Materi 3

DATA MINING

3 SKS | Semester 6 | S1 Sistem Informasi | UNIKOM | 2015
Nizar Rabbi Radliya | nizar.radliya@yahoo.com

Nama Mahasiswa	
NIM	
Kelas	

Kompetensi Dasar

Memahami definisi set data, tipe data, kualitas data, serta similaritas dan dissimaliritas.

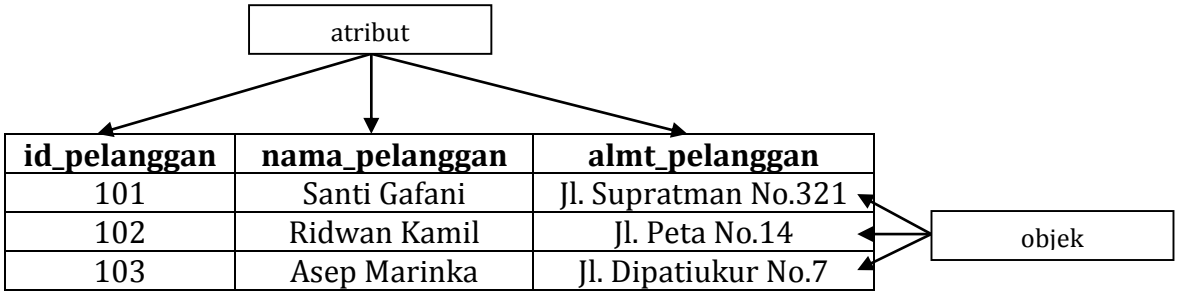
Pokok Bahasan

Set Data:
1. Definisi dan tipe data
2. Kualitas data
3. Similaritas dan dissimilaritas

I. Definisi Set Data

Set data (*data set*/himpunan data) merupakan kumpulan objek dan atributnya. Nama lain dari objek yang sering digunakan diantaranya *record, point, vector, pattern, event*, *observation*, *case*, *sample*, *instance*, entitas. Objek digambarkan dengan sejumlah atribut yang menerangkan sifat atau karakteristik dari objek tersebut. Atribut juga sering disebut variabel, *field*, fitur, atau dimensi. Atribut adalah sifat/properti/karakteristik objek yang nilainya bisa bermacam-macam dari satu objek dengan objek lainnya, dari satu waktu ke waktu yang lainnya.

Sebagai contoh seorang pelanggan merupakan objek, dimana objek pelanggan tersebut memiliki beberapa atribut seperti id pelanggan, nama, alamat dan lain-lain. Setiap pelanggan memungkinkan memiliki nilai atribut yang berbeda dengan pelanggan lainnya, serta memungkinkan perubahan nilai atribut dari waktu ke waktu.



Gambar 1. Perbedaan Atribut dan Objek

II. Tipe Data

Tipe atribut dapat dibedakan dari nilai beserta sifatnya. Ada empat sifat yang dimiliki atribut secara umum, yaitu:

1. Pembeda (*distinctness*): = dan ≠
2. Urutan (*order*): <, >, ≤, ≥
3. Penjumlahan, Pengurangan (*addition*): + dan –
4. Perkalian, Pembagian (*multiplication*): * dan /

Umumnya tipe atribut ini ada dua yaitu kategori (kualitatif) dan numerik (kuantitatif). Dari kedua tipe tersebut dibagi lagi menjadi beberapa sub tipe yang disesuaikan dengan sifat nilai yang dimilikinya.

Tabel 1. Tipe Atribut

Tipe Atribut		Penjelasan	Contoh
Kategoris (Kualitatif)	Nominal	Nilai atribut berupa nominal memberikan nilai berupa nama. Dengan nama inilah sebuah atribut membedakan dirinya pada data yang satu dengan yang lain (=, ≠).	Kode Pos, Jenis Kelamin.
	Ordinal	Nilai atribut bertipe ordinal mempunyai nilai berupa nama yang mempunyai arti informasi terurut (<, >, ≤, ≥).	Indek Nilai (A, B, C, D, E)
Numerik (Kuantitatif)	Interval	Nilai atribut dimana perbedaan diantara dua nilai mempunyai makna yang berarti (+, -).	Tanggal
	Rasio	Nilai atribut dimana perbedaan diantara dua nilai dan rasio dua nilai mempunyai makna yang berarti (*, /)	Panjang, berat, tinggi

Atribut nominal dan ordinal merupakan tipe kategoris, nilainya kualitatif; dimana nilai tersebut sebenarnya simbolik; tidak mungkin dilakukan operasi aritmatika. Sedangkan interval dan rasio merupakan tipe numerik, nilainya kuantitatif; dimana nilai tersebut dapat dilakukan operasi aritmatika; bisa direpresentasikan dengan nilai integer atau kontinu.

