mining.

Data Preparation, pada tahapan KDD ini adalah bertujuan untuk mempersiapkan data yang diperlukan untuk proses mining sebuah data. Tujuan yang lainnya adalah agar data yang digunakan bernar-benar sesuai dengan permasalahan yang ada dan yang akan dipecahkan, dalam hal penyelesaiannya dengan penggunaan data preparation ini dapat dijamin kebenarannya dan dalam format yang sesuai atau tepat dengan apa yang diinginkan atau yang ditujukan. Pada tahapan data preparation ini merupakan tahapan yang paling banyak mengkonsumdi resources yang meliputi manusia, biaya dan waktu yang tersedia. Di dalam tahapan ini terdapat data selection yang bertujuan sebagai pengidentifikasi semua sumber informasi internal dan eksternal dan memilih sebagian saja dari data yang akan diperlukan guna pengembangan dan penggunaan aplikasi data mining.

Disamping data selection yang ada pada tahapan data preparation, data preprocessing juga berguna dalam hal ini bertugas untuk meyakinkan kualitas dari data yang telah dipilih pada tahapan sebelumnya, tetapi dalam keadaan yang sebenarnya sering adanya permasalahan dalam penggunaannya yaitu meliputi *Noisy data* dan *Missing Values*. Ada lagi metode yang juga digunakan untuk membantu dalam pelaksanaak data preparation yaitu *Data Transformation* yang bertugas dan berfungsi untuk mengubah data ke dalam model analitis serta memodelkan data agar sesuai dengan analisa yang diharapkan dan format data yang diperlukan oleh algoritma data mining.

Ada bebrapa hal yang memang sangat perlu diperhatikan terlebih dahulu dalam menggunakan data transformation, antara lain adalah ***Categorial*** dan ***Quantitative***. ***Categorial*** adalah semua nilai yang mungkin ada, bersifat terbatas yang berdasarkan nominal dan ordinal. Yang dimaksud nominal di sini adalah tanpa adanya urutan sebagai contohnya adalah status perkawinan atau jenis kelamin. Sedangkan yang dimaksud dengan ordinal adalah dengan adanya urutan sebagai contohnya adalah rating loyalitas customer atau yang lainnya. ***Quantitative*** adalah semua nilai yang mungkin dapat diukur perbedaannya. Dalam hal ini ada dua hal yang dapat dijadikan pertimbangan, antara lain yaitu continuous dan discrete. Continuous adalah nilai-nilai bilangan real seperti gaji bulanan, rata-rata transaksi dalam satu periode waktu. Sedangkan discrete adalah nilai-nilai bilangan bulat seperti jumlah pegawai, jumlah transaksi dalam satu periode waktu.

Tahapan selanjutnya adalah Data mining, dalam tahapan ini adalah melakukan proses pencarian pengetahuan terhadap data yang ditransformasikan pada tahap sebelumnya. Dalam metode Data Mining telah di bagi menjadi 3 kategori yang meliputi *Automated Exploration/Discovery, Prediction/Classification, Explaination/Description.* Pada *Automated Exploration/Discovery* merupakan sebuah clustering misalnya dapat dipakai untuk mendapatkan segmentasi pasar baru dari suatu produk. Kategori yang kedua adalah *Prediction/Classification* adalah belajar suatu model predictive, ANN, GA, Regression dapat dipakai untuk peramalan nilai penjualan. Sedangkan pada kategori yang ketiga adalah *Explaination/Description* yang memiliki fungsi menjelaskan atau mendeskripsikan. Induksi Decision Tree dan Ekstraksi Rule dapat digunakan untuk menjelaskan perilaku customer berdasar demographic dan sejarah pembeliannya

Pada tahapan *Analysis Of Result* berfungsi untuk menginterpretasikan dan mengevaluasi output dari tahap mining: patterns. Pendekatan analisa yang digunakan akan bervariasi menurut operasi data mining yang digunakan, tetapi biasanya akan melibatkan teknik visualisasinya.

Tahapan yang terakhir adalah *Assimilation of Knowledge* yang berfungsi untuk menggunakan hasil mining yang telah dievakuasi ke dalam perilaku organisasi dan system informasi perusahaan.

Tahapan Proses KDD (menurut Daniel L. Silver) meliputi *Data Consolidation (and Cleansing), Data Selection and Preprocessing, Data Mining, Interpretation and Evaluation*.Pada tahapan Data Mining terdapat **Evaluasi**: dilakukan secara validasi statistik dan pengujian signifikan-nya; dilakukan review secara kualitatif hasil mining oleh pakar di bidangnya. **Interpretasi**: Decision tree dan rule dapat dibaca langsung; Clustering harus digambarkan dan ditabelkan. Pemakaian **visualization tools**: analisis sensitif; histogram untuk distribusi nilai; animasi dan plotting time-series. Sedangkan pada tahapan *Interpretation and Evaluation* terdapat **Evaluasi**: dilakukan secara validasi statistik dan pengujian signifikan-nya; dilakukan review secara kualitatif hasil mining oleh pakar di bidangnya. **Interpretasi**: Decision tree dan rule dapat dibaca langsung; Clustering harus digambarkan dan ditabelkan. Pemakaian **visualization tools**: analisis sensitif; histogram untuk distribusi nilai; animasi dan plotting time-series.

