Similarity Approach Based to Customer Behavior for Trade Business Metrics

Marischa Elveny\textsuperscript{1,}\textsuperscript{*}, Mahyuddin KM Nasution\textsuperscript{2}, Muhammad Zarlis\textsuperscript{3}, and Elviawaty Muisa Zamzami\textsuperscript{4}

\textsuperscript{1,2,3,4} Faculty of Computer Science and Information Technology (Fasilkom-TI), Universitas Sumatera Utara, Medan, Indonesia

E-mail: marischaelveny@usu.ac.id

Abstract. Data mining is a process of mining data in very large amounts of data. The purpose of this data mining is to extract and identify data for the sake of certain information related to a large database. To find user behavior based on competing merchants, the similarity of several objects is needed. Similarity involves the process of characterizing each object or describing in detail the features of the object. In the process it produces the naming and weighting of the characteristics of objects so as to provide a way to measure their similarity. The way is determined by placing a comparison of what is the same and what is different from the two objects. The results achieved in the form of similarity methods to look for customer behavior in a competitive merchant. The purpose of this study is set out in a simple knowledge base in the form of information to find relationships and models. With this research it is expected to produce new standard models and paradigms that can be used in accordance with the needs of industry and society.

1. Introduction

Data mining is a process of extracting or extracting large amounts of data and information, which have not been known before, but can be understood and useful from large databases and are used to make a very important business decision. [1] Data mining describes a collection of techniques with the aim of finding unknown patterns in the data that has been collected. [2] Merchant is a seller of goods / services that has a business entity (physical store) or an online store cooperation with the Bank in providing payment receipt services via e-money of the bank concerned [3]. There are 1,383 customers and 448 competitive merchants spread across North Sumatra Province [4]. Semantic innovation or information innovation in information mining, can solve the problem of large data, which requires techniques that involve similarities to group information or records in specific gatherings [5]. Similarities is looking for similarities between input and source [6]. Similarity is used to look for relationships or relationships between customer behavior at competitive merchants. The similarity of two items is estimated dependent on the idea of closeness, used to give a restricted worth [6]. There are many enhanced of similarity, be that as it may, it includes an arrangement of segments from objects requiring a preferential approach. This study expects to uncover something in like manner, where each component of the objects is sorted [7]. According to previous studies, measurements of similarity between two objects were carried out to avoid duplication of similar objects [9]. However, the characteristics of objects can be measured with a variety of similarities. The existence of similarity gives meaning
to the measurement results. Measurements such as Jaccard coefficient, mutual information, dice coefficient, and overlap coefficient are measurements that conceptually involve the same vector [10]. Based on the description above, this research was conducted to obtain a new similarity method used to look for the relationship or similarity of customer behavior at competitive merchants.

2. Methods

2.1. State of The Art

In the research methodology, the first step taken is to collect data from a mobile wallet based on merchants and users registered on a competitive payment channel. This is intended to see customer behavior when transacting at a merchant. The captured data is classified and grouped using taxonomies and is grouped based on the location of the merchant. After that the method approach is done using the jaccard coefficient, Cosine and K-Nearest Neighbord as generators of similarity values. The next step is weighted. Weighting is a decision-making technique in a process that involves a variety of factors together by giving weight to each of these factors in order to obtain a value that results in similarity of behavior.

![Diagram](image)

Figure 1. Research Methodology

2.2. Data collection

The research began by gathering data and information about the number of customers, and the number of merchants in all regions in North Sumatra. Data is captured using a mobile wallet that has been registered at various merchants and customers throughout North Sumatera. Mobile wallet is the use of electronic money (e-money) transaction services that use mobile phones or gadgets that are mobile. The following picture is a mobile wallet that is used to capture merchant and customer data which can be seen in Figure 2. From the results of data collection, there are 1383 customers spread across the North Sumatra region with presentations which can be seen in Figure 3 as follows:
2.3. Profiling
From the existing data grouped, classified and mapped its characteristics. At the profiling stage, data is mapped and classified according to the taxonomy of the merchant ecosystem. After that the data is regrouped based on the location of the merchant to get the vector needed in its weighting.

In the merchant ecosystem there are 3 things that are ranked highest in the hierarchy, namely Type, Identity and location. Furthermore, the type, identity and location are regrouped based on objects that are connected to each other in order to communicate information accordingly.
Within types, groups are reproduced based on multi-channels to look for similarities between objects. Taxonomy separates patterns in the literature according to various levels of coverage and levels of abstraction to reduce the problem of overlapping patterns. The hierarchical structure helps reduce the complexity of the business model model and enhances its application in the context of improvements in e-metric data users.

