T-18

DETEKSI OUTLIER BERBASIS KLASTER DENGAN ALGORITMA *SHARED NEAREST NEIGHBOR*

Alvida Mustika Rukmi[1]

[1]Jurusan Matematika ITS Surabaya
alvida@matematika.its.ac.id

Abstrak

Deteksi outlier merupakan salah satu bidang penelitian yang penting dalam mendeteksi perilaku yang tidak normal seperti deteksi intrusi jaringan, diagnosa medis, dan lain-lain. Bermacam-macam metode telah dikembangkan baik berdasarkan teknik seperti statistic-based, distance-based, density-based, clustering-based,subspace-based, dan lain-lain.Pengklasteran obyek data merupakan salah satu cara untuk mendapatkan data, terutama data berdimensi tinggi. Obyek-obyek data yang mempunyai kemiripan (similarity)tinggi, akan berada dalam satu klaster, dan sebaliknya, jika menunjukkan ketidakmiripan (dissimilarity), akan berada pada klaster berbeda.Pendekatan deteksi outlier berbasis klaster adalah dengan mengesampingkanklaster-klaster kecil yang jauh dari klaster yang lain. Pendekatan ini dapat digunakandengan menggunakan sebarang teknik klasterisasi, namun memerlukan threshold berapajumlah minimum ukuran klaster dan jarak antara klaster kecil dengan klaster yang lebih besar,dimana menganggap obyek-obyek dalam klaster yang kecil sebagai kandidat outlier. Algoritma shared nearest neighbor (SNN) pada proses pengklasteran, menetapkan ketetanggaan dari sebuah data berdasarkan nilai ε, yakni jari-jari(radius) daerah ketetanggaan, dan menempatkan data-data yang mempunyai ’k- tetangga ’ sama berada dalam satu klaster jika jumlah shared nearest neighbor, MinT, melebihi ambang batas yang ditentukan.Untuk mendapatkan klaster-klaster yang memuat data dengan kemiripan tinggi diperoleh dari hasil pengujian yang memerlukan ketepatan dalam menentukan nilai k, MinT, Eps.

Kata kunci : deteksi oultlier , shared nearest neighbor, pengklasteran,

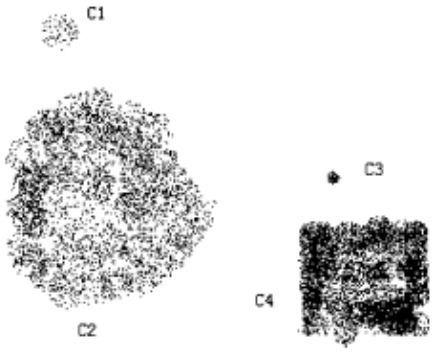
PENDAHULUAN

Deteksi outlier

Deteksi outlier merupakan bidang penelitian yang penting sebagai bagian dari berbagai aplikasi. Diantara cara deteksi outlier pada sebuah data set adalah berbasis klaster. Salah satu pendekatan deteksi outlier berbasis klaster adalah dengan mengesampingkan klaster-klaster kecil yang jauh dari klaster yang lain. Pendekatan ini dapat digunakan dengan menggunakan sebarang teknik pengklasteran, namun memerlukan ambang batas berapa jumlah minimum ukuran klaster dan jarak antara klaster kecil dengan klaster yang lebih besar. Pendekatan lain adalah dengan menentukan derajat di mana sebuah obyek berada pada sebarang klaster. Sebagai perwakilan klaster dapat digunakan centroid untuk mengitung jarak antara obyek dengan klaster. Ada beberapa cara untuk mengukur jarak sebuah obyek ke sebuah klaster. Caranya adalah mengukur jarak sebuah obyek terhadap centroid terdekat. Atau dapat juga dengan mengukur jarak relatif

obyek dengan centroid terdekat. Jarak relatif adalah rasio jarak obyek terhadap centroid dibagi dengan jarak rata-rata semua titik terhadap centroid klaster di mana ia berada.

Algoritma shared nearest neighbors sebagai salah satu metoda pengklasteran, dapat digunakan dalam pendeteksian outlier, di mana sebuah outlier didefinisikan sebagai sebarang obyek yang tidak berada pada klaster yang"cukup besar".



Gambar 1. Klaster pada dataset DS1

Pada Gambar 1 ditunjukkan data dua dimensi yang terdiri dari 4 klaster C1, C2,C3, dan C4. Dari sudut pandang klaster, obyek-obyek data pada C1 dan C3 dapat dianggap sebagai outlier karena tidak terdapat pada klaster yang besar yaitu C2 dan C4. C2 dan C4 disebut klaster besar karena C2 dan C4 merupakan klaster yang dominan pada data set, yaitu memuat sebagian besar obyek pada data set.