Tahapan Proses KDD (menurut Jia Weihan) meliputi Data cleaning, Data Integration, data selection, Data Transformation, Data Mining, *Pattern Evaluation*, *Knowledge Presentation*. **Data Cleaning**: menghilangkan noise dan data yang inkonsisten. **Data Integration**: menggabungkan berbagai macam sumber data. **Data Selection**: memilih data yang relevan (dari database) dengan "*analysis task*". **Data Transformation**: transformasi atau konsolidasi data ke dalam bentuk yang lebih baik untuk mining, dengan mewujudkan operasioperasi summary dan aggregation. **Data Mining**: mengekstrak patterns dari data dengan menerapkan "intelligent methods". **Pattern Evaluation**: mengidentifikasi sejumlah pola yang sungguh-sungguh menarik dan bakal menjadi pengetahuan berdasarkan sejumlah pengukuran ketertarikan (**interestingness measures**) seperti *rule support* dan *rule*

*confidence* untuk rule extraction. **Knowledge Presentation**: penggunaan teknik-teknik visualisasi dan representasi untuk menyajikan pengetahuan yang telah diperoleh kepada user.

**INPUT DAN OUTPUT DARI 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. Sebagai suatu rangkaian proses, data mining dapat dibagi menjadi beberapa tahap yang diilustrasikan di Gambar di bawah ini. 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. Intergrasi 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.

**TUGAS DAN FUNGSI DATA MINING**

1. Pembuatan laporan, Pembuatan laporan merupakan salah satu kegunaan data warehousing yang paling umum dilakukan. Dengan menggunakan query sederhana didapatkan laporan perhari,perbulan, pertahun atau jangka waktu kapanpun yang diinginkan.
2. On-Line Analytical Processing (OLAP), Dengan adanya data warehouse, semua informasi baik detail maupun hasil summary yang dibutuhkan dalam proses analisa mudah didapat. OLAP mendayagunakan konsep data multi dimensi dan memungkinkan para pemakai menganalisa data sampai mendetail, tanpa mengetikkan satupun perintah SQL. Hal ini dimungkinkan karena pada konsep multi dimensi, maka data yang berupa fakta yang sama bisa dilihat dengan menggunakan fungsi yang berbeda. Fasilitas lain yang ada pada sofware OLAP adalah fasilitas rool-up dan drill-down. Drill-down adalah kemampuan untuk melihat detail dari suatu informasi dan roll-up adalah kebalikannya.
3. Proses informasi eksekutif, data warehousing dapat membuat ringkasan informasi yang penting dengan tujuan membuat keputusan bisnis, tanpa harus menjelajahi keseluruhan data. Dengan menggunakan data warehousing segala laporan telah diringkas dan dapat pula mengetahui segala rinciannya secara lengkap, sehingga mempermudah proses pengambilan keputusan. Informasi dan data pada laporan data warehousing menjadi target informatif bagi pengguna.

Data mining bertujuan untuk mengekstrak pengetahuan dan wawasan melalui analisis data dalam jumlah besar dengan menggunakan teknik permodelan yang canggih. Data mining mengkonversi data menjadi pengetahuan dan informasi yang harus ditindaklanjuti. Data yang akan di analisis bisa saja berada di dalam data mart yang terorganisir dengan baik atau berada didalam data warehouse, atau mungkin juga diambil dari berbagai sumber data yang terstruktur.

Sebuah prosedur data mining memiliki banyak tahapan, biasanya melibatkan proses manajemen data yang luas sebelum pada akhirnya menerapkan statistik ataupun mesin algoritma pembelajaran  dan pengembangan model yang tepat. Saat ini sudah banyak software (data mining tools) yang dapat mendukung seluruh prosedur data mining. Sebuah data mining model terdiri dari seperangkat aturan, persamaan ataupun “transfer function” yang komplek, dimana ini dapat digunakan untuk mengidentifkasi pola data yang berguna, memahami dan memprediksi perilaku. Model data mining dapat dikelompokkan menjadi dua kelas utama sesuai dengan tujuannya, dua kelas utama itu adalah sebagai berikut:

**Supervised Model / Predictive Models**

Di dalam permodelan Supervised, atau Predictive,  atau Directed, atau Targeted, tujuannya adalah untuk meramalkan sebuah kejadian atau memperkirakan sebuah nilai dari atribut numerik yang berkelanjutan, di dalam model ini terdapat sebuah input field  atau disebut dengan atribut dan sebuah output atau disebut juga dengan target field. Input Field juga disebut dengan predictor karena input field digunakan oleh model untuk mengenali sebuah fungsi prediksi untuk output field. kita dapat membayangkan bahwa Predictor sebagai bagian dari fungsi X, dan target Field adalah bagian dari Y yang merupakan hasilnya.

