Product Recommendation System Design Using Cosine Similarity and Content-based Filtering Methods

Cut Fiarni[1], Herastia Maharani[2], Nathania Calista[3]

***Abstract*— The wide variety of products offered by a company, combined with consistent demands of specific products from customers, create a certain problem for an organization when they want to market a new product. Organizations need information that can help them promote the most suitable product based on their customer's characteristics. The organizations also need to suggest alternative products for customer if the requested product is unavailable. In this paper, a Recommender System that could suggest either new or alternatif products to customer based on their characteristic and transaction history is designed. This proposed system adopts Cosine Similarity method to calculate product similarity score and Content-based Filtering to calculate customer recommendation score and used as a model for the proposed system. Subsequently, these models are used to classify customers as well as products according to their transaction behavior and consequently recommends new products more likely to be purchased by them. Based on the testing results of the proposed system, it can be concluded that the chosen methods can be utilized to recommend products and customer of products. It is shown that Precision and Recall of product similarity scores and customer recommendation for product scores are 100% and 93.47%.**

***Keywords*—Product Recommendation, Recommender System, Cosine Similarity, Content-Based Filtering.**

I. INTRODUCTION

Changes in the current industrial sector and increasingly high competition between companies encourage companies to innovate in implementing right strategies to retain loyal customers and to bring in new customers. Companies are required to be able to adapt their products to customers' needs and characteristics. This is what causes many variations of products to be sold at different prices and quality. Errors in product marketing and selling can cause customer dissatisfaction which can lead to a decrease in company performance.

To prevent this from happening, companies usually need information regarding segmentation, targeting, and understanding of product positions based on sales transactions that occur, or commonly known as Segmentation, Targeting, Positioning (STP) strategies. Research related to data mining utilization in modeling information system tools that are able to help the formulation of marketing and sales strategies for goods is an interesting subject of research but is still relatively rare. Researches have been conducted to model relationships between the sale of goods/products in certain areas and characteristics of their products. The results of this model were applied to a decision support system design that aimed to assist decision making regarding product marketing strategies in an area [1].

In this paper, a case study was conducted at PT. X which was a company engaged in retail textile dye. PT. X implements a B2B business model whose customers are companies engaged in textiles. In process of marketing new products, marketing department of PT. X always offers new products to all customers who handle mass marketing at various prices for each customer. Meanwhile, in a process of providing alternative products to customers, marketing only sees color similarities without looking at other characteristics of a product. Of three STP strategies, the most possible and must be carried out first in designing marketing strategies is a segmentation process, which is dividing or grouping customers according to their characteristics based on sales transactions that have occurred. To assist PT. X in marketing and offering products to customers, in this paper a recommendation system is designed. This recommendation system is able to group products and customers according to their characteristics, provide product recommendations to customers, and provide potential customer recommendations. In designing a recommendation system, the most important factor is a recommendation system modeling by implementing a right method, according to transaction data characteristics.

II. METHOD

In designing a product recommendation system, system modeling is the most important factor. Therefore, the focus of this research is as follows.

1. Identifying a model for customer needs introduction about a product and similarity of alternative products to the products needed by the customer.
2. Identifying a model for customer character recognition and segmentation recognition based on its activities by conducting a sales transaction data processing. This aims to facilitate the company to market new products, both to old customers and new customers.

As explained in the previous section, the product that will be recommended in this proposed system is dye for textile materials. For those who are involved in or working in the textile industry and fashion, knowledge and understanding of the types and properties of textile fibers is a basic capital that must be owned. Characteristics and properties of a textile material are largely determined by the characteristics and properties of the constituent fibers [2]. In addition, processing

[1,2] *Lecturer, Information System Department Institut Teknologi Harapan Bangsa, Jln. Dipatiukur No. 80-84 Bandung 40132 INDONESIA (phone: 022-2506636; e-mail: cutfiarni@ithb.ac.id, herastia@ithb.ac.id)*