Definisi-definisi pada Algoritma Shared Nearest Neighbor

Definisi 1. Ketetanggaan (Jarvick dan Patrick, 1973)
ε-*neighbor* dari sebuah titik p pada himpunan titik ,D, dinyatakan dengan Nε (p), dimana Nε (p)={q $\in$ D, jarak(p,q) $\leq \varepsilon$ }.

Definisi 2. Ketetanggaan dua obyek. (Qiang Yang dkk, 2004)
Jika o_1 dan o_2 berupa dua obyek berdimensi-d dan K berupa himpunan atribut kunci. Jika δ berupa ambang batas kemiripan dengan $0< \delta \leq |K| \leq d$, maka o_1 dan o_2 adalah *neighbor* (tetangga) jika keduanya mempunyai sedikitnya δ atribut yang bernilai sama dalam K.

Definisi 3. Graf kemiripan (Qiang Yang dkk, 2004).
Graf kemiripan dari dataset DB, G=(V,E), berupa graf tidak berarah dimana $v_1, v_2, .., v_n \in V$ = DB adalah himpunan node (vertex) dan E himpunan edge sehingga $\{v_1,v_2\} \in E$ jika obyek v_1 dan v_2 mempunyai kemiripan.Graf tak berarah G=(V,E) adalah lengkap jika node-nodenya adalah pasangan adjacent

Definisi 4. Klaster Inti (Qiang Yang dkk, 2004)
$C^r \subseteq$DB adalah klaster inti jika memenuhi tiga syarat berikut :

1. $| C^r | \geq \alpha$, dimana α adalah ambang batas yang menentukan ukuran minimal klaster inti
2. .Dua obyek sebarang dalam C^r adalah mirip.
3. Tidak ada C', dimana $C^r \subseteq C' \subseteq$ DB, yang memenuhi syarat (2). C^r adalah klaster inti maksimal jika C^r berupa klaster inti dengan kardinalitas (penggumpalan) maksimal. Klaster Inti C^r adalah grup (klaster) padat yang mempunyai jumlah maksimal pasangan obyek yang sama.

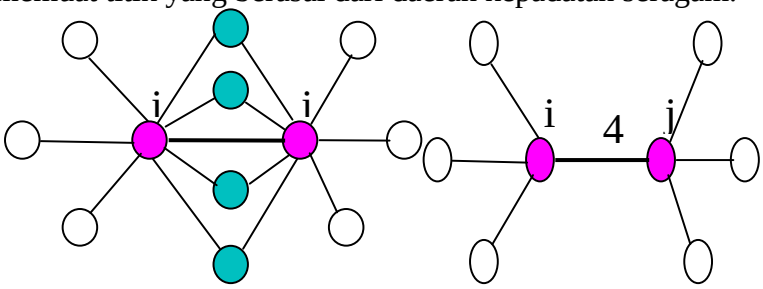
Definisi 5. Klaster. (Qiang Yang dkk, 2004)
Sebuah klaster inti akan membentuk sebuah klaster jika $C^r \subseteq$DB berupa klaster inti dan jika θ berupa ambang batas dengan $0 \leq \theta \leq 1$. C adalah klaster jika C = { v $\in$ DB | v $\in$ C^r atau C^r memuat sedikitnya θ x | C^r | obyek yang sama dengan v}.

Algoritma Shared Nearest Neighbor
Algoritma*shared nearest neighbor* (SNN) pada proses pengklasteran dikenal sebagai cara untuk mengatasi masalah pengukuran jarak dalam data berdimensi tinggi (Jarvis dan Patrick, 1973) yang dikembangkan oleh (Gupta, 1999).

Algoritma SNN memerlukan 3 input parameter :
- k , yakni jumlah daftar neighbor setiap titik
- ε, yakni jari-jari(radius) daerah neighbor,Nε(p)dari sebuah titik p
- MinT, yakni jumlah minimal titik interior dalam daerah Nε(p)

Parameter ini dapat mengendalikan resolusi pengklasteran dan memudahkan user untuk mengendalikan berapa banyak titik yang diklaster atau kebalikannya, berapa banyak titik yang dikategorikan sebagai outlier.
Pada algoritma ini, pasangan titik diletakkan dalam klaster yang sama jika :
- menggunakan bersama lebih dari ε-*nearest neighbor*
- saling berada dalam daftar *k-nearest neighbor*

Definisi kemiripan digunakan algoritma SNN untuk mengenali titik utama dan membangun klaster di sekeliling titik utama. Klaster-klaster tidak memuat semua titik, namun hanya memuat titik yang berasal dari daerah kepadatan seragam.



Gambar 2. SNN antara titik i dan j mempunyai jumlah SNN sebesar 4 .