Model ini menggunakan Input Field yang dianalisis sehubungan dengan pengaruhnya terhadap Target Field. Pengenalan Pola (pattern recognition) selalu “diawasi” / “supervised” oleh target field. Hubungan dibentuk antara Field Input dan Outputnya”. “Fungsi” pemetaan  input-output di generate oleh si modelnya, yang mengasosiasikan si predictor dengan si output dan memungkinkan prediksi pada nilai output dari nilai yang diberikan pada Input Field. Predictive Model selanjutnya dikategorikan kedalam model Klasifikasi (classification) dan model Estimasi (estimation)

**Classification Model atau Propensity Model (Model Klasifikasi atau Model Kecenderungan)**

Pada model ini, Target Groups (Kelompok sasaran) atau Class-nya sudah diketahui dari awal. tujuannya adalah untuk mengklasifikasi informasi kedalam kelompok-kelompok yang telah ditetapkan sebelumnya. dengan kata lain, model yang dihasilkan dapat digunakan sebagai mesin penilai yang menilai setiap informasi yang di proses masuk kategori kelompok mana. pekerjaan tersebut juga memperkirakan score kecenderungan pada setiap informasi, score kecenderungan menunjukkan kemungkinan terjadinya sebuah peristiwa yang sudah diklasifikasi sebelumnya.

**Estimation Model (model estimasi)**

Model ini mirip dengan model klasifikasi, tapi memiliki perbedaan utama, model ini digunakan untuk memprediksi sebuah nilai pada suatu field yang berkesinambungan berdasarkan nilai-nilai yang diamati dari input attributes.

**Unsupervised Model**

Pada unsupervised model atau undirected model, tidak ada namanya Output Field, yang ada cuma Input Field.  Pattern Recognition (pengenalan pola) tidak diarahkan dan tidak dipandu oleh target attribute tertentu. Model ini bertujuan untuk menemukan pola data didalam sekumpulan input field.

Unsupervised Model meliputi:

###### Cluster Model

Didalam model ini, “pengelompokan” sebelumnya tidak ada, sebaliknya kita ingin supaya algortma menganalis pola data yang di input dan mengidentifikasi data tersebut sesuai dengan kelompok alaminya (kita sebut saja kelompok alami ini dengan Cluster).  ketika ada data baru, maka data-data tersebut akan dikelompokkan kedalam cluster yang sudah terbentuk sebelumnya.

**Association Model dan Sequence Model**

Model ini tidak melibatkan prediksi langsung dari satu field, bahkan semua field yang terlibat memiliki peran ganda, karena field-field ini bertindak sebagai input dan output pada saat yang sama. Association Model mendeteksi asosiasi antara persitiwa diskrit, produk atau atribut. Model Sequence Mendeteksi asosiasi dari waktu ke waktu.

**TEXT MINING**

Text mining adalah salah satu bidang khusus dari *data mining*. Sesuai dengan buku *The Text Mining Handbook*, text mining dapat didefinisikan sebagai suatu proses menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan tools analisis yang merupakan komponenkomponen dalam *data mining* yang salah satunya adalah kategorisasi. Tujuan dari *text mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Jadi, sumber data yang digunakan pada *text mining* adalah kumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi terstruktur. Adapun tugas khusus dari *text mining* antara lain yaitu pengkategorisasian teks *(text categorization)* dan pengelompokan teks *(text clustering)*.

*Text mining* bisa dianggap subjek riset yang tergolong baru. Text mining dapat memberikan solusi dari permasalahan seperti pemrosesan, pengorganisasian / pengelompokkan dan menganalisa unstructured text dalam jumlah besar.

Dalam memberikan solusi, text mining mengadopsi dan mengembangkan banyak teknik dari bidang lain, seperti Data mining, Information Retrieval, Statistik dan Matematik, Machine Learning, Linguistic, Natural Languange Processing, dan Visualization. Kegiatan riset untuk text mining antara lain ekstraksi dan penyimpanan text, preprocessing akan konten text, pengumpulan data statistik dan indexing dan analisa konten.

Permasalahan yang dihadapi pada text mining sama dengan permasalahan yang terdapat pada data mining, yaitu jumlah data yang besar, dimensi yang tinggi, data dan struktur yang terus berubah, dan data *noise.* Perbedaan di antara keduanya adalah pada data yang digunakan. Pada data mining, data yang digunakan adalah *structured data*, sedangkan pada text mining, data yang digunakan text mining pada umumnya adalah *unstructured data*, atau minimal *semistructured*. Hal ini menyebabkan adanya tantangan tambahan pada text mining yaitu struktur text yang complex dan tidak lengkap, arti yang tidak jelas dan tidak standard, dan bahasa yang berbeda ditambah translasi yang tidak akurat.

**Koleksi Dokumen**

Salah satu elemen kunci dari *text mining* adalah kumpulan dokumen yang berbasis teks. Pada prakteknya, *text mining* ditujukan untuk menemukan pola dari sekumpulan dokumen yang jumlahnya sangat besar dan bisa mencapai jumlah ribuan bahkan sampai jutaan. Koleksi dokumen bisa statis, dimana dokumen tidak berubah, atau dinamis, dimana dokumen selalu di*update* sepanjang waktu.

