A Hybrid Recommendation System for E-Commerce based on Product Description and User Profile

Tessy Badriyah, Erry Tri Wijayanto, Iwan Syarif, Prima Kristalina
Informatics Department (Program Studi D4 Teknik Informatika)
Electronic Engineering Polytechnic Institute of Surabaya (EEPIS)
Kampus Politeknik Elektronika Negeri Surabaya (PENS) Keputih Sukolilo Surabaya 60111, INDONESIA
tessy@pens.ac.id, errytri03@gmail.com, iwanarif@pens.ac.id, prima@pens.ac.id

***Abstract*—E-commerce is an online trading system that eases transactions for both sellers and consumers without having to meet in person. The prevalence of e-commerce has increased competition amongst sellers, hence the users of e-commerce has to increase their performance, one of them by using recommendation system. This research develops a hybrid recommendation system for e-commerce that implements Content-based Filtering and Collaborative Filtering methods, which will compute the simmilarities of product description and user profile. In experiment results, it was found that the recommendation has similarity with product description and the preference of user profile with the average of precision value is 67.5% and recall value is 71.47%.**

Keywords—recommendation systems, content-based filtering, collaborative filtering, hybrid

I. INTRODUCTION

The recommendation system is a tool and a technique that provides suggestions about things that can be utilized by the user. In an online store or e-commerce service, the suggestions given can be offered as services or products.

With the addition of recommendation features, the level of convenience from shopping in online stores will increase. Users do not need to look at the entire product catalog to find the product they want, because the system can detect user preferences and provide potential choices of products that are preferred by the user.

The purpose of this research is to build a recommendation system on e-commerce that has the following contributions:

1. This research uses Text Mining TF-IDF (term frequency-inverse document frequency) method to generate tags automatically based on product description
2. The system created combines generate tag results automatically with the user profile to get relevant recommendations

The rest of the paper is organized as follows. The next section provides a brief overview of some related research work related to recommendation systems. The third section explains about two methods that form the basis of the research, which are Content Based Filtering and Collaborative Filtering along with TF-IDF method to generate tags automatically. Section 4 explains in detail how the method we propose gives recommendations to users. Section 5 describes results of different experiments and discussion of the results. The final section provides conclusion.

II. RELATED WORKS

The Implementation of Collaborative Filtering was first performed by (Goldberg, Nichols, Oki, & Terry, 1992). The system built revolved around a small community that has a limited scope. However, this cannot be done for a wider community where individuals do not know each other.

Likewise, with other systems that use the Collaborative method as done by Schafer, J.B. et.al. (2007), that use the adjusted cosine similarity method. While the research conducted by GroupLens research system (Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994), (Konstan et al., 1997)] provides collaborative filtering solutions for news and movie sites. While research related to Content-based filtering is done by Robin van Meteren and Maarten van Someren, (2010) who built recommendation system using PRES (Personalized Recommender System) which can give suggestion to user by making dynamic hyperlink for websites that contain collection of articles about home improvement.

Instead of the method of content-based filtering and collaborative filtering being done separately, a research conducted by Vibhor Kant and Kamal K. Bharadwaj, (2012) and Songjie Gong (2012) combines the two methods. In their research, the quality of recommendation result is improved by combining two methods and fuzzyfication of the similarity value. The result of the hybrid combination provides satisfactory results. That is what motivate us to conduct research by combining two methods, Collaborative Filtering and Content based Filtering techniques in our own way.

III. IMPLEMENTING A RECOMMENDER SYSTEM IN E-COMMERCE

This section describes the two methods that underlie the basis of this research: Content based filtering and Collaborative filtering method along with TF-IDF method which used to generate tags automatically.

A. Content based Filtering

The flow process of the content-based filtering method begins with the source of information, which is the collection of items to be recommended. From them the description of each item is taken, and will be analyzed by the Content Analyzer. In the content analysis process (Content Analyzer), each item description will be analyzed using text mining to generate the words that often appear (stop word), this will then be used as features in each item respectively.

Content Analyzer produces a collection of data in the form of structured items that are stored in Represented Items. The next process is Profile Learner which learns the user profile. User is learned by retrieving data from Represented Items, and then user profile will be generated and stored in the Profiles data store.

The next process is Filtering Component, which filters items that exist based on a particular user profile. The user profile data is retrieved from the Profiles data store, while the item data is retrieved from the Represented Items data store. Data user profile and data items are compared and filtered so as to generate a list of recommended items for the user.

