CLUSTERING DATA NON-NUMERIK DENGAN PENDEKATAN ALGORITMA K-MEANS DAN *HAMMING DISTANCE* STUDI KASUS BIRO JODOH

Darlis Heru Murti, Nanik Suciati, Daru Jani Nanjaya
Jurusan Teknik Informatika,
Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember
Kampus ITS, Jl. Raya ITS, Sukolilo – Surabaya 60111, Telp. + 62 31 5939214, Fax. + 62 31 5913804
Email: darlis@its-sby.edu, nanik@its-sby.edu, darunanjaya@gmail.com

ABSTRAK

Clustering *adalah salah satu metode populer yang telah digunakan di berbagai bidang penelitian mulai dari kecerdasan buatan, teknologi jaringan syaraf, pengenalan pola, hingga pengolahan gambar. Salah satu teknik yang digunakan dalam* clustering *adalah dengan menggunakan algoritma k-means. Namun sayangnya, algoritma k-means hanya bisa digunakan untuk* dataset *yang atributnya bernilai numerik. Padahal dalam kenyataannya, suatu* database *bisa terdiri atas data-data yang bernilai numerik maupun non-numerik. Dalam penelitian ini akan dibahas mengenai penggunaan algoritma k-means pada suatu* clustering *data non-numerik (*categorical*), dengan dibantu* Hamming Distance *sebagai alat untuk mengukur* distance *dari masing-masing atribut* categorical*-nya. Kasus yang diambil adalah pada* dataset *suatu biro jodoh yang mana akan menjadi menarik karena dengan* clustering *ini dapat diketahui bagaimana pola pembentukan grup-grup yang memiliki karakteristik hampir sama di dalam keanggotaan suatu biro jodoh. Pada penelitian ini juga akan diberikan implementasi penggunaan* clustering *dalam pencarian individu di suatu data biro jodoh.*

Metodologi yang digunakan dalam penelitian ini meliputi beberapa tahapan. Tahapan pertama adalah persiapan data, yaitu data-data keanggotaan biro jodoh. Tahapan selanjutnya adalah proses modifikasi data dari non-numerik menjadi numerik. Kemudian tahap perhitungan distance *antar-data. Lalu tahapan* clustering *pada data yang teleh bernilai jarak. Dan diakhiri dengan tahapan ringkasan dari hasil proses-proses tersebut.*

Uji coba dan evaluasi dilakukan dengan menggunakan dataset *nyata yaitu data biro jodoh Grasco, Sakinah Surabaya, Libe, dan O'Diva. Dari uji coba tersebut didapatkan bahwa* clustering *dapat dilakukan pada atribut-atribut* categorical *yang ditransformasikan terlebih dahulu ke dalam bentuk numerik. Selain itu, kesamaan (*similarity*) dan karakterisktik dari masing-masing keanggotaan biro jodoh bisa diketahui.*

***Kata Kunci* :** *Data Mining, Clustering, Unsupervised Learning, K-Means.*

1. PENDAHULUAN

Clustering bisa diartikan sebagai suatu proses pemilahan sebuah set data menjadi kelompok-kelompok *cluster* yang terpisah dan masing-masing memiliki kesamaan. *Clustering* bisa juga berarti kumpulan dari metode-metode *data mining* yang unsupervised, yang tujuannya adalah untuk memilah-pilah suatu *dataset* keseluruhan menjadi sejumlah *cluster-cluster* yang ukurannya lebih kecil.

Dalam penelitian ini akan ditawarkan metode *clustering* dengan menggunakan algoritma k-means untuk *dataset* yang atributnya adalah berupa non-numerik (*categorical*). Sehingga basis data yang pada umumnya terdiri atas data-data dengan atribut non-numerik dapat dilakukan *clustering*.

Tujuan dari penelitian ini adalah untuk mengembangkan suatu aplikasi *clustering* pada data-data non-numerik dengan menggunakan algoritma k-means dan *Hamming Distance* dengan contoh kasus data biro jodoh dan untuk mengimplementasikan fasilitas pencarian individu dalam suatu basis data biro jodoh.