**Representasi feature pada dokumen**

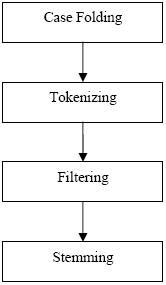
Algoritma yang digunakan pada text mining, biasanya tidak hanya melakukan perhitungan hanya pada dokumen, tetapi pada juga feature. Empat macam feature yang sering digunakan:

1. Character, merupakan komponan individual, bisa huruf, angka, karakter spesial dan spasi, merupakan block pembangun pada level paling tinggi pembentuk semantik feature, seperti kata,term dan concept.Pada umumnya, representasi character-based ini jarang digunakan pada beberapa teknik pemrosesan teks.
2. Words.
3. Terms merupakan single word dan frasa multiword yang terpilih secara langsung dari corpus. Representasi term-based dari dokumen tersusun dari subset term dalam dokumen.
4. Concept, merupakan feature yang di-generate dari sebuah dokumen secara manual, rule-based, atau metodologi lain. Pada tugas akhir ini, concept di-generate dari argument atau verb yang sudah diberi label pada suatu dokumen.

**Text Preprocessing**

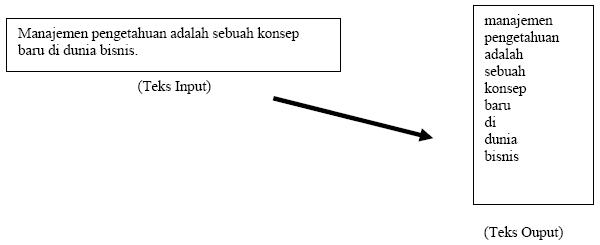
Struktur data yang baik dapat memudahkan proses komputerisasi secara otomatis. Pada text mining, informasi yang akan digali berisi informasi-informasi yang strukturnya sembarang. Oleh karena itu, diperlukan proses pengubahan bentuk menjadi data yang terstruktur sesuai kebutuhannya untuk proses dalam *data mining*, yang biasanya akan menjadi nilai-nilai numerik. Proses ini sering disebut Text Preprocessing. Setelah data menjadi data terstruktur dan berupa nilai numerik maka data dapat dijadikan sebagai sumber data yang dapat diolah lebih lanjut.

**Ekstraksi Dokumen**

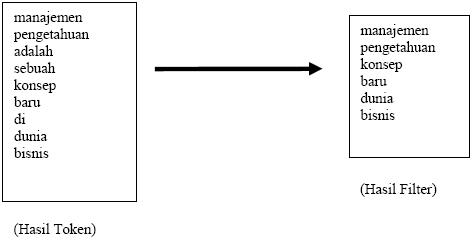


Teks yang akan dilakukan proses *text mining*, pada umumnya memiliki beberapa karakteristik diantaranya adalah memiliki dimensi yang tinggi, terdapat *noise* pada data, dan terdapat struktur teks yang tidak baik. Cara yang digunakan dalam mempelajari suatu data teks, adalah dengan terlebih dahulu menentukan fitur-fitur yang mewakili setiap kata untuk setiap fitur yang ada pada dokumen. Sebelum menentukan fitur – fitur yang mewakili, diperlukan tahap pre processing yang dilakukan secara umum dalam *text mining* pada dokumen, yaitu case folding, tokenizing, filtering, stemming, tagging dan analyzing.

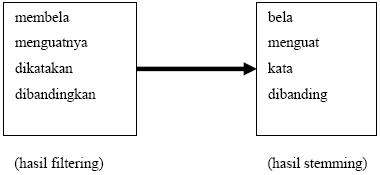
*Case folding* adalah mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf ‘a’ sampai dengan ‘z’ yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter. Tahap *tokenizing / parsing* adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya. Contoh dari tahap ini adalah sebagai berikut:



Tahap *filtering* adalah tahap mengambil kata - kata penting dari hasil token. Bisa menggunakan algoritma *stoplist* (membuang kata yang kurang penting) atau *wordlist* (menyimpan kata penting). *Stoplist / stopword* adalah katakata yang tidak deskriptif yang dapat dibuang dalam pendekatan bag-of-words. Contoh stopwords adalah “yang”, “dan”, “di”, “dari” dan seterusnya. Contoh dari tahapan ini adalah sebagai berikut:



Tahap *stemming* adalah tahap mencari *root* kata dari tiap kata hasil *filtering.* Pada tahap ini dilakukan proses pengembalian berbagai bentukan kata ke dalam suatu representasi yang sama. Tahap ini kebanyakan dipakai untuk teks berbahasa inggris dan lebih sulit diterapkan pada teks berbahasa Indonesia. Hal ini dikarenakan bahasa Indonesia tidak memiliki rumus bentuk baku yang permanen. Contoh dari tahapan ini pada teks berbahasa inggris adalah sebagai berikut:



**Kategorisasi Teks**

Kategorisasi 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.

