Recommendation System for Wedding Service Organizer using Content-Boosted Collaborative Filtering Methods

Rahman Indra Kesuma* and Amirul Iqbal

Department of Informatics Engineering, Institut Teknologi Sumatera, Lampung Selatan, Indonesia

* Email: rahman.indra@if.itera.ac.id

Abstract. The utilization of smartphones that grow continuously is as a result of the changing lifestyle from the peoples in the growing digital world era. This condition can be seen on the large penetration of smartphones in half of the world society. The high smartphones utilization can increase the chance of many businesses, especially in online businesses that are used for promotions and transactions. The efficient and effective process become the main issues in technology development. One example is to meet the needs of wedding services in online business, which most users still have difficulties or need more time in finding the product as desired. Therefore, a recommendation system concept is applied in this research that is able to help the process of product promotion or searching specifically related to new products and involve new users. The content-boosted collaborative filtering method is used in these systems that implement two previous methods, namely content-based that is used to form a full product rating matrix by a user, while collaborative filtering is used to make recommendations. Based on the experimental results that the system can recommend with 69% accuracy and helpful especially for newly added items or newly registered users.

1. Introduction

The increasing of smartphones usage is a result of the rapid development of the world into digital technology that is resulting the changes in all aspects of people's lives. This condition is proven by the relatively high usage of smartphones technology in Indonesia until 2017, with the numbers of smartphone ownership in urban areas are 70.96% and in rural areas are 42.06% [1]. This phenomenon is also perceived in the world society, where the penetration of this technology revolves around 43% of the world population and it is predicted to continually increase up to 61% in 2025 [2].

The number of smartphone users that continually grow should be used by many companies to gather the peoples for certain concerns. The examples are company that commits to the business either micro, small, medium or large businesses. This condition makes e-commerce growth well because of the ease of access something by using smartphone devices [3]. The term of e-commerce that refers to business activities or transactions using information technology media and internet is currently growing [4]. This growth can be seen in the number of internet users who make online transactions through e-commerce in 2018, which come to 1.7 billion users from 4 trillion internet users in the world [5].

The phenomenon of the growing of online businesses usage has been perceived to meet several aspects of society needs. One of them is for the wedding preparations for the bride and groom who will celebrate their wedding. Most brides want the best event on their wedding, with the right choice of variety wedding necessary. However, it is known that many brides do not have enough time to surveying of wedding organizer, while in an area there are certainly many choices available [6]. The availability of a website from the many wedding organizer, obvious also has not been able to facilitate the brides to
recommendation process cannot be done manually because it involves transaction history from all customers and products. So the recommendation system is widely used in online businesses to do automatically.

Many previous studies focus on the topic of recommendation systems that are used to increase promotion or product sales of an online business. In this chapter, the previous study describes three different approaches, namely content-based, collaborative filtering, and hybrid methods. The content-based filtering method approach is used in research of [12], which provides recommendations by using information extraction from user profiles, product descriptions, and other matters relating to transactions that happen. This method is used to provide recommendations of a product contained in its online business, by analyzing the similarity of the user's profile with the component constituent product vector.

Based on the analysis, a product is considered to be liked by users, and then this product will be recommended to users. However, from the results that have been obtained, it is found that there is a
major drawback of this method that is the inability of the recommendation system to provide recommendations for new products or products that have never been seen by the user.

Another method approach for building a recommendation system is collaborative filtering. This method is used in [11] to create a recommendation mechanism in Netflix with the data set used containing 17 thousand films and more than 500 thousand users. One method used in the implementation of this research is item-based collaborative filtering, in which this method finds a similarity between products by considering the transactions record between products against existing users. This method is considered an easy method to be implemented and efficient in producing solutions. However, there are many challenges and limitations contained in this method, which is often known as sparsity and scalability. Sparsity is a condition where the recommendation system is faced with products that are not popular or new stored in the system, so this product will most likely never be included in the product that will be recommended. While scalability is a condition in which computing will continually grow because this method has a dependency on the relationship between products and customers. So that increasing the number of products and customers, can result in computing costs that also increase.

Problems that arise from the two previous methods pose challenges for optimizing the performance of the recommendation system. Optimization of existing methods for recommendation systems usually uses a hybrid method approach. Content-Boosted Collaborative Filtering method used in [9] is one of the methods that apply this approach. This method uses two recommendation methods namely content-based (CB) and collaborative filtering (CF). The basic approach used is to apply CB to change the user's assessment matrix of products, which are incomplete into a complete assessment matrix. Next, use CF to make recommendations. The results of the study found that this hybrid method could solve the sparsity problems faced in the two previous methods, but there was still a need for improvements to be able to optimize the performance of this hybrid method. One of them is by using a predictive method that is good enough to build a complete assessment matrix of the user and product relationship.