![Figure 4. Taxonomy](image)

### 2.4. User Data

As a vector generator in carrying out its weighting, data is taken by selecting several determinants based on user data on the merchant, namely:

| Username code | Merchant               | Kelurahan | sub-district   | districts     |
|---------------|------------------------|-----------|----------------|---------------|
| Ahmah Muhajir | Abu Bakr               | Sei Agul  | West Medan     | Medan city    |
| Dhita Septi Adlyan | Abu Bakr               | Independent | New Medan     | Medan city    |
| Lifi Rukmana  | Abu Bakr               | The famous base | Medan Johor   | Medan city    |
| Novera Hijjah Hutaso | Abu Bakr               | Silalas   | West Medan     | Medan city    |
| Yos Daniel Marihot | Abu Bakr               | Pioneer   | Medan Timur    | Medan city    |
| Eli Nur Purba Light | Assorted fried snacks | Dwi Kora   | Medan Helvetia | Medan city    |
| Princess Riennika A | Assorted fried snacks | Padang Bulan | Medan Selayang | Medan city    |
| Pascal Moses Sarapi | Chicken Suparman       | Kesawan   | West Medan     | Medan city    |
| amelia warp   | Chicken Suparman       | Padang Bulan | New Medan     | Medan city    |
| Muhammad Saifudin | Chicken Suparman     | Army Glugur I | Medan Timur    | Medan city    |
|                | Chicken Porridge       |           |                |               |


3. Result and Discussion

3.1. Method Approach

1. Jaccard Coefficient

Jaccard Coefficient is one method used to calculate the similarity between two objects (items). The jaccard coefficient (measuring similarity) and jaccard distance (measuring difference) are measurements of asymmetric information on binary (and non-binary) variables. Like cosine distance and matching coefficient, in general the calculation of this method is based on vector space similarity measure.

\[
J(X, Y) = \frac{\sum_p 1x_i y_i}{\sum_p 1x_i^2 + \sum_p 1y_i^2 - \sum_p 1x_i y_i}
\]

(1)

Where x is the value of the key and y is the value of the document.

2. Cosine Coefficient

The Cosine Coefficient the technique is a strategy used to figure likeness (level of similarity) between two items. When all is said in done, the figuring of this technique depends on vector space comparability measure. This cosine similarity method calculates the similarity between two objects (for example D1 and D2) expressed in two vectors using keywords (keywords) from a document as a measure.

\[
\text{CosSim} = \frac{q_i d_i}{\|q_i\| \|d_i\|} = \frac{\sum_j 1(q_{ij} d_{ij})}{\sqrt{\sum_j 1(q_{ij})^2 \cdot \sum_j 1(d_{ij})^2}}
\]

(2)

Information:

- \(q_{ij}\) = the weight of the term \(j\) in the archive \(i\) = \(tf_{ij} \cdot idf_j\)
- \(d_{ij}\) = the weight of the term \(j\) in the archive \(i\) = \(tf_{ij} \cdot idf_j\)

3. K-NN (K-Nearest Neighbor)

K-Nearest Neighbor is an algorithm that is often used for text and data classification. The use of K-Nearest Neighbor has the nature of self learning where if there are more documents, then more resources can be used to compare. K-Nearest Neighbor means finding the closest neighbor to the set to be classified.

\[
D_{euc}(P, Q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}
\]

(3)

Information:

- \(P\) and \(Q\) = point on the n-dimensional vector space.
- \(P_i\) and \(Q_i\) = scalar quantity for i-dimension in n-dimensional vector space.

To measure the distance between and then do the value 1 for the maximum value because the square root if the distance is usually greater than 1, and if the distance is less than 1 will produce a very important value for similarity, then the formula:

\[
D_{euc}(P, Q) = \frac{1}{1 + \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}}
\]

(4)

3.2. Weighting

Weighting is a decision-making technique in a process that involves a variety of factors together by giving weight to each of these factors. The characteristics and weighting of each object form a vector space that is the target of the optimization of the measurement process by reducing its complexity. For the reason that each space has its own measurement complexity.
For each \( a, b \in X \), there is a form that is expressed as \( a \cap b \in X \), this shows the relationship between \( a \) and \( b \), of course \( a \cap b \) is also a vector in the vector \( X \) space, which has weight and direction. The differences \( a \) and \( b \) are the distances that express \( d(a, b) \) as they indicate the existence of each of \( a \) and \( b \) in vector \( X \) space, i.e.

\[
a = b \quad \text{or} \quad a \cap b > 0
\]  

(5)

3.3. Similarity

Similarity (similarity) is a measurement to determine the similarity of one object to another object. Similarity involves the process of characterizing each object by describing in detail the features of the object.