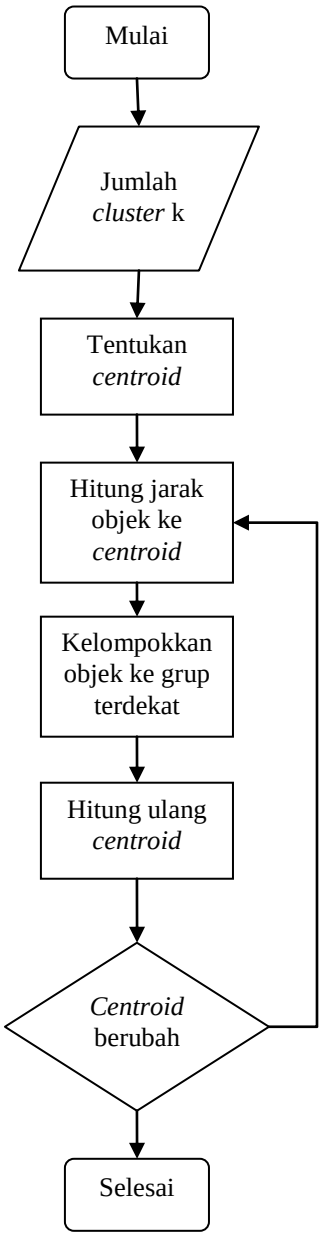
2. ALGORITMA K-MEANS

Algoritma K-Means merupakan salah satu algoritma *clustering* yang paling umum digunakan dalam berbagai aplikasi, yaitu sebuah algoritma untuk mengelompokkan suatu data berdasarkan nilai atributnya menjadi sebanyak *k cluster.*

Alur algoritma K-Means adalah sebagai berikut:

1. Menentukan jumlah *k cluster* yang akan dibentuk.
2. Menentukan sejumlah *k* data untuk dijadikan *cluster center*.
3. Menentukan keanggotaan suatu *cluster* dengan cara mengumpulkan data-data ke *cluster center* terdekat.
4. Mengkalkulasi nilai tengah suatu *cluster* untuk dijadikan *cluster center* yang baru. Proses ini beriterasi ke langkah 3 apabila *cluster center* baru yang terbentuk tidak sama dengan *cluster center* sebelumnya, hingga *cluster center* tidak berubah lagi.

Apabila digambarkan, maka alur diagram dari algoritma k-means adalah sebagai berikut :



Gambar 1. Diagram alur algoritma k-means

3. *HAMMING DISTANCE*

Ada banyak cara untuk mengukur jarak antar suatu obyek dengan obyek lainnya. Berbagai macam bentuk jarak (*distance*), antara lain adalah *Euclidean Distance*, *City Block (Manhattan) Distance*, *Chebyshev Distance*, *Minkowski Distance*, *Canberra Distance*, dan *Hamming Distance* [TEK-2004]. Masing-masing bentuk jarak tersebut mempunyai formula tersendiri.

Euclidean Distance sangat sering digunakan dalam proses penghitungan jarak antara dua objek. *Euclidean Distance* menghitung akar pangkat dua dari perbedaan kordinat dari sepasang objek, dengan formula sebagai berikut :

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2}$$

Contoh:

Jika A (0,3,4,5) dan B (7,6,3,-1), maka *euclidean distance* antara A dan B adalah :

$d_{BA} = ((0-7)^2 + (3-6)^2 + (4-3)^2 + (5+1)^2)^{1/2}$
$= (49+9+1+36)^{1/2}$
$= 9{,}747$

City Block Distance yang juga dikenal sebagai *Manhattan Distance*, *Boxcar Distance*, *Absolute Value Distance*, merepresentasikan jarak antara dua objek secara absolut, dengan formula sebagai berikut:

$$d_{ij} = \sum_{k=1}^{n} \left|x_{ik} - x_{jk}\right|$$

Contoh:

Jika A (0,3,4,5) dan B (7,6,3,-1), maka *city block distance* antara A dan B adalah :

$d_{BA} = |0-7| + |3-6| + |4-3| + |5+1|$
$= 7+3+1+6$
$= 17$

Chebyshev Distance disebut juga *Maximum Value Distance* menghitung *max* dari besaran absolut perbedaan antara kordinat dari sepasang objek. Formulanya adalah sebagai berikut :

$$d_{ij} = \max_{k} \left|x_{ik} - x_{jk}\right|$$

Contoh:

Jika A (0,3,4,5) dan B (7,6,3,-1), maka *chebyshev distance* antara A dan B adalah :

d_{BA} = maks { |0-7|, |3-6|, |4-3|, |5+1| }
= maks {7,3,1,6}
= 7

Minkowski Distance merupakan penjabaran dari *Euclidean, City Block, dan Chebyshev. Minkowski* menggunakan bilangan *order* λ. Apabila λ=1, maka merupakan *City Block Distance.* Apabila λ=2, maka merupakan *Euclidean Distance*. Apabila λ=∞, maka merupakan *Chebyshev Distance*. Formula dari *Minkowski Distance* adalah sebagai berikut :

$$d_{ij} = \sqrt[\lambda]{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^{\lambda}}$$

Contoh:

Jika A (0,3,4,5) dan B (7,6,3,-1), maka *minkowski distance* dengan order 3 antara A dan B adalah :

$d_{BA} = ((0-7)^3 + (3-6)^3 + (4-3)^3 + (5+1)^3)^{1/3}$
$= (-343-27+1+216)^{1/3}$
$= (-533)^{1/3}$
$= -5,348$

(Jarak antara B dan A (d_{AB}) adalah 5,348, berbeda dengan jarak d_{BA}. Biasanya yang digunakan adalah yang bernilai positif).