3. System Description
Generally, the recommendation process are performed by the system can be seen in Figure 1. The process begins by collecting detailed information related to content items from the wedding organizer or vendor as well as the user's profile. Besides, a history of rating transactions from users of items is also collected to form an actual user rating matrix. Furthermore, the process will start with the recommendation method which is a hybrid approach method that uses two recommendation methods, namely CB and CF [9].

The fundamental approach of the method used is to use CB to transform incomplete user rating matrix into a complete user rating matrix, which then uses CF to make recommendations. The step by step from this method for providing recommendations is as follows:

![Figure 1. System Overview.](image-url)
• Forms a user rating matrix, for each registered user. The rating matrix contains the value of $v_{u,i}$ which means the rating value given by the user $u$ to the product $i$, considering with the following formula

$$v_{u,i} = \begin{cases} r_{u,i} : \text{If user } u \text{ rate item } i \\ c_{u,i} : \text{Otherwise,} \end{cases}$$

(1)

where $r_{u,i}$ is the actual rating data from the user $u$ to the product $i$, whereas $c_{u,i}$ is a prediction of the rating value using the CB method.

• The CB method then provide the prediction processing as a problem of categorizing the text. Information from the product or user profile is used as document text and user ratings of the product are used as class labels. The Naive Bayes Text Classifier concept is used in two classification processes:
  - Rating history from the users to a product, will be used to predict the rating value on the product that has never been rated by that user. The process is performed by considering the similarity value of the product description between the unrated products and the rated products.
  - If CB is faced with active users who are newly registered in the system, then the history of other users will be used to predict the rating value of a product that will be given by the new user. The process is performed by looking at the similarity of the new user profile with the profile of users who have already provided an rating in the system.

• After the rating matrix of each user for each product is successfully generated (full user rating matrix), the recommendation process then start to the CF method phase. This process begins by giving weights to all users by considering similarities with other users ($p_{a,u}$).

$$p_{a,u} = \frac{\sum_{i=1}^{n_{a}} (v_{a,i} - \overline{v}_a) \cdot (v_{u,i} - \overline{v}_u)}{\sqrt{\sum_{i=1}^{n_{a}} (v_{a,i} - \overline{v}_a)^2 \cdot (v_{u,i} - \overline{v}_u)^2}},$$

(2)

where $v_{a,i}$ is the rating data given to product $i$ by user $a$. Furthermore, $\overline{v}_a$ is the average rating given by the user $a$. Whereas $m$ is the total number of products used as a comparison between users.

• Next calculate Harmonic Mean Weighting ($hm_{i,j}$) and Hybrid Correlation Weight ($hw_{a,u}$) to calculate the correlation weights between users, using the following formula

$$hm_{i,j} = \frac{2m_mm_j}{m_i+m_j}, \quad m_u = \begin{cases} n_u : \text{If } n_u < \beta \\ \beta : \text{Otherwise,} \end{cases}$$

(3)

$$hw_{a,u} = hm_{a,u} + sg_{a,u}, \quad sg_{a,u} = \begin{cases} q : \text{If } q < \beta \\ \beta : \text{Otherwise,} \end{cases}$$

(4)

where $n_a$ is the number of products that have been rated by the user $u$, $\beta$ is a parameter of the minimum number of products that must have been rated as consideration for giving a good weight in the calculation (in this study use a value of 15), while $q$ is the number of products that have been rated by the user $a$ and $u$ users.

• Next is calculating Self Weighting ($sw_a$) which is used as a weight to increase the level of confidence in the predicted value from the results of CB processing for a user. This value is given by

$$sw_a = \begin{cases} \frac{n_a}{\beta} \times max : \text{If } n_a < \beta \\ max : \text{Otherwise,} \end{cases}$$

(5)

by considering the number of products that have been rated by a user ($n_a$), $\beta$ is the same parameter as in equation (3) and $max$ which is the weight value which represents the highest level of confidence from the predicted value.
After obtaining all the required values, the last calculation is to calculate the predicted value \( p_{a,i} \) of a user \( a \) to product \( i \), that use the following formula

\[
p_{a,i} = \begin{cases} 
\bar{v}_a + \frac{\sum_{u=1}^{n} \hat{w}_{a,u} (c_{a,i} - \bar{v}_a)}{\sum_{u=1}^{n} \hat{w}_{a,u}} & \text{if } a \text{ is a new user} \\
\bar{v}_a + \frac{\sum_{u=1}^{n} \hat{w}_{a,u} (v_{u,i} - \bar{v}_u)}{\sum_{u=1}^{n} \hat{w}_{a,u}} & \text{otherwise}
\end{cases}
\]