Sementara berdasarkan jumlah nilainya, atribut dapat dibedakan menjadi dua, yaitu:

1. Diskret

Sebuah atribut dapat bernilai diskret jika mempunyai nilai dalam himpunan jumlah yang terbatas. Jenis ini bisa ditemukan pada atribut kategoris yang hanya mempunyai beberapa variasi nilai (domain), seperti indek nilai yang hanya mempunyai lima

kemungkinan nilai (A, B, C, D, E). Contoh lainnya adalah jenis kelamin (pria, wanita), benar/salah, ya/tidak, 0/1.

2. Kontinu

Sedangkan atribut yang bernilai kontinu akan mempunyai jangkauan nilai real. Seperti variabel panjang, tinggi, berat dimana nilainya biasanya menggunakan representasi *floating point* (desimal). Namun, meskipun menggunakan representasi real, ukuran presisi jumlah angka di belakang koma tetap digunakan.

III. Karakteristik Set Data

Ada tiga karakteristik umum set data yang mempunyai pengaruh besar dalam data mining, yaitu dimensionalitas, sparsitas, resolusi. Berikut adalah penjelasan dari ketiga karakteristik tersebut:

1. Dimensionalitas (*dimensionality*)

a. Dimensionalitas dapat diartikan sebagai jumlah atribut yang dimiliki oleh objek-objek dalam data set.

b. Data dengan jumlah dimensi yang sedikit (rendah) punya kecendrungan berbeda secara kualitatif dengan data dalam kontek yang sama, tetapi dengan jumlah dimensi yang lebih banyak (tinggi).

c. Kesulitan yang berhubungan dengan data dimensi tinggi sering disebut sebagai *curse of dimensionality*.

d. Untuk itu pada tahap *preprocessing* (proses awal) perlu dilakukan pengurangan dimensi (*dimensionality reduction*)

2. Sparsitas (*sparsity*)

a. Untuk set data dengan fitur asimetrik (jumlah fitur yang terisi nilai tidak sama antara satu data dengan data yang lain), banyak atribut data mempunyai nilai 0 di dalmnya; dalam banyak kasus, kurang dari 1% mempunyain nilai bukan 0.

b. Dalam praktiknya, tentu ini menguntungkan karena komputasi menjadi lebih ringan (cepat) dan kapasitas penyimpanan juga lebih sedikit.

3. Resolusi (*resolution*)

a. Untuk data yang digambarkan dalam bentuk grafik yang memerlukan koordinat spasial, karakteristik resolusi yang digunakan juga akan berpengaruh.

b. Pola dalam data bergantung pada level resolusi.

c. Jika resolusi terlalu baik (tidak ada perbedaan/halus), pola mungkin tidak akan kelihatan, jika resolusi terlalu kasar atau sempit, pola juga akan hilang.

IV. Jenis Set Data

Jenis data dapat dibedakan menjadi tiga kelompok, yaitu data record, data berbasis grafik (*graph data*), dan data terurut (*ordered data*). Pada perkuliahan ini kita hanya akan menggunakan data record untuk proses data mining.

Kebanyakan metode data mining mengasumsikan bahwa set data yang diproses adalah kumpulan baris data (*record*/*entries*/*objects*), dimana setiap barisnya terdiri atas sejumlah fitur (atribut) yang tetap. Dalam set data berbentuk data record, tidak ada hubungan antara baris data dengan baris data yang satu dengan baris data yang lainnya dan juga dengan set data yang lain. Setiap baris data berdiri sendiri sebagai sebuah data individu.

