Integrated Girvan-Newman and K-means Algorithm for Customer Segmentation in E-commerce

Ihsan Satriawan
School of Electronic Egineering and Informatics
Institute of Technology Bandung
Bandung, Indonesia
ihsan.satriawan.20[at]gmail.com

G.A. Putri Saptawati
School of Electronic Egineering and Informatics
Institute of Technology Bandung
Bandung, Indonesia
putri[at]informatika.org

***Abstract*—Customer segmentation become one of the ways for a company to be able to provide better service to customers. By segmenting customers, company can be more understand behavior of customers. In fact, the approach which has been used to obtain customer segmentation is still inadequate, because the information generated is merely classify customers based on criteria established at the beginning, like the RFM value of every customer. This study proposes an additional process before doing customer segmentation, which is the process of detecting community formed by interaction between customers. This additional process called a community detection. With this additional processing, customer segmentation is expected to produce better information.**

***Keywords*—Data Mining, Clustering, Customer Segmentation, Community Detection**

I. Introduction

E-commerce merupakan salah satu sektor perdagangan yang sedang berkembang pesat. Menurut riset yang dilakukan oleh perusahaan konsultan A.T. Kearney pada tahun 2015, menunjukkan bahwa Indonesia memiliki potensi pasar online antara 25 - 30 milyar dollar.

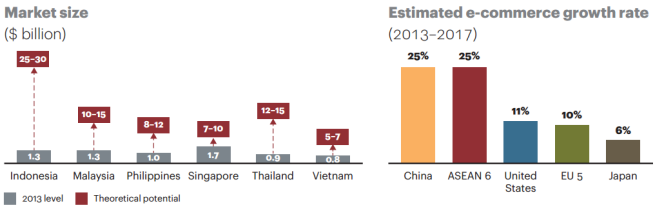

Fig. 1. ASEAN's Market Potency

Dengan besarnya nilai potensi pasar online di Indonesia, tidak aneh jika begitu banyaknya perusahaan di ranah e-commerce, seperti Tokopedia, Bukalapak, Hijup, Matahari-Mall, dll. Hal tersebut mengakibatkan persaingan yang terjadi begitu kompetitif, sehingga membuat perusahaan harus dapat memberikan pelayanan yang lebih baik dibandingkan yang lain. Segmentasi pelanggan merupakan salah satu cara yang dapat dilakukan oleh perusahaan dalam rangka memahami perilaku pelanggannya sehingga dapat meningkatkan pelayanan yang sudah ada.

Salah satu teknik yang dapat digunakan untuk memproses data yang dimiliki perusahaan untuk menghasilkan segmentasi pelanggan adalah data mining. Data mining merupakan teknik untuk menemukan pola dari data dan mendapatkan informasi yang berguna.

Clustering merupakan salah satu metode data mining untuk mengelompokkan objek berdasarkan karakteristik yang dimiliki. Dengan demikian clustering dapat dimanfaatkan perusahaan untuk melakukan segmentasi pelanggan, dengan melakukan pengelompokkan pelanggan berdasarkan karakteristik tertentu seperti demografi, pola pembelian, dll.

Deteksi komunitas menggunakan pendekatan graf dalam melakukan pengelompokkan pelanggan. Proses pengelompokkan pelanggan pada deteksi komunitas berdasarkan sekumpulan simpul dengan hubungan didalam kelompok yang erat dan hubungan yang jarang di antara kelompok.

Pada penelitian ini, kami sudah melakukan pendekatan lain, yakni melakukan integrasi proses deteksi komunitas dan clustering dalam menghasilkan segmentasi pelanggan. Pendekatan ini dilakukan dengan melakukan pengelompokkan pelanggan berdasarkan pola interaksi yang terjadi antar pelanggan, hal ini untuk menemukan komunitas yang terjadi di pelanggan. Pelanggan yang telah terdeteksi komunitas nya menjadi masukan proses clustering untuk menemukan segmen nya masing-masing, nilai RFM menjadi nilai dasar pembentukan segmen untuk setiap pelanggan.