Pada kategorisasi teks, diberikan sekumpulan kategori (label) dan koleksi dokumen yang berfungsi sebagai data latih, yaitu data yang digunakan untuk membangun model, dan kemudian dilakukan proses untuk menemukan kategori yang tepat untuk dokumen test, yaitu dokumen yang digunakan untuk menentukan akurasi dari model. Misalkan ada sebuah dokumen *x* sebagai inputan, maka output yang dihasilkan oleh model tersebuat adalah kelas atau kategori *y* dari beberapa kategori tertentu yang telah didefinisikan sebelumnya (*y1,…,yk*). Adapun contoh dari pemanfaatan kategorisasi teks adalah pengkategorisasian berita ke dalam beberapa kategori seperti bisnis, teknologi, kesehatan dan lain sebagainya; pengkategorisasian email sebagai spam atau bukan; pengkategorisasian kilasan film sebagai film favorit, netral atau tidak favorit; pengkategorisasian *paper* yang menarik dan tidak menarik; dan penggunaan dari kategorisasi teks yang paling umum adalah kategorisasi otomatis dari *web pages* yang dimanfaatkan oleh portal Internet seperti Yahoo. Kategorisasi otomatis ini memudahkan proses *browsing* artikel berdasarkan topik tertentu yang dilakukan oleh *user*. Salah satu algoritma kategorisasi yang sering digunakan adalah algoritma Naive bayes multinomial. Algoritma ini merupakan algoritma yang menerapkan metode probabilistic learning method.

**WEB MINING**

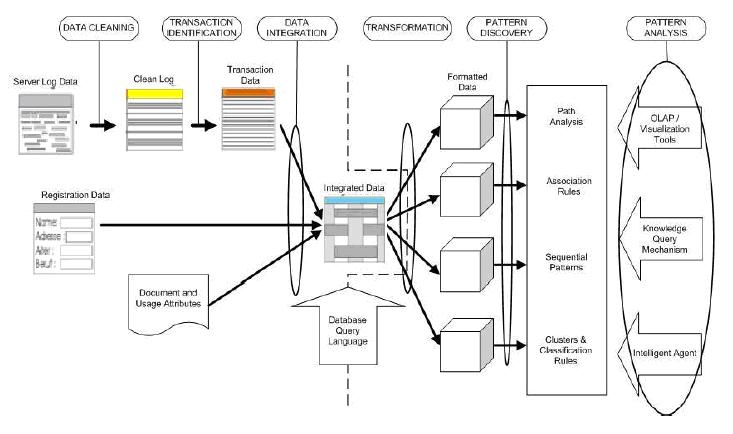
Web mining adalah ekstraksi pola-pola penting dan bermanfaat namun tersimpan secara implisit pada kumpulan data yang relatif besar pada layanan world wide web. Web mining teridiri atas tiga bagian yaitu: web content mining, web structure mining, dan web usage mining [2].

Web content mining adalah suatu proses otomatis untuk menemukan informasi yang berguna dari dokumen atau data. Pada prinsipnya teknik ini mengekstraksi kata kunci yang terkandung pada dokumen. Isi data web antara lain dapat berupa teks, citra, audio, video, metadata, dan hyperlink. Ada dua strategi yang umum digunakan: pertama langsung melakukan mining terhadap data, dan kedua melakukan pencarian serta mengimprove hasil pencarian seperti layaknya search engine.

Web struncture mining dikenal juga  sebagai web log mining adalah teknik yang digunakan untuk menemukan struktur link dari hyperlink dan membangun rangkuman website dan halaman  web. Salah satu manfaatnya adlah untuk  menentukan pagerank pada suatu halaman web.

Web usage mining adalah teknik untuk mengenali perilaku pelanggan dan struktur web melalui informasi yang diperoleh dari log, click stream, cookies, dan query. Berbagai tool yang sudah ada antara lain WebLogMiner yang melakukan mining terhadap data log. Teknik yang lebih canggih digunakan untuk melakukan OLAP. Manfaat web usage mining adalah untuk kustomosasi halaman berdasarkan profil pengguna, menentukan ketertarikan pelanggan terhadap produk tertentu, dan menentukan target market yang sesuai.

Tahapan *web mining* dapat diilustrasikan dengan gambar berikut:



Gambar. Tahapan pada *web mining*

Pada gambar tahapan pada *web mining* dibagi menjadi tiga kelompok yaitu *preprocess*, *process*, dan, *post process*. Tahapan *preprocess* meliputi data *cleaning*, *transaction identification*, *integration*, dan *transformation*. Pada tahap *process* diterapkan sejumlah formulasi statistik antara lain untuk mengurangi jumlah atribut dengan cara membuang atribut yang tidak berpengaruh *(information gain)*. Pada tahap ini dapat juga dilakukan teknik *clustering*, asosiasi, dan klasifikasi. Pada tahap *post processing* dilakukan analisis lebih lanjut untuk mengolah hasil *mining* pada tahapan sebelumnya. Hal ini perlu dilakukan sebab sering sekali hasil yang diperoleh pada tahap *process* tidak memberikan sesuatu yang dapat digunakan secara langsung, sehingga diperlukan teknik lainnya seperti visualisasi grafik dan analisis statistik lainnya.