Langkah-langkah Algoritma Shared Nearest Neighbor
Langkah-langkah pengerjaan dengan algoritma SNN adalah sebagai berikut :
1. Bangun matriks kemiripan
2. Sparsifikasi matrik kemiripan
3. Bentuk klaster
   Bangun graf shared nearest neighbor dari jumlah titik yang mempunyai kemiripan SNN sebesar $\geq \varepsilon$ pada setiap titik.Temukan kepadatan SNN setiap

titik. Titik-titik utama adalah semua titik yang mempunyai kepadatan SNN ≥ MinT.Bentuk klaster dari titik utama. Jika dua titik utama berada pada dalam radius ε, maka keduanya ditempatkan dalam klaster yang sama.

4. Temukan outlier

PEMBAHASAN

Pengklasteran dengan Algoritma SNN

Pengklasteran ini mengambil dataset berupa koleksi buku berbahasa Indonesia sebanyak 500 buah. Pengklasteran koleksi buku didasarkan pada judul buku yang menggambarkan topik muatan buku. Judul buku merupakan data teks, dimana data berupa kata dasar dan menimbang setiap kata berdasarkan bobotnya pada teks.

Pre-processing yang dilakukan yakni :

a. Tokenisasi judul buku dari masing-masing buku
   Hasil tokenisasi masing-masing judul buku disimpan dalam Tabel 1. Setiap baris(record) menyimpan IDBuku, judul buku dan token pembentuk judul buku sebanyak k. Judul buku yang kurang dari k token, maka sisa kolom terakhir bernilai null.
b. Penghilangan kata depan, sandang kata sambung, jenis kata yang tidak termasuk kata significant, tanda baca (punctuation),dan spasi, tidak terpilih sebagai token.Relasi terhadap kata sinonim diabaikan. Kata “analisis” dan “analisa” adalah sinonim, namun tetap dianggap dua kata yang berbeda. Polisemi (identifikasi kata yang ambigu) juga diabaikan, contoh : Jaguar bisa menyatakan nama binatang atau merk mobil.

Tabel 1. Tabel Buku-Token (diasumsikan k = 6)

Books							
BookID	**Title**	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**
8001	Adat dan upacara perkawinan daerah Bengkulu	daerah	Adat	perkawinan	upacara	Bengkulu	
8002	Adat dan upacara perkawinan daerah Jambi	daerah	Adat	perkawinan	upacara	Jambi	
8003	Adat dan upacara perkawinan daerah Istimewa Aceh	daerah	Adat	perkawinan	upacara	Istimewa	Aceh
8004	Adat dan upacara perkawinan daerah Istimewa Yogyakarta	daerah	Adat	perkawinan	upacara	Istimewa	Yogyakarta

Tabel Kata-Frekuensi dibentuk untuk menyimpan daftar semua kata yang dihasilkan dari proses tokenisasi yang dipaparkan oleh Tabel 1. Frekuensi kemunculan masing-masing kata i pada judul buku koleksi dinyatakan df_i. Sedangkan frekuensi munculnya kata i pada sebuah judul buku, tf_i, diasumsikan 1. Frekuensi kemunculankata hasil tokenisasi disimpan dalam Tabel 2.

Tabel 2. Tabel Kata-Frekuensi

IDKata	Kata	Freq(dfi)
69703	Manajemen	125
69451	Indonesia	73
69547	Kimia	59
69480	teknik	48
69499	bahasa	46

IDKata	Kata	Freq(dfi)
……..	……	…..
71000	pintu	1
71001	jendela	1
71003	raya	1
71004	jembatan	1

Matrik kemiripan pada Klaster

Matrik kemiripan pada klaster digunakan untuk mengetahui kedekatan topik buku-buku pada sebuah klaster hasil pengklasteran dengan algoritma SNN. Pengukuran cosinus digunakan untuk menghitung kesamaan antara data teks. *Entry* pada matrik kemiripan diperoleh dari ukuran kemiripan berdasarkan jarak menggunakan persamaan kosinus beikut :

dengan

d_i : vektor data ke-i
d_j : vektor data ke-j
n : banyaknya data
$c_{i,k}$: skalar pada vektor data ke-i
$c_{j,k}$: skalar pada vektor data ke-j

Dalam membangun matrik kemiripan, nilai kemiripan dua buku diperoleh berdasarkan persamaan kosinus. Misal : Buku1 berjudul "*Manajemen pemasaran : suatu pendekatan strategis dengan orientasi global*" dan Buku2 berjudul " *Manajemen pemasaran global : konsep dan aplikasi*". Kata pada Buku1 ={manajemen, pemasaran, pendekatan, strategis, oerientasi, global}. Kata pada Buku2 = {manajemen, pemasaran, global, konsep, aplikasi}. Matrik Buku-Kata dibentuk dari token-token sebagai berikut :
T1=manajemen,T2=pemasaran,T3=global,T4=konsep,T5=aplikasi,T6= pendekatan,
T7= strategis,T8= oerientasi