[3] *Alumnus, Information System Department Institut Teknologi Harapan Bangsa, Jln. Dipatiukur No. 80-84 Bandung 40132 INDONESIA (e-mail: nathaniacalist96@gmail.com)*

that can be carried out for a particular type of textile material is also largely determined by characteristics and properties of constituent fibers, starting from equipment selection, work procedures to chemicals types that can be used. These fiber's basic properties will not disappear during a textile processing process. The textile processing processes carried out only aim to repair, improve, add, and optimize fiber's basic properties, so that it becomes a quality textile material according to its intended use. In this paper an analysis of transaction data that has been owned by the company is conducted to test the most appropriate method for modeling customer recommendations for a product to be sold and product alternatives based on characteristics of company's needs. Based on problems analysis' results, needs related to design for alternative product recommendation and customer segmentation applications are as follows.

1. Ability to calculate similarity scores between products to provide alternative product recommendations (product similarity).
2. Ability to calculate customer recommendation scores to provide a list of potential customers who can buy a product (customer recommendation).

A. Cosine Similarity

In general, a similarity function is a function that accepts input of two objects then calculates similarity between the two objects and returns in a form of real numbers. The value returned by the similarity function generally ranges in interval [0...1]. In this method, similarities between two *n* dimensional vectors are calculated by looking for cosine value of angle between the two. The cosine similarity formula is given as follows [3].

$$similarity\ (x, y) = \cos(\theta) = \frac{\sum_{i=1}^{n} x.y}{\|x\|\|y\|} \quad (1)$$

with

x.y = vector dot product of *x* dan *y*, calculated by $\sum_{k=1}^{n} x_k y_k$,
$||x||$ = vector length *x*, calculated by $\sum_{k=1}^{n} x_k^2$,
$||y||$ = vector length *x*, calculated by $\sum_{k=1}^{n} y_k^2$.

The greater the similarity function result, the two objects evaluated are considered to be more similar. In a function that produces a value in a range [0...1], value of 1 means that two objects are exactly the same, while a value of 0 means that two objects are completely different.

B. Profile Matching

In a profile matching method, a similarity calculation is done by comparing between one value profile with several other competency value profiles. In this method, the difference between the comparison and the required competency requirements is sought. The difference from these competencies is called the gap, with a smaller gap having higher values. Following are some stages and formulations of calculations using a profile matching method [4].

1) Weighting: At this stage, a value of each aspect is determined by using weighted values that have been determined for each aspect itself.

2) Grouping of Core and Secondary Factors: Core factors are aspects (competencies) that stand out or are most needed. To calculate the core factor values, (2) is used [5].

$$NCF = \frac{\sum NC}{\sum IC} \quad (2)$$

with

NCF = average core factor value,
NC = total value of core factor,
IC = number of core factor items.

Secondary factors are items other than aspects that have been defined as core factors. To calculate the value of the secondary factor, (3) is used [5].

$$NSF = \frac{\sum NS}{S} \quad (3)$$

with

NSF = average value of secondary factor,
NS = total number of secondary factor values,
IS = number of secondary factor items.

3) Calculation of Total Value: From values of core factors and secondary factors calculated from each aspect, a total value of each aspect predicted to affect performance of each profile is calculated. To calculate total value of each aspect, this following formula is used [5].

$$N = X\%\ NCF + X\%\ NSF \quad (4)$$

with

N = total value of each aspect,
NCF = average core factor value,
NSF = average value of secondary factor,
X% = processed percentage value.

4) Ranking: The final result of profile matching process is a ranking of submitted candidates. Ranking detemination refers to calculation results given in (5) [5].

$$Ranking = 70\%\ NCF + 30\%\ NSF \quad (5)$$

with

NCF = average core factor value,
NSF = average value of secondary factor.

C. Content-based Filtering

Content-based filtering method is based on a descriptions of items and user tendency's profil. In content-based recommender system, keywords are used to describe items while used profiles are built to indicate items liked by users [6]. In other words, this algorithm tries to recommend items that are the same as those that users like in the past or check current trends. Specifically, various candicate items are compared to items that were previously classified by the user and the most suitable item is recommended. This approach is based on information retrieval and information filtering.