***Information Gain***

Tahapan *process* merupakan tahapan utama dalam *web mining*. Pada tahap ini, atribut yang akan diolah harus diminimalisasi terlebih dahulu dengan tujuan untuk membuang atirbut yang tidak perlu sehingga hanya atribut yang mempunyai relevansi kuat yang akan diproses, sehingga efisiensi *space* dan waktu dapat dicapai dan kualitas data yang dihasilkan lebih baik. Salah satu cara untuk mereduksi atribut adalah dengan *information gain*.

*Information gain* adalah pengukuran yang dilakukan untuk menseleksi atribut. Pengukuran ini dipelopori oleh Claude Shannon pada teori informasi [1] dituliskan sebagai:

formula1

Dalam hal ini pi adalah probabilitas sebuah *tuple* pada D masuk ke kelas Ci dan diestimasi dengan | Ci,D|/|D|. Fungsi log diambil berbasis 2 karena informasi dikodekan berbasis bit.

Misal S adalah himpunan beranggotakan s data. Misalkan label atribut yang mendefinisikan kelas memiliki m nilai berbeda yang mendefinisikan m kelas Ci (untuk i=1,2,..,m) berbeda. Misalkan Si adalah jumlah sampel S yang masuk ke dalam kelas Ci. Berdasarkan (1) maka informasi yang diharapkan perlu untuk mengklasifikasi suatu sampel yang diberikan adalah:

formula2

Yang dalam hal ini pi adalah peluang suatu sampel sembarang masuk ke kelas Ci dan ditaksir dengan si/s.

Misalkan atribut A yang bukan atribut suatu kelas, memiliki v buah nilai yang berbeda yaitu {a1, a2, …,av}. Atribut A dapat digunakan untuk memisahkan S ke dalam v subset Sj yaitu {S1, S2, …,Sv}, yang dalam hal ini Sj memuat sampel-sampel di dalam S yang memiliki nilai aj dari A. Jika atribut A dipilih sebagai atribut uji (atribut yang terbaik untuk pemisahan) maka *subset*-*subset* tersebut akan berhubungan dengan cabang-cabang yang turun dari simpul yang memuat himpunan S. Misalkan Sij adalah jumlah sampel dari kelas Ci di dalam suatu *subset* Sj. Entropi atau informasi harapan berdasarkan pemisahan ke dalam *subset-subset* A dihitung dengan:

*formula3*

Bentuk formula4bertindak sebagai pemberat dari *subset* j dan merupakan jumlah sampel di dalam *subset* tersebut (yang memiliki nilai aj dari A) dibagi dengan jumlah total sampel di dalam S. Semakin kecil nilai entropi maka semakin murni pemisahan *subset*. Untuk suatu *subset* Sj informasi harapan berdasarkan (2) adalah:

*formula5*

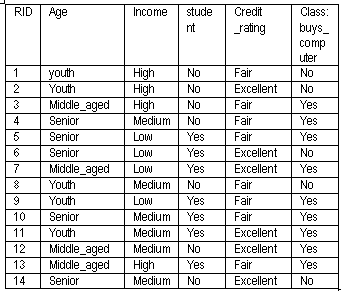
Yang dalam hal ini pij = sij / |sj| dan merupakan peluang dari suatu sampel di dalam Sj termasuk kelas Ci.

Selanjutnya, *Information Gain* atribut A diperoleh dengan perhitungan:

*Gain(A) = I(s1,s2,…,sm) – E(A) ……..(4)*

Dengan penjelasan lain, Gain(A) adalah reduksi yang diharapkan di dalam entropi yang disebabkan oleh pengenalan nilai atribut dari A. Atribut yang memiliki nilai *information gain* terbesar dipilih sebagai uji atribut untuk himpunan S. Selanjutnya suatu simpul dibuat dan diberi label dengan label atribut tersebut, dan cabang-cabang dibuat untuk masing-masing nilai dari atribut.

Berikut diberikan sebuah contoh untuk menghitung information gain:

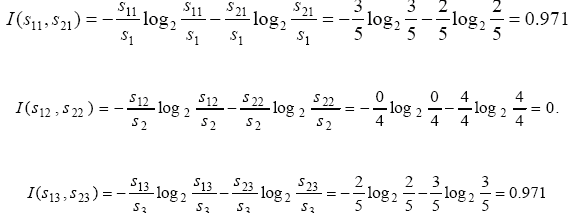


Gambar. Contoh pereduksian atribut

Jika atribut buys\_computer diambil sebagai label kelas, maka ada sebanyak m=2 kelas berbeda {C1=no, C2=yes}. Jumlah record pada data set adalah sebanyak 14 record sehingga s=14. Dengan demikian ada sebanyak s1=5 sampel dari S di kelas C1 dan sebanyak s2=9 sampel dari S di kelas C2. Dengan demikian informasi harapan untuk mengelompokkan kelas tersebut adalah:

