Customer Community Identification from Community Detection Graph in E-Commerce

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

G.A. Putri Saptawati
School of Electronic Engineering 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*—Community Detection, Graph, Community Identification**

I. Introduction

E-commerce is one of the sectors in trading which is rapidly growing. Research by consulting firm A.T. Kearney in 2015, showed that Indonesia had potential in E-commerce between USD 25 - 30 billion.

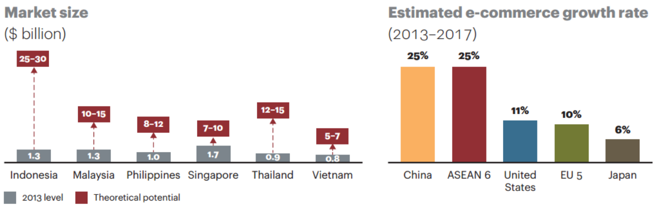

Fig. 1. ASEAN's Market Potency

Based on those potential value, no wonder if Indonesia has many E-commerce companies like Tokopedia, Bukalapak, Hijup, MatahaMall, etc. This condition make every company must give better service than other and make their customer satisfied if want to survive. Company must be able to understand their customer behavior when bought product and arrange marketing strategy based on that information [1].

Data mining is crucial for extracting and identifying useful information from a large amount of data [3]. E-Commerce company operate databases in a long way, such that all transactions are stored in chronological order. A record-of-transaction database typically contains the transaction date and the products bought in the course of a given transaction. Usually, each record also contains an customer ID, particularly when the purchase was made using a credit card or a frequentbuyer card [3].

Networks have become ubiquitous as data from many different disciplines can be naturally mapped to graph structures [2]. An interesting feature that real networks present is the clustering or community structure property, under which the graph topology is organized into modules commonly called community detection. Informally, Community detection is to partition the set of network nodes into multiple groups such that the nodes within a group are connected densely, but connections between groups are sparse [2]

In this research, we have use another approach to extract useful information from data e-commerce. We modelled data e-commerce as graph and apply community detection method to get customer community and extract useful information from community.

II. Related Work

There are some research about community detection. The Girvan-Newman (GN) algorithm proposed by Girvan and Newman [4] exploits the concept of edge betweenness, which is a measure of the centrality and influence of an edge in a network. Community detection in weighted graph based on twitter data use GN algorithm by Mairisha [7]. Based on the time complexity of some of the methods for large networks, a detection method with a small time complexity was required. The best method that was found was the detection method first proposed by Newman and Girvan [5]

However, community in networks depends on what we define relation each node. We still need to identify each community. In order to perform this task, another procedure is required, a different property must be used to identify the communities. In many cases, a possible choice is to use the semantics of a community as a unique property. For example, with the TF-IDF technique apply by Uchida et al [6], an attempt was made to characterize the communities detected in a blogosphere. Specific topics discussed in each community were found, and it was shown that the communities can be identified by such specific topics. Thus, it can be considered that semantics is a useful property for identifying a community

Research by Uchida et all, use TF-IDF to identify communities detected without apply feature reduction and term weighted. it's potential error in choose term to identify topic in entries/document [9]. Feature reduction and term weighted is part of feature selection. Feature selection becomes very important, because it features the words chosen results reduction features and patterns formed determine the precision of word weighting results grouping news [9], [10]

III. Methodology

In this research, we apply Term Frequency-Inverse Document Frequency (TF-IDF) and Term Contribution (TC) to identify customer community. The steps of the research process as shown in figure 2

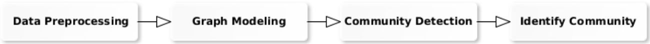

Fig. 2. Overall methodology

A. Data Preprocessing

This step is selects related data that we acquired from e-commerce to be used in case study of discovering customer community 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. Prepared data is shown at Table I

TABLE I. Sample Data Preprocessing Result

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

B. Graph Modeling

We construct a network regarding customers as nodes and interaction each customers (column "Total Frequency Interaction" in Table I) as undirected edges ,although interaction is directed, we consider the network undirected, making each edge bidirectional. For make each edge bidirectional, we apply symmetrizing, equation 1 show that formula [11]

$$\bar{A} = (A + A^T)/2 \tag{1}$$

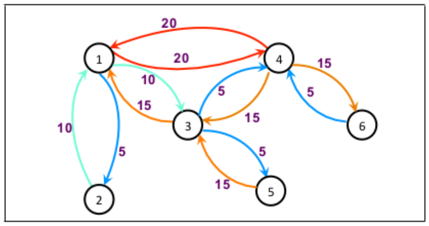

Fig. 3. Example of a weighted directed network

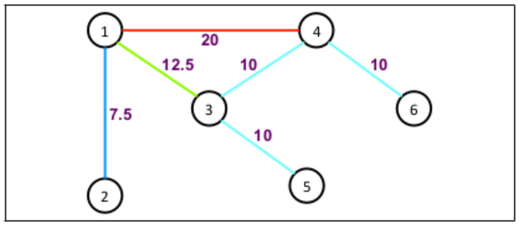

Fig. 4. Symmetrized version of the weighted directed network shown in Fig 5

C. Community Detection

The GN algorithm is a divisive hierarchical clustering algorithm exploiting the concept of edge betweenness [4]. 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.