To create a user profile, the system focuses on two information types, namely user tendency model and user interaction history on the recommendation system. An important issue of content-based filtering is that this system can study trends in user actions related to one source of content and use it on all content types or not. In its application, content-based filtering methods can also be combined with other methods such as demographic filtering or collaborative filtering, as done in [7]-[9].

D. K-Means Algorithm

K-means algorithm is one algorithm with partitional, because K-means is based on determining an initial number of groups by defining its initial centroid value [10].

Data that has a short distance to a centroid will create a cluster. This process continues until there is no change in each group. Following are steps included in the K-means algorithm [10].

1. Determining *k* as the number of formed clusters.
2. Determining initial *k* (cluster center point) randomly.
3. Calculating distance of each object to each centroid from each cluster.
4. Allocating each object into the nearest centroid.
5. Conducting iteration, then determining new centroid position using (2).
6. Repeating step 3 when the new centroid position is different.

III. System Analysis and Modeling

In this section there is analysis of the most appropriate methods to be implemented in the proposed system modeling. These methods were tested on sales transaction data with a total of 144 products and a total of 37 customers in the range of years from 2014 to 2016.

A. Data Preparation and Cleaning

This paper uses sales transaction data with a total of 144 products and a total of 37 customers in the range of years from 2014 to 2016 using Customer Reports - Products 2016 - 2017), Customer Report Lists 2014, and Products Catalog.

B. Product Similarity Modeling

In analyzing product characteristics similarity, two calculation methods were used on 144 dye stuff products to determine exact methods to be used, namely the cosine similarity method and profile matching method. Based on customer needs analysis for a product and cleaning and merging process from Product Catalog brand, it was found that the characteristics used to calculate product similarity score were characteristics of fabric color (color code) fabric type, dyeing technic, capability of printing fabric processing, light test value (light), and fabric washing test value (washing). Furthermore, product data was calculated for its precision and recall scores for both methods. The goal was to obtain the most appropriate method to be applied in modeling product similarity in a proposed system. Calculation result summary of the 144 products on the method is shown in Table I.

TABLE I
PRECISION CALCULATION RESULT AND RECALL DETERMINATION OF PRODUCT SIMILARITY

Product	Precision		Recall	
	CS	PM	CS	PM
Black SRL	30%	50%	50%	83%
Black B320	40%	10%	80%	20%
Blue BG	30%	40%	38%	50%
⋮	⋮	⋮	⋮	⋮
Average	33%	30%	59%	50%

Description: CS = Cosine Similarity
PM = Profile Matching

Based on precision and recall tests results in Table I, it can be seen that cosine similarity method has a predominant value compared to profile matching method. In addition, the value of precision and recall tests in both methods is quite low because not all samples have the same product similarities numbers. In addition, in determining product similarity, expert only focuses on the similarity of colors and fabric types, while in calculating a predetermined model, product similarity value is calculated with other characteristics, namely match with product treatment activity and the results of test scores given to the product. Therefore, it can be concluded that the more appropriate method used to find product similarities is to use cosine similarity.

C. Customer Recommendation

In conducting customer recommendation analysis, in terms of marketing new products, data processing was carried out on Customer Reports - Products and Product Catalogs using the content-based filtering method along with the K-means clustering method as a comparison.

Furthermore, the data was processed so that customer profile data was obtained by a formula as follows [6].

$$User\ Profiles = \sum_{i=1}^{n} A\ B_i \quad (6)$$

with

A = number of product purchases,
B = product characteristics,
n = number of product characteristics,
i = characteristic of product number-i.

In calculating this user profile, two calculation methods were carried out, namely calculations without normalization and calculations with normalized data with the aim of eliminating and reducing data redundancy and ensuring data dependencies. In addition, the calculation were carried out twice. The first calculation was done by combining all fabric types and the second calculation is to separate fabric type with the aim of ensuring the model gives the right customer profile value.