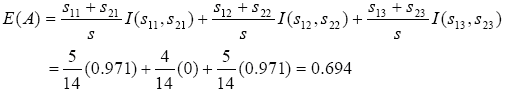
formula6

Misalkan atribut A yang akan ditinjau dengan age, dan A memiliki 3 nilai berbeda, sehingga v=3 {a1=‘<30‘, a2=’30-40‘,a3=‘>40‘}. Selanjutnya atribut A tersebut dapat digunakan untuk memisahkan S ke dalam 3 subset yaitu {S1 untuk ‘<30‘, S2 untuk ’30-40‘, S3 untuk ‘>40‘}. Dengan demikian S11=3 adalah jumlah sampel dari kelas C1=no di dalam subset S1 untuk ‘<30‘. Selanjutnya S21=2 adalah jumlah sampel dari kelas C2=yes di dalam subset S1 untuk ‘<30‘.Selanjutnya S12=0 adalah jumlah sampel dari kelas C1=no di dalam subset S2 untuk ’30-40‘. Selanjutnya S22=4 adalah jumlah sampel dari kelas C2=yes di dalam subset S2 untuk ’30-40‘. Selanjutnya S13=2 adalah jumlah sampel dari kelas C1=no di dalam subset S3 untuk ’>40‘.Selanjutnya S23=3 adalah jumlah sampel dari kelas C2=yes di dalam subset S3 untuk ’>40‘. Sehingga s=s11+s21+s22+s13+s23 = 14 sampel. Kalau sj adalah jumlah sampel data pada masing-masing subset Sj, maka informasi harapan untuk masing-masing subset Sj adalah sebagai berikut:



Yang dalam hal ini, s1=s11+s21 = 3+2 dan s2=s12+s22 = 0+4 = 4 dan s3=s13+s23 = 2+3 = 5.

Entropi yang merupakan informasi harapan berdasarkan pemisahan kedalam subset-subset A dihitung sebagai:



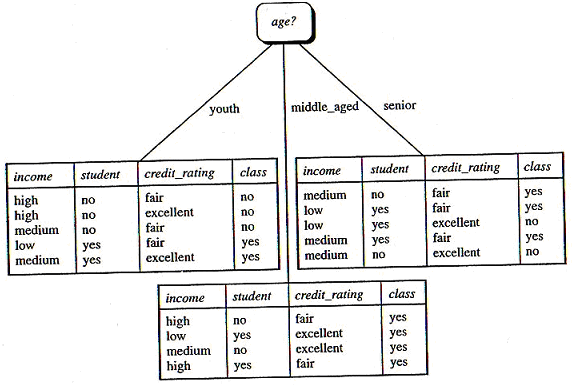
Dengan demikian, *informatian gain* untuk atribut age adalah:

Gain(age) = I(s1,s2) – E(age)

= 0.94-0.694 = 0.246

Cara yang sama dilakukan terhadap atribut income, student, dan credit\_rate dan diperoleh nilai masing-masing attribut adalah Gain(income)=0.029, dan Gain(student)=0.151, dan Gain(credit\_rate)=0.046.

Didefinisikan sebuah nilai ambang *(threshold)* sebagai nilai batas untuk menentukan atibut-atribut apa saja yang akan digunakan. Atribut yang akan digunakan adalah atribut yang memiliki nilai Gain(atribut) >= *threshold*. Sebagai contoh, bila nilai *threshold* yang diberikan adalah 0.1 maka atribut yang akan digunakan adalah age dan student, berurutan memiliki gain sebesar 0.246 dan 0.151. Hasilnya seperti *decicion tree* gambar berikut.



Gambar. Decision Tree

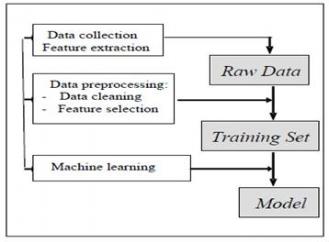
**MULTIMEDIA MINING**

Multimedia mining adalah salah satu bagian ilmu data mining untuk menambang/menggali pengetahuan dari informasi atau data multimedia. Akan tetapi multimedia mining tidak sekedar perluasan dari data mining, karena merupakan upaya interdisipliner yang memanfaatkan keahlian dalam multimedia retrieval, pengolahan data multimedia, computer vision, machine learning, dan kecerdasan buatan.  Multimedia mining berkaitan dengan ekstraksi pengetahuan implisit, relasi data multimedia, atau pola lain yang tidak secara eksplisit disimpan dalam file multimedia.

Sistem multimedia mining dapat secara otomatis mengekstraksi informasi semantik sebagai pengetahuan dari file multimedia.  Umumnya, sistem database multimedia mengelola koleksi besar objek multimedia, seperti image, video, audio dan data hypertext.

Secara umum, file multimedia dari database harus  mengalami praproses sesuai dengan format data yang sesuai, selanjutnya mengalami berbagai transformasi dengan ekstraksi fitur untuk menghasilkan fitur penting dari file multimedia.  Dengan fitur yang dihasilkan, mining dapat dilaksanakan dengan menggunakan teknik data mining untuk menemukan pola signifikan untuk kemudian dievaluasi dan diinterpretasikan untuk mendapatkan pengetahuan yang diinginkan.