II. Related Work

Terdapat beberapa penelitian yang pernah dilakukan terkait segmentasi pelanggan. Segmentasi pelanggan menggunakan model RFM yang dikembangkan untuk mengukur loyalitas pelanggan, diteliti oleh Bunnak, Thammaboosadee, dan Kiattisin, mereka menggunakan algoritma K-Means dan Decision tree untuk segmentasi pelanggan[1]. Customer segmentation using decision tree to identify VIP customer in mobile communication industry by Zhang Yihua[2]. Terdapat beberapa penelitian terkait deteksi komunitas yang pernah dilakukan. The Girvan-Newman (GN) algorithm proposed by Girvan and Newman [3] exploits the concept of edge betweenness, which is a measure of the centrality and influence of an edge in a network. Deteksi komunitas pada graf berbobot berdasarkan data twitter menggunakan algoritma GN yang dilakukan oleh Mairisha [4]

Berdasarkan penelitian yang sudah ada, masih belum ada penelitian yang membahas secara menyeluruh proses segmentasi pelanggan yang dilakukan dengan menambahkan infor-

masi terkait komunitas yang terbentuk dari interaksi antar pelanggan

III. METHODOLOGY

In this paper, we apply customer segmentation combine with community detection. The steps of the research process as shown in Fig. 2.

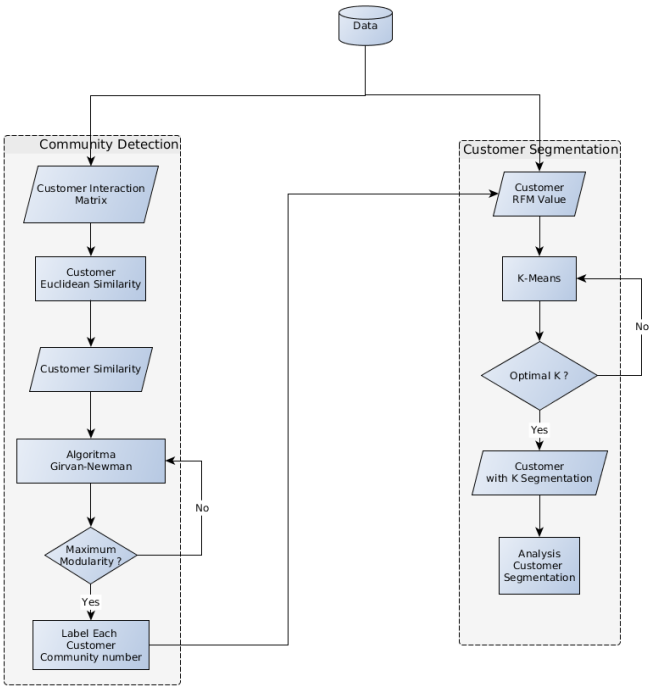

Fig. 2. Overall methodology

A. Data and Data Preprocessing

This step selects related dataset to be used in case study of customer segmentation and then pre-processes data which is an important step. Data preprocessing eliminates irrelevant data by some methods such as data integration, data transformation, and data reduction.

B. Customer Similarity

Pada tahap ini dilakukan proses pencarian kesamaan dari dua objek. pada penelitian kali ini menggunakan euclidean similarity yang merupakan ukuran kesamaan yang umum digunakan [5]. Nilai similarity antar objek ini yang menjadi nilai masukan pada tahap berikutnya.

$$d(p1, p2) = \sqrt{\Sigma(s_{p1} - s_{p2})^2} \tag{1}$$

$$\frac{1}{1 + d(p1, p2)} \tag{2}$$

C. Community Detection by GN Algorithm

The GN algorithm is a divisive hierarchical clustering algorithm exploiting the concept of edge betweenness [3]. Three methods were proposed for the calculation of edge betweenness. Among them, the shortest-path method typically shows the best results. The edge betweenness of an edge is informally the number of shortest paths between pairs of nodes that pass through it. Since communities are loosely connected by a few intergroup edges, all shortest paths between different communities must pass through one of these few edges. Then, those edges connecting communities will have high edge betweenness. Thus, the communities are detected by eliminating such edges repeatedly.