	T1	T2	T3	T4	T5	T6	T7	T8
B1	1	1	1	0	0	1	1	1
B2	1	1	1	1	1	0	0	0

dimana nilai skalar masing-masing bernilai 1, karena nilai tf_i =1.Kemiripan kedua buku tersebut diukur dengan persamaan cosinus :

$$\text{sama}(B1,B2) = \frac{1.1+1.1+1.1+0.1+0.1+1.0+1.0+1.0}{\sqrt{1^2+1^2+1^2+1^2+1^2+1^2}\,x\sqrt{1^2+1^2+1^2+1^2+1^2}}$$

$$= \frac{3}{\sqrt{6}x\sqrt{5}} = 0{,}67$$

Nilai kemiripan antara dua buku adalah bobot pada matrik kemiripan. Matrik ini bersifat simetri, dimana setiap baris dan kolom ke-i menyatakan buku ke-i., sehingga elemen (i,i) mempunyai kemiripan bernilai 1. Sedangkan elemen (i,j) menyatakan nilai kemiripan antara buku ke-i dan buku ke-j.

Sparsifikasi Matrik Kemiripan

Sparsifikasi matriks kemiripan dengan cara mempertahankan k neighbor yang paling mirip saja. Hal ini berkoresponden dengan cara mempertahankkan k link terkuat dari graf kemiripan

Nilai parameter yang akan dimasukkan ke dalam sistem pengklasteran dengan algoritma SNN berkenaan dengan sparsifikasi matrik kemiripan adalah parameter k. Nilai k menentukan berapa buku tetangga terdekat (nearest neighbor) dari sebuah buku yang akan dimasukkan pada daftar k-NN berdasarkan nilai kemiripan antar dua buku. Contoh : ambil nilai k = 10, maka daftar 10-NN berisi 10 buku tetangga terdekat dari sebuah buku. Misal Tabel 3 menggambarkan daftar k-NN dari Buku-2 :

Tabel 3. Daftar 10-NN dari b_2

Buku	b_2	b_1	b_5	b_8	b_{11}	b_{12}	b_3	null	null	Null
Nilai Mirip	1	0,94	0,87	0,54	0,41	0,33	0,12	0	0	0

Jika buku-2 mempunyai kurang dari 10 buku tetangga terdekat, katakan 7 buku tetangga terdekat, maka sisa 3 kolom akhir diisi dengan null. Hal ini terjadi jika buku-2 hanya mempunyai nilai kemiripan lebih besar dari nol dengan 7 buku saja. Sebaliknya, sebuah buku mempunyai tetangga terdekat lebih dari 10 buku, maka diambil 10 buku tetangga terdekat saja.

Pembentukan klaster

Setelah pembangunan matrik kemiripan, setiap kolom i menunjukkan nilai kemiripan buku i terhadap buku lainnya.Matrik menunjukkan bahwa dua buku (b_i,b_j) mempunyai kemiripan yang paling dekat jika bobot (b_i,b_j)menunjukkan nilai tertinggi.Kemudian nilai kemiripan setiap kolom diurut menurun, sehingga akan diperoleh urutan buku-buku melajur ke bawah berdasarkan nilai kemiripan terurut menurun. Isi daftar k-NN dari b_i adalah buku-buku yang merupakan tetangga terdekat yaitu mempunyai nilai kemiripan k tertinggi terhadap b_i. Matrik k-NN disusun dari daftar k-NN semua buku pada koleksi. Jadi, setiap kolom ke-i matrik k-NN memuat isi daftar k-NN setiap b_i. Jika jumlah buku = n, maka ukuran matrik k-NN adalah [k,n]

Tabel 4. Contoh daftar 10-NN dari b_2

1-NN	2-NN	3-NN	4-NN	5-NN	6-NN	7-NN	8-NN	9-NN	10-NN
2	1	5	8	11	12	3	4	6	7

Matrik SNN dibangun berdasarkan matrik k-NN. Setiap elemen (i,j) matrik SNN menyatakan banyaknya buku tetangga yang sama di antara daftar k-NN dua buku.

Matrik SNN digunakan untuk membentuk klaster-klaster. Parameter Eps dan MinT dimasukkan, dimana Eps adalah ambang batas minimal SNN dan MinT adalah ambang batas minimal banyaknya buku untuk membentuk klaster. Buku-buku tersebut akan membentuk sebuah klaster.Beberapa buku yang tidak masuk kategori dimasukkan sebagai *outlier*, jika kata – kata penyusun buku tidak mempunyai keterkaitan dengan buku lain. Hal ini ditandai dengan nilai df_i kata kuncinya rendah, misal nilai df_i = 1 atau df_i =2.

