DATA MINING

3 SKS | Semester 6 | S1 Sistem Informasi

SIMILARITY

Nizar Rabbi Radliya
nizar@email.unikom.ac.id

SIMILARITY & DISSIMILARITY

Kemiripan (similarity) adalah ukuran numerik dimana dua objeknya mirip,
nilai 0 jika tidak mirip dan nilai 1 jika mirip penuh.

Ketidakmiripan (dissimilarity) adalah ukuran numerik dimana dua objek yang berbeda, jangkauan nilai 0 sampai 1 atau bahkan sampai ∞.

SIMILARITY & DISSIMILARITY – DATA SATU ATRIBUT

Istilah ketidakmiripan = ukuran jarak (*distance*) antara dua data.

Jika s = ukuran kemiripan dan d = ukuran ketidakmiripan,

Jika interval/range nilainya adalah [0,1],

Maka dapat dirumuskan bahwa s+d=1.

Atau s = $\frac{1}{1+d}$ atau s = e^{-d}.

SIMILARITY & DISSIMILARITY – DATA SATU ATRIBUT

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}.

TIPE DATA

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, NIM, 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

SIMILARITY & DISSIMILARITY – DATA 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	$s = 1 - d$	$d = \lvert x - y \rvert / (n - 1)$
Interval dan Rasio	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = \frac{d - \min(d)}{\max(d) - min(d)}$	$d = \lvert x - y \rvert$

SIMILARITY & DISSIMILARITY – DATA SATU ATRIBUT

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,

{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 = 0.5

SIMILARITY & DISSIMILARITY – DATA SATU ATRIBUT

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.

SIMILARITY & DISSIMILARITY – DATA SATU ATRIBUT

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.

DISSIMILARITY – DATA MULTI ATRIBUT

Jarak Euclidian

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

Jarak Manhattan/City Block

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

Jarak Chebyshev

D(x,y) = max (|x - y|)

DISSIMILARITY – DATA MULTI ATRIBUT

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

Jarak Euclidian

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

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

DISSIMILARITY – DATA MULTI ATRIBUT

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

Jarak Manhattan/City Block

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

Manhattan	P1	P2	P3
P1	0	3	1
P2	3	0	4
P3	1	4	0

DISSIMILARITY – DATA MULTI ATRIBUT

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

Jarak Chebyshev

D(x,y) = max (|x - y|)

Chebyshev	P1	P2	P3
P1	0	3	1
P2	3	0	3
P3	1	3	0

SIMILARITY – DATA MULTI ATRIBUT

Simple Matching (SMC) & Jaccard Coefficients (J)
Dapat digunakan untuk menghitung similaritas dua vektor biner.

M01 = Jumlah atribut dimana p adalah 0 dan q adalah 1
M10 = Jumlah atribut dimana p adalah 1 dan q adalah 0
M00 = Jumlah atribut dimana p adalah 0 dan q adalah 0
M11 = Jumlah atribut dimana p adalah 1 dan q adalah 1

SMC = number of matches / number of attributes
SMC = (M11 + M00) / (M01 + M10 + M11 + M00)

J = number of 11 matches / number of not-both-zero attributes values
J = (M11) / (M01 + M10 + M11)

Contoh:

Hitung Similaritas dari dua vektor berikut:

p = 1 0 0 0 0 0 0 0 0 0

q = 0 0 0 0 0 0 1 0 0 1

SIMILARITY – DATA MULTI ATRIBUT

Cosine Similarity
Dapat digunakan untuk menghitung similaritas dua vektor dokumen (tipe kontinyu).

Jika d1 dan d2 adalah dua vektor dokumen maka similaritas antara dua vektor tsb:
Cos (d1,d2) = (d1.d2) / | |d1| | | |d2| |

Contoh:

Hitung Similaritas dari dua vektor berikut:

d1 = 3 2 0 1

d2 = 1 0 0 2

SIMILARITY – DATA MULTI ATRIBUT

Extended Jaccard Coefficient (Tanimoto)
Dapat digunakan untuk menghitung similaritas dua vektor tipe kontinyu.

T (p,q) = (p.q) / $||p||^2$ **+** $||q||^2$ $- \mathbf{p}.\mathbf{q}$

Contoh:

Hitung Similaritas dari dua vektor berikut:

p = 3 2 0 1

q = 1 0 0 2

SIMILARITY – FITUR CAMPURAN

Siswa	Tinggi	Berat	Jenis Kelamin	Alamat	Sekolah
Rina	160	50	0	1	2
Ruli	159	62	1	1	3
Adit	168	67	1	2	2
Siti	175	70	0	3	1

Berapa nilai Kemiripan untuk vektor pertama (Rina) dan kedua (Ruli)?

$$s(x, y) = \frac{\sum_{i=1}^{r} S_i(x,y)}{\sum_{i=1}^{r} w_i}$$

NEXT

TEKNIK ASOSIASI