Canberra Distance menghitung penjumlahan dari pecahan-pecahan perbedaan kordinat antara dua buah titik. Setiap pecahan bernilai antara 0 hingga 1. Formulanya adalah sebagai berikut :

$$d_{ij} = \sum_{k=1}^{n} \frac{\left|x_{ik} - x_{jk}\right|}{\left|x_{ik}\right| + \left|x_{jk}\right|}$$

Contoh:

Jika A (0,3,4,5) dan B (7,6,3,-1), maka *canberra distance* antara A dan B adalah :

$$d_{BA} = \frac{|0-7|}{0+7} + \frac{|3-6|}{3+6} + \frac{|4-3|}{4+3} + \frac{|5+1|}{5+1}$$
$= 1+1/3+1/7+1$
$= 2,476$

Hamming Distance digunakan untuk menghitung jumlah perbedaan dari dua deret bilangan biner yang mempunyai panjang yang sama sesuai dengan posisi dari setiap digit biner. *Hamming Distance* memiliki formula sebagai berikut :

$$d_{ij} = q + r$$

Dimana *q* adalah jumlah variabel dengan nilai 1 pada objek ke-*i* tapi bernilai 0 pada objek ke-*j*. Sedangkan *r* adalah jumlah variabel dengan nilai 0 pada objeck ke-*i* tapi bernilai 1 pada objek ke-*j*.

Contoh:

Buah A (bentuk bulat, manis, berbiji, berair) dan buah B (bentuk tidak bulat, manis, tidak berbiji, dan tidak berair), maka dapat direpresentasikan sebagai deret biner sebagai berikut : A (1,1,1,1) dan B (0,1,0,0). *Hamming distance* antara A dan B dapat dicari dengan cara:

q = 3 (bernilai satu di A tapi bernilai 0 di B)
r = 0 (bernilai satu di B tapi bernilai 0 di A)
d_{BA} = 3 + 0
= 3

Hamming Distance sangat berguna dalam pencarian jarak antar data, khususnya dalam data biro jodoh.

3. *CLUSTERING* BIRO JODOH

Untuk dapat melakukan *clustering* pada data suatu biro jodoh, harus dilakukan suatu proses normalisasi terlebih dahulu terhadap data-data dalam biro jodoh. Hal ini dilakukan karena data-data asli pada biro jodoh umumnya adalah data-data yang non-numerik. Sehingga jika tetap menggunakan data-data non-numerik maka akan kesulitan untuk menentukan sebuah *mean* dari suatu *cluster*. Untuk itu, suatu proses transformasi perlu untuk dilakukan. Setelah dilakukan proses normalisasi (*binning*), selanjutnya dilakukan penentuan nilai tengah dari suatu *cluster*, yang biasa disebut sebagai *cluster center*. Penentuan *cluster center* ini dilakukan secara acak. Jumlah *cluster center* sesuai dengan jumlah *cluster* yang akan dibentuk.

Kemudian dilakukan penghitungan jarak antar data dan *cluster center* yang telah dipilih agar dapat diketahui jarak terpendek antar suatu data dengan *cluster center*, sehingga dapat diketahui pula ke *cluster* mana suatu data menjadi anggota. Apabila *cluster-cluster tersebut* telah terbentuk, maka *cluster center* dikalkulasi ulang lagi, hingga didapatkan *cluster-cluster center* yang baru. Lalu data-data biro jodoh ditentukan lagi ke *cluster* mana data-data itu menjadi anggota. Hal ini dilakukan iteratif hingga *cluster center* tidak berubah lagi.