(6)

Considering all parameters explained in the previous step, where \( c_{a,i} \) is the content-based prediction value of user \( a \) for product \( i \), \( v_{u,i} \) is the value of the rating matrix of user \( u \) for product \( i \), \( \bar{v}_u \) is the average value of all products rated by user \( u \), \( n \) which contains the number of nearest neighbours of user \( a \), as well as the values of \( p_{a,u} \), \( \hat{w}_{a,u} \) and \( \hat{w}_a \) which are the results of equations (2), (4) and (5) respectively.

Finally, for each user \( a \), take a number of products \( i \) that have not been rated, by considering the biggest value of \( p_{a,i} \) from equation (6), to make recommendations to users \( a \).

4. Domain Description
In this research, the implementation of the proposed hybrid method is in the domain of brides/groom rating to the decoration products of the six wedding event vendors. The data set used in the training and the testing process was obtained from a survey that has been done by researchers toward 14 users and 25 decoration products from these vendors. Users who have different characteristics are built profiles that can describe every user based on the attributes of age, job, ethnicity, wedding theme, favorite music genre, favorite color, wedding budget, number of guests and the status of the wedding organizer usage. While the product profile included in the content-based predictor processing is built using the product description and price attributes.

In the learning and the testing process of the proposed hybrid method of 14 users who have rated the product, a simulation of 4 users is applied as a newly registered user (testing) and 10 users are considered as old users (training). Besides, from the pair of 10 user ratings of 25 stored products (250 ratings) 150 rating data are used randomly to form the actual user rating matrix.

5. Result and Discussion
In the implementation of the Content-Based Predictor method, two phases of data preprocessing are used, namely the elimination of useless characters with the Regular Expression function and the removal of words that are not related to the prediction context by using the Indonesian Stopword function. Furthermore, the processing of multi-class text classification is used using the Multinomial Naive Bayes Text Classifier method at the CB stage. The multi-class in question describes the rating that might be given by the user to the item, which is in the range of values from 0 to 5.

Furthermore, the implementation of the Collaborative Filtering method uses two key parameters in the calculation process, namely the minimum product rate that provides a good consideration in the calculation process and the weight value that represents the highest level of confidence from the predicted CB value. In this research, the minimum value of product rate is 15, because the results of experiments that have been conducted on 4 minimum product rate values (5, 10, 15, 20, 25) show that the parameter value of 15 has the best results in the accuracy of the prediction. These results can be seen in Figure 2. Whereas for the weight value parameter of the CB confidence level prediction, a value of 2 is used, because it considers good results from previous studies [9].
In the quality calculation of the recommendation results, another approach to testing predictions is used which does not look at the matching level of the rating prediction results with the actual rating given by the user. The quality calculation applied uses the approach that check the status of goodness in products predictions that users will indeed add the product in the list under consideration. The approach is prediction accuracy by measuring usage prediction, where the evaluation through a comparison of the status of the product used/considered (actual data) by the user with the results of recommendations given by the system [13]. The results to be obtained from these comparisons form a table with 4 possible outcomes, which can be seen in Table 1.

**Table 1. The Possible Result from A Recommendation**

| Recommended Product   | Unrecommended Product |
|-----------------------|-----------------------|
| Used/Considered Product | True-Positive (TP)     |
| Unused/Unconsidered Product | False-Negative (FN) |
|                       | False-Positive (FP)    |
|                       | True-Negative (TN)     |

Furthermore, based on the possible results of these recommendations, the values of accuracy, precision, and recall are calculated. The formulation and definition of the 3 quality assessments are as follows [14]:

- **Accuracy** is the level of sensibility of a system can recommend or not recommend a product correctly, which is obtained by the following formula (7),

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

(7)

- **Recall or Sensitivity** is the proportion of Recommended Product conditions that correspond to Used Products to the overall Used Product, which is the formula as follows

$$\text{Recall} = \frac{TP}{TP + FN}.$$

(8)

- **Precision or Confidence** is the proportion of Recommended Product conditions that correspond to Used Products to the overall Recommended Product. This value can be obtained by applying this formula (9)

$$\text{Precision} = \frac{TP}{TP + FP}.$$

(9)

Based on the explanation of determining parameters and calculating the quality of recommendations, and by applying the equation (10)

$$\text{Recommendation Product} = \begin{cases} 1 & : \text{if the rating value} \geq 3 \\ 0 & : \text{Otherwise} \end{cases},$$

(10)

to classify the rating values into the recommended/used product group, the experimental results are CBCF ability in providing recommendations for wedding organizer or vendors in this study is 69% accuracy, precision 74%, and 84% recall.
The authors realize that the results obtained are not very good but by considering the data of newly registered users in providing recommendations, which if pure CF or pure CB are used (usually using a profile product in providing recommendations) then the recommendation process cannot be performed. However, the recommendation process can still perform with CBCF, which can provide the recommendation for the newly registered users with the capability that is 54% accuracy, 57% precision, and 85% recall.