Pembentukan matrik kesamaan digunakan untuk menampilkan kedekatan antar buku

berdasarkan nilai bobot pada setiap sel matrik.. Matrik ini bernilai satu pada semua sel diagonalnya (i, i), karena menyatakan buku itu sendiri. Pada Gambar 2, (b_1,b_2) mempunyai bobot kesamaan nilai 0,77 yang menunjukkan nilai kesamaan tertinggi pada baris ke-1. Artinya, bahwa buku ke-1 (berjudul ”Manajemen strategi : daya saing dan globalisasi konsep ”) mempunyai kesamaan tertinggi dengan buku ke-2 (berjudul “"Manajemen strategi : konsep dan kasus “) dibandingkan dengan 7 buku lain. Kesamaan buku ke-1 terhadap buku ke-2 ditunjukkan dengan penggunakan bersama tiga kata yaitu : manajemen, konsep, dan strategi.

1	0,077	0,007	0,048	0,011	0,049	0,064	0,047	0,044
0,076	1	0,008	0,055	0,012	0,057	0,077	0,054	0,051
0,007	0,008	1	0,008	0,008	0,008	0,01	0,008	0,008
0,048	0,055	0,008	1	0,013	0,059	0,082	0,056	0,052
0,011	0,012	0,008	0,013	1	0,911	0,019	0,019	0,012
0,049	0,056	0,008	0,059	0,917	1	0,085	0,058	0,054
0,064	0,077	0,01	0,082	0,019	0,085	1	0,079	0,072
0,047	0,054	0,008	0,056	0,013	0,058	0,08	1	0,051
0,044	0,05	0,008	0,052	0,012	0,054	0,072	0,051	1

Gambar 2. Matrik kesamaan

PENGUJIAN

Pengujian ini dimaksudkan sebagai alat evaluasi terhadap pemberian nilai parameter selama proses pengklasteran. Jika nilai Eps = 2, kata kunci dibentuk dari dua kata yang digunakan bersama pada judul buku antara 2 buku. Asumsi bahwa kedekatan dua judul buku ditunjukkan oleh penggunaan bersama dua kata antar dua buku. MinT adalah parameter yang digunakan sebagai ambang batas jumlah minimal buku untuk membentuk sebuah klaster.

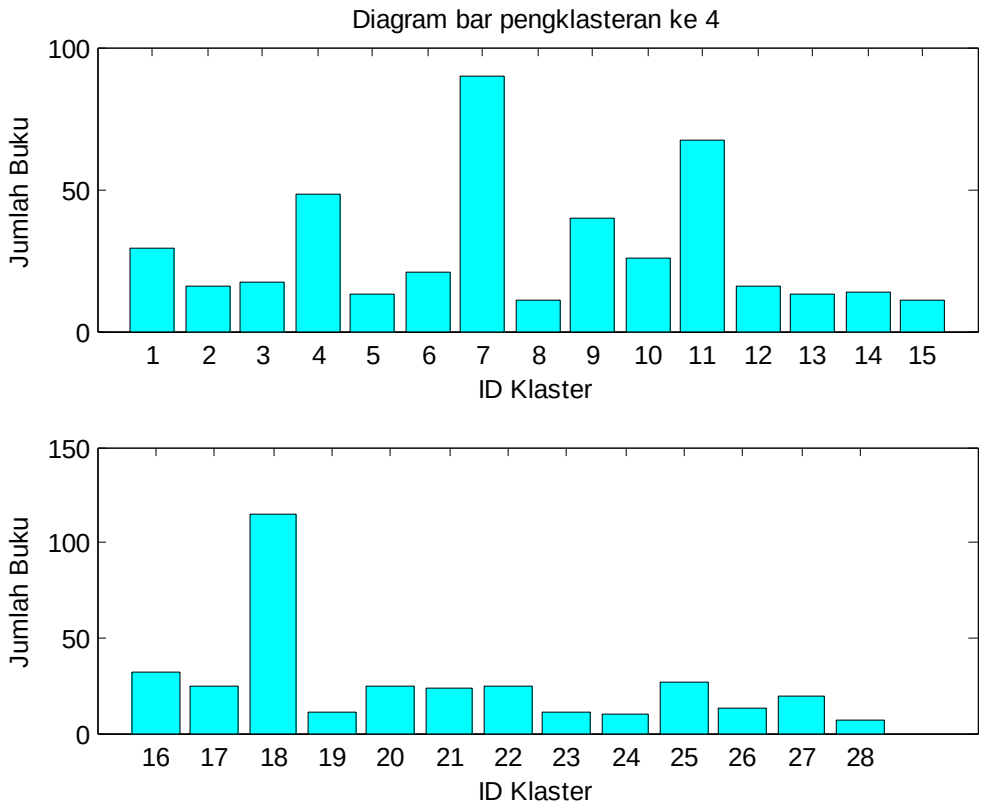
Diagram bar pada Gambar 7 menggambarkan hasil pengklasteran. Sumbu-x menyatakan identitas klaster, setiap batang bar menunjukkan sebuah klaster dan sumbu-y menyatakan jumlah buku pada setiap klaster. Sedangkan lingkaran *pie* pada Gambar 7 menunjukkan perbandingan data terklaster terhadap data outlier