The conducted calculation yielded results that there was no difference between the results of content-based filtering method without data normalization and with data normalization. This was because the number of filtering attributes was the same in the calculation of data normalization

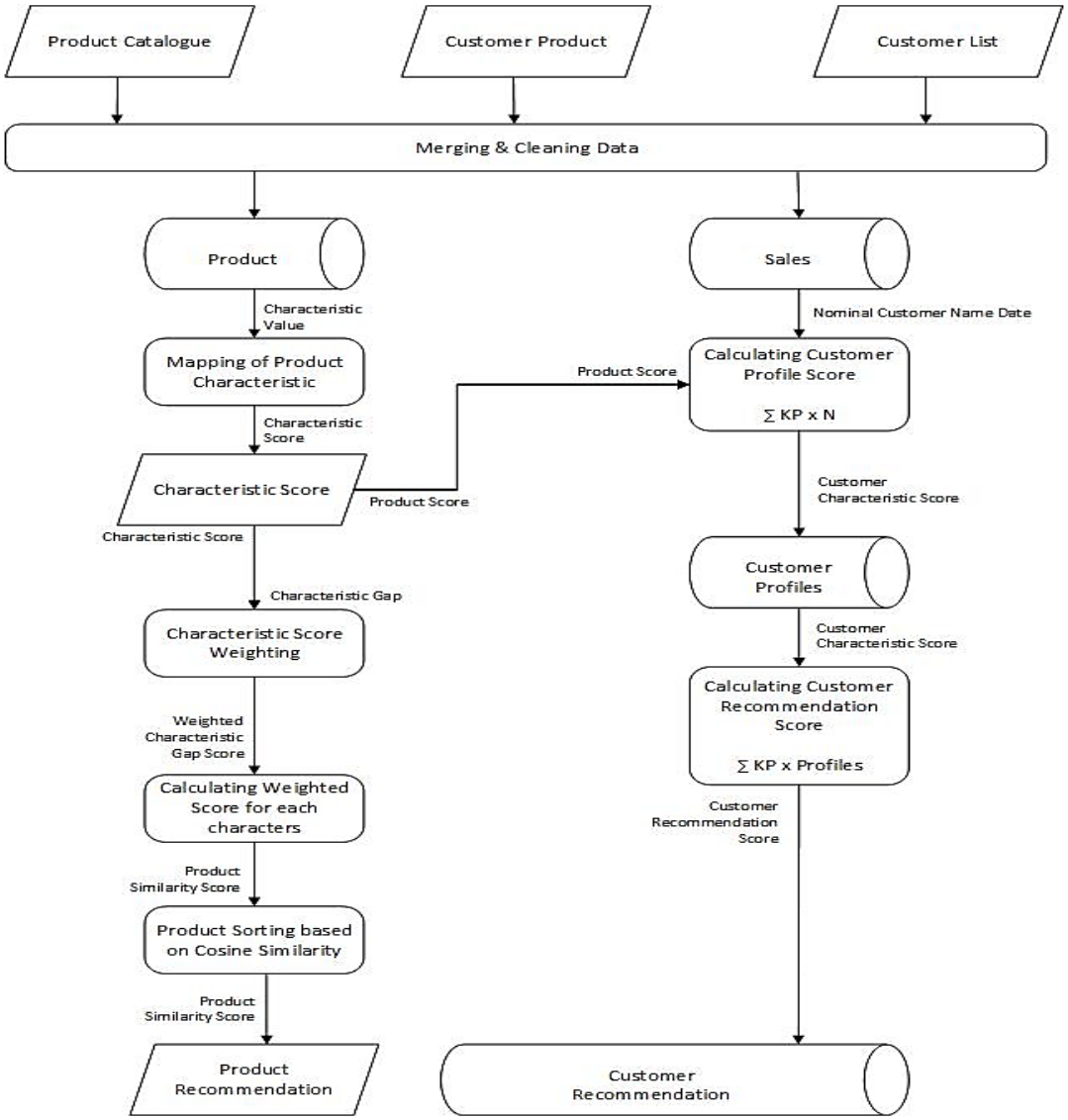

Fig. 1 Proposed system architecture scheme.

and without data normalization. Furthermore, there was a significant difference between a calculation with and without fabric separation, i.e. the calculation with fabrics separation resulted in a more accurate recommendation value than the calculation without fabric separation. This was because the calculation with fabrics separation had more filtering attributes than the calculation without fabric separation. Therefore, the utilized user profile results were user profile carried out with fabric types separation.