1) The Process of Determine Product Profile

An important part of content-based filtering is the process of determining product profiles. Product profile is a set that contains some data that describes a product. Some product profiles will form product profile matrices. Generally, product profile determination is done as follows:

- The set of product categories. Some buyers may be more interested in products within certain categories.
- The set of product specifications. Some buyers may prefer products with certain specifications.
- The set of years of product manufacturing. Some buyers may be more interested in finding old products or more looking for the latest products.

In contrast to the product profile determination methods mentioned above, this research uses the Text Mining TF-IDF method to generate tags automatically derived from the description of a product.

In general, the TF-IDF technique aims to know the value of the number of related words between documents. And in the context of this research, the document in question is the product description content. It is known that the formula of calculating TF-IDF is as follows:

$$TFIDF_{d,t} = FREQ_{d,t}\ (1 + \log(\frac{N}{DFREQ_t})) \qquad (1)$$

Where,

- $FREQ_{d,t}$ = number of term *t* in the document *d*
- N = Total number of document used
- $DFREQ_t$ = number of documents where term t appears

B. Collaborative Filtering

Collaborative filtering makes recommendations based on other people - some people collaborate to come up with recommendations. It works as follows - suppose the task is to recommend a product to user A. Then it will be searched among other users of the site to find one that is similar to the user A in the product she/he enjoys. Collaborative filtering can be user based (user based collaborative filtering) or item based (item based collaborative filtering).

There are several ways of calculating similarities between items. It can be done using cosine similarity algorithm (Desphande and Karypis, 2004) and correlation similarity (Sarwar et al., 2001) which will be described below.

1) Cosine Similarity

In this case, two items are considered as two vectors in user-space dimension m. Similarity between items is calculated by the cosine calculation of angles between two vectors. According to the principle of arithmetic, two vectors are said to be the same if they form an angle of 0 ° (zero degrees), or the cosine is equal to 1 (one). Formally, similarity between items A and B, denoted by sim(A,B) is given by

$$sim(A,B) = \cos(\theta) = \frac{A.B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \qquad (2)$$

Where:

$sim(A,B)$ is a similarity measure between vector A and B

$A_i B_i$ are components of vector A and B respectively

1) Correlation Similarity

In correlation similarity, the similarity between two items is calculated using a statistical technique called Pearson Correlation. To calculate the correlation value between two items, the rating value that does not have a pair on the same user is excluded from the calculation. For example, a user set is denoted by U, which assigns ratings to items A and B. Then the correlation value of both items can be expressed as:

$$sim(A,B) = \mathrm{corr(A,B)} = \frac{\sum_{u\in U}(R_{u,A}-R_A)(R_{u,B}-R_B)}{\sqrt{\sum_{u\in U}(R_{u,A}-R_A)^2}\sqrt{\sum_{u\in U}(R_{u,B}-R_B)^2}} \qquad (3)$$

Where:

$R_{u,A}$ denotes the rating of user u on item A

R_A is the average rating of the A-th item.

$R_{u,B}$ denotes the rating of user u on item B

R_B is the average rating of the B-th item.

IV. The Recommendation Process

Fig 1 illustrates a main diagram of the recommendation system in detail. The picture describes how the recommendation system produces a list of product recommendations displayed to user.

From Fig 1 it can be seen that there are several processes and components involved such as rating the product, searching for product features, user profile formation, product profile formation, and then matching process until finally a list of products that have similar features is generated or we may call it a list of product recommendations.

Briefly, during the process of Recommendation system, each product has all the scores and tags of all products, which will then create the Product Profile.

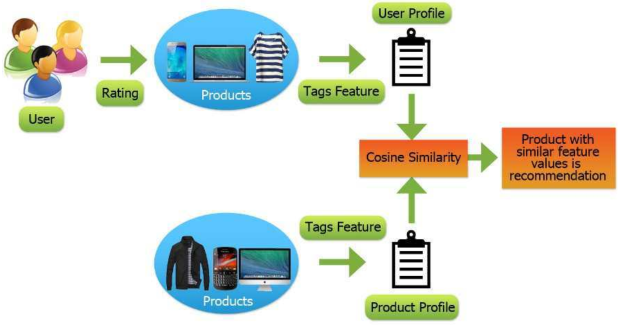

Fig. 1. The main Diagram of Recommendation System

After user rates a product, a User Profile can be made. From the Product and User Profile, we can do a matching process which aims to calculate the degree of distance between two Profiles. The profiles are in the form of matrices and we use cosine distance formula to calculate the similarity. Each product has a degree of distance to the user, afterwards the product recommendations are displayed to users from the highest distance to the smallest.