Untuk menentukan berapa banyak komunitas yang harus terbagi di dalam jaringan, digunakan sebuah pengukuran kualitas dari pembagian jaringan yang bernama modularity. 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, equation (3) merupakan formula modularity GN [3].

$$Q = \frac{1}{2m}\Sigma_{vw}[A_{vw} - \frac{(k_v k_w)}{2m}]\sigma(c_v c_w) \tag{3}$$

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). Perhitungan modularity dilakukan di setiap pembentukan komunitas baru. Jika terbentuk x komunitas, maka terdapat x nilai modularity. Jumlah komunitas yang memiliki nilai modularity tertinggi berarti pembentukan komunitas tersebut sesuai dengan kenyataan yang sebenarnya

D. Customer RFM Model

In this step, Customer RFM Model is applied by defining the scaling of R, F, and M variable. The variable is : Recency (R) is customer's last purchase, Frequency (F) is the total number of purchase during a spesific period and Monitary (M) is the amount of money used to purchases in during a spesific period. The RFM model usually used in retail company. For example the RFM value of customer in a supermarket, R is the latest time of a customer purchase, F is how many times a customer made a purchase, and M is berapa banyak uang yang sudah pelanggan keluarkan [6]. Untuk nilai RFM setiap pelanggan, akan dilakukan proses normalisasi menjadi nilai dengan rentang 1-5. Each RFM value customer become normalization process with range value [1-5] Fig 3 is example of RFM Model before transformation and Fig 4 example of RFM Model after transformation

E. Customer Segment by K-Means

In this step, customer segment is applied by use clustering K-Means algorithm for find segment each customer. K-Means Clustering is the simplest clustering algorithm. K-means grouped the objects into K clusters. K is the number

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

Fig. 3. RFM Model before Transformation

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

Fig. 4. RFM Model after Transformation

of clusters that will be generated, defined by the user. The quality of cluster is measured by silhouete. Silhouette can be used to measure the separation distance between the resulting clusters. This measure has a range of [-1, 1]. Step in K-Means Clustering Algorithm is :

1) Decide the number of clusters k
2) Initialize the center of the clusters
3) Attribute the closest cluster to each data point
4) Set the position of each cluster to the mean of all data points belonging to that cluster
5) Repeat steps 3-4 until convergence

F. Analysis Customer Segment

This step refers to the representation and applying the obtained model to the real usage, which will be discussed in the next section.

IV. Experimental Result

As described in the previous section, we will organize the experiment results follows with the step of methodology in the previous section.

A. Data and Data Preprocessing

This research used database from one e-commerce muslimah in Indonesia for last 5 years (2011-2016). The database contains two parts as follows:

- Data Interaction 88,103 records
- Transaction of customer purchases are total 128,628 records

After making a selection of data, the records which include missing values and inaccurate values are removed, and eliminated the redundant attributes. Next, the data is transformed into appropriate formats. Finally, the dataset which are characterized by the following three fields: ID-Sender, ID-Receiver, Total Frequency Interaction

TABLE I. Sample Data Preprocessing Result

ID-Sender	ID-Receiver	Total Frequency Interaction
79424	78112	2
64554	43874	2
48249	21061	18

B. Customer Similarity

In this steps applied formula (2) for data preprocessing result to get similarity each customer based on interaction, table II show sample customer similarity

TABLE II. Sample Customer Similarity

ID-Customer	ID-Customer	Similarity
38	28996	0.333333333333
21061	48249	0.0526315789474
26797	39713	0.5
33892	86097	0.333333333333

C. Community Detection by GN Algorithm

Customer similarity become input for community detection process based on GN Algorithm. Fig 5 show graph based on customer interaction

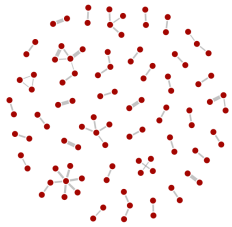

Fig. 5. Graph Customer Interaction

Untuk setiap iterasi proses deteksi komunitas, modularity menjadi nilai ukur kualitas dari jumlah komunitas yang terbentuk. Table III show modularity score untuk setiap jumlah komunitas yang terbentuk. Fig 6 show graph result community detection

TABLE III. Modularity Score

Modularity Score	Number of Communities
0.938736	41
0.940306	42
0.844997	47
0.796996	51
0.707456	56

D. Customer RFM Model

This step uses data obtained in the previous step applied with the defined the scales of R, F, M attributes as described in the previous section.