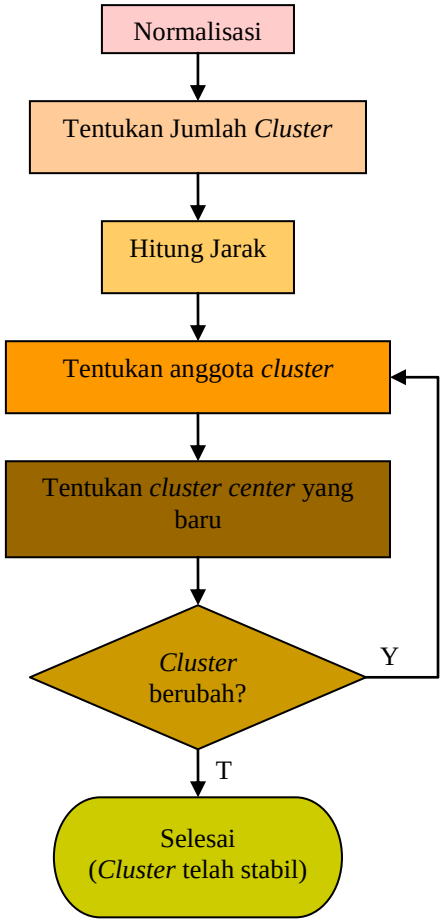
4. PROSES *BINNING*

Proses *Binning* disebut juga proses normalisasi, yaitu proses untuk mentransformasi nilai-nilai dari data-data non-numerik menjadi data-data yang bisa dikalkulasi. Proses kalkukasi ini diperlukan dalam pengimplementasian algoritma k-mean untuk proses *clustering* suatu data.

Data biro jodoh mengalami proses *binning*, dan nilainya berubah menjadi nilai biner sesuai dengan jumlah kategori yang dimiliki setiap atribut. Aturan untuk proses *binning* dalam penelitian ini adalah sebagai berikut :

1. Untuk atribut dengan dua kategori, setiap kategori di-set dengan biner 1 atau 0 sesuai dengan nilai atribut.
   Contoh: untuk atribut 'jenis kelamin' yang memiliki kategori 'pria' dan 'wanita', maka nilai dari atribut tersebut adalah 0 (untuk 'pria') atau 1 (untuk 'wanita').
2. Untuk atribut dengan lebih dari dua kategori, nilai dari atribut itu akan bernilai biner dengan jumlah *digit* biner sesuai dengan jumlah kategori dalam atribut tersebut. Dan pengesetan *digit* 1 akan sesuai dengan kondisi *true* dari setiap kategori atribut.
   Contoh: untuk atribut agama yang terdiri atas lima pilihan (Islam, Kristen, Katolik, Hindu, Buddah), maka nilai atribut untuk atribut yang bernilai 'Islam' adalah 10000 dan untuk atribut yang bernilai 'Hindu' adalah 00010.

Nilai ini yang nantinya akan dikalkukasi, misalnya dalam mencari jarak terdekat antara data dan *cluster center* atau ketika melakukan pencarian *cluster center* – *cluster center* baru dalam proses iterasi.



Gambar 2. Diagram alur proses *Clustering* Biro Jodoh

5. PROSES MENGHITUNG JARAK

Proses menghitung jarak digunakan untuk menentukan jarak terdekat antara sebuah data dengan suatu *cluster center*. Proses ini dilakukan dengan menggunakan data hasil proses *binning*. Data-data tersebut dihitung jaraknya terhadap setiap *cluster* yang ada dengan cara membandingkan nilai dari masing-masing atribut pada masing-masing data. Misalkan terdapat sebuah data mempunyai 5 atribut dengan nilai (0,0100,1000,001,010) dan sebuah *cluster center* dengan atribut (1,0010,0100,001,010). Maka jarak antar keduanya adalah :

Jarak = ((0 XOR 1) + (0100 XOR 0010)/4 +
¼ (1000 XOR 0100) +
⅓ (001 XOR 001) +
⅓ (010 XOR 010))
Jarak rata-rata = (1 + 2/4 + 2/4 + 0 + 0) / 5
Jarak rata-rata = 2/5
Jarak rata-rata = 0,4

Keterangan: *Hamming Distance* dapat pula diimplementasikan sebagai operasi *bitwise* XOR antara dua objek, karena hasil dari operasi antar dua data yang berbeda yang akan bernilai 1, dan antar dua data yang sama akan bernilai 0.

6. PROSES *CLUSTERING* DATA

Inti dari proses *clustering* ada di sini. Apabila telah diketahui jarak-jarak antara setiap data dengan setiap *cluster center*, maka data-data selanjutnya dapat dikelompok-kelompokkan berdasarkan jaraknya terhadap suatu *cluster center*. Suatu data akan dijadikan anggota suatu *cluster* yang mempunyai jarak paling minimum dengan data tersebut. Kemudian, setelah terbentuk *cluster-cluster*, *cluster center* akan dicari lagi dengan mencari "nilai tengah" dari *cluster* yang bersangkutan. Nilai tengah ini bisa berupa *mean*, *median*, atau *modus* dari atribut-atribut data yang menjadi anggota *cluster* tersebut. Jika *cluster center* yang baru telah didapatkan, maka akan dibandingkan dengan *cluster* yang lama. Jika berbeda (berubah), proses iterasi berjalan kembali. Namun jika ternyata *cluster center* bernilai sama, maka proses iterasi selesai dan *cluster* dinyatakan sebagai *cluster* yang stabil.

