Integrasi Algoritma Girvan-Newman dan K-Means untuk Segmentasi Pelanggan berdasarkan Deteksi Komunitas pada Graf di E-Commerce

TESIS

Oleh

Ihsan Satriawan

23513008

PROGRAM STUDI MAGISTER INFORMATIKA

SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA

INSTITUT TEKNOLOGI BANDUNG

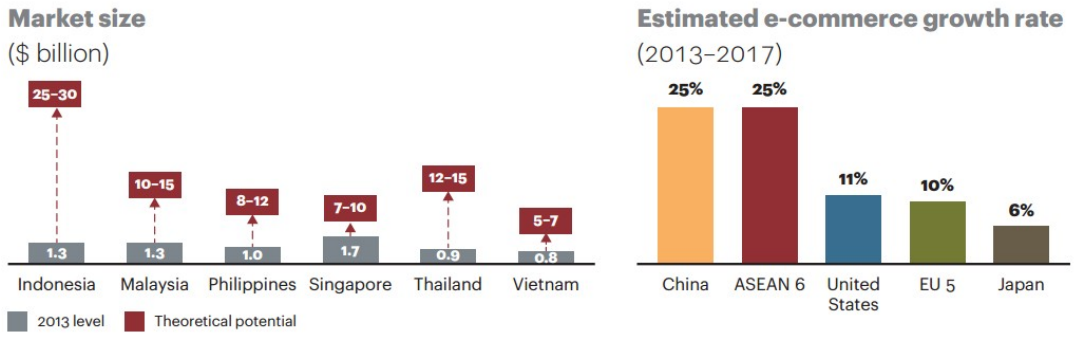
2016

BAB I

Pendahuluan

I.1. Latar Belakang

E-commerce merupakan salah satu sektor perdagangan yang sedang berkembang pesat ditengah gencarnya penetrasi internet dimasyarakat Indonesia. Gambar 1.1, menunjukkan bahwa Indonesia memiliki potensi pasar yang besar dibandingkan negara ASEAN lainnya dengan pesatnya penetrasi internet serta cukup besarnya nilai pasar *e-commerce* di Indonesia, tidak aneh jika bermunculan pemain-pemain disektor ini seperti Hijup, Bukalapak, Tokopedia, Kaskus, Elevenia, MatahariMall, dll. Dengan banyaknya pemain di sektor ini. mengakibatkan ketat nya persaingan yang terjadi antar pemain. Kondisi persaingan yang begitu kompetitif membuat perusahaan menyadari betapa pentingnya peran pelanggan, karena pelanggan yang menjadi alasan tetap bertahannya suatu perusahaan.



Gambar 1.1. ASEAN market potencial (sumber : A.T. Kearney analysis)

Oleh karena itu, perusahaan harus dapat memahami perilaku dari pelanggannya sehingga dapat memberikan pelayanan yang lebih baik. Salah satu cara untuk dapat memahami pelanggan adalah dengan membagi pelanggan berdasarkan karakteristik tertentu ke dalam beberapa kelompok (Miranda dan Henriques, 2013) (Tsiptsis dan Chorianopoulos, 2009). Dengan demikian pelanggan yang bergabung dalam sebuah

kelompok merupakan pelanggan-pelanggan yang mirip atau berhubungan satu sama lain serta berbeda atau tidak berhubungan dengan pelanggan dalam kelompok yang lain.

Dengan mengelompokkan pelanggan ke dalam segmentasi yang berbeda, maka dapat mempermudah marketing dalam merancang strategi penjualan dengan merancang strategi sesuai dengan karakteristik dari setiap segmentasinya. Terdapat peneliti yanng memanfaatkan algoritma K-Means untuk mendapatkan segmentasi pelanggan dengan RFM sebagai dasar pembentukan segmentasinya. RFM sendiri merupakan metode yang digunakan untuk menganalisa karakteristik pelanggan, RFM model terdiri dari *Recency* menunjukkan waktu terakhir pelanggan membeli barang*, Frequency* merupakan jumlah pembelian yang telah dilakukan oleh pelanggan dalam rentang waktu tertentu, dan *Monetary* merupakan jumlah uang yang dikeluarkan pelanggan untuk membeli barang pada rentang waktu tertentu (Bhunnak dkk, 2015).

Namun dengan melakukan segmentasi pelanggan hanya berdasarkan karakteristik pelanggan saja, hasil yang didapat kurang maksimal, hal ini dikarenakan bisa jadi terdapat beberapa pelanggan yang sebenarnya tergabung dalam satu kelompok berdasarkan interaksi nya, namun dikarenakan karakteristik pembelian nya berbeda, maka pelanggan tersebut terpisah ke dalam kelompok yang berbeda. Dengan demikian perlu ada nya proses tambahan untuk menangani hal tersebut.

Deteksi komunitas merupakan pengelompokan dengan melakukan pemisahan atau pemecahan ke dalam sejumlah kelompok berdasarkan karakteristik tertentu yang diinginkan. Struktur komunitas pada graf dikenal juga sebagai *cluster, groups,* atau *module* (Palla dkk, 2005) yang semua nya memiliki karakteristik dasar yang sama, yakni sekumpulan simpul dengan hubungan didalam kelompok yang erat dan hubungan yang jarang di antara kelompok (Girvan dan Newman, 2001). Dengan demikian deteksi komunitas pada graf dapat digunakan untuk membantu perusahaan memahami pelanggan. Algoritma Girvan-Newman merupakan algoritma yang populer digunakan untuk mendeteksi komunitas pada suatu graf (Moon dkk, 2014) (Lunagariya dkk, 2014)

(Bocu dan Tabirca, 2010) (Kameyama dkk, 2007).

Penelitian ini akan membahas integrasi algoritma Girvan-Newman dengan K-Means untuk melakukan segmentasi pelanggan. Algoritma Girvan-Newman digunakan untuk mencari komunitas pelanggan yang memiliki hubungan erat dengan e-commerce lalu kemudian akan dilakukan segmentasi pelanggan menggunakan K-Means dari pelanggan yang memiliki hubungan erat dengan e-commerce tersebut untuk mengetahui karakteristik dari kelompok pelanggan tersebut.

I.2. Rumusan Masalah

Berdasarkan latar belakang masalah yang ditelah diuraikan, maka permasalahan utama dalam penelitian dapat dirumuskan dalam *research question* sebagai berikut :

1. Apakah algoritma Girvan-Newman dan K-Means dapat diintegrasikan untuk melakukan segmentasi pelanggan ?
2. Bagaimana integrasi algoritma Girvan-Newman dan K-Means untuk melakukan segmentasi pelanggan ?
3. Bagaimana hasil atau kinerja integrasi algoritma Girvan-Newman dan K-Means dalam melakukan segmentasi pelanggan ?

I.3. Tujuan Penelitian

Tujuan dari penelitian ini adalah mendapatkan segmentasi pelanggan dan mengetahui hasil dan kinerja integrasi algoritma Girvan-Newman dan K-Means dalam melakukan segmentasi pelanggan.

I.4. Batasan Masalah

Dalam upaya agar penelitian ini tetap fokus pada tujuan utamanya maka perlu ditambahkan batasan persoalan yang ingin diselesaikan. Berikut adalah batasan yang ditetapkan dalam penelitian ini:

1. Komunitas yang dimaksud dalam penelitian ini merupakan komunitas yang terdapat dalam graf
2. Graf yang digunakan adalah graf tidak berarah dan berbobot
3. Data interaksi yang digunakan merupakan data interaksi penggunaan point oleh pelanggan

I.5. Metodologi

Metodologi penelitian ini adalah eksperimen dengan tahapan pengerjaan sebagai berikut :

1. Studi Pustaka

   Tahap studi pusataka bertujuan untuk memperkaya pemahaman mengenai konsep dari hal-hal yang berkaitan dengan deteksi komunitas, algoritma Girvan-Newman, algoritma K-Means, Segmentasi, RFM.

2. Pengumpulan Data

   Data yang digunakan merupakan data yang dapat merepresentasikan interaksi antar pelanggan yang dapat direpresentasikan ke dalam bentuk graf dan data yang merepresentasikan karakteristik pelanggan.

3. Praproses Data

   Tahapan praproses data menggunakan teknik yang sudah ada menyesuaikan dengan bentuk data.

4. Implementasi Metode

   a) Melakukan pembentukan komunitas menggunakan algoritma Girvan-Newman

   b) Melakukan segmentasi pelanggan menggunakan algoritma K-Means berdasarkan komunitas yang terbentuk.

5. Analisis Hasil dan Kesimpulan

   (a) Analisis hasil pembentukan komunitas dan segmentasi pelanggan menggunakan algoritma Girvan-Newman dan K-Means

6. Pembuatan Laporan

   Penulisan laporan penelitian berdasarkan sumber referensi yang menjadi teori dasar penelitian dan hasil penelitian berupa tahap-tahap dari setiap kegiatan yang dilakukan.

I.6. Sistematika Penulisan

Penulisan laporan tesis ini dilakukan secara sistematika dengan dibagi ke dalam beberapa bab seperti di bawah ini :

1. Bab I Pendahuluan

   Bab ini menjelaskan tentang beberapa hal, yakni : latar belakang tesis, rumusan masalah, tujuan penelitian, batasan masalah, metode penelitian dan sistematika penulisan

2. Bab II Tinjauan Pustaka

   Bab tinjauan pustaka menjelaskan beberapa teori atau konsep dasar yang mendukung pengembangan tesis, yakni konsep graf, deteksi komunitas, algoritma Girvan-Newman, algoritma K-Means, RFM, Segmentasi.

