Journal of Theoretical and Applied Information Technology
30th June 2022. Vol.100. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

IMPLEMENTATION OF RECOMMENDATION SYSTEM IN E-COMMERCE USING APPROXIMATE NEAREST NEIGHBOR

[1] HANDY TANTYO, [2]TUGA MAURITSIUS

[1,2] Information System Management Department

BINUS Graduate Program-Master of Information Systems Management

Bina Nusantara University, Jakarta 11480

E-mail: [1] handy.tantyo@binus.ac.id, [2] tmauritsus@binus.edu

ABSTRACT

At least 20 e-commerce businesses in Indonesia have stopped operating due to human resource problems. The recommendation system is an important feature in e-commerce which was initially done manually. However, a computing field has emerged that can replace human labor in recommendation systems, reduce human errors and work more efficiently. This paper aims to implement an efficient and scalable recommendation system using Machine Learning techniques. The content-based filtering used in this paper uses the Approximate Nearest Neighbor algorithm with different indexes that provide similar product recommendations. Data is collected from Tokopedia product information, which is divided into five categories. The optimal Nprobe is found at 6% which is the standard Nprobe for each index. The index that has the best recall@R parameter, fastest prediction time and training time respectively are IDMap,Flat, IVF1024_HNSW32,Flat and IDMap,Flat.

Keywords: *E-commerce, Recommendation System, Machine Learning, Content Based Filtering, Approximate Nearest Neighbor*

1. INTRODUCTION

Information technology and digital economy introduce new opportunities for all economic sectors. The development of information technology and its products support the creation of digital economy [1].

171.17 million out of 264.16 million Indonesian population (more than 60%) are internet literate or capable of using the internet. Online shopping using e-commerce is one the most frequent online activity in Indonesia [2]. E-commerce has advantages for increasing competitiveness in organization, especially in business management, finance, accounting and many more business fields [3]. Even more, Indonesian e-commerce transactions in 2022 are estimated to increase by more than 50% from 2020 or more than 400% when compared to 2018 [4].

However, there are more than 20 e-commerce startups collapsed in the last 2 years (2017-2019) [5]. The major problem is lack of competition with other e-commerce firms caused by recruiting too many workers for doing manual work, including recommendation system. In addition, recommendation system has an important role to increase the company's revenue. Like the world's e-commerce giant, Amazon, around 35% of their revenue comes from the recommendation engine [6].

With this manual process, the recommendations might be subjective and not based on the market traffic. Meanwhile, if the company has implemented machine learning techniques, usually the modeling requires users-items relationships information. However, this method is very limited with sparse data and can lead to cold start issue, which occurs when the input does not have any information related to other products or users, such as new products [7]. When the cold start issue occurred, the product will not get recommendation and made the page will not show the recommendation's widget or page.

Machine