7. PROSES PENCARIAN INDIVIDU

Proses pencarian individu merupakan bentuk implementasi dari *clustering biro jodoh.* Dalam proses ini, *user* dapat memasukkan parameter-parameter pencarian data untuk kemudian menghasilkan *result* data-data individu yang memiliki karakteristik yang mendekati.

Metodenya adalah sebagai berikut :

1. Terlebih dahulu melakukan proses normalisasi terhadap data pencarian, hingga terbentuk data *binning*. Proses normalisasi menggunakan metode yang sama dengan proses normalisasi seperti pada proses *binning*, yang mana akan mengubah nilai-nilai atribut dari data pencarian ke dalam bentuk deretan bilangan biner.
2. Hasil *binning* ini kemudian digunakan untuk proses penghitungan jarak, yaitu jarak antara data pencarian dan masing–masing *cluster center* dari setiap *cluster*.
3. Menentukan *cluster* terdekat, yaitu adalah *cluster* dengan jarak terdekat. *Cluster* ini dianggap memiliki persamaan karakteristik paling tinggi. *Cluster* terpilih adalah *cluster* yang nantinya akan dijadikan acuan.
4. Dengan menggunakan *cluster* acuan, dilakukan penghitungan jarak antara data pencarian dan setiap data yang merupakan anggota dari *cluster* acuan tersebut. Selanjutnya akan didapatkan tingkat kesamaan dari setiap data dalam *cluster* acuan dengan data pencarian, sehingga dapat diketahui data-data mana yang paling sesuai dengan yang dicari.

8. UJI COBA *CLUSTERING*

Proses uji coba dilakukan dengan menggunakan data nyata dari media-media biro jodoh. Data awal (*raw data*) merupakan data biodata dari anggota-anggota biro jodoh. Data-data ini diletakkan ke dalam tabel *profil*, yang berisi data-data dari anggota biro jodoh, termasuk di dalamnya adalah biodata, keterangan fisik dan nonfisik, serta informasi-informasi lainnya. Atribut yang menyertai setiap data adalah sebanyak 18 atribut.

Data-data tersebut kemudian terlebih dahulu dinormalisasikan dengan menggunakan aturan proses *binning* sebagaimana telah disebutkan sebelumnya sehingga menjadi sebuah data yang bernilai biner tanpa mengubah struktur basis data. Sebagian hasil dari proses *binning* terhadap data-data asli adalah sebagai berikut :

Data No 1

Atribut	Nilai awal	Hasil Binning
Jenis Kelamin	wanita	0
Umur	24	1000
Tinggi	142	1000
Berat	50	1000
Berat Badan	ideal	00001
Rambut	lurus	10000
Warna Kulit	putih	10000
Warna Rambut	hitam	10000
Suku/Keturunan	Jawa	0001000000000
Status Pernikahan	lajang	1000
Jumlah Anak	0	1000
Agama	Islam	100000
Pendidikan	S1	0000100
Pekerjaan	Guru	0000010000
Merokok	Tidak	0
Berkacamata	Tidak	0
Berolahraga	Ya	1
Berilbab	Ya	1

Data No 135

Atribut	Nilai awal	Hasil Binning
Jenis Kelamin	pria	1
Umur	28	0100
Tinggi	169	0010
Berat	70	0100
Berat Badan	gemuk	10000
Rambut	lurus	10000
Warna Kulit	sawo matang	01000
Warna Rambut	hitam	10000
Suku/Keturunan	Jawa	0001000000000
Status Pernikahan	lajang	1000
Jumlah Anak	0	1000
Agama	Islam	100000
Pendidikan	S1	0000100
Pekerjaan	swasta	010000000
Merokok	Tidak	0
Berkacamata	Tidak	0
Berolahraga	Ya	1
Berilbab	Tidak	0

Gambar 3. Hasil proses *binning* pada salah satu data

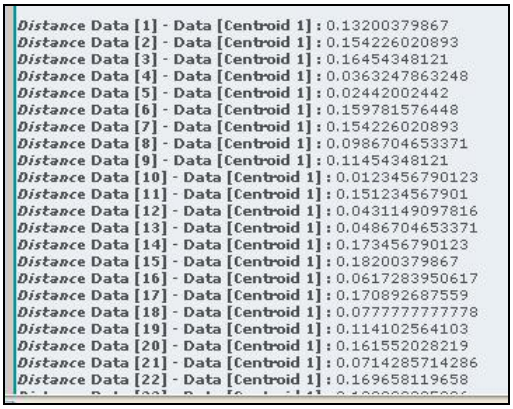
Sebagaimana aturan dalam algoritma k-means, jumlah *cluster* harus telah ditentukan sebelumnya. Dalam hal ini, jumlah *cluster* ditentukan oleh *user* melalui parameter masukan di awal program. Pada uji coba kali ini digunakan jumlah *cluster* yang akan dibentuk adalah sebanyak tiga buah *cluster*.