3. Bab III Analisis dan Perancangan Solusi

   Bab ini menjelaskan tentang analisis masalah segmentasi pelanggan pada deteksi komunitas menggunakan algoritma Girvan-Newman pada graf tidak berarah dan berbobot serta analisis integrasi Girvan-Newman dan K-Means untuk melakukan segmentasi pelanggan

4. Bab IV Implementasi dan Pengujian

   Bab ini menjelaskan hasil dan analisis dari pembentukan segmentasi pelanggan dan pengaruh algoritma Girvan-Newman pada K-Means terhadap hasil segmentasi pelanggan pada deteksi komunitas graf tidak berarah dan berbobot.

5. Bab V Kesimpulan dan Saran

   Pada bagian ini akan dijelaskan kesimpulan dari hasil penelitian yang telah dilakukan, serta saran untuk melakukan pengembangan dari hasil penelitian.

BAB II

Studi Literatur

II.1. Graf

Teori graf adalah cabang ilmu yang mempelajari sifat–sifat graf. Secara informal, suatu graf adalah himpunan benda-benda yang disebut vertex (node) yang terhubung oleh edge-edge (arc). Biasanya graf digambarkan sebagai kumpulan titik (melambangkan vertex) yang dihubungkan oleh garis-garis (melambangkan edge). Definisi yang lebih formal adalah suatu graf G yang dapat dinyatakan sebagai G = <V,E>. Graf G terdiri atas himpunan V yang berisikan vertex/node pada graf tersebut dan himpunan dari E yang berisi edge pada graf tersebut. Himpunan E dinyatakan sebagai pasangan dari vertex (Munir, 2003).

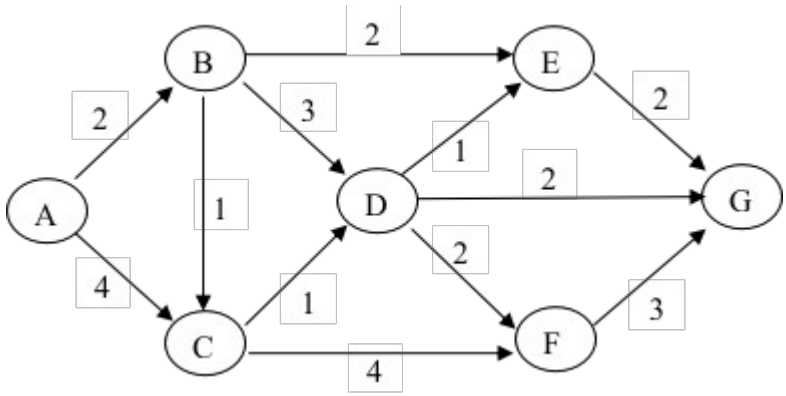
Terdapat jenis-jenis graf, yaitu graf berarah dan berbobot, graf tidak berarah dan berbobot, graf berarah dan tidak berbobot, dan graf tidak berarah dan tidak berbobot. Berikut ini adalah gambaran dan penjelasan dari jenis-jenis graf yang telah disebutkan:

1. graf berarah dan berbobot, tiap edge mempunyai anak panah dan mempunyai bobot. Gambar II.1 merupakan contoh graf yang memiliki arah dan bobot.

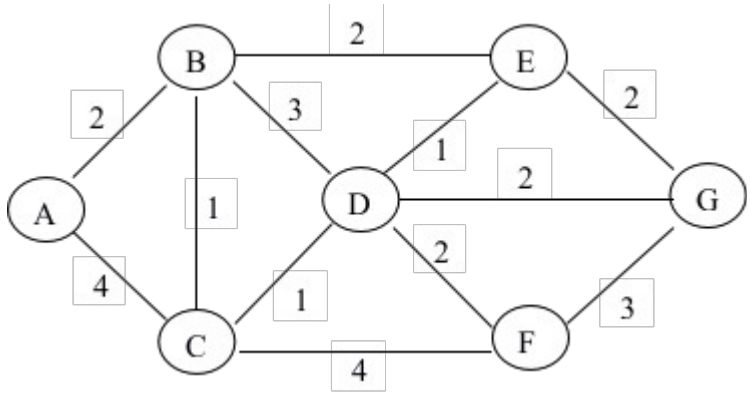
2. graf tidak berarah dan berbobot, tiap edge tidak mempunyai anak panah tetapi mempunyai bobot. Gambar II.2 merupakan contoh graf yang tidak berarah dan berbobot.

3. graf berarah dan tidak berbobot, tiap edge mempunyai arah yang tidak berbobot. Gambar II.3 merupakan contoh graf berarah dan tidak berbobot.

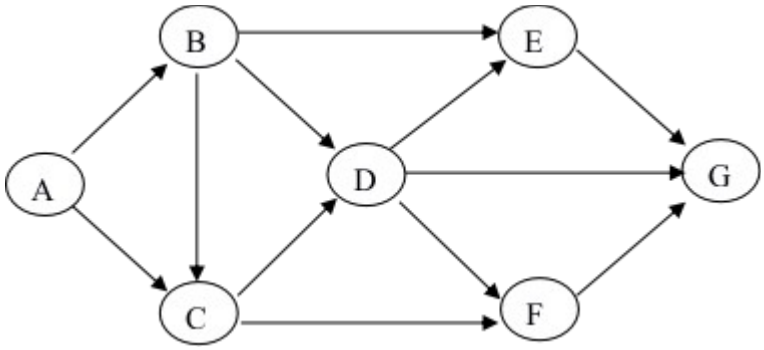
4. graf tidak berarah dan tidak berbobot, tiap edge tidak mempunyai arah dan tidak juga berbobot. Gambar II.4 merupakan contoh graf yang tidak berarah dan tidak berbobot.



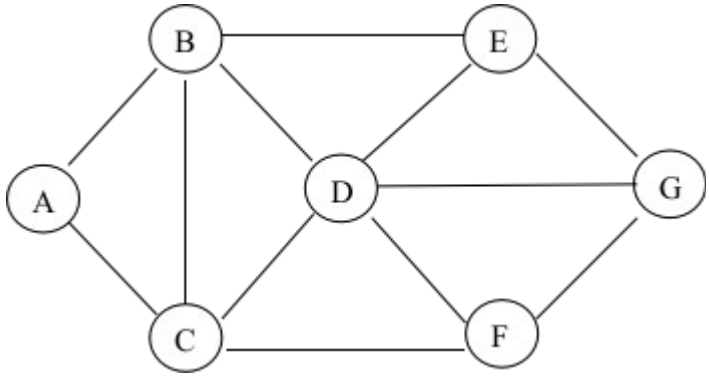
Gambar II.1. Graf berarah dan berbobot



Gambar II.2. Graf tidak berarah dan berbobot



Gambar II.3. Graf berarah dan tidak berbobot



Gambar II.4. Graf tidak berarah dan tidak berbobot

II.2. Euclidean Similarity

Euclidean similarity merupakan ukuran kesamaan yang umum digunakan dalam mencari kesamaan dari dua objek. Pada prinsipnya, perhitungan *euclidean similarity* mengukur jarak antar dua objek yang dihitung menggunakan teorema pythagoras (Segaran, 2007)

$$d(p1,p2)=\sqrt{\sum (s_{p1}-s_{p2})^2} \quad (2.1)$$

$$\frac{1}{1+d(p1,p2)} \quad (2.2)$$

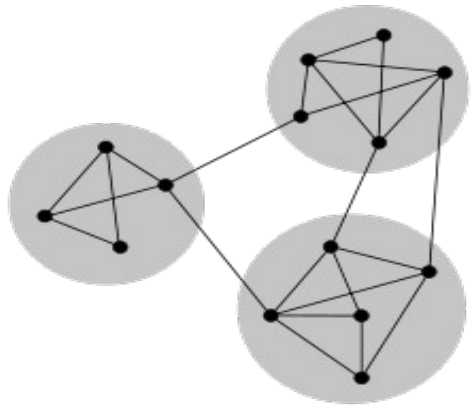
II.3. Deteksi Komunitas

Komunitas merupakan sub graf yang ada di dalam suatu jaringan. Komunitas terdiri dari sekumpulan *node* yang memiliki keterhubungan yang erat didalamnya, tetapi hubungan antar komunitas merupakan hubungan yang lemah. Deteksi komunitas merupakan pengelompokan dengan melakukan pemisahan ke dalam sejumlah kelompok berdasarkan karakteristik tertentu yang diingkan. Dalam terminologi data mining, deteksi komunitas termasuk ke dalam analisis kelompok (*cluster analysis*). *Cluster analysis* merupakan proses pengelompokkan data (objek) yang didasarkan hanya pada informasi yang ditemukan dalam data yang menggambarkan objek tersebut

dan hubungan di antaranya. Tujuannya adalah agar objek-objek yang bergabung dalam sebuah kelompok merupakan objek-objek yang mirip satu sama lain dan berbeda dengan objek dalam kelompok yang lain. Tujuan pengelompokan dibagi menjadi dua, yakni pengelompokan untuk pemahaman, dan pengelompan untuk penggunaan.

Pengelompokan untuk pemahaman merupakan pengelompokan yang dibentuk berdasarkan struktur yang ada pada data. Pada umumnya pengelompokan ini merupakan proses awal yang nantinya akan dilanjutkan dengan peringkasan, pelabelan kelas pada setiap kelompok. Pengelompokan untuk penggunaan merupakan pengelompokan yang bertujuan untuk mencari prototipe kelompok yang paling representatif terhadap data.