Dalam sistem basis data, umumnya ada sejumlah tabel yang saling berhubungan menggunakan suatu kunci (kunci utama, kunci tamu). Akan tetapi dalam set data record, diasumsikan bahwa hanya ada satu tabel yang berisi sejumlah baris data. Oleh karena itu, biasanya set data yang diolah dalam data mining adalah keluaran dari sistem data warehouse yang menggunakan *query* untuk melakukan pengambilan data dari sejumlah tabel dalam sistem basis data. Ada beberapa contoh set data yang masuk dalam jenis data record, diantaranya data matrik (*matrix data*), data keranjang belanja (*market basket data*), dan data dokumen.

1. Data Matrik

Jika set data berisi kumpulan data yang mempunyai sejumlah atribut (fitur) numerik yang sama, set data tersebut dapat dipandang sebagai vektor (data) dalam wilayah multidimensi, dimana masing-masing dimensi menyatakan satu atribut yang berbeda. Sekumpulan data matrik dapat diinterpretasikan sebagai mantrik M x N, dimana M adalah jumlah baris (satu baris menyatakan satu record/objek) dan N adalah jumlah kolom (dimana satu kolom menyatakan satu atribut/fitur). Contoh data matrik dapat dilihat pada tabel 2 dibawah ini.

Tabel 2. Contoh Data Matrik

Tinggi	Berat	Sepatu	Celana
168	60	38	30
175	85	42	35
170	77	39	39

2. Data Keranjang Belanja/Transaksi

Data keranjang belanja (data transaksi) adalah set data yang setiap recordnya berisi sejumlah item dan jumlah item untuk sebuah transaksi bisa berbeda dengan transaksi yang lain. Contohnya bisa dilihat pada kasus keranjang belanja di pasar atau supermarket, dimana setiap pembeli melakukan pembelian barang yang jumlah dan jenisnya bisa berbeda dengan pembeli yang lain. Contoh data transaksi dapat dilihat pada tabel 3 dibawah ini.

Tabel 3. Contoh Data Transaksi

TID	Item
1	Susu, Bedak, Sabun
2	Susu, Mentega
3	Bedak, Gula, Sabun, Roti

3. Data Dokumen

Setiap dokumen merupakan satu vektor 'term'. Tiap term merupakan satu komponen (atribut) dari vektor tersebut. Nilai dari setiap komponen menyatakan berapa kali kemunculan term tersebut dalam suatu dokumen. Contoh dari data dokumen dapat dilihat pada tabel 4 dibawah ini.

Tabel 4. Contoh Data Dokumen

	Sistem	Informasi	Data	Proses
Makalah 1	18	19	15	9
Makalah 2	6	6	4	8
Makalah 3	14	0	2	12

V. Kualitas Data

Permasalahan kualitas data ditinjau dari aspek pengukuran data dan pengumpulan data.

1. Kesalahan Pengukuran (*Measurement Error*)

Kesalahan ini mengacu pada permasalahan hasil dari proses pengukuran. Problem yang umum terjadi adalah nilai yang dicatat berbeda dari nilai sebenarnya untuk beberapa tingkat. Pada atribut kontinu, beda numerik dari hasil pengukuran dengan nilai sebenarnya disebut dengan *error*. Yang termasuk dalam jenis kesalahan pengukuran adalah:

a. *Noise*

Merupakan komponen random dari suatu error pengukuran. Noise berkaitan dengan modifikasi dari nilai asli. Contoh: distorsi atau penyimpangan dari suara orang saat berbicara di telepon yang jaringannya buruk.

b. Bias

Suatu variasi pengukuran dari kuantitas yang sedang diukur dengan pengurangan antara mean dan nilai kuantitas yang diketahui.

c. *Precision*

Kedekatan dari pengukuran berulang (dari kuantitas yang sama) satu dengan yang lainnya. Diukur dengan standar deviasi.

d. *Accuracy*

Kedekatan pengukuran terhadap nilai sebenarnya dari kuantitas yang sedang diukur.

Contoh kasus:

Terdapat berat standar laboratorium suatu benda adalah 1 gram dan kita akan menghitung *precision* dan bias dari skala benda dari hasil pengukuran yang baru. Kita melakukan pengukuran sebanyak lima kali dan memperoleh {1.015, 0.990, 1.013, 1.001, 0.986}.