To decide how much community divide in networks, measure of quality of divide in networks, which call the modularity. The modularity is, up to a multiplicative constant, the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random, equation 2 show formula from modularity GN [4]

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

Good community detection on graph have big modularity score, otherwise poor community detection have small modularity score. Modularity score calculate on every create new community. If formed x community, than there is x modularity score. The number of communities that have the highest modularity score indicates the formation of the community in accordance with the actual reality.

D. Identify Community

Inspired from use TF-IDF to summarization document and TC and Z-score as feature reduction in domain text mining [8], we apply this concept to identify customer community. In our case, we associate term as product and document as community and customer.

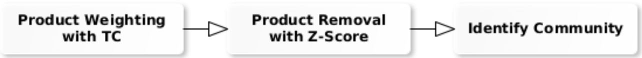

Fig. 5. Flow Identify Community

1) Term Frequency-Inverse Document Frequency: TF-IDF stands for "Term Frequency, Inverse Document Frequency". It is a way to score the importance of words (or "terms") in a document based on how frequently they appear across multiple documents [14]. TF-IDF contain two value, term frequency (tf) and inverse document frequency (idf), where tf is number of occurence of term in document and idf is number of document containing term. Value of idf follow equation 3 and equation 4 shown formula for calculate TF-IDF

$$idf(t) = \log \frac{D}{df(t)} \tag{3}$$

$$TFxIDF(t,d) = tf(t,d)idf(t,d) \tag{4}$$

2) Term Contribution: TC influences the value of word frequency (tf) as a component score. TC score calculate from term frequency (TF) and Inverse Document Frequency (IDF) [12]

$$TC(t) = \Sigma_{i,j \cap i \neq j} f(t,d_i) X f(t,d_j) \tag{5}$$

where $f(t,d)$ represents the $tf * idf$ weight of term in document d whereas i and j are document i document j and i not equal j. $\Sigma f(t,d)$ is sum TF-IDF from term t in document d

3) Z-Score: Assuming that the important words are selected based set of words that have a high score and dominant among other, it can be determined by the difference in distance between score. Z-score calculating a value of the average deviation and density of data [13]. This can be made possible as a solution for the elimination or word removal. This is more flexible and can be used for different datasets. The optimal parameters can be identified at a middle value (zero). Words that have a score greater than zero (≥ 0) is an important word and relevant, otherwise, the word is not an important word and irrelevant and will be removed.

$$Z = \frac{x - \bar{x}}{\sigma} \tag{6}$$

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.

By knowing the characteristics of each segment is shown in Table VII and customer segmentation result show on Table VI, the company can provide different treatment to certain customers in the segment. For example in the Loyal Customer Segment, there are some members of that segment is incorporated in the dominant community of its members incorporated herein Profit Customer Segment, so the company can provide different treatment for these customers to be able to increase transactions, thus increasing segment type, of Loyal Customer Segment into Profit Customer Segment.

TABLE II. Character Segment

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 useful for company to develope specific strategic marketing.

References

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

[2] Malliaros, F. D., and Vazirgiannis, M., ”Clustering and community detection in directed networks: A survey”. Physics Reports 2013.

[3] Ahmed, R. E., Shehab, M., Morsy, S., and Mekawie, N., ”Performance study of classification algorithms for consumer online shopping attitudes and behavior using data mining”. 2015 Fifth International Conference on Communication Systems and Network Technologies

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

[5] Kameyama, S., Uchida, M., Shirayama, S., ”A New Method for Identifying Detected Communities Based on Graph Substructure” 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology

[6] M. Uchida, N. Shibata and S. Shirayama,”Identification and Visualization of Emerging Trends from Blogosphere”, Proceedings of ISWSM, pp. 305-306 (2007)

[7] 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

[8] Munggaran. M. Rizky., ”Identification of Trending Events From News Using Modified K-means Clustering Technique and Term Contribution Technique”, Masters Program Thesis, Institut Teknologi Bandung, 2016

[9] Dai, Xiangying, och Yunlian Sun. ”Event Identification within News Topics .” IEEE , 2010: 978-1-4244-6837-9/10.

[10] Dai, Xiang-Ying, Qing-Cai Chen, Xiao-Long Wang, och Jun Xu. ”Online Topic Detection And Tracking Of Financial News Based On Hierarchical Clustering .” Proceeding of International Conference on Machine Learning and Cybernetics, 2010: pp. 3341-3346.

[11] Gopalakrishnan, K., Balakrishnan, B., and Jordan, R., ”Clusters and Communities in Air Traffic Delay Networks”, 2016 American Control Conference (ACC)

[12] Liu, Tao, Shengping Liu, Zheng Chen, och Wei-Ying Ma. ”An Evaluation on Feature Selection for Text Clustering.” Proceedings of the Twentieth International Conference on Machine Learning, 2003.

[13] Olga Vechtomova, Stephen Robertson, and Susan Jones, ”Query expansion with long-span collocates,” Information Retrieval, vol. 6, no. 2, pp. 251-273 , 2003.

[14] Lan, Man, och Chew Lim Tan. ”Supervised and Traditional Term Weighting Methods for Automatic Text Categorization .” IEEE PAMI, 2007: VOL. 10, NO. 10.