Hal penting dalam deteksi komunitas adalah menemukan komunitas dari objek berdasarkan kesamaan konten dan hubungan antar objek. Hubungan ini dapat dicirikan dengan kesamaan antar objek, seperti kesamaan konten, jenis kelamin, usia, jumlah *products* yang dibeli, dan lain-lain. Hubungan tersebut dapat digambarkan kedalam bentuk graf. Graf mewakili relasi antar objek, dimana *node* dianggap sebagai objek dan sisi dianggap sebaga relasi antar dua objek. Pada kenyataannya, *node-node* yang berada dalam satu komunitas yang sama memiliki karakteristik yang mirip. Gambar II.5 menunjukan contoh komunitas dalam jaringan.



Gambar II.5. Contoh komunitas dalam jaringan

II.4. Algoritma Girvan-Newman

Algoritma Girvan-Newman pertama kali dikemukakan oleh Michelle Girvan dan Mark Newman, merupakan metode yang digunakan untuk mendeteksi komunitas dalam suatu graf. Algoritma Girvan-Newman menggunakan konsep dari *hierarchical clustering* dengan metode *divisive.* Untuk memaksimalkan hasil dari algoritma ini digunakan nilai kualitas yang dapat dihitung menggunakan perhitungan *network modularity*.

Ide dasar dibalik algoritma Girvan-Newman adalah dengan mengidentifikasi sisi-sisi yang menghubungkan simpul-simpul yang berbeda komunitas untuk kemudian secara progresif menghapusnya sehingga struktur komunitas muncul. Dalam algoritma Girvan-Newman, pengidentifikasian sisi-sisi tersebut dilakukan melalui perhitungan *edge betweenness centrality.* Langkah algoritma Girvan-Newman unruk mendeteksi komunitas adalah sebagai berikut (Girvan dan Newman, 2001) :

1. Hitung *edge betweenness* untuk semua sisi dalam jaringan.
2. Hapus sisi dengan nilai *betweenness* tertinggi.
3. Hitung ulang *edge betweenness* untuk sisi yang dipengaruhi oleh penghapusan sisi pada langkah 2.
4. Ulangi langkah 2 sampai semua sisi terpotong.

Pada algoritma Girvan-Newman perhitungan *betweenness* menjadi perhitungan yang paling penting dalam pembentukan komunitas. Perhitungan *betweenness* akan dijelaskan pada sub-bab berikutnya.

II.4.I. Betweenness

Pada algoritma Girvan-Newman perhitungan nilai *betweenness* merupakan pehitungan paling penting, hal ini dikarenakan proses algoritma Girvan-Newman melakukan perhitungan *betweenness* secara berulang-ulang. Pada algoritma ini, perhitungan

betweenness yang dipakai adalah nilai *betweenness* dari sisi (*edge betweenness*). *Betweenness* dari suatu sisi adalah nilai yang dimiliki oleh suatu sisi dimana nilai tersebut sama dengan jumlah jalur terpendek yang melaluinya. Konsep *betweenness* yang digunakan dalam algoritma Girvan-Newman adalah jembatan diantara komunitas pasti memiliki nilai *edge betweenness* yang tinggi. Normalnya untuk graf $G:=(V,E)$ dengan n vertikal (dengan sisi), *betweenness* yang menggunakan jalur terpendek $C_{B-F}(v)$ untuk vertex v adalah (Izquierdo dkk, 2006) :

$$C_{B-F}(v)=\sum_{i=2}^{n}(w_i.\sum_{s\neq v\neq t\in v}(\frac{\sigma_{st}(i,v)}{\sigma_{st}(i)})) \quad (2.1)$$

Keterangan :

w_i = bobot yang ditentukan untuk jalur dengan panjang *I*

$\sigma_{st}(i)$ = jumlah dari jalur dengan panjang *i* dari *s* ke *t*

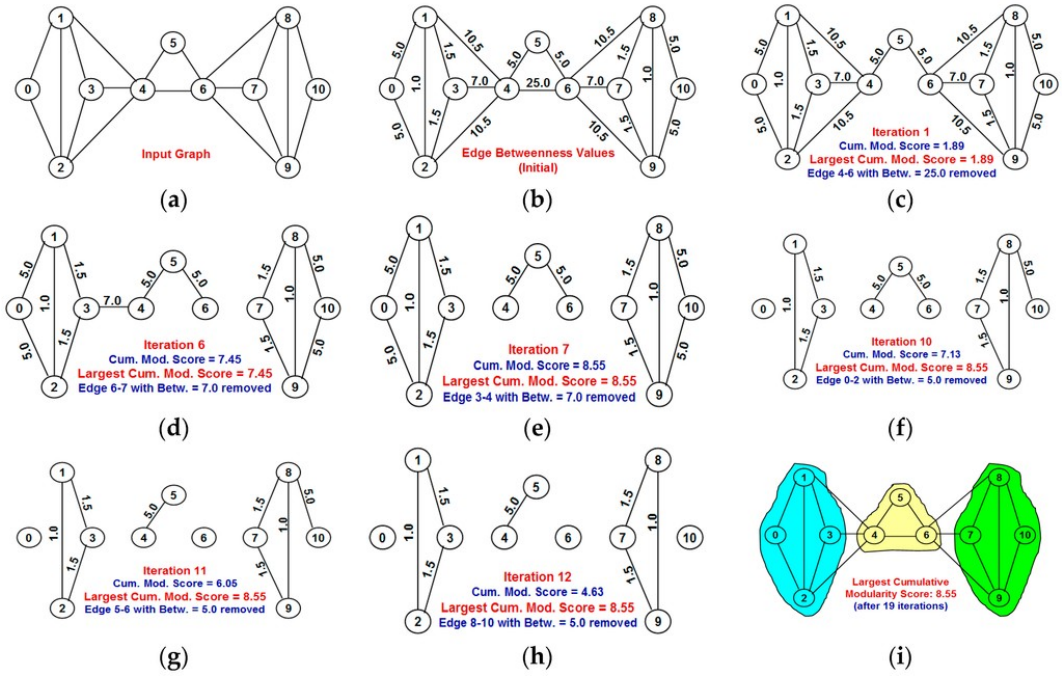
$\sigma_{st}(i,v)$ = jumlah jalur dengan panjang *i* dari *s* ke *t* yang melewati vertex *v*

Algoritma *betweenness* di atas untuk melakukan perhitungan *betweenness* memerlukan waktu $O(n^3)$, sehingga dalam menghitung nilai *betweenness* dalam jaringan yang besar memerlukan komputasi yang mahal. Brandes mampu menghitung *betweenness* dengan waktu $O(nm+n^2\log n)$ pada graf berbobot (Alahakoon, 2011). Penelitian ini perhitungan nilai *betweenness* dilakukan menggunakan algoritma Brandes dengan menggunakan *edge betweenness* dan graf berbobot. Gambar II.6 menunjukkan pseudocode yang digunakan (Brandes, 2007) dan gambar II.7 menunjukkan ilustrasi konsep *edge betweenness*.

input: directed graph $G = (V, E)$
data: queue Q, stack S (both initially empty) and for all $v \in V$:
$dist[v]$: distance from source
$Pred[v]$: list of predecessors on shortest paths from source
$\sigma[v]$: number of shortest paths from source to $v \in V$
$\delta[v]$: dependency of source on $v \in V$
output: betweenness $c_B[v]$ for all $v \in V$ (initialized to 0)

```
for s ∈ V do
    ▼ single-source shortest-paths problem
        ▼ initialization
            for w ∈ V do Pred[w] ← empty list
            for t ∈ V do dist[t] ← ∞;   σ[t] ← 0
            dist[s] ← 0;   σ[s] ← 1
            enqueue s → Q
        while Q not empty do
            dequeue v ← Q;   push v → S
            foreach vertex w such that (v, w) ∈ E do
                ▼ path discovery  // — w found for the first time?
                    if dist[w] = ∞ then
                        dist[w] ← dist[v] + 1
                        enqueue w → Q
                ▼ path counting  // — edge (v, w) on a shortest path?
                    if dist[w] = dist[v] + 1 then
                        σ[w] ← σ[w] + σ[v]
                        append v → Pred[w]
    ▼ accumulation  // — back-propagation of dependencies
        for v ∈ V do δ[v] ← 0
        while S not empty do
            pop w ← S
            for v ∈ Pred[w] do δ[v] ← δ[v] + (σ[v]/σ[w]) · (1 + δ[w])
            if w ≠ s then c_B[w] ← c_B[w] + δ[w]
```

Gambar II.6 Pseudocode *edge betweenness*



Gambar II.7 Ilustrasi konsep edge betweenness

II.5. Modularity

Modularity merupakan properti atau sifat yang dimiliki oleh graf dan dapat digunakan sebagai penilaian standar dalam menentukan kualitas komunitas yang ada pada graf tersebut. Komunitas dapat ditentukan dengan ukuran *modularity*, yaitu ukuran kekuatan pembagian suatu jaringan menjadi satu atau beberapa komunitas. Semakin besar nilai *modularity* artinya semakin kuat hubungan simpul dengan lingkungannya dan membentuk *dense network*. *Dense network* akan membentuk komunitas, sedangkan lawannya *sparse network* akan menjadi penghubung antar komunitas atau tidak terhubung sama sekali.

Modularity banyak digunakan sebagai tolak ukur pembagian jaringan. Nilai *modularity* yang tinggi dari suatu partisi jaringan merepresentasikan bahwa partisi tersebut memiliki sisi intra-komunitas yang lebih erat tetapi inter-komunitasnya yang jarang. *Modularity* nilainya selalu lebih kecil dari 1 dan bernilai 0 jika semua simpul dalam jaringan berada dalam komunitas yang sama. Nilai *modularity* yang besar menandakan struktur komunitas yang kuat.