Jawab:

Mean = (1.015 + 0.990 + 1.013 + 1.001 + 0.986) / 5 = 1.001

Bias = 1.001 – 1 = 0.001

Precision=

$$\sqrt{\frac{(1.015-1.001)^2+(0.990-1.001)^2+(1.013-1.001)^2+(1.001-1.001)^2+(0.986-1.001)^2}{4}}$$

2. Kesalahan Pengumpulan (Collection Error)

Mengacu pada kesalahan-kesalahan (*error*) seperti hilangnya objek data atau nilai atribut, atau lingkup objek data yang tidak tepat. Yang temasuk dalam kesalahan pengumpulan diantaranya:

a. *Outliers*

Merupakan objek data dengan sifat yang berbeda sekali dari kebanyakan objek data dalam data-set. Misalkan, terdapat data penelitian tentang tinggi anak siswa SMA yakni 160cm sampai 180cm. Tetapi dalam data tersebut terdapat anak yang mempunyai tinggi

140cm. Data anak dengan tinggi 140cm tersebut yang disebut data outlier, karena berbeda sangat jelas.

Terdapat beberapa hal yang mempengaruhi munculnya data outlier antara lain:

1) Kesalahan dalam pemasukan data
2) Kesalahan dalam pengambilan *sample*
3) Memang ada data-data ekstrim yang tidak dapat dihindarkan keberadaannya.

b. Missing value

Merupakan nilai dari suatu atribut yang tidak ditemukan. Asalannya terjadinya *missing value* adalah:

1) Informasi tidak diperoleh (misal, orang-orang menolak untuk memberikan data umur dan berat bedan)
2) Atribut yang mungkin tidak bisa diterapkan ke semua kasus (misal, pendapatan tahunan tidak bisa diterapkan pada seseorang yang pengangguran)

Penanganan *missing values* adalah dengan:

1) Mengurangi objek data
2) Memperkirakan *missing values*
3) Mengabaikan *missing values* pada saat analisis
4) Mengganti dengan semua nilai yang mungkin (tergantung probabilitasnya)

c. *Duplicate data*

Set data mungkin terdiri dari objek data yang ganda (duplikat), atau hampir selalu terjadi duplikasi antara satu dengan yang lainnya. Persoalan utama ketika menggabungkan data dari sumber-sumber yang bervariasi (heterogen). Contoh: orang yang sama dengan alamat email yang lebih dari satu. Pembersihan data (*data cleaning*) merupakan proses yang berkaitan dengan permasalahan data yang duplikat.

VI. Similaritas dan Dissimilaritas

Kemiripan (*similarity*) adalah ukuran numerik dimana dua objeknya mirip, nilai 0 jika tidak mirip dan nilai 1 jika mirip penuh. Sementara ketidakmiripan (*dissimilarity*) adalah derajat numerik dimana dua objek yang berbeda, jangkauan nilai 0 sampai 1 atau bahkan sampai ∞.

4.1. Kemiripan dan Ketidakmiripan Data Satu Atribut

Istilah ketidakmiripan juga dapat disebut sebagai ukuran jarak (*distance*) antara dua data. Jika s adalah ukuran kemiripan dan d adalah ukuran ketidakmiripan, serta jika interval/range nilainya adalah [0,1], maka dapat dirumuskan bahwa s+d=1. Sebenarnya ukuran kemiripan dan ketidakmiripan tidak harus selalu dalam interval [0,1], tetapi boleh juga menggunakan interval seperti [0,10], [0,100], bahkan menggunakan nilai negative seperti [-1,1], [-10,10] dan sebagainya. Transformasi nilai s dan d tidak hanya terbatas pada formula s+d=1, karena ada juga yang menggunakan s = $\frac{1}{1+d}$ atau s = e^{-d}.

Pada metode tertentu dalam klasifikasi, ada juga yang mengharuskan agar nilai interval ketidakmiripan data harus ditransformasi dalam interval yang ternormalisasi [0,1]. Sebagai contoh, ada data dengan nilai ketidakmiripan {10, 12, 25, 30, 40} dengan intervalnya [10,40]. Jika akan ditransformasi ke dalam interval [0,1], kita bisa menggunakan formula x = $\frac{x-\min(x)}{\max(x)-min(x)}$ sehingga nilai-nilai ketidakmiripan tersebut ditransformasi menjadi {0, 0.667, 0.5, 0.6667, 1}.