Based on the explanation of the system flow recommendations in Fig 1, the important parts of recommendation system will be described in the following section.

A. *How to obtain Product's tags (features)*

Product profiles are built based on features in the form of product tags. These product tags are obtained from the process of mining product description content as described in Fig 2. The method for obtaining product tag features uses the general text mining process that is tokenizing, filtering, stemming and analyzing.

The process carried out at the analyzing stage in the text mining process used the calculation of the Term Frequency Inversed Document Frequency (TF-IDF) value of each word in each product description using the TF-IDF formula described earlier in formula (1). After that, on each product description was taken 4 (four) words with the highest TF-IDF value. In this case, for the threshold value the number of words taken is assigned it means 4 tags is defined for each product.

B. *Represents the Product Profile*

After getting the tags (features) on each product where taken 4 tags that have the highest TF-IDF value, then the next process is a formation of Product Profile.

Fig 2 shows a process diagram forming a Product Profile. Combined vector products and product tags form the product profile. The product profile can have a value of 1 or 0, which will have a value of 1 when a tag appears on a product and is 0 for otherwise.

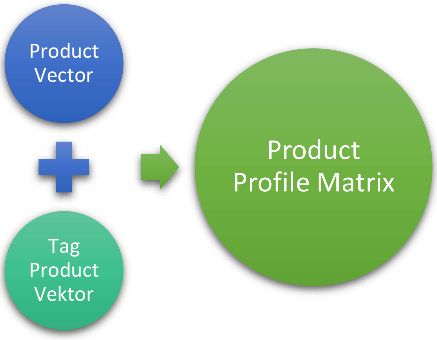

Fig. 2. Process Diagram forming a Product Profile

Table 1 gives an example of creating a product profile if it is known that there are 3 products, namely Product A, Product B and Product C. The combined features of both products are 'pc', 'handphone', 'laptop', 'superior', and 'application'.

TABLE I. THE EXAMPLE OF PRODUCT PROFILE MATRIX

	pc	handphone	laptop	superior	application
Product A	1	0	0	1	1
Product B	1	1	1	0	0
Product C	0	0	1	1	0

Furthermore, the product profile will be calculated in distance with the user profile using cosine distance. In the following section will explain how to obtain a user profile.

C. *Represents the User Profile*

In this study, not only create vectors that describe the product which we previously call it the Product Profile. But also create vectors with the same components to describe the user preferences or we call it the User Profile.

The user profile is obtained by using the rating given by the user on a product as a matrix. Table II below represents an example of user rating of a product called a user rating matrix.

TABLE II. THE EXAMPLE OF USER RATING MATRIX

	Product A	Product B	Product C
User 1	3	5	1
User 2	2	2	3

Next, how to form user profiles is illustrated by the diagram shown in Fig 3.

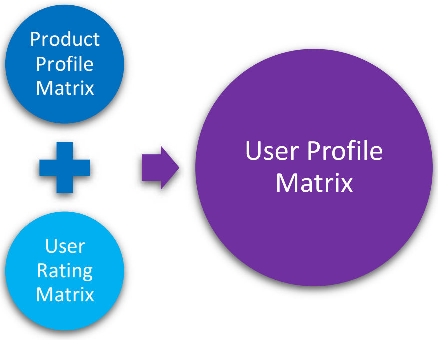

Fig. 3. Process Diagram forming User Profile

The value of user rating to a product is worth between 1 and 5. This value is not a boolean number, so we need to normalize the value of this rating by subtracting the average value of a user. This allows a negative score for a product with a rating value below the average rating and a positive value for a product with a rating rating above the average rating.

Based on the existing Product profile matrix in Table I and User rating matrix in Table II, then we can get the User Profile matrix in Table III.

TABLE III. THE EXAMPLE OF PROFILE USER MATRIX

	pc	handphone	laptop	superior	application
User 1	1	2	0	-1	0
User 2	-1	-1	-1/2	-1/2	-1

D. *Finding Recommended Product ranking*

Based on the two vectors that have been discussed in the previous section, the Product Profile and the User Profile vectors, we can estimate the degree of a user will be preferable with any products. How to calculate the degree is by calculating the cosine distance between the Product Profile vector and the User Profile vector. The formula calculates the cosine distance between two vectors using TF-IDF method as described in formula 2.