II.5.I. Modularity Girvan dan Newman

Sudah banyak peneliti yang mengusulkan algoritma untuk membentuk suatu komunitas, seperti *spectral dichotomy method based on graf theory* , dan *K-L algorithm*, algoritma Girvan-Newman (Girvan dan Newman, 2001) berdasarkan penghapusan sisi yang dilakukan secara bertahap. Masalah umum yang seringkali dihadapi oleh metode tersebut adalah belum dapat memberikan informasi yang objektif terkait berapa banyak komunitas yang harus terbagi di dalam jaringan. Terkait masalah tersebut, Newman dan Girvan mengusulkan untuk mengukur kualitas dari pembagian jaringan dengan *modularity* (*modularity* GN).

Modularity mengukur tingkat penyimpangan jumlah sisi yang menghubungkan simpul-simpul dari komunitas di jaringan dari nilai rata-rata dalam jaringan acak dengan urutan tingkat simpul yang sama. Berikut ini rumus *modularity* yang diusulkan oleh Girvan

dan Newman :

$$Q=\frac{1}{2m}\sum vw[A_{vw}-\frac{(k_v k_w)}{2m}]\sigma(c_v,c_w) \quad (2.2)$$

Keterangan :

A_{vw} merupakan elemen dari matriks *adjacency* : bernilai 1 jika vertex *v* dan *w* terhubung dan bernilai 0 jika tidak terhubung

c_v merupakan komunitas yang didalamnya terdapat vertex *v.*

c_w merupakan komunitas yang didalamnya terdapat vertex *w.*

Fungsi $\delta(c_v,c_w)$ merupakan simbol Kronecker delta, akan bernilai 1 apabila $c_v=c_w$ dan 0 untuk lainnya.

m adalah jumlah sisi yang ada pada graf.

k_v merupakan derajat dari simpul *v*.

k_w merupakan derajat dari simpul *w.*

Jumlah sisi yang dimungkinkan terbentuk antara simpul *v* dan *w* dihitung melalui $\frac{k_v k_w}{2m}$

Pembagian graf ke dalam komunitas yang baik berkorespondensi dengan nilai *modularity* yang besar, sebaliknya pembagian graf ke dalam komunitas yang buruk berkorespondensi dengan nilai *modularity* yang kecil. *Modularity* inilah yang akhirnya menjadi standar yang akurat dalam mengukur kualitas pembagian struktur komunitas pada graf (deteksi komunitas).

II.6. RFM Model

Konsep RFM merupakan teknik marketing yang digunakan untuk menganalisa perilaku pelanggan seperti kapan terakhir kali *pelanggan* melakukan pembelian *products* (*recency*), seberapa sering pelanggan melakukan pembelian (*frequency*), dan berapa banyak uang yang sudah pelanggan keluarkan (*monetary*). Pendekatan tersebut berguna membagi pelanggan menjadi beberapa kelompok (*customer segmentation*) untuk mengetahui pelanggan mana saja yang akan meninggalkan layanan ataupun mengetahui pelanggan mana saja yang menjadi sumber penjualan terbesar (Tsiptsis dan Chorianopoulos, 2009). Gambar II.8 menunjukkan contoh data RFM model.

CustomerID	Recency (Day)	Frequency (Number)	Monetary (TL)
1	3	6	540
2	6	10	940
3	45	1	30
4	21	2	64
5	14	4	169
6	32	2	55
7	5	3	130
8	50	1	950
9	33	15	2430
10	10	5	190
11	5	8	840
12	1	9	1410
13	24	3	54
14	17	2	44
15	4	1	32

Gambar II.8 Contoh data RFM model

Kolom *renceny* menunjukkan berapa jarak hari dari pelanggan terakhir membeli *products* sampai rentang hari yang ditentukan. Berdasarkan gambar II.5, untuk *pelangganID* 2 nilai *recency*-nya menunjukkan nilai 6, itu menandakan *customerID* 2 melakukan transaksi 6 hari sebelumnya. Kolom *frequency* menunjukkan seberapa sering pelanggan melakukan transaksi. Untuk *customerID* 2, nilai *frequency*-nya menunjukkan nila 10, ini menandakan bahwa *customerID* selama ini sudah melakukan

10 pembelian. Sedangkan Kolom *monetary* menunjukkan sebarap banyak uang yang telah *pelanggan* keluarkan dalam transaksi selama ini.

Ketiga nilai tersebut dapat digunakan untuk mengidentifikasi nilai dari pelanggan. Nilai pelangan dapat digunakan untuk melihat perilaku pelanggan yang akan tetap melakukan pembelian dimasa yang akan datang, berikut ini penjelasannya:

1. Semakin kini waktu dari seorang pelanggan memiliki produk dari suatu perusahaan, maka semakin banyak kemungkinan pelanggan tersebut akan melakukan pembelian kembali pada masa yang akan datang.

2. Semakin sering seorang pelanggan membeli produk dari suatu perusahaan, maka semakin banyak kemungkinan pelanggan tersebut akan melakukan pembelian kembali pada masa yang akan datang.

3. Semakin banyak uang dari seorang pelanggan yang digunakan untuk membeli produk dari suatu perusahaan, maka semakin banyak kemungkinan pelanggan tersebut akan melakukan pembelian kembali pada masa yang akan datang.

II.7. K-Means

K-Means merupakan salah satu metode *clustering* yang mengelompokkan data ke dalam bentuk satu atau lebih *cluster*/kelompok. Data yang memiliki karakteristik sama/mirip dikelompokkan ke dalam satu *cluster*/kelompok dan data yang memiliki karakteristik berbeda dikelompkkan dengan *cluster*/kelompok yang lain (Huang dan Song, 2014). Langkah-langkah melakukan *clustering* dengan metode K-Means sebagai berikut :

1. Pilih jumlah *cluster k*

2. Inisialisasi *k* pusat *cluster,* pusat *cluster* diberi nilai awal dengan angka random

3. Alokasikan semua data/objek ke *cluster* terdekat. Kedekatan dua objek ditentukan

berdasarkan jarak kedua objek tersebut. Demikian juga kedekatan suatu data ke *cluster* tertentu ditentukan jarak antara data dengan pusat *cluster*. Pada tahap ini perlu dihitung jarak tiap data ke tiap pusat *cluster*. Jarak paling dekat antara satu data dengan satu *cluster* tertentu akan menentukan suatu data masuk kedalam *cluster* mana. Untuk menghitung jarak semua data ke setiap titik pusat *cluster* dapat menggunakan teori jarak Euclidean yang dirumuskan sebagai berikut

$$D(i,j)=\sqrt{(X_{1i}-X_{1j})^2+(X_{2i}-X_{2j})^2+\ldots+(X_{ki}-X_{kj})^2} \quad (2.3)$$

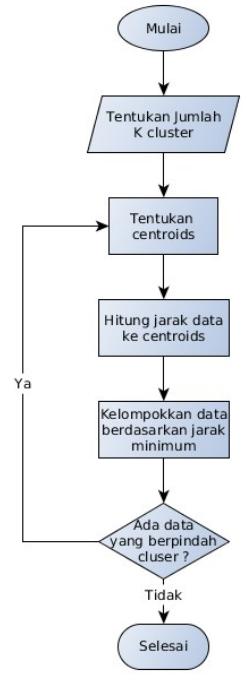
dimana :

$D(i,j)$ = Jarak data ke *i* ke pusat *cluster j*

X_{ki} = Data ke *i* pada atribut data ke *k*

X_{kj} = Titik pusat ke *j* pada atribut ke *k*

Gambar II.6 menunjukkan *flowchart* dari proses K-Means.



Gambar II.9 *Flowchart* K-Means

Kelemahan dari metode K-Means, diawal perlu ditentukan jumlah *k cluster* yang akan dibentuk oleh metode K-Means. Hal ini menjadi hal krusial ketika belum diketahui jumlah *cluster* yang ingin dibentuk. Oleh karena itu perlu ada mekanisme tambahan untuk mengevaluasi *cluster* yang terbentuk, salah satu nya menggunakan *silhouette index*. Penjelasa *silhouette index* akan dijelaskan pada sub bab berikutnya.

II.7.I. Silhouette Index

Silhouette index digunakan untuk mengetahui kualitas dan kekuatan *cluster*, sebarapa baik suatu objek ditempatkan dalam suatu *cluster*. *Silhouette index* melakukan pengujian kualitas *cluster* dengan memperhitungkan jarak suatu objek pada *cluster* tertentu dengan objek pada *cluster* yang lain. Langkah perhitungan *silhouette index* sebagai berikut (Ding dkk, 2014):

1. Hitung rata-rata jarak dari suatu objek misalkan *i* dengan semua objek lain yang berada dalam satu *cluster*

$$a(i)=\frac{1}{|A|-1}\sum j\in A, j\neq i\, d(i,j) \quad (2.4)$$

dengan *j* adalah objek lain dalam satu *cluster* A dan *d(i,j)* adalah jarak antar objek *i* dengan *j*.