Untuk fitur yang menggunakan tipe ordinal, misalnya sebuah atribut yang mengukur kualitas produk dengan skala {rusak, jelek, sedang, bagus, sempurna}, skala tersebut harus ditransformasikan ke dalam nilai numerik, misalnya {rusak=0, jelek=1, sedang=2, bagus=3, sempurna=4}. Kemudian, ada dua produk P1 dengan kualitas bagus dan P2 dengan kualitas jelek. Jarak (ketidakmiripan) antara P1 dan P2 dapat dihitung dengan cara D(P1,P2) = 3-1 = 2, atau jika dalam interval [0,1] menjadi $\frac{3-1}{4}$ = 0.5, sedangkan nilai kemiripannya adalah 1-0.5 = 5.

Untuk atribut bertipe numerik (interval dan rasio), nilai ketidakmiripan didapat dari selisih absolut di antara dua data. Misalnya atribut usia, jika P1 adalah usia 45 dan P2 usia 25, sedangkan jangkauan nilai usia dalam data adalah [5,75], nilai ketidakmiripan P1 dan P2 adalah D(P1,P2) = 45-25 = 20, atau jika dalam interval [0,1] menjadi $\frac{20-5}{75-5}$ = 0.21, sedangkan nilai kemiripannya adalah 1-0.21 = 0.79.

Tabel 5. Formula Kemiripan dan Ketidakmiripan Dua Data Dengan Satu Atribut

Tipe Atribut	Kemiripan	Ketidakmiripan
Nominal	s = $\begin{cases} 1\ jika\ x = y \\ 0\ jika\ x \neq y \end{cases}$	d = $\begin{cases} 0\ jika\ x = y \\ 1\ jika\ x \neq y \end{cases}$
Ordinal	s = 1 – d	d = |x – y|/(n – 1)

Interval dan Rasio	s = -d, s = $\frac{1}{1+d}$, s = e^{-d} , s = $\frac{d-\min(d)}{\max(d)-min(d)}$	D = |x – y|

4.2. Ketidakmiripan Data Multiatribut

Terdapat banyak cara untuk menghitung jarak (ketidakmiripan) yang dapat digunakan untuk menghitung dua data dari beberapa atribut untuk setiap data (dari dua objek), diantaranya:

1. Jarak Euclidian

$$D(x,y) = \sqrt{\sum_{j=1}^{n} |x - y|^2}$$

2. Jarak Manhattan/City Block

$$D(x,y) = \sum_{j=1}^{n} |x - y|$$

3. Jarak Chebyshev

$$D(x,y) = \max_{j=1} N\left(|x_j - y_j|\right)$$

Sebagai contoh kita akan melakukan pengukuran jarak antardata dengan jarak Euclidean pada data tabel 6 di bawah ini.

Tabel 6. Contoh Data Dua Dimensi

Point	x	y
P1	1	1
P2	4	1
P3	1	2

Tabel 7. Hasil Pengukuran Jarak Euclidean

Euclidean	P1	P2	P3
P1	0	3	1
P2	3	0	3.16
P3	1	3.16	0

VII. Daftar Pustaka

[1] Astuti, F.A. 2013. Data Mining. Yogyakarta: Andi.
[2] Kusrini & Taufiz, E.L. 2009. Algoritma Data Mining. Yogyakarta: Andi.

[3] Prasetyo, E. 2012. Data Mining: Konsep dan Aplikasi Menggunakan MATLAB. Yogyakarta: Andi.
[4] Prasetyo, E. 2014. Data Mining: Mengolah Data Menjadi Informasi Menggunakan MATLAB. Yogyakarta: Andi.

VIII. Materi Berikutnya

Pokok Bahasan	Pemrosesan Awal Data
Sub Pokok Bahasan	1. Agregasi (*aggregation*) 2. Penarikan contoh (*sampling*) 3. Diskretisasi dan binerisasi (*discretization and binarization*) 4. Pemilihan fitur (*feature subset selection*) 5. Pembuatan fitur (*feature creation*) 6. Transformasi atribut (*attribute transformation*)