Jumlah klaster yang banyak mengindikasikan jumlah kata yang menjadi kata kunci juga banyak. Akan lebih banyak sebuah buku berada pada klaster berlainan jika lebih dari satu tokennya menjadi kata kunci. Berikut ini diagram bar dan lingkaran pie dari tiga hasil pengklasteran, yaitu :

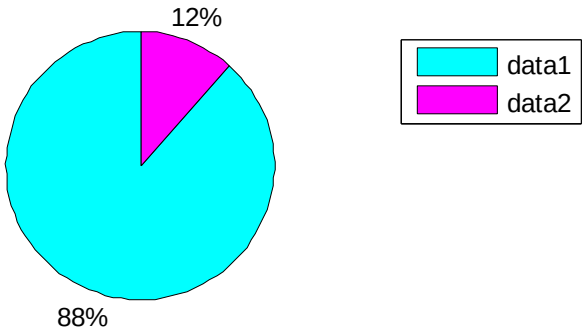
- Pengklateran ke 4 dengan nilai parameter yang dimasukkan :

  K= 40; Eps=5 ;MinT=10

  Jumlah buku yang tertinggi pada klaster 18 menunjukkan nilai > 100. Hal ini disebabkan nilai Eps yang kecil akan menarik buku-buku dengan 1 kata kunci berfrekuensi tinggi ke dalam satu klaster. Pada Tabel Books , kata ’manajemen’ mempunyai frekuensi tinggi = 123, sehingga buku-buku yang memuat kata ’manajemen’ akan berada pada klaster yang sama.



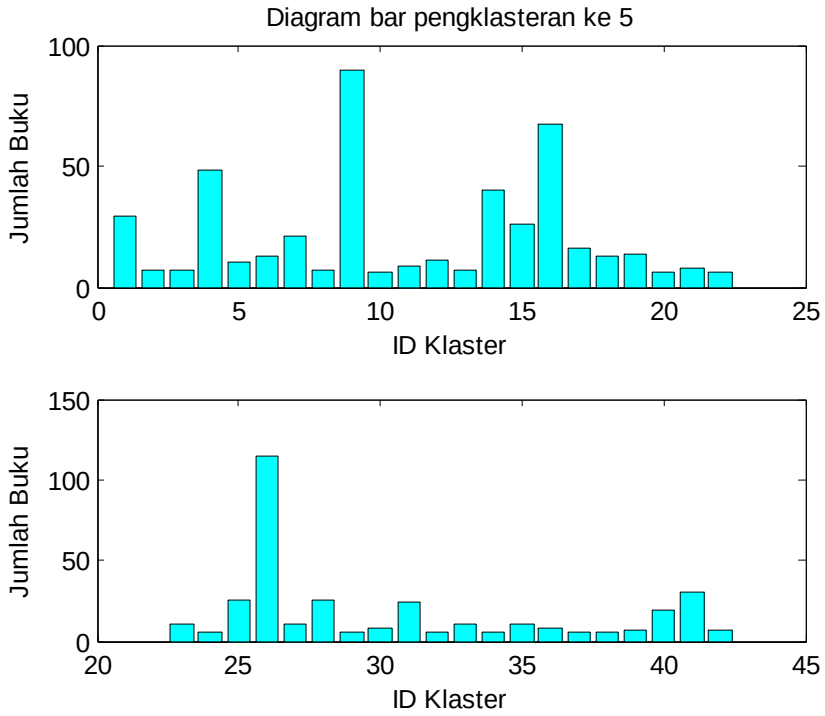
Perbandingan jumlah buku terklaster terhadap buku ’outlier’



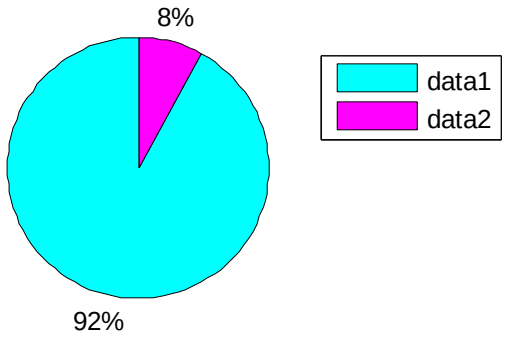
Gambar 7. Diagram Bar dan lingkaran pie pengklasteran ke 4