Proses aplikasi multimedia mining dapat dilihat pada Gambar 1 berikut (Kotsiantis et al, 2004).

[](http://altien.files.wordpress.com/2009/12/mm_mining1.jpg)

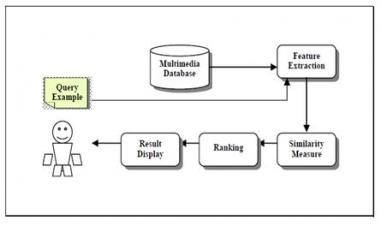
Gambar.  Proses multimedia mining

Pengumpulan data adalah titik awal dari sebuah pembelajaran sistem, sehingga kualitas data mentah dicapai pada kinerja secara keseluruhan.  Disini akan dilakukan ekstraksi fitur yang sesuai dengan tipe data yang ada (teks, gambar, audio, video), misalkan untuk data teks dengan tokenisasi, dan data audio dilihat dari pitch atau frekuensi audio.  Kemudian, tujuan dari pra-proses data adalah untuk menemukan fitur penting atau seleksi  fitur dari data mentah. Pra-proses data meliputi pembersihan data, normalisasi, transformasi, seleksi fitur, dll.  Proses pembelajaran bisa cepat, jika informatif fitur dapat diidentifikasi pada tahap pra-proses.  Hasil dari pra-proses data adalah training set. Jika diberikan sebuah training set, sebuah model pembelajaran harus memilih untuk belajar dari itu. Kemudian dengn machine learning dapat diperoleh model yang diinginkan dengan cara klasifikasi atau kluster.

**MULTIMEDIA RETRIEVAL**

Jika multimedia mining adalah bagian ilmu dari data mining, maka multimedia retrival adalah bagian dari ilmu temu kembali informasi yang menfokuskan pada temu kembali informasi dari data multimedia.  Secara umum, teknik-teknik yang dilakukan sama pada temu kembali informasi untuk data bukan multimedia.  Berbeda dengan data yang hanya berupa numeric atau teks, karena data multimedia bisa apa saja (teks, audio, image dan video) maka multimedia retrieval adalah model temu kembali informasi berbasis konten (content-base infromastion retrieval).  Disini yang bisa menjadi query tidak hanya teks atau angka saja, bisa audio atau gambar tergantung jenis data yang akan dicari.

Wei dan Li, 2004, menyajikan arsitektur untuk data multimedia berdasarkan content-base infromastion retrieval, sebagai berikut :

[](http://altien.files.wordpress.com/2009/12/mm_retriev.jpg)

Gambar. Arsitektur content-base information retrieval untuk data multimedia

Dalam content-base information retrieval system, konten media dalam database diekstrak dan dideskripsikan oleh vector fitur multi-dimensi, atau disebut deskriptor. Vektor fitur media merupakan fitur dataset. Untuk mengambil data yang diinginkan, pengguna mengirimkan contoh query untuk system temu kembali. Sistem kemudian merepresentasikan contoh-contoh ini dengan vector fitur. Jarak (yaitu, dalam arti ukuran kesamaan) antara vektor fitur dari contoh query dan yang ada di media dalam fitur dataset kemudian dihitung dan dirangking. Temu kembali dilakukan dengan menerapkan suatu skema pengindeksan untuk memberikan cara yang efisien untuk mencari database media. Akhirnya, sistem merangking hasil pencarian dan kemudian mengembalikan hasil pencarian teratas yang paling mirip dengan contoh query.

Untuk content-base information retrieval system, seorang perancang harus mempertimbangkan empat aspek: fitur ekstraksi dan representasi, dimensi reduksi fitur, pengindeksan, dan spesifikasi query.

**PERBEDAAN MULTIMEDIA MINING DAN MULTIMEDIA RETRIEVAL**

Yang dilakukan dalam multimedia retrieval adalah bagaimana mengekstraksi pengetahuan implisit, relasi data multimedia, atau pola lain yang tidak secara eksplisit disimpan dalam file multimedia (discovering), dibanding pada multimedia retrieval yang hanya untuk merangking query yang dicari untuk disajikan ke user (searching).

Multimedia retrieval hanya menekankan pada temu kembali data/informasi yang sesuai dengan query yang diminta.  Sedangkan multimedia mining mengelola data/informasi yang sesuai dengan proses analisa yang diterapkan agar diperoleh informasi yang lebih banyak (pengetahuan) berdasarkan query yang diminta.

Karena data perlu  dianalisa berdasarkan teori dalam data mining, maka dalam multimedia mining harus melakukan cleaning data dan seleksi fitur untuk membuat analisanya menjadi lebih cepat dan sederhana, artinya data yang “tidak baik” dibuang saja dan fitur yang tidak signifikan mempengaruhi tidak perlu diikutkan dalam analisa.  Sedangkan pada multimedia retrieval, hal ini tidak dilakukan.