Gambar 4. Menentukan jumlah *cluster* yang akan dibentuk

Lalu, berdasarkan jumlah *cluster* yang telah diinputkan oleh *user*, secara acak dipilih dari data-data sebanyak jumlah *cluster* dan akan digunakan sebagai *cluster center* tahap pertama. Karena data-data yang dijadikan *cluster center* merupakan hasil dari pemilihan secara acak, maka *cluster center* – *cluster center* yang terbentuk pada tahap pertama ini mungkin tidak

akan sama dengan *cluster center – cluster center* yang terbentuk pada uji coba selanjutnya, meskipun jumlah *cluster* yang diinputkan adalah sama.

Setelah proses penghitungan jarak selesai dilakukan, di mana setiap data telah dihitung jaraknya dengan setiap *cluster center*, maka akan terbentuk tabel jarak seperti tampak di bawah ini :



Gambar 5. Jarak yang didapat dari proses penghitungan jarak

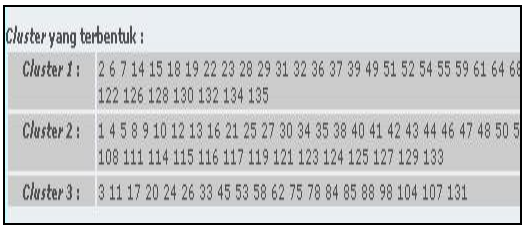
Kemudian dari hasil tersebut akan ditentukan setiap data tersebut termasuk dalam *cluster* mana dari sejumlah *cluster* yang telah terbentuk. Hal ini dilakukan dengan menggunakan ketentuan bahwa suatu data akan bergabung kepada *cluster* di mana *cluster center*-nya memiliki jarak terdekat dengan data.

Sebagai contoh, dari proses penghitungan jarak, didapat bahwa jarak antara *Data 1* dan masing-masing *cluster center* adalah sebagai berikut.

Cluster 1 : 0.247876814543
Cluster 2 : 0.0980531813865
Cluster 3 : 0.13200379867

Maka dapat ditentukan bahwa *Data 1* akan menempati *cluster 2* karena nilai jarak tersebut lebih kecil dibandingkan jarak antara *Data 1* dengan *cluster center 1* maupun *cluster center 3.* Proses penentuan ke *cluster* mana suatu data akan masuk juga dilakukan juga untuk setiap data yang lain, sehingga setiap data akan tergabung ke dalam suatu *cluster* yang jarak antara *cluster center* dan data memberikan hasil yang terdekat.

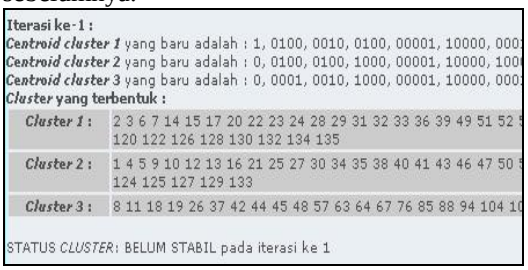
Setelah proses tersebut selesai maka akan terbentuk *cluster-cluster* awal beserta anggota-anggotanya. Contoh hasil bentukan *cluster-cluster* awal dengan memasukkan inputan jumlah *cluster* sebanyak 3 adalah seperti yang terlihat pada gambar di bawah :



Gambar 6. Contoh *cluster-cluster* awal yang terbentuk

Sebagaimana pembentukan *cluster center – cluster center* pada tahap pertama, pembentukan *cluster-cluster* awal ini juga mempunyai kemungkinan untuk berbeda dengan *cluster-cluster* yang terbentuk pada uji coba selanjutnya, meskipun *user* juga menggunakan input jumlah *cluster* yang sama.

Setelah tercipta *cluster center – cluster center* yang baru, akan dilakukan pembandingan antar *cluster center* yang baru dengan *cluster center* yang lama. Proses pembandingan ini akan memberikan dua kemungkinan: *cluster center – cluster center* yang baru sama persis dengan *cluster center – cluster center* sebelumnya atau *cluster center – cluster center* yang baru masih berbeda dari *clsuter center – cluster center* sebelumnya. Suatu *cluster* dianggap belum stabil jika nilai dari tiap-tiap *cluster center*-nya masih berbeda, jika dibandingkan dengan *cluster center* sebelumnya.