2. Hitung rata-rata jarak dari objek *i* tersebut dengan semua objek di *cluster* lain, dan diambil nilai terkecilnya.

$$d(i,C)=\frac{1}{|A|}\sum j\in C\, d(i,j) \quad (2.5)$$

dengan *d(i,C)* adalah jarak rata-rata objek dokumen *i* dengan semua objek pada *cluster* lain C dimana A ≠ C.

$$b(i)=min\, C\neq A\, d(i,C) \quad (2.6)$$

3. Nilai *silhouette index* nya adalah :

$$s(i)=\frac{b(i)-a(i)}{max(a(i),b(i))} \quad (2.7)$$

II.8. Penelitian Terkait

Berikut ini merupakan beberapa penelitian yang pernah dilakukan mengenai segmentasi pelanggan serta deteksi komunitas pada graf, yaitu:

Tabel II.1 Penelitian Terkait

No	Judul	Penulis, Tahun	Deskripsi Penelitian	Metode
1	Applying Data Mining	Bunnak,	Pengelompokkan pelanggan	clustering

	Techniques and Extended RFM Model in pelanggan Loyalty Measurement	Thammaboosadee, dan Kiattisin. (2015)	berdasarkan data transaksi pembelian	(K-Means Clustering) dan decision tree
2	Vip pelanggan Segmentation Based on Data Mining in Mobile-communications Industry	Yihua (2010)	Pengelompokkan pelanggan berdasarkan karakteristik gender dan pekerjaan	clustering (K-Means Clustering) dan decision tree
3	Building clusters for CRM startegies by mining airlines pelanggan data	Miranda, H. S., dan Henriques, R. (2013)	Melakukan perbandingan metode dalam pengelompokkan pelanggan berdasarkan demografi dan aktifitas menggunakan penerbangan	k-means, SOM and H-SOM
4	Community structure in social and biological networks	Girvan, M., dan Newman, M. (2001)	Melakukan deteksi komunitas pada graf dengan konsep edge betweenness	Girvan-Newman
5	Integrasi Konsep Coupling Degree Untuk Perhitungan Modularitas Dalam Analisis Kualitas Struktur Komunitas Pada Weighted Graph	Mairisha, M. (2016)	Melakukan deteksi komunitas pada graf berbobot data twitter	Girvan-Newman

Berdasarkan penelitian-penelitian tersebut masih belum terdapat proses secara menyeluruh terkait segmentasi pelanggan yang dilakukan berdasarkan komunitas yang terbentuk dari interaksi antar pelanggan, kemudian melakukan analisa karakteristik pelanggan berdasarkan nilai RFM.

Bab III

Analisis dan Perancangan

III.1. Analisis Permasalahan

Setiap perusahaan yang sudah memutuskan untuk beroperasi dengan menjalankan bisnis nya dalam pasar yang luas, harus menyadari bahwa sulit untuk memberikan melayani kebutuhan seluruh pelanggan. Hal ini dikarenakan pada dasar nya setiap pelanggan memiliki kebutuhan dan perilaku yang berbeda, sehingga menyebabkan pasar yang terbentuk bersifat heterogen. Dengan demikian perusahaan perlu melakukan identifikasi pasar untuk memilah-milah pelanggan ke beberapa kelompok atau biasa disebut segmentasi pelanggan. Untuk melakukan segmentasi pelanggan, yang harus dilakukan adalah memilih kriteria dasar yang paling tepat untuk membagi pelanggan. Pada umum nya kriteria dasar yang digunakan untuk membagi pelanggan sebagai berikut (Tsiptsis dan Chorianopoulos, 2009):

1. Geografis

   Segmentasi berdasarkan wilayah, misal melakukan segmentasi untuk pelanggan yang tinggal di daerah barat atau timur.

2. Demografis

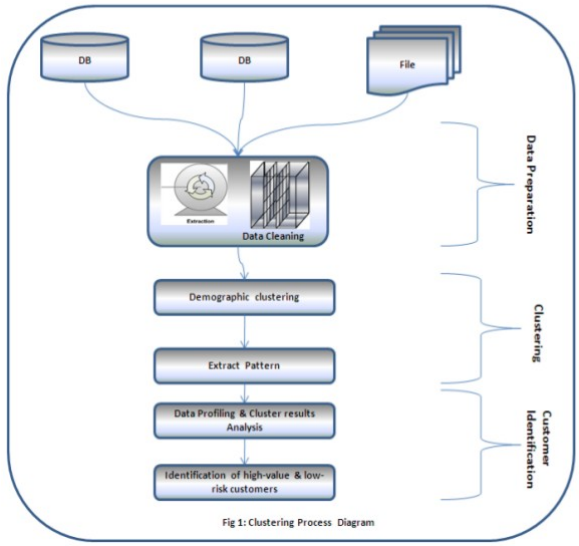
   Demografis bisa berupa usia, jenis kelamin, pendidikan, dll. Karena memiliki kriteria tersendiri, misal nya segmentasi pelanggan dengan usia diatas 30 tahun atau segmentasi pelanggan dengan jenis kelamin wanita.

Di kondisi pasar yang kompetitif saat ini, pendekatan ini tidak cukup. Sebaliknya, organisasi harus memiliki pandangan yang lengkap dari pelanggan mereka dalam rangka untuk mendapatkan keuntungan kompetitif. Mereka juga harus fokus pada

kebutuhan pelanggan mereka, ingin, sikap, perilaku, preferensi, dan persepsi, dan untuk menganalisa data yang relevan untuk mengidentifikasi segmen yang mendasari. Identifikasi kelompok dengan karakteristik unik akan memungkinkan organisasi untuk mengelola dan target mereka lebih efektif dengan, antara lain, penawaran produk disesuaikan dan promosi

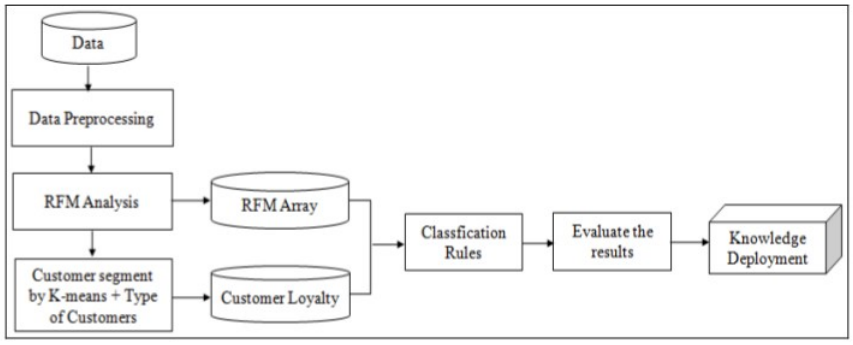
III.2. Solusi Permasalahan

Dari sumber literatur, terdapat pendekatan yang telah digunakan untuk menemukan segmentasi pelanggan pada *e-commerce*. Salah satunya adalah pendekatan dengan memanfaatkan *data mining* dalam proses segmentasi pelanggan yang dijelaskan pada literatur (Rajagopal, 2011), (Bhunnak dkk, 2015). Metode *data mining* yang dimanfaatkan dalam proses segmentasi pelanggan adalah *demographic clustering process* (Rajagopal, 2011). Rajagopal melakukan segmentasi pelanggan dengan melakukan dua tahapan, yakni tahap pertama melakukan data *cleansing,* lalu pada tahap kedua melakukan identifikasi profile dari segmentasi pelanggan yang terbentuk dari *demographic clustering process.* Gambar III.1 menggambarkan proses segmentasi pelanggan yang dilakukan oleh Rajagopal.



Gambar III.1. Proses Segmentasi Pelanggan Rajagopal

Terdapat juga pendekatan lain yang dilakukan oleh peneliti lain, seperti yang dilakukan oleh Panwad Bhunnak. Bhunnak melakukan segmentasi pelanggan dengan memanfaatkan metode K-Means berdasarkan nilai RFM untuk setiap pelanggan (Bhunnak dkk, 2015). Namun peneliti menambahkan sebuah nilai tambahan untuk setiap data pelanggan yang akan diolah, yakni melakukan identifikasi tipe pelanggan berdasarkan nilai RFM setiap pelanggan. Gambar III.2 menggambarkan proses segmentasi pelanggan yang dilakukan oleh Bhunnak.



Gambar III.2. Proses Segmentasi Pelanggan Bhunnak

Ketika akan mengimplementasikan dua pendekatan diatas, terdapat beberapa kelemahan, antara lain:

1. Segmentasi pelanggan yang dihasilkan belum memperhatikan kelompok/komunitas pelanggan yang terbentuk berdasarkan kedekatan antar pelanggan, karena data pelanggan yang digunakan sebagai dasar pembentuk segmentasi belum melewati proses deteksi kelompok/komunitas pelanggan.

Pada penelitian ini proses segmentasi pelanggan yang diusulkan adalah pemanfaatan proses deteksi komunitas dalam proses segmentasi pelanggan berdasarkan nilai RFM setiap pelanggan. Pemanfaatan deteksi komunitas ini bertujuan untuk membentuk

kelompok pelanggan yang memiliki interaksi kuat didalam kelompok yang sama dan interaksi kurang kuat dengan pelanggan di kelompok yang berbeda. Proses segmentasi pelanggan akan dilakukan berdasarkan nilai RFM setiap pelanggan yang berada dalam kelompok yang sama, sehingga segmentasi pelanggan yang dihasilkan dapat lebih akurat karena proses segmentasi pelanggan menggunakan pelanggan yang memiliki interaksi kuat satu sama lainnya.