- Pengklateran ke 5 dengan nilai parmeter yang dimasukkan :

  K= 40; Eps=5 ;MinT=5
  Dibandingkan pengklasteran ke 4, pengklasteran ke 5 menghasilkan lebih banyak klaster, karena MinT lebih kecil. Selain itu, lingkaran pie dari keduanya juga menampilkan nilai yang berbeda. Telah disebutkan sebelumnya bahwa MinT yang lebih kecil akan menghasilkan outlier yang lebih sedikit. Sedangkan pengaruh dari nilai Eps yang kecil sama seperti pada pengklasteran ke 4.



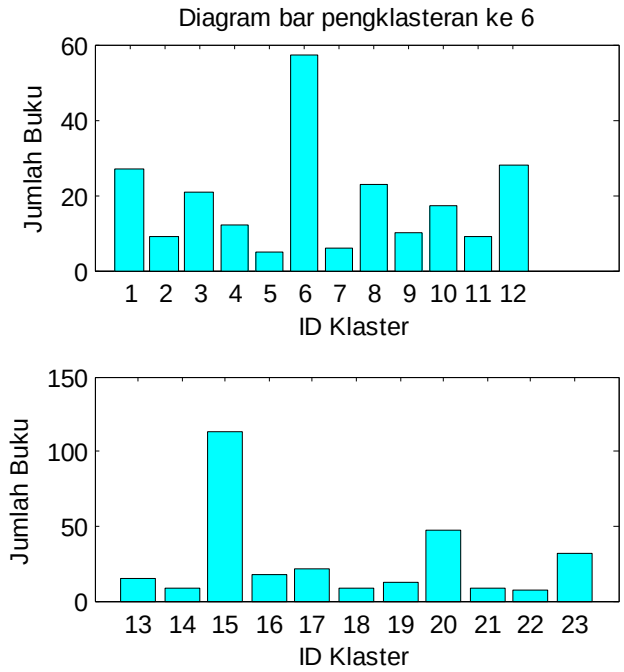
Perbandingan jumlah buku terklaster terhadap buku ’outlier’



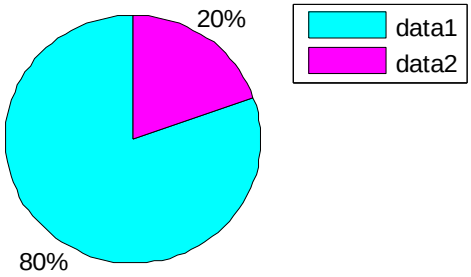
Gambar 8 Diagram Bar dan lingkaran pie pengklasteran ke 5

- Pengklateran ke 6 dengan nilai parmeter yang dimasukkan :

  K= 40; Eps=10 ;MinT=5



Perbandingan jumlah buku terklaster terhadap buku 'outlier'



Gambar 9 Diagram Bar dan lingkaran pie pengklasteran ke 6

Diagram bar pengklasteran ke 6 menampilkan jumlah klaster yang terbentuk lebih sedikit dari pengklasteran ke 5, walaupun nilai MinT sama. Hal ini disebabkan lebih banyak buku yang termasuk dalam outlier karena perbedaan nilai Eps.

Tabel 5 berikut ini adalah beberapa hasil dari nilai parameter berbeda untuk membentuk sebuah klaster dari seluruh koleksi buku dengan nilai k=8 :

Tabel 5. Tabel Jumlah Klaster dan Outlier terhadap nilai Eps dan MinT yang berbeda

Eps	MinT	Jumlah Klaster	% Outlier
2	5	25	0,4
2	3	8	0,45
1	5	18	0,16
1	10	18	0,16

Pada Tabel 6 berikut ini menyatakan hasil pengklasteran dengan mengambil nilai parameter k=25 terhadap nilai Eps dan MinT yang berbeda-beda .

Tabel 6. Tabel Jumlah Klaster dan Outlier terhadap nilai Eps dan MinT yang berbeda

Pengklasteran ke-	Eps	MinT	Jumlah Klaster	%Outlier
1	4	7	36	0,07
2	5	7	34	0,09
3	4	9	28	0,1
4	5	10	28	0,12
5	5	5	42	0,09
6	10	5	23	0,2
7	3	9	31	0,09
8	3	6	35	0,06
9	2	8	29	0,05
10	2	5	32	0,03

Nilai outlier menunjukkan persentasi jumlah buku yang tidak terklaster terhadap jumlah seluruh buku. Pengambilan nilai Eps = 10 memberikan hasil outlier yang besar, yakni 20 %. Sedangkan nilai Eps = 5 memberikan hasil outlier lebih kecil, yakni berkisar 9%-11% dan nilai Eps = 2 memberikan hasil outlier yang lebih kecil lagi, yakni berkisar 3 %-5%. Hal ini menunjukkan bahwa semakin kecil nilai Eps akan memberi hasil nilai persentasi outlier yang semakin kecil.