E. Customer Segment by K-Means

In this process, customer are classified by K-Means based on RFM value and use K range of [3-8]. With silhouete score, K cluster with best silhouete score yang dipilih untuk menjadi

Fig. 6. Graph Result Community Detection

TABLE IV. RESULT K-MEANS WITH SILHOUETE SCORE

Number of Cluster	Silhouete Score
3	0.407802895718
4	0.405678928295
5	0.40253003193
6	0.415421594014
7	0.408311013086
8	0.402908225812

jumlah segment yang dibentuk. Table IV show silhouete score each K cluster.

6 Cluster have best silhouete score, than customer segment divided into 6 segment. Centroid each cluster digunakan untuk mengetahui karakteristik setiap cluster yang terbentuk. Table V show centroid for 6 cluster.

TABLE V. CENTROID CLUSTER

Cluster	Recency	Frequency	Monetary
1	2.15384615	1.61538462	1.53846154
2	1.2	4.8	4.86666667
3	4.52941176	2.35294118	2.52941176
4	2.10526316	3.15789474	3.21052632
5	3.52941176	4.41176471	4.29411765
6	4.69230769	1.07692308	1.07692308

TABLE VI. RESULT CUSTOMER SEGMENTATION

ID-Customer	Cluster	Community
29167	5	7
26797	2	14
33548	2	14
65182	2	33

selain berhasil mengelompokkan pelanggan kedalam segmen nya masing-masing terdapat juga informasi terkait komunitas untuk setiap anggota segment yang terbentuk. Informasi tersebut dapat membantu e-commerce dalam menyusun strategi pemasaran untuk setiap segmen, karena dengan memiliki informasi komunitas setiap anggota segmen tersebut, maka dapat menyusun strategi yang lebih spesifik.

F. Analysis Customer Segment

Dengan mengetahui karakteristik setiap segment yang ditunjukkan Table VII and result customer segmentation show on Table VI, perusahaan dapat memberikan penanganan yang berbeda kepada pelanggan pada segment tertentu. Misalnya pada Loyal Customer Segment, terdapat beberapa anggota segment tersebut yang tergabung dalam komunitas yang dominan anggota komunitas nya tergabung didalam Profit Customer Segment, sehingga perusahaan dapat memberikan penanganan yang berbeda untuk pelanggan tersebut untuk dapat meningkatkan transaksi, sehingga dapat meningkatkan segment, dari Loyal Customer Segment into Profit Customer Segment.

TABLE VII. CENTROID CLUSTER

Cluster	Recency	Frequency	Monetary	Description
1	Low	Low	Low	New Customer
2	Low	High	High	Profit Customer
3	High	Low	Low	Churn Customer
4	Low	Medium	Medium	Loyal Customer
5	Medium	High	High	Profit Customer
6	High	Low	Low	Churn Customer

V. CONCLUSION

This research attemps to try combine community detection and clustering process applying to customer segmentation. This research take RFM model and K-Means Clustering, and use GN algorithm for community detection then the result can be identified characteristics each segment with knowledge about community each segment. It is will be useful for company untuk menyusun strategi yang lebih spesifik.

REFERENCES

[1] Bunnak., Thammaboosadee., and Kiattisin, "Applying Data Mining Techniques and Extended RFM Model in Customer Loyalty Measurement" Journal of Advances in Information Technology Vol. 6, No. 4, November 2015

[2] Zhang Yihua, "Vip Customer Segmentation Based on Data Mining in Mobile-communication Industry". IEEE 978-1-4244-6005-2/10, 2010

[3] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," Physical Review E, vol. 69, no. 2, p. 026113, 2004.

[4] Mairisha, M, "Integration of Coupling Degree Concept for Calculating Modularity in Quality Analysis of Community Structure Based on Weighted Graph" Masters Program Thesis, Institut Teknologi Bandung, 2016

[5] Segaran, T, "Programming Collective Intelligence", O'Reilly Press 2007

[6] Tsiptsis, K., and Chorianopoulos, A, "Data Mining Techniques in CRM", Wiley 2009