III.3. Rancangan Solusi yang Diusulkan

Proses solusi yang diusulkan adalah proses segmentasi pelanggan berbasis nilai RFM dengan memanfaatkan proses deteksi komunitas yang dapat menghasilkan segmentasi pelanggan yang lebih akurat. Proses solusi yang diusulkan terdiri dari dua tahap, yaitu:

1. Tahap Deteksi Komunitas

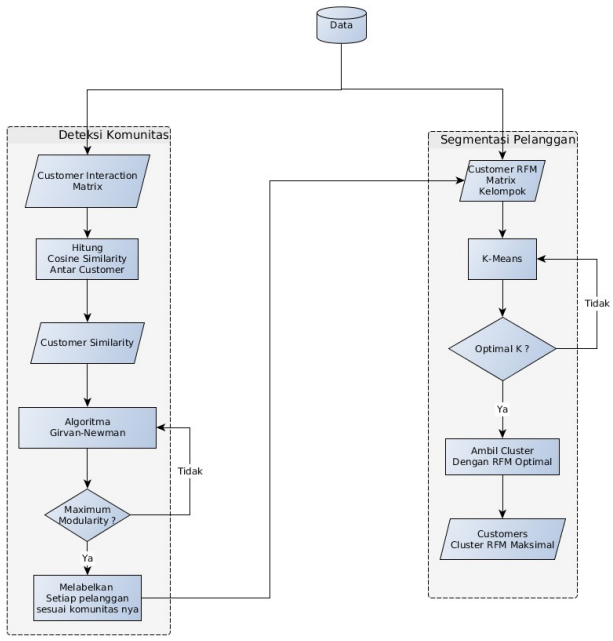
   Tahap deteksi komunitas bertujuan untuk melakukan pengelompokkan pelanggan suatu *e-commerce*. Tahap ini bertujuan untuk mengidentifikasi pelanggan ke dalam komunitas/kelompok yang memiliki interaksi kuat satu sama lain. Masukan tahap ini adalah data interaksi pelanggan, sedangkan keluaran tahap ini adalah jumlah komunitas/kelompok dengan pelanggan yang menjadi anggota didalam nya.

2. Tahap Segmentasi Pelanggan

   Tahap segmentasi pelanggan bertujuan melakukan pembagian segmentasi pelanggan untuk kelompok/komunitas yang terbentuk pada tahapan sebelumnya. Proses segmentasi pelanggan dilakukan dengan melihat setiap nilai RFM dari pelanggan lalu kemudian dikelompokkan berdasarkan kemiripan nilai RFM nya. Masukan tahapan ini adalah daftar pelanggan untuk setiap komunitas/kelompok yang terbentuk dari proses tahap deteksi komunitas dengan nilai RFM nya masing-masing. Keluaran tahap ini adalah daftar pelanggan yang menjadi anggota

segmentasi masing-masing.

Masing-masing tahapan terdiri dari proses yang dapat dilihat lebih detail pada gambar III.3.



Gambar III.3. Kerangka proses segmentasi pelanggan yang diusulkan

III.3.1. Tahap Deteksi Komunitas

Sebelum masuk ke tahap deteksi komunitas, maka data mentah yang akan digunakan perlu dilakukan pengolahan dalam tahap praproses. Tahap praproses merupakan tahap untuk melakukan transformasi data dari data mentah menjadi data yang siap digunakan untuk masuk ke tahap berikutnya, tabel III.1 merupakan contoh data mentah yang harus ditransformasi

Tabel III.1. Contoh Data Mentah

Email Pengirim	Email Penerima	Point	Tanggal
123@gmail.com	222@yahoo.com	2	01-01-2016
333@yahoo.com	111@yahoo.com	4	02-01-2016
123@gmail.com	222@yahoo.com	5	01-01-2016
...	…	…	...

Proses yang dilakukan berdasarkan langkah-langkah berikut :

1. Mengubah email baik untuk penerima ataupun penerima menjadi sebuah user id dengan lookup ke tabel user yang tersedia
2. Melakukan agregasi data berdasarkan berapa kali email penerima pernah melakukan pengiriman point ke email penerima

Tabel III.2 merupakan contoh dataset yang akan digunakan setelah melakukan langkah-langkah praproses diatas.

Tabel III.2. Contoh Dataset yang digunakan

id_pengirim	id_penerima	frekuensi
41290	41601	4
41290	13183	2
29338	68613	8
13183	41290	4
...	...	...

Berdasarkan tabel III.2 diatas, maka dapat dilihat bawah id 41290 pernah berinteraksi dengan id 13183 sebanyak 2 kali dalam bentuk mengirimkan point dan id 13183 pernah berinteraksi dengan id 41290 sebanyak 4 kali dalam bentuk mengirimkan point

Setiap pelanggan suatu *e-commerce* memiliki ketertarikan yang berbeda satu sama lain, namun pelanggan bisa dikelompokkan ke dalam kelompok berdasarkan ketertarikan yang sama antar pelanggan. Pengelompokkan pelanggan bisa didasari dari banyak hal, antara lain dari demografi, kelamin atau interaksi antar pelanggan. Pada penelitian ini, pelanggan akan dikelompokkan berdasarkan interaksi melalui penggunaan point reward yang dimiliki setiap pelanggan. Untuk memudahkan pengolahan data, data interaksi, pelanggan yang berupa pemanfaatan point reward, yang pernah digunakan oleh setiap pelanggan diubah ke dalam bentuk tabel yang merupakan hasil keluaran dari tahap praproses, seperti berikut:

Tabel III.3 Contoh interaksi pelanggan

id_pengirim	id_penerima	frekuensi
41290	41601	4
41290	13183	2
29338	68613	8
13183	41290	4
…	…	...

Untuk mengelompokkan pelanggan berdasarkan interaksi penggunaan point yang pernah digunakan, maka perlu dihitung kemiripan (*similarity*) antar pelanggan. Pada penelitian ini perhitungan *similarity* akan menggunakan perhitungan *euclidean similarity*. Berdasarkan tabel III.1, maka bisa didapatkan matriks *similarity* pelanggan

sebagai berikut:

Tabel III.2 Contoh matriks *similarity* pelanggan

	41290	41601	13183	29338	68613
41290	0	0.20000000000000001	0.33333333333333331	0	0
41601	0.20000000000000001	0	0	0	0
13183	0.33333333333333331	0	0	0	0
29338	0	0	0	0	0.1111111111111111
68613	0	0	0	0.1111111111111111	0

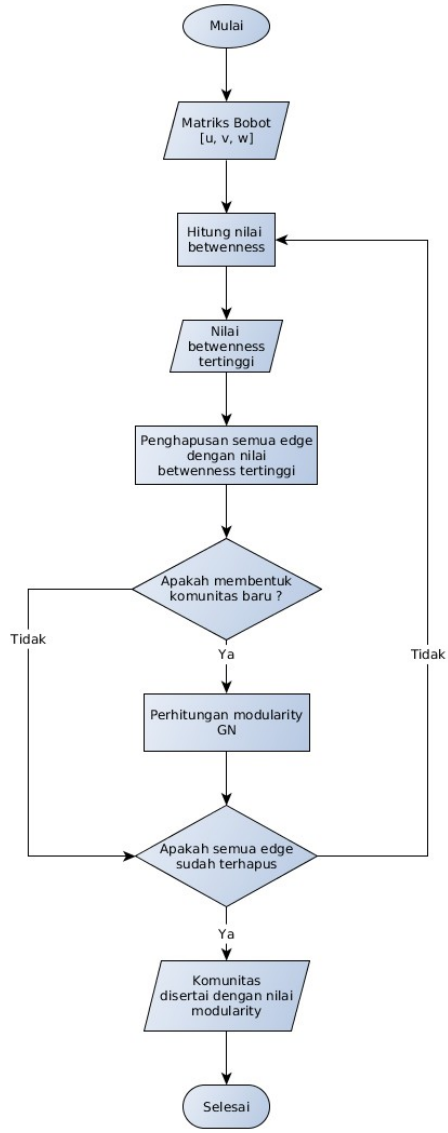
Untuk mencari *similarity* antar pelanggan, maka perlu mencari nilai jumlah frekuensi untuk setiap pelanggan, lalu dihitung menggunakan formula (2.2). Pada tabel IV.2 berlaku interaksi bolak-balik, sehingga untuk mendapatkan *similarity* nya harus menggunakan kedua nilai tersebut. Dengan menggunakan formula (2.2) kita dapat mengetahui *similarity* pelanggan id 41290 dengan pelanggan id 13183

euclidean_similarity = 1 / (1 + euclidean_distance(2, 4))

euclidean_similarity = 0.33333333333333331

pelanggan *similarity* matriks yang terbentuk akan menjadi masukan untuk algoritma pendeteksi komunitas, pada penelitian ini menggunakan algoritma Girvan-Newman sebagai algoritma deteksi komunitas. Matriks bobot yang merupakan pelanggan

similarity matriks awalnya akan dihitung nilai *betweeness*-nya. Kemudian sisi yang memiliki nilai *betweeness* tertinggi akan dipotong karena sisi tersebut kurang memiliki kemiripan dengan entitas yang ada disekitarnya, atau dengan kata lain sisi tersebut sering dilalui oleh jalur terpendek (berdasarkan konsep *edge betweeness*). Proses pemotongan sisi dilakukan sampai semua sisi terpotong. Penyertaan pengukuran *modularity* pada algoritma pendeteksian komunitas dimaksudkan untuk mengevaluasi dengan memberikan penilaian terhadap setiap struktur komunitas yang dihasilkan dari algoritma deteksi komunitas disetiap iterasi. Pengukuran *modularity* yang digunakan pada penelitian ini adalah *modulariy* Girvan dan Newman (GN). Perhitungan *modularity* dilakukan disetiap pembentukan komunitas baru. Jumlah komunitas yang memiliki nilai *modularity* tertinggi berarti pembentukan komunitas tersebut sesuai dengan kenyataan yang sebenarnya. Berikut ini merupakan *flowchart* penyertaan *modularity* GN pada algoritma Girvan-Newman :



Gambar III.5 *Flowchart* algoritm Girvan-Newman dengan penyertaan perhitungan *modularity* GN

Jumlah komunitas dengan nilai *modularity* tertinggi akan dipilih menjadi keluaran tahap deteksi komunitas, sehingga setelah melewati tahap ini, pelanggan akan terkelompokkan kedalam komunitasinya. Untuk setiap komunitas yang terbentuk perlu dianalisa untuk dapat mengetahui karakteristik dari setiap komunitas yang terbentuk.

