# 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*—In business today, it is very important to be able to identify customer behavior, and arrange strategy based that information. In this research, we propose graph modeling, customer community for e-commerce data, and use concept TF-IDF (Term Frequency-Inverse Document Frequency) in text mining domain to identify customer community. In this research, we associate term as product which is bought by customer and document as community or customer. With this approach, we can get information from customer community, where contains typical product from each community. This information can useful for arrange promotion product to customer community.**

***Keywords*—*Community Detection, Graph, Community Identification, TF-IDF***

## 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.

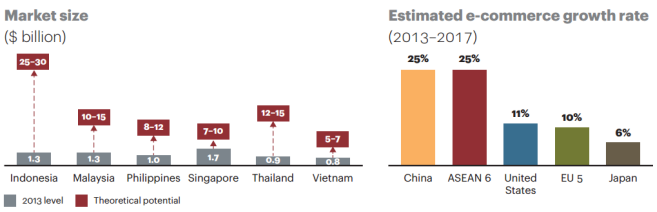


Fig. 1. ASEAN's Market Potency

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 [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

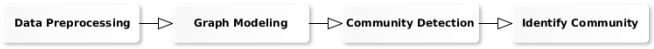


Fig. 2. Overall methodology

### 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]

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

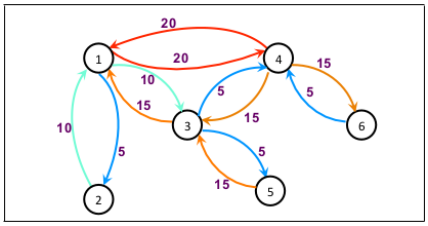


Fig. 3. Example of a weighted directed network

### 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

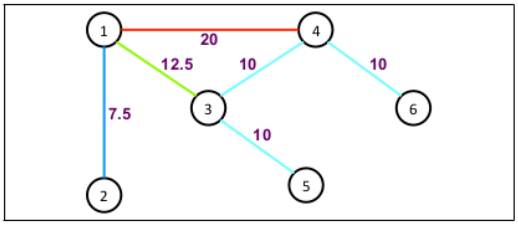


Fig. 4. Symmetrized version of the weighted directed network shown in Fig 5

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 which is bought by customer and document as community or 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 [15]. TF-IDF contain two value, term frequency $(tf)$ and inverse document frequency $(idf)$, where $tf$ is number of occurence of term in document (in our case, number of occurance of product bought in community or customer) and $idf$ is number of document containing term (in our case, number of community or customer bought product). 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 [14]. 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 $(\geq 0)$ 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.

### A. Data Preprocessing

This research used database from one e-commerce muslimah in Indonesia for last 1 years (2015-2016). After making a selection of data, the records which include missing values and inaccurate values are removed, and eliminated the redundant attributes. The dataset contains 88,103 records data interaction and 178,276 records data transaction. Data interaction is interaction customer to customer for transfer point reward.

### B. Graph Modeling

We build graph object with node is set of customer and undirected edge is interaction each customer with mean frequency interaction between customer as the edge weight. Generated graph is visualize as show in Fig 6

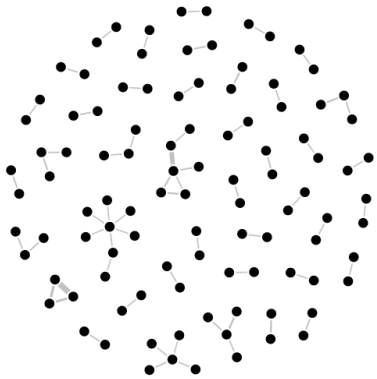

Fig. 6. Flow Identify Community

### C. Community Detection

For each community detection process iteration, modularity score become quality measure. Table II show modularity score for each number of community formed. Because 42 community have highest modularity score, than graph divide into 42 subgraph or community. Fig 7 show graph result community detection. We used community library, a Python implementation of Girvan-Newman community detection algorithm for weighted graphs [13]

TABLE II. Modularity Score

| Modularity Score | Number of Communities |
|---|---|
| 0.732606 | 42 |
| 0.701448 | 46 |
| 0.641020 | 50 |
| 0.447219 | 61 |
| 0.428557 | 62 |

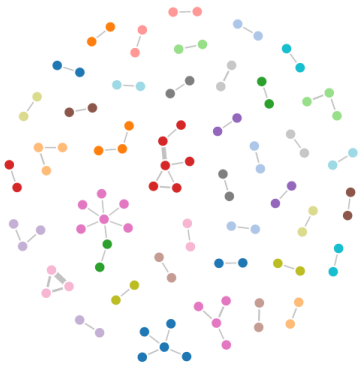

Fig. 7. Graph Result Community Detection

### D. Identify Community

First, we apply equation 4 to our graph data, where in our case product bought frequency as word frequency $(tf)$ and number of community or customer bought product as number of document containing term $(idf)$, after that we calculate TC score follow equation 5 for each product bought in every community, and then calculate Z-score to removed not an importand product.

There are 15,413 distinct product bought on all community, to simplify, we use Z-score to remove not an importand

product. Based on product with z-score more than 0, there are remain 1,075 distinct product to process in next step. After that we calculate TC score for each community, and get product with highest TC score . Table III show result 3 community with their characteristic product.

TABLE III. CHARACTERISTIC PRODUCT COMMUNITY

| Community | Characteristic Product |
|---|---|
| 6 | 'Inner Zipper Reguler', 'Dity shawl', 'Amanda Pants 1', 'Long Vest', 'Ciput Classic' |
| 13 | 'Jilbab Paris Polos 3', 'Scarf Gadih', 'Ciput Classic', 'Fuji Top Pusako', 'Jilbab Paris Jump 03' |
| 32 | 'Pearl Glamour Peafowl', 'Long Shirt', 'Nadin Dress', 'Pashmina Hoodie Mutiara Double', 'Inner NV' |

## V. CONCLUSION

This research attemps to try identify customer community in graph customer e-commerce. We use Girvan-Newman algotihm for community detecton. Community detection result can identified using concept tf-idf in text mining and to improve identify, we apply feature reduction with Z-score. The result is with apply tf-idf in graph customer e-commerce, we can get characteristic product in each community. It can support marketing to arrange promotion product to each community.

## 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] Jahanbakhsh, K. (2016, April). A Python implementation of Girvan-Newman community detection algorithm for weighted graphs. Retrieved July 10, 2016, from https://github.com/kjahan/community/

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

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