Then we calculate cosine similarity of the example of Product Profile matrix in Table I and User Profile matrix in Table III. The result is ***User 1*** will get product recommendation from three product that is Product A, B and C, which product will be recommendation according to the order of preference.

The result of vector calculation of ***User 1*** with Product A, B and C is 1.000, 1.732 and -0.707. The greatest result is a highly recommended product to the user. So the product recommendation if sorted is Product B, A, and C. If we applied into many amount of product data and tag features, the results of the recommendations will show the maximum results.

V. EXPERIMENTS AND ANALYSIS

This section will describe the experiments and analyze the recommendations made. These include experiments to obtain product tags, experiments to obtain product recommendations and analysis of the accuracy of the recommendation system by calculating the precision and recall values along with the required execution time calculations. From the results of this analysis we can get conclusions about the performance of the system.

A. *The experiment to obtain Product's Tags (features)*

In each process of searching the Product's tags (features), it will search 20 words that potentially be tags which describe the product. From those 20 words are calculated the value of TF-IDF it can be obtained 4 words with the highest TF-IDF values. These four words become the Product's tags that will be used in content-based filtering. The output of this process is four product's tag that can be shown in the product details and display it on the product detail page.

B. *The Experiment to obtain Product Recommendation*

This following will illustrate the experiment to obtain product recommendations. Table IV exemplifies the id_product and rating provided by the user.

TABLE IV. THE RATING TABLE

Product id	Rate
142	3
127	4
163	1
161	2
131	5

Following the recommendation process as described in the previous section, the similarity value between the User Profile vector and the Product Profile vector can be shown in Table V.

The output of Product Recommendation process in Table V is the list of products that can be seen in the product recommendation page.

C. *The Analysis of Recommendation Results*

This section will describe the performance analysis of the recommendations by measuring the accuracy of the recommendations by calculating the precision and recall values along with the calculation of the execution time.

1) *Measuring Performance of Recommendation*

Performance measurements of the recommendations are made by calculating the precision and recall values. In general, the formulas used to calculate precision and recall values are:

$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|} \quad (4)$$

$$recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|} \quad (5)$$

TABLE V. THE RECOMMENDATION TABLE

No.	Product id	Similarity
1	142	0.7218155795421
2	127	0.58838906332371
3	132	0.40684150846918
4	131	0.40684150846918
5	130	0.25372911280874
6	161	0.22529395361466
7	129	0.20123343429659
8	133	0.10936599690032
9	153	0.04812103863614
10	140	0.025154179287073

Table VI below gives the result of calculating precision and recall values.

TABLE VI. THE RESULTS OF PRECISION AND RECALL

No.	a	b	(a+b)	c	(a+c)	Precision (%)	Recall (%)
1.	5	5	10	1	6	50	83.33
2.	6	4	10	2	8	60	75
3.	9	1	10	4	13	90	69.23
4.	7	3	10	5	12	70	58.33
			Rata-rata			67.5	71.47

Where:

- a is the number of products found and relevant
- b is the number of products found but not relevant
- c is the number of products not found but relevant
- d is the number of products not found and not relevant

The average Precision value of the above data is 67.5% and the recall value is 71.47% of the 0% -100% scale. From that value, it can be seen that the precision value is lower than the recall value based on the user-rated tag on the product.

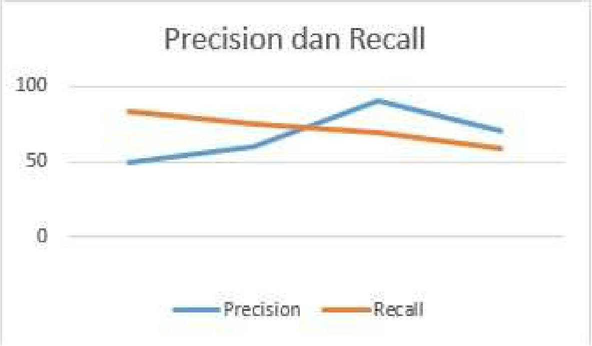

Fig. 4. Precision and Recall

2) *Measuring Execution Time*

The next test is to calculate the execution time to run the program during the process of making product features in the form of tags and the process of forming the recommendations. The calculation of execution time needs additional applications contained in the web browser, namely page load time. Based on the page load time calculation in Fig 5, the average execution time required is 7.7 seconds in creating the tag features. The high time of this execution is caused by the process of reading the content of the product description is long enough so that it depends on the number of characters on the content description. This long execution time does not interfere the user because the algorithm is put in the system admin section.