III.3.2. Tahap Segmentasi Pelanggan

Setiap pelanggan suatu *e-commerce* selain memiliki ketertarikan yang berbeda, memiliki perilaku (*behaviour*) yang berbeda juga. Pelaku e-commerce perlu mengidentifikasi perilaku pelanggan-nya agar memahami pelanggan sehingga dapat membuat strategi yang sesuai. Salah satunya dengan melakukan segmentasi pelanggan. Pada penelitian ini akan menggunakan RFM model sebagai dasar untuk melakuan segmentasi pelanggan serta menggunakan K-Means untuk mendapatkan kelompok berdasarkan nilai RFM setiap pelanggan.

Untuk setiap pelanggan, nilai R akan diperoleh berdasarkan kapan terakhir kali melakukan transaksi, nilai F akan diperoleh berdasarkan berapa kali pelanggan melakukan transaksi dalam rentang waktu tertentu, dan nilai M diperoleh berdasarkan jumlah uang yang sudah dibayarkan pelanggan selama melakukan transaksi dalam rentang waktu tertentu . Untuk nilai RFM setiap pelanggan, akan dilakukan proses normalisasi menjadi nilai dengan rentang 1-5. Gambar III.4 dan gambar III.5 merupakan gambaran proses nomalisasi yang dilakukan.

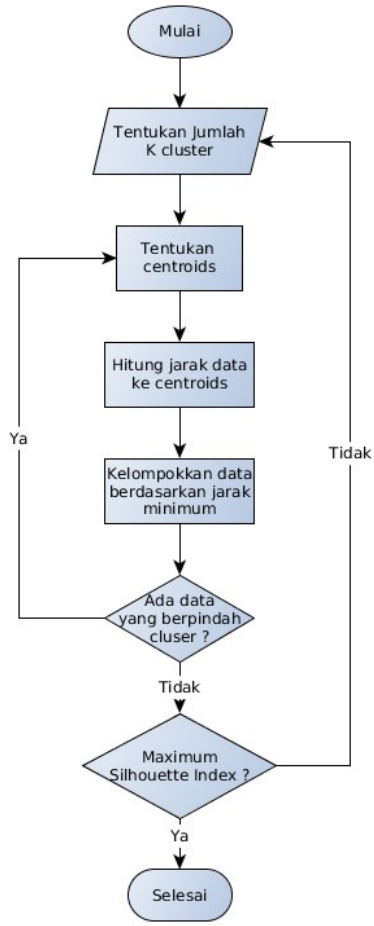
CustomerID	Recency (Day)	Frequency (Number)	Monetary (TL)
1	3	6	540
2	6	10	940
3	45	1	30
4	21	2	64
5	14	4	169
6	32	2	55
7	5	3	130
8	50	1	950
9	33	15	2430
10	10	5	190
11	5	8	840
12	1	9	1410
13	24	3	54
14	17	2	44
15	4	1	32

Gambar III.4 Contoh nilai RFM sebelum normalisasi

CID	Rec.	R	CID	Freq.	F	CID	Mon.	M
12	1	5	9	15	5	9	2430	5
1	3	5	2	10	5	12	1410	5
15	4	5	12	9	5	8	950	5
7	5	4	11	8	4	2	940	4
11	5	4	1	6	4	11	840	4
2	6	4	10	5	4	1	540	4
10	10	3	5	4	3	10	190	3
5	14	3	7	3	3	5	169	3
14	17	3	13	3	3	7	130	3
4	21	2	14	2	2	4	64	2
13	24	2	4	2	2	6	55	2
6	32	2	6	2	2	13	54	2
9	33	1	15	1	1	14	44	1
3	45	1	3	1	1	15	32	1
8	50	1	8	1	1	3	30	1

Gambar III.5 Contoh nilai RFM setelah normalisasi

Nilai RFM setiap pelanggan yang telah dinormalisasi tersebut yang akan menjadi masukan untuk proses *clustering* menggunakan algoritma K-Means. Proses *clustering* diawali dengan menentukan k jumlah *cluster* lalu tentukan nilai k pusat *cluster* dan kemudian alokasikan semua data ke *cluster* terdekat dengan menghitung jarak setiap data ke pusat *cluster* tersebut. Penelitian ini menggunakan *euclidean distance* untuk menghitung jaraknya. Setelah mengalokasikan semua data ke *cluster* yang terdekat, hitung kembali pusat *cluster* dengan anggota yang sekarang, pusat *cluster* didapat dari rata-rata dari semua data dalam *cluster* tersebut. Ulangi proses tersebut sampai pusat *cluster* tidak berubah lagi. Namun penentuan nilai k *cluster* menjadi kelemahan dari algoritma K-Means, sehingga perlu ada tahap evaluasi k *cluster* yang terbentuk. Penelitian ini menggunakan *silhouette index* untuk melalakukan evaluasi untuk *cluster* yang terbentuk, sehingga akan dipilih k *cluster* dengan nilai *silhouette index* tertinggi. Berikut ini *flowchart* K-Means dengan penyertaan nilai *silhouette index* :



Gambar III.6 *Flowchart* k-means dengan penyertaan *silhouette index*

Jumlah k *cluster* dengan nilai *silhouette index* tertinggi akan dipilih menjadi keluaran tahap segmentasi pelanggan, dengan demikian pelanggan akan ter-*cluster* ke dalam segmentasi nya. Untuk setiap *cluster* yang terbentuk dianalisa untuk dapat mengetahui mana *cluster* dengan nilai RFM terbaik.

Bab IV

Implementasi dan Pengujian

IV.1. Implementasi

Fungsi utama dalam program yang diimplementasikan dalam penelitian ini adalah untuk melakukan praproses, deteksi komunitas, dan segmentasi pelanggan terhadap data masukan berupa data interaksi antar pelanggan serta data transaksi pelangga. Sehingga menghasilkan keluaran berupa pelanggan yang sudah tersegmentasi. Program yang dibangun akan mengimplementasi sesuai proses yang diusulkan pada bab III.

Untuk keperluan implementasi proses yang diusulkan maka dibangun sebuah alat bantu berupa *software* yang dapat menjalankan proses yang diusulkan, berikut ini merupakan beberapa *software* yang digunakan :

1. PostgresSQL

   Sebagai basis data yang digunakan untuk menyimpan data mentah serta menyimpan hasil segmentasi pelanggan

2. Python versi 2.7.0

   Sebagai bahasa pemrograman yang digunakan untuk melakukan proses pengolahan data sehingga dapat menghasilkan segmentasi pelanggan sesuai dengan yang diusulkan

3. Library Scipy

   Library yang terdapat dalam ekosistem python, pada penelitian ini digunakan untuk menghitung similarity antar pelanggan

4. Library Pandas

Library yang terdapat dalam ekosistem python, pada penelitian ini digunakan untuk membantu dalam pengolahan data menjadi bentuk yang lebih mudah diolah

5. Library Scikit-Learn

   Library yang terdapat dalam ekosistem python, pada penelitian ini digunakan untuk implementasi algoritma K-Means untuk segmentasi pelanggan

IV.2. Pengujian

Pengujian merupakan bagian dari penelitian untuk melakukan pengujian terhadap *software* yang telah diimplementasikan. *Software* digunakan untuk melakukan deteksi komunitas dan segmentasi pelanggan terhadap data mentah yang tersedia untuk dievaluasi secara fungsionalitasnya. Dalam menguji fungsionalitas, *software* yang telah dibangun dievaluasi berdasarkan keluaran yang dihasilkan, apakah bentuknya sesuai dengan spesifikasi dan dapat dinyatakan cukup untuk menjawab tujuan penelitian ini.

IV.2.1. Data

Untuk melakukan pengujian terhadap proses yang diusulkan dalam penelitian ini, diperlukan data interaksi antar pelanggan serta data transaksi pelanggan. Data interaksi digunakan sebagai dasar pembentukan komunitas yang terjadi didalam pelanggan dengan melihat nilai *similarity* antar pelanggan. Sedangkan data transaksi pelanggan digunakan untuk membentuk nilai RFM sebagai nilai dasar pembentukan segmentasi pelanggan. Data interaksi antar pelanggan dan data transaksi yang digunakan dalam penelitian ini adalah data penggunaan point antar pelanggan dan data transaksi pelanggan di salah satu e-commerce fashion muslim Indonesia.

Poin yang dijadikan dasar data interaksi ini merupakan sebuah nilai tukar uang dalam e-commerce tersebut. Setiap pelanggan mendapatkan point setelah melakukan

pembelanjaan di e-commerce tersebut atau mendapatkan kiriman point dari pelanggan yang lain. Kondisi mendapat serta mengirimkan point antar pelanggan ini lah yang dijadikan sebagai data interaksi yang digunakan pada penelitian kali ini. Data transaksi yang digunakan terdiri dari kapan transaksi tersebut terjadi, berapa uang yang dibayarkan pelanggan pada transaksi tersebut.

Data yang diperoleh disimpan dalam sebuah tabel di aplikasi basis data untuk kemudahan pengelolaan dan manipulasi. Data mentah ini akan melalui tahap praproses untuk diolah menjadi masukan dalam proses yang diusulkan untuk mendapatkan segmentasi pelanggan.

IV.2.2. Hasil

Data mentah yang telah didapat, pertama kali dilakukan praproses sesuai yang dijelaskan pada bab III.3. Tabel IV.1 merupakan data yang didapat setelah melakukan praproses terhadap data mentah yang didapat

Tabel IV.1. Data hasil praproses

ID_Pengirim	ID_Penerima	Frekuens
79424	78112	2
64554	43874	2
48249	21061	18
…	…	...