6. Conclusion and Future Work

Based on the results and discussion that have been presented previously, the conclusion of this study are (i) CBCF with the Minimum Product Rate parameter value of 15 in this study has a better capability than other values, which CBCF can provide recommendations with an accuracy of 69%, 74% precision, and 84% recall; (ii) CBCF with the Minimum Product Rate parameter value of 15 in this study has a better capability than other values, which CBCF can provide recommendations with an accuracy of 69%, 74% precision, as well as 84% recall; and finally (iii) CBCF's ability for this study is not very good in providing recommendations to all existing Users, but CBCF can provide recommendations on the condition of Newly Registered Users with an ability of 54% accuracy, 57% precision and 85% recall (which if pure CB or pure CF are used for this condition then can’t be given recommendations).

Considering the results, discussion and conclusions that have been obtained, a lot of subsequent work to get better CBCF performance, the improvement plan includes (i) try other algorithms to do text classification by considering better algorithms than multinomial naive bayes, such as support vector machines or variants of artificial neural networks; (ii) training data should contain a variety of user-profiles and products, thus making the CBCF learning process better, and more able to provide recommendations with a variety of conditions encountered; or maybe can (iii) try other learning strategies for conducting CBCF training (such as K-Fold Cross Validation) so that all the data obtained has been treated as training data or testing data.

Acknowledgments

This research is supported by internal grant from Department of Research, Community Service and Education Quality Assurance, Institut Teknologi Sumatera (Hibah Penelitian ITERA Smart 2019) with the contract number B/317/IT9.C1/PT.01.03/2019.

References

[1] Soemartono H K 2018 Potret Zaman Now Pengguna dan Perilaku Internet Indonesia! (Jakarta: Buletin APJII) p 1-7
[2] Anonymous 2018 The Mobile Economy 2018 (London: The GSM Association)
[3] Millward S 2014 Indonesia Diproyeksi Lampaui 100 Juta Pengguna Smartphone di 2018 Accessed on March 3rd, 2019
[4] Maulana S M, Susilo H and Riadi 2015 Implementasi e-commerce sebagai media penjualan online (studi kasus pada toko pastbrik kota malang) J Administrasi Bisnis 29(1) 1-9
[5] Kemp S 2018 Digital in 2018: World’s Internet Users Pass The 4 Billion Mark Accessed on March 6th, 2019
[6] Kristin D M and Lisanti Y 2014 Wedding organizer order management ComTech: Comp, Math and Eng App 5(2) 839-850
[7] Najiyah I and Suharyanto S 2017 Sistem informasi wedding planner berbasis web J Ilmu Pengetahuan dan Teknologi Komp 3(1) 79-86
[8] Ricci F, Rokach L and Shapira B 2011 Recommender System Handbook ed Ricci F, Rokach L, Shapira B and Kantor P B (New York: Springer Science+Business Media) pp 1-29
[9] Melville P, Mooney R J and Nagarajan R 2002 Proc. Eighteenth National Conference on Artificial Intelligence (Edmonton) (Cambridge: AAAI Press/MIT Press) p 187
[10] Sofian J and Solihin H H 2018 Pembangunan aplikasi m-wedding sebagai solusi usaha wedding organizer J Tekno dan Info 8(2) 1-7
[11] Molina L E and Bhulai S 2018 Recommencement System for Netflix (Amsterdam: Vrije Universiteit Amsterdam)
[12] Badriyah T, Fernando R and Syarif I 2018 Proc. Konferensi Nasional Sistem Informasi (Pangkal Pinang) (Pangkal Pinang: STMIK Atma Luhur) p 554
[13] Shani G and Gunawardana A 2011 Recommender System Handbook ed Ricci F, Rokach L, Shapira B and Kantor P B (New York: Springer Science+Business Media) pp 257-297
[14] Powers D M W 2011 Evaluation: from precision, recall and f-measure to roc, informedness, markedness & correlation J of Machine Learning Tech 2(1) 37-63