The user profiles were also tried through cluster calculations using the K-means clustering method with the aim of dividing the data into several groups so that data that possessing the same characteristics were grouped into one group. Customers grouping into clusters based on transaction history could help in recommending potential customers to get a specific product offer, as applied to [11]. This calculation was carried out four times, consisting of K-means clustering, K-means clustering with normalization data, K-means clustering with weighting, and K-means clustering with weighting and data normalization. Results of cluster calculations are shown in Table II.

From testing results on Table II, it can be seen that there is no homogeneity in one cluster and heterogeneity between clusters, so it can be concluded that the K-means clustering algorithm is not appropriate for searching customer recommendations. Furthermore, customer recommendation calculations were performed through the calculation of user prediction by multiplying the product information characteristics with user profiles using the following formula [6].

$$User\ Prediction = \sum_{i=1}^{n} A_i B_i \quad (7)$$

with

A = score of user profiles characteristics,
B = product characteristics,
n = number of product characteristics,
i = characteristic of product number-i.

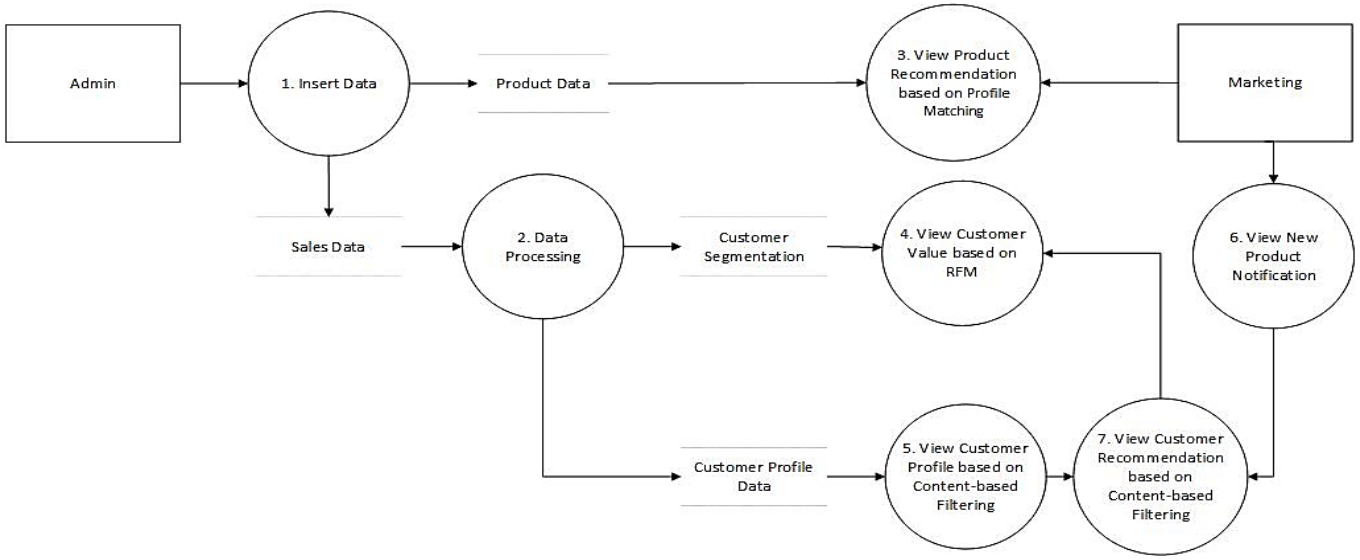

Fig. 2 DFD level 1.