Nilai MinT menunjukkan ambang batas minimal jumlah buku dalam sebuah klaster. Karena penetapan MinT dilakukan setelah memasukkan nilai Eps, maka pengamatan hasil terhadap pengambilan nilai MinT yang berlainan berdasarkan nilai Eps yang sama. Dengan input nilai Eps = 5 untuk MinT = 10 membentuk 28 klaster, MinT= 7 membentuk 34 klaster, dan MinT= 5 membentuk 42 klaster. Begitu juga terhadap pengambilan nilai MinT untuk nilai Eps = 3, menunjukkan hal yang sama.

KESIMPULAN

Penentuan nilai parameter pada algoritma SNN akan mempengaruhi hasil pengklasteran. Parameter k merupakan input awal dimana penambahan jumlah kata yang akan digunakan pada pengklasteran akan mengakibatkan jumlah buku yang terklaster juga meningkat. Parameter Eps digunakan untuk pembentukan kata kunci. Pembentukan kata kunci dari kata-kata hasil tokenisasi menentukan pembentukan klaster dengan memasukkan buku-buku yang memuat katakunci. Semakin sedikit kata kunci yang dapat digunakan semakin sedikit jumlah klaster. sebuah klaster hanya memuat buku-buku bertopik sama, sehingga jumlah buku yang termasuk ke dalam sebuah klaster hanya sedikit. Hal ini menyebabkan data outlier sangat banyak

Pemberian nilai Eps yang besar, akan membentuk klaster-klaster yang memuat buku-buku dengan kemiripan topik yang tinggi. Hal ini mengakibatkan masing-masing klaster mempunyai nilai ketepatan yang tinggi karena relevansi buku-buku terhadap

kata kunci juga tinggi, sehingga hanya sedikit buku-buku yang tidak relevan yang muncul pada setiap klaster

Nilai persentasi outlier yang kecil berarti menunjukkan lebih banyak buku yang dapat diklaster. Hal ini disebabkan nilai Eps yang kecil akan 'menarik' buku-buku yang mempunyai kata berfrekuensi rendah pada judul bukunya untuk masuk pada sebuah klaster yang anggotanya mempunyai kemiripan dengan buku-buku tersebut.

Sedangkan pengaruh terhadap pengambilan nilai MinT dengan nilai Eps tetap,menunjukkan bahwa semakin besar nilai MinT akan membentuk jumlah klaster yang semakin sedikit.

Penggunaan algoritma SNN dapat menampilkan data outlier dari hasil pengklasteran.

DAFTAR PUSTAKA

He, Z., Xu, X., Deng, S.2003.*Discovering Cluster-based Local Outliers, Pattern Recognition Letter*, hal :1641-1650.

Breunig, M. M.., Kriegel, H. P., Ng, R. T., Sander, J.2000. *LOF: Identifying Density based Local Outliers*. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, hal : 93-104.

Ertoz L., Steinbach M., dan Kumar V.2003. *Finding Clusters of different Sizes, Shapes, and Densities in Noisy, High Dimensional Data*. Second SIAM International Conference on Data Mining, San Fransisco, USA

Ertoz L., Steinbach M., dan Kumar V.2002. *Finding topics in document, a Shared Nearest Neighbor Approach Clustering and Information Retrieval*. Kluwer Academic Publisher.

Heidelberg.2005.*High Dimensional Shared Narest Neighbor Clustering Algorithm*. Lecture Notes on Compuer Science, Publisher Springer Berlin vol.3614, hal. 494-50

Jain A.K dan Dubes R.C .1998. *Algirthms for Clustering Data*. Prentice Hall.

Jarvis R.A. dan Patrick E. 1973. *Clustering Using a Similarity Measure based on Shared Nearest Neighbor*. Proceeding IEEE Transaction on Computer, vol C-22, hal. 1025-1034.

Karphys G., Han E.H., dan Kumar V. 1999. *" CHAMELEON : A Hierarchical Clustering Algorithm Using Dynamic Modeling"*. Proceedings IEEE Transaction on Computer, vol. 32, hal. 68-75.

Moosinghe H.D.K., dan Pang-Ning T. 2006. *Outlier Detection Using Random Walks*. Department of Computer Science & Engineering Michigan State University.

Qing Zang dkk. 2004. *Cluster Cores-based Clustering for High Dimensional Data*. Department of Computer Science, Hongkong university of Science & Technology