Untuk setiap id pelanggan yang ada di data hasil praproses, akan dilakukan proses perhitungan kemiripan (*similarity*) menggunakan perhitungan *euclidean similarity*

sesuai yang ditunjukkan pada subbab III.3.1. Tabel IV.2 menunjukkan hasil dari perhitungan *similarity* antar pelanggan

Tabel IV.2. Data *Similarity* antar Pelanggan

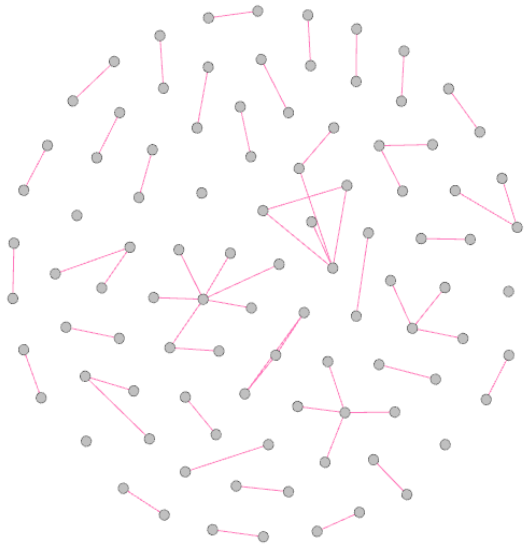
ID_Pelanggan	ID_Pelanggan	*Similarity*
8	15	0.333333333333
16	58	0.0526315789474
69	38	0.5
18	7	0.333333333333
…	…	...

Data *similarity* ini yang akan menjadi masukan ke dalam algoritma Girvan-Newma untuk dilakukan proses deteksi komunitas. Dari data *similarity* yang masuk ke dalam proses deteksi komunitas, terbentuklah 43 komunitas dengan nilai *modularity* sebesar 0.941531, tabel IV.3 merupakan hasil dari proses deteksi komunitas

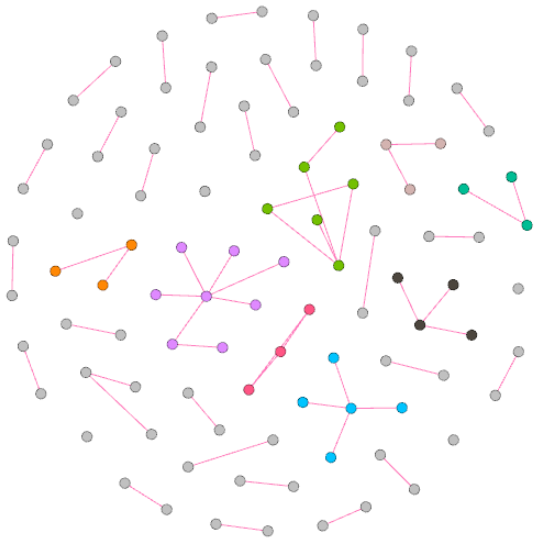
Tabel IV.3. Data Hasil Deteksi Komunitas

ID_Pelanggan	Komunitas
28996	0
38	0
16078	6
29167	6

36736	6
…	...



Gambar IV.1. Graph Interaksi Pelanggan



Gambar IV.2. Graph Interaksi Pelanggan setelah deteksi komunitas

Untuk setiap user id pelanggan yang telah terdeteksi masuk kekomunitas mana saja, maka dilakukan proses pembentukan nilai RFM nya. Nilai RFM didapat sesuai yang dijelaskan pada subbab III.3.2. Tabel IV.4 merupakan nilai RFM untuk setiap user id pelanggan yang sudah terdeteksi komunitas nya.

Tabel IV.4. Nilai RFM Pelanggan

ID_Pelanggan	Recency	Frequency	Monetary
29167	4	5	5
19758	3	3	3
28996	4	2	3
29244	5	3	2
…	…	…	...

Nilai RFM ini yang akan dijadikan sebagai kriteria pembentukan segmentasi pelanggan, kemudian sesuai yang sudah dijelaskan pada subbab III.3.2, pembentukan segmentasi menggunakan K-Means, percobaan dilakukan menggunakan rentang jumlah cluster antara 3 sampai 9 cluster, dengan rincian sebagai berikut:

Tabel IV.5. Hasil Percobaan dengan K-Means

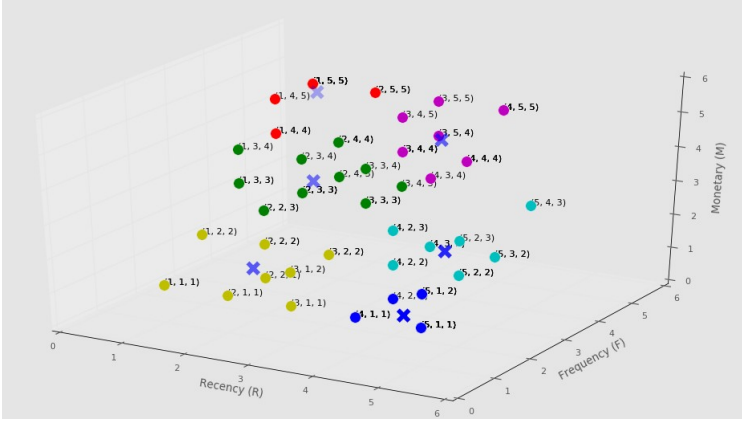
Jumlah Cluster	Nilai Silhouette
3	0.425606275001
4	0.450230267368
5	0.454735783457

6	0.455861425693
7	0.447197289496
8	0.445071577548
9	0.438796134647

Terlihat dari tabel IV.5 bahwa dengan jumlah cluster 6 memiliki nilai silhouette tertinggi dibandingkan nilai silhouette dengan jumlah cluster yang berbeda, sehingga akan dibentuk segmentasi pelanggan dengan jumlah 6 segmen. Hasil cluster dari K-Means memiliki sebuah nilai centroid untuk setiap cluster yang terbentuk, nilai ini dapat digunakan untuk mengetahui karakteristik setiap cluster sesuai nilai dasar pembentuknya, tabel IV.6 menunjukkan centroid untuk setiap cluster yang terbentuk.

Tabel IV. 6. Centroid setiap cluster

Cluster	Recency	Frequency	Monetary	Jumlah anggota
1	4.70588235	1.05882353	1.23529412	17
2	2.1	3.15	3.3	20
3	1.1875	4.8125	4.875	16
4	4.53846154	2.46153846	2.30769231	13
5	3.4375	4.3125	4.3125	16
6	2.09090909	1.54545455	1.54545455	11



Gambar IV.3. Visualisasi Segmentasi Pelanggan

IV.3. Analisis Hasil

DAFTAR REFERENSI

Alahakoon, T., Tripathi, R., Kourtellis, N., Simha, R., dan Lamnitchi, A. (2011): K-Path Centrality: A New Centrality Measure in Social Networks, IEEE

Bocu, R., dan Tabirca, S. (2010): Protein Communities Detection Optimization Through an Improved Parallel Newman-Girvan Algorithm, 9th RoEduNet IEEE International Conference 2010

Brandes, U. (2008): On variants of shortest-path betweenness centrality and their generic computation, ScienceDirect

Bunnak., Thammaboosadee., dan Kiattisin. (2015): Applying Data Mining Techniques and Extended RFM Model in pelanggan Loyalty Measurement. Journal of Advances in Information Technology Vol. 6, No. 4, November 2015

Ding., Xu., Liu., dan Wu.(2014): T-S Model Identification Based on Silhouette Index and Improved Gravitational Search Algorithm , IEEE

Girvan, M., dan Newman, M. (2001): Community structure in social and biological networks, Proc. of the National Academy of Science, 2002

Huang, Song, (2014): Clustering Analysis on E-commerce Transaction Based on K-means Clustering, JOURNAL OF NETWORKS, VOL. 9, NO. 2, FEBRUARY 2014

Izquierdo., Luis, R., Hanneman, A. (2006): Introduction To The Formal Analysis Of Social Networks Using Mathematica

Kameyama, S., Uchida, M., dan Shirayama, S. (2007): A New Method for Identifying Detected Communities Based on graf Substructure, 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology – Workshops

Lunagariya, D., Somayajulu, D., dan Krishna, P. (2014): SE-CDA: A Scalable and Efficient Community Detection Algorithm, 2014 IEEE International Conference on Big Data

Mairisha, M. (2016): Integrasi Konsep Coupling Degree Untuk Perhitungan Modularitas Dalam Analisis Kualitas Struktur Komunitas Pada Weighted Graph, Tesis Program Magister, Institut Teknologi Bandung.

Miranda, H. S., dan Henriques, R. (2013): Building clusters for CRM startegies by mining airlines pelanggan data, IEEE

Moon, S., Lee, J., dan Kang, M. (2014): Scalable Community Detection from Networks by Computing Edge Betweenness on MapReduce, IEEE

Munir, R. (2003). Matematika Diskrit Edisi Kedua. Bandung : Informatika

Palla, G., Derenyi, I., Farkas, I., dan Viscek, T. (2005): Uncovering the overlapping community structure of complex networks in nature and society, Nature 435, June 2005

Rajagopal, S. (2011): Customer Data Clustering Using Data Mining Technique, International Journal of Database Management Systems (IJDMS) Vol.3, No.4, November 2011

Segaran, T. (2007): *Programming Collective Intelligence*, O'Reilly Press

Tsiptsis, K., dan Chorianopoulos, A. (2009): Data Mining Techniques in CRM, Wiley

Yihua, Z. (2010): Vip Customer Segmentation Based on Data Mining in Mobile-communications Industry, IEEE