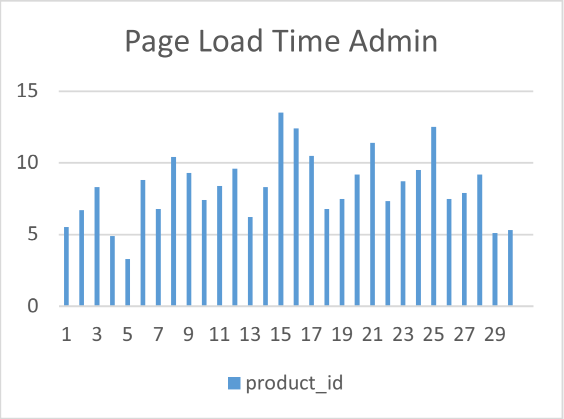

Fig. 5. Page Load Time Diagram

VI. CONCLUSIONS

In this paper, we develop a hybrid recommendation system based on product profile and user profile. Our research uses Text TF-IDF (term frequency-inverse document frequency) method to generate tags automatically based on product description. It is part of content-based filtering in the Recommendation System. After that, Product Profile in the form of Product's tags then combined with User Profile by using cosine similarity method. Cosine similarity method commonly used in Collaborative Filtering that exist in Recommendation System. Thus, we combine two methods in the Recommendation System: Content based Filtering and Collaborative Filtering method.

The advantage of determining product profile through the creation of tags automatically is the process becomes more efficient and dynamic. Efficient because the Product's tag do not need to be input manually by the administrator. While the more dynamic because the tag results obtained will adjust to the content of the product description. If the product description changes, the tags will change automatically.

The use of product features in the form of tags more descriptive than determine product based on the category, year of production and product specifications as commonly used in the Content based filtering method. More descriptive because a category may be owned by one product and may also belong to another product. As well as the use of year of production, certainly in the same year there are so many products produced. So that, it could be less relevant to find the recommended products.

In the experiment results, it was found that the recommendation has similarity with product description and the preference of user profile with the average of precision value is 67.5% and recall value is 71.47%. The size of precision and recall precision also depend heavily on what is really meant by "relevant products" and how to ensure whether if a document is relevant or not. It is very difficult to achieve ideal recall-precision levels because they are based on a very flexible and dynamic measure of relevance.

ACKNOWLEDGMENT

The authors would like to thank the Ministry of Research, Technology and Higher Education of Indonesia for funding this work as part of Overseas Conference Assistance Program (Program Bantuan Seminar Luar Negeri).

REFERENCES

[1] Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. Commun. ACM, 35(12), 61-70. doi:10.1145/138859.138867

[2] Schafer, J. B., Konstan, J., & Riedl, J. (1999). Recommender systems in e-commerce. Paper presented at the Proceedings of the 1st ACM conference on Electronic commerce, Denver, Colorado, USA

[3] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: an open architecture for collaborative filtering of netnews. Paper presented at the Proceedings of the 1994 ACM conference on Computer supported cooperative work, Chapel Hill, North Carolina, USA.

[4] Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Riedl, J. (1997). GroupLens: applying collaborative filtering to Usenet news. Commun. ACM, 40(3), 77-87. doi:10.1145/245108.245126

[5] Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. Paper presented at the Proceedings of the 10th international conference on World Wide Web, Hong Kong, Hong Kong.

[6] Lemire, Daniel and Maclachlan, Anna. Slope One Predictors for Online Rating-Based Collaborative Filtering. California : SIAM Data Mining (SDM'05), 2005.

[7] GroupLens. datasets. grouplens. [Online] 2014. http://grouplens.org/datasets/movielens/.

[8] Robin van Meteren dan Maarten van Someren.(2010). "Using Content-Based Filtering for Recommendation". New York: Springer

[9] Elaine Cecilia Gatto dan Sergio Donizetti.(2010). "Using Content-based Filtering in a System of Recommendation in the Context of Digital Mobile Interactive TV". New York: Springer

[10] Z. Qiu, M. Chen, and J. Huang, "Design of Multi-mode E-commerce Recommendation System," 2010 Third Int. Symp. Intell. Inf. Technol. Secur. Informatics, no. 807018, pp. 530–533, Apr. 2010.

[11] Vibhor Kant dan Kamal K. Bharadwaj.(2012). "Enhancing Recommendation Quality of Content-based Filtering through Collaborative Predictions and Fuzzy Similarity Measures". New York: Springer