Based on Figure 6, it explains that the set A (customer) represents the object by assigning
any pair of parts $a, b \in A$, thus the relationship $h = (a \rightarrow b)$. As an object, set $A$ consistent of $a_i$ (behavior), $i = 1 ... , n$ or point in the form of set $A$ which can be separated into subspaces of $A$. Size from $A$ is $|A| = n$.

If set’s $A = \{a_i - i = 1 ..., n\}$ represents the object by which $a_i$ is a component of $A$, then each component has a relationship with the other components. For example, $|A| = n$, we obtain n (n - 1) relationships or there is a set section of relationships $H = \{h_j - j = 1 ...., m\}$ for $a_i \in A$, $i = 1 ..., n$. Each relationship can be described in binary $\{0,1\}$ to confirm fellowship of the two section in $A$. So, each component owned more than one relationship with the other.

The concept of similarity is related to symmetrical shapes, and one of the symmetric shapes is a right triangle, but when $a, b$ and $c$. Therefore, the measurement of parameters of similarity is aimed at the side of the square.

Based on equations (1), (2), (3), (4), (5), then:

**Description 1.** A representation of an object with relation $(a \rightarrow b)$

**Description 2.** If $A = \{a_i - i = 1 ..., n\}$ is a make a metaphor of an object, then respectively component has a relationship with other components.

**Description 3.** A distance where $c = 1 - b^2 (b^2 + b^2)$. Based on the equation $c = 0$, then $c^2$ and $b^2 + b^2$ is not the same or $c^2 < b^2 + b^2$, so 0 is defined as proximity.

### 3.4. Results Graph

![Figure 7. Results Graph](image)

The results of information 1,2 and 3 can be seen in figure 7. This shows that there is a strong symmetrical relationship between vectors in the vector space. In theory, each result of measuring similarity is at a certain position between 0,1 and each similarity based on the formula of each method differs from one another, which appears on a straight and curved line. However, the existence of a similarity can still be sought using derivatives of the formulation which will be carried out in further research, which is seen in the following dotted lines, based on e-metric data to look for the similarity of customer behavior towards a merchant.

### 4. Conclusion

Based on the results of this study, several methods have been implemented using Jaccard coefficients that appear on the orange colored lines, where the value of Jaccard is always...
consistent with showing a straight line. cosine that appears on the blue line and K-NN that appears on the green line, shows that there are similarities between these two methods, where the lines appear curved. This research has been carried out on semantic web data. Advanced green color research studies with dashed lines that will be applied to customer e-metric data in mobile payments.

Acknowledgment
This research was supported by the faculty of computer science and information technology (Fasilkom-TI), Sumatera Utara University.

References
[1] Pandit, R., Sengupta, S., Naskar, SK, Dash, NS, & Sardar, MM. 2019. Improving Semantic Similarity with Cross-Lingual Resources: A Study in Bangla - A Low Resourced Language. Informatics, 6, 19.
[2] Cheng, J. and Zhang, L. 2019, Jaccard coefficient-based bi-clustering and fusion recommender systems for solving sparsity data, Lecture Notes in Computer Science (including subseries of Lecture Notes in Meanings of Intelligence and Lecture Notes in Bioinformatics) 11440 LNAI, 369-380. DOI: 10.1007 / 978-3-030-16145-3 29.
[3] R. Syah, M.K.M. Nasution, M. Elveny, H Arbie. 2020. Optimization Model for Customer Behavior With MARS and KYC System. Journal of Theoretical and Applied Information Technology 98(13).
[4] Eickhoff, M., Muntermann, J., & Weinrich, T. 2017. What do FinTechs actually do? A Taxonomy of FinTech Business Models. In Thirty Eighth International Conference on Information Systems (6ICIS 2017), Seoul, South Korea.
[5] MK M, Nasution. 2020, Determining the existence of a Similarity, ResearchGate, DOI: 10.13140 / RG.2.2.27255.16800.
[6] H. Gimpel, D. Rau, M. Röglinger. 2018. Understanding FinTech start-ups - a taxonomy of consumer-oriented service offerings. Electronic Markets.28, 245–264.