Gambar 7. *Cluster-cluster* baru yang masih belum stabil

Sebaliknya, suatu *cluster* dianggap stabil apabila *cluster center* sudah tidak berubah lagi jika dibandingkan dengan *cluster center* sebelumnya. Iterasi akan selesai jika telah terbentuk *cluster* yang stabil sebanyak jumlah *cluster* yang telah diinputkan sebelumnya.

Pada uji coba *Proses Clustering* dengan menggunakan inputan jumlah *cluster* sebanyak tiga buah, didapat *cluster-cluster* stabil seperti terlihat pada gambar berikut ini.

Gambar 8. *Cluster-cluster* baru yang telah stabil

Dari hasil tersebut (dengan terbentuknya *cluster-cluster* yang telah stabil), maka dapat diketahui karakteristik dari masing-masing *cluster*. Karakteristik ini didapat dengan cara melihat atribut mana yang paling sering muncul dalam suatu *cluster*. Karakteristik yang terbentuk adalah seperti di berikut ini.

Untuk *cluster 1*, karakteristiknya adalah sebagai berikut :

Kebanyakan anggotanya adalah : Berjenis kelamin pria, berumur antara 26 - 35 tahun, bertinggi badan antara 161 - 170 cm dengan berat badan antara 61 - 70 kg, mempunyai bentuk badan ideal dengan warna kulit kuning langsat, berambut lurus warna hitam. Berasal dari suku/keturunan Jawa dan beragama Islam. Status pernikahannya lajang dengan jumlah anak tidak ada. Ciri-ciri lainnya adalah mempunyai pekerjaan sebagai Tidak Bekerja dengan latar pendidikan S1. Selain itu, tidak merokok, tidak berkacamata, aktif berolahraga, dan tidak berjilbab.

Untuk *cluster* 2, karakteristiknya adalah sebagai berikut :

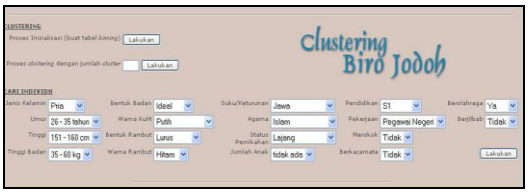
Kebanyakan anggotanya adalah : Berjenis kelamin wanita, berumur antara 26 - 35 tahun, bertinggi badan antara 151 - 160 cm dengan berat badan antara 35 - 60 kg, mempunyai bentuk badan ideal dengan warna kulit putih, berambut lurus warna hitam. Berasal dari suku/keturunan Cina dan beragama Islam. Status pernikahannya lajang dengan jumlah anak tidak ada. Ciri-ciri lainnya adalah mempunyai pekerjaan sebagai Tidak Bekerja dengan latar pendidikan S1. Selain itu, tidak merokok, tidak berkacamata, aktif berolahraga, dan tidak berjilbab.

Sedangkan untuk *cluster* 2, karakteristiknya adalah sebagai berikut :

Kebanyakan anggotanya adalah : Berjenis kelamin wanita, berumur antara 26 - 35 tahun, bertinggi badan antara 151 - 160 cm dengan berat badan antara 35 - 60 kg, mempunyai bentuk badan ideal dengan warna kulit kuning langsat, berambut lurus warna hitam. Berasal dari suku/keturunan Jawa dan beragama Islam. Status pernikahannya pernah bercerai dengan jumlah anak tidak ada. Ciri-ciri lainnya adalah mempunyai pekerjaan sebagai Tidak Bekerja dengan latar pendidikan S1. Selain itu, tidak merokok, tidak berkacamata, aktif berolahraga, dan tidak berjilbab.

9. UJI COBA PENCARIAN INDIVIDU

Pencarian individu diujicobakan dengan memasukkan beberapa parameter nilai atribut kepada sistem, yaitu sebanyak 18 parameter sesuai dengan data birojodoh.



Gambar 9. Form yang digunakan dalam proses *Pencarian Individu*

Apabila sebelumnya telah dilakukan proses *binning* dan telah dibentuk *cluster* sebanyak tiga buah, lalu pada pencarian individu dimasukkan parameter sebagai berikut:

Jenis kelamin: Wanita
Umur: 26 - 35 tahun
Tinggi badan: 151 - 160 cm
Berat badan: 35 - 60 kg
Bentuk badan: Ideal
Rambut: Lurus
Warna kulit: Putih
Warna rambut: Hitam
Suku/keturunan: Jawa
Status pernikahan: Lajang
Jumlah anak: tidak ada
Agama: Islam
Pendidikan terakhir: S1
Pekerjaan: Pegawai Negeri
Merokok: Tidak
Berkacamata: Tidak
Aktif berolahraga: Ya
Berjilbab: Tidak

Maka dari parameter-parameter tersebut akan didapatkan hasil sebagai berikut (diurutkan berdasarkan tingkat kesamaan antara data pencarian dan data-data dalam *cluster* yang terpilih) :

Gambar 10. Hasil dari proses pencarian individu

Dari hasil tersebut diketahui bahwa data pencarian termasuk ke dalam *cluster* ke-2 dari total 3 *cluster* dan menghasilkan 52 *result. Result* diurutkan berdasarkan tingkat kesamaannya. Dua data pertama yang didapat

dari hasil uji coba Proses Pencarian Individu adalah :

1. Data nomor 40 - (100%)
Ciri-ciri fisik : wanita (29 tahun) dengan status lajang.
Berasal dari suku/keturunan Jawa. Beragama Islam.
Tubuh ideal (155cm/40kg) berkulit putih, dengan rambut lurus warna hitam.
Pendidikan terakhirnya adalah S1 dan pekerjaannya sekarang adalah PNS
Merokok : Tidak.
Berkacamata : Tidak.
Aktif berolahraga : Ya.
Berjilbab : Tidak.
Jumlah anak (apabila ada) : 0 orang.

2. Data nomor 133 - (94.44%)
Ciri-ciri fisik : wanita (32 tahun) dengan status lajang.
Berasal dari suku/keturunan Jawa. Beragama Islam.
Tubuh ideal (158cm/44kg) berkulit putih, dengan rambut lurus warna hitam.
Pendidikan terakhirnya adalah S1 dan pekerjaannya sekarang adalah wiraswasta
Merokok : Tidak.
Berkacamata : Tidak.
Aktif berolahraga : Ya.
Berjilbab : Tidak.
Jumlah anak (apabila ada) : 0 orang.

Pada *result* nomor 1, terdapat data yang mempunyai tingkat kesamaan 100% atau dengan kata lain, atribut-atribut data tersebut bernilai sama persis dengan apa yang dicari. Kemudian disusul oleh data 133 yang berada di posisi kedua dalam daftar *result.* Data ini memiliki tingkat kesamaan 94,44% (perbedaannya ada pada atribut pekerjaan, yaitu *wiraswasta* bukan *pegawai negeri*).

Result yang didapat tidak mesti sama persis dengan data pencarian. Jadi, ada kemungkinan tidak terdapatnya *result* yang benar-benar memiliki tingkat kesamaan 100%. Pada *result* seperti ini setiap data pada *cluster* tersebut masih memiliki *error* pada salah satu atau lebih atribut yang dicari.

Hasil pencarian ini juga tergantung dari proses-proses sebelumnya, yaitu penentuan jumlah *cluster*, dan tentunya proses *clustering* itu sendiri.

10. KESIMPULAN

Kesimpulan yang dapat diambil dari uji coba yang telah dilakukan, antara lain sebagai berikut :

1. Proses *clustering*, yang menggunakan algoritma k-means, dapat juga digunakan untuk data-data yang beratribut non-numerik (*categorical*) yang telah ditransformasi ke dalam bentuk biner (numerik) sehingga dapat diolah. Proses transformasi data seperti ini disebut juga sebagai proses normalisasi atau *binning* (membinerkan). Contoh dalam hal ini adalah data suatu biro jodoh, yang kebanyakan nilai kategori dari atribut-atribut datanya adalah berupa *categorical*.
2. Pencarian suatu data (individu yang diinginkan/didamba berdasarkan atribut-atribut kriteria) dapat dilakukan dengan mencari jarak antara data yang dicari dengan masing-masing *cluster center* dari *cluster-cluster* yang telah stabil. Lalu dicari *cluster center* mana yang mempunyai jarak terdekat dengan data untuk dijadikan sebagai *cluster* acuan. Hasil dari proses pencarian individu tidak selalu akan memberikan hasil yang memiliki tingkat kesamaan 100% dengan data yang dicari, ada kalanya data atau bahkan semua data dalam *cluster* acuan tersebut memiliki tingkat kesamaan di bawah 100% dengan data yang dicari.

11. DAFTAR PUSTAKA

1. Fayyad, Usama, Chaudhuri, Surajit, Bradley, Paul, "Data mining and its Role in Database Systems", 2000.
2. Nguyen Thi Minh Hai, "Parallel Clustering Algorithms for Categorical and Mixed Data", Multi-Media Systems Laboratory, Japan Advanced Institute of Science and Technology, 2004.
3. Frank Keller, "Clustering Connectionist and Statistical Language Processing", Computer Linguistik Universität des Saarlandes, 2002.
4. Oracle Corporation, "Oracle 9i, Data mining Concept", 2002.