TABLE II
RESULTS OF K-MEANS CLUSTERING

Category	Cluster	Homogeneity in 1 Cluster	Heterogeneitiy between Cluster
K-means Clustering	*k* = 2 - 18	X	X
K-means Clustering Normalization	*k* = 2 – 18	X	X
Weighted K-means Clustering	*k* = 2 - 18	X	X
Weighted K-means Clustering Normalization	*k* = 2 - 18	X	X

After a recommendation value for each customer in each product was obtained, the value was summarized so that the highest recommendation from each customer was obtained for particular product shown in Table II. Based on test results and recommendations from content-based filtering method and the K-means clustering calculation, it is obtained that:

1. there is no difference between the results of the content-based filtering method calculation without data normalization and with data normalization;
2. there is a significant difference between a calculation with fabrics separation and without fabric separation, namely a alculations with fabrics separation produces a recommendation value that is more accurate than a calculation without fabrics separation; and
3. calculations using K-means clustering method are not appropriate for identifying customer groups.

D. System Architecture of Product and Costumer Recommendation

Based on explanatory factor analysis results that had been done on methods in the previous sections, namely cosine similarity and profile matching methods for product recommendation modeling and with the content-based filtering method along with the K-means clustering method for customer recommendation modeling, it could be concluded that the most appropriate method for proposed system was to integrate cosine similarity and content-based filtering methods. For more details, the proposed system architecture is shown in Fig. 1.

In Fig. 1 it can be seen that there are three reports on transaction data entered into the proposed system, namely Product Catalog, Customer - Product, and Customer List. These three reports are included in the proposed system to experience cleaning and merging processes with Product Catalog undergoing a cleaning process. This process produces two databases, namely a Product database and a Sales database. The next process of Product database is a score mapping process for each product characteristic so that scores are generated for each characteristic in each product. Then a calculation of gap between scores of main products and other products was carried out. The results of this gap calculation were weighted and calculated so that it resulted a Product Recommendation List had been sorted based on the highest to lowest product similarity score. Scores for each characteristic of this product were also utilized to be combined with the sales quantity data (Sales database) of each customer. Once combined, the user profiles were calculated based on product characteristics to produce a Customer Profiles database. This database was used to perform customer recommendation calculations.

IV. IMPLEMENTATION AND TESTING

This section describes the design of the proposed system modeling that was designed in the previous section and its testing. The model used in designing this system consists of Data Flow Diagrams (DFD) and Entity Relationship Diagrams (ERD).

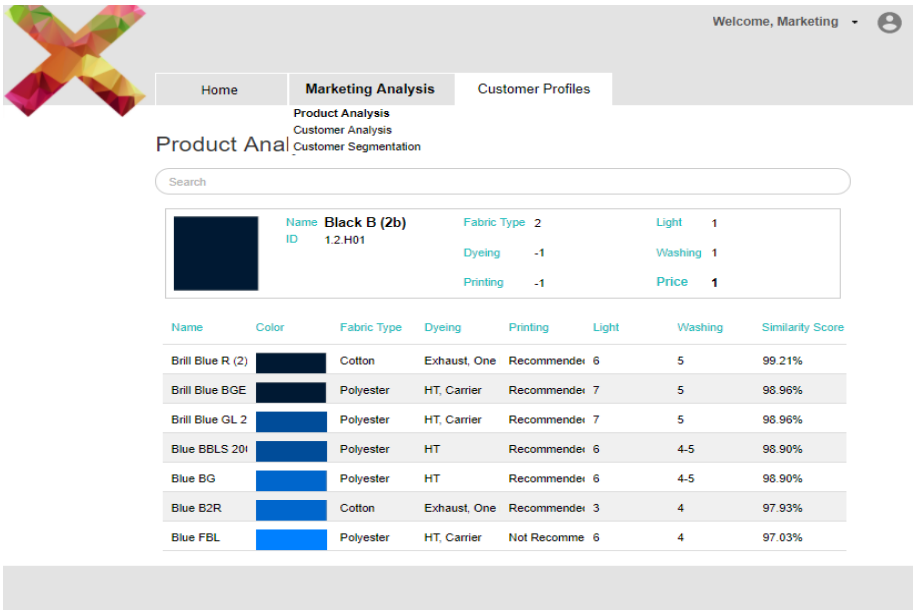

Fig. 3 Product analysis page.

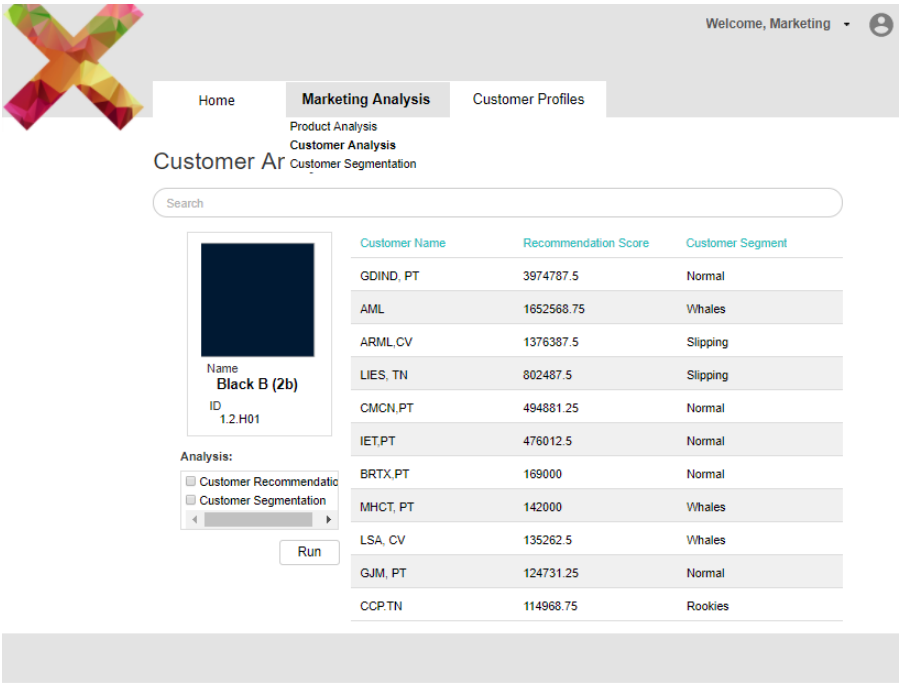

Fig. 4 Customer analysis page.

A. System Designing

The entire process involved in the system and the data used are depicted in DFD level 1 at Fig. 2. In this proposed system, a database system is constructed consisting of ten tables that are connected to each other, namely customer_profiles, segmentation, product, sales, and other additional tables.

B. User Interface

The system is specifically designed for marketing user. This user can see product similarity and customer recommendation processing results. Fig. 3 and Fig. 4 are interfaces of the designed marketing user system. Product Analysis contains an analysis of similarity score of selected products with other products. On this page, the similarity score is sorted from the largest to the smallest similarity score. Also attached are specifications for each product that has similarities to the selected product. The Product Analysis page is shown in Fig. 3. Customer Analysis includes an analysis of customer recommendation scores and customer segments that are on the selected product. Before doing the analysis, marketing can check what you want to see and how the data is sorted. The Customer Analysis page is shown in Fig. 4.

C. Testing

At this stage, the method was tested based on average result of precision and recall testing values for each main feature of proposed system. After the utilized method was known, then a test was performed on three algorithms that had been designed with a calculation model that had been done using tests of precision, recall, accuracy, and F-measure with the following steps.

1. Matching the algorithm results with the results of calculation model method.
2. Determining values of TP, FP, TN, and FN [12].
3. Calculating values of precision, recall, accuracy and F-measure testing [12].

Testing results of two main models of the proposed recommendation system are shown in Table III below. Seen in Table III, from proposed system testing for modeling results, the product recommendation feature has a test value of precision, recall, accuracy, and F-measure of 100%. While testing on customer recommendation feature modeling produces test values for precision, recall, accuracy, and F-measure of 100%, 93.47%, 93.47%, and 96.24%. This shows that the designed proposed algorithm is in accordance with the calculation model that has been done, so that the proposed algorithm is worth applying.

TABLE III
RESULT OF ALGORITHM TESTING

No	Testing	PS	CR
1	Precision	100%	100%
2	Recall	100%	93.47%
3	Accuracy	100%	93.47%
4	F-measure	100%	96.24%

Description: PS = Product Similarity
CR = Customer Recommendation

V. CONCLUSIONS AND SUGGESTION

Based on the analysis, design, and testing of product recommendation systems and customer segmentation, some conclusions can be drawn as follows.

Data mining techniques can be applied in system modeling that has the ability to provide recommendations for new products and alternatives. However, it is necessary to analyze and explore profound data because each data has different characteristics and the company needs specific information. Meanwhile, from the testing results of comparative methods, the most appropriate method to be adopted in this proposed system is cosine similarity to calculate product similarity score. This method can help companies to choose alternative products to match the product demand by customers.

On the other hand, to provide customer recommendations that are in accordance with company characteristics and purchase transactions that occur, the most appropriate method is content-based filtering. This method calculates customer recommendation scores that can help companies select customers in marketing, especially marketing new products, so that marketing can be more effective.

A suggestion that can be further developed from this research is to develop a concept of Segmentation, Targeting, Positioning (STP) from the marketing strategy. One marketing strategy is to understand customer characteristics, namely by grouping customer characteristics that are known from values provided by the customer to company by using a Recency, Frequency and Monetary (RFM) method which is calculated annually based on recorded sales transactions.

REFERENCES

[1] E.M. Sipayung, C. Fiarni, and R. Tanudjaya, "Decision Support System for Potential Sales Area of Product Marketing Using Classification And Clustering Methods," *Proc. of 8th International Seminar on Industrial Engineering and Management*, 2015, pp. DSS.33-39.

[2] A.R. Horrocks and S.C. Anand, *Handbook of Technical Textiles*, Cambridge, England: Woodhead Publishing Limited in association with the Textile Institute Abington Hall, 2000.

[3] P. Kotler, *Manajemen Pemasaran*, Milenium Ed., Translation by B. Molan and H. Teguh, Jakarta, Indonesia, 2000.

[4] Kusrini, *Konsep dan Aplikasi Sistem Pendukung Keputusan*, Yogyakarta, Indonesia: Andi Offset, 2007.

[5] L. Zhiqiang, S. Werimin, and Y. Zhenhua, "Measuring Semantic Similarity between Words Using Wikipedia," *Proc. of 2009 International Conference on Web Information Systems and Mining*, 2009, pp. 251-255.

[6] C.C. Aggarwal, *Recommender Systems*: *The Textbook*, Cham, Switzerland: Springer, 2016.

[7] H. Maharani and F.A. Gunawan, "Sistem Rekomendasi Mobil Berdasarkan Demographic dan Content-Based Filtering," *Jurnal Telematika ITHB*, Vol. 9, No. 2, pp. 64-68, 2014.

[8] P. Lenhart and D. Herzog, "Combining content-based and collaborative filtering for personalized sports news recommendations," *Proc. of the 3rd Workshop on New Trends in Content-Based Recommender Systems*, 2016, pp. 3-10.

[9] Z. Lu, Z. Dou, J. Lian, X. Xie and Q. Yang, "Content-Based Collaborative Filtering for News Topic Recommendation," *Proc. of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 217-223.

[10] H-P. Kriegel, E. Schubert, and A. Zimek, "The (Black) Art of Runtime Evaluation: Are We Comparing Algorithms or Implementations?" *Knowledge and Information Systems*, Vol. 52, No. 2, pp. 341–378, 2016.

[11] E.M. Sipayung, H. Maharani, and B.A. Paskhadira, "Designing Customer Target Recommendation System Using K-Means Clustering Method," *International Journal of Information Technology and Electrical Engineering (IJITEE)*, Vol. 1, No. 1, pp. 1-7, 2017.

[12] D.M. Powers, "Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness, and Correlation," *Journal of Machine Learning Technologies*, Vol. 2, No. 1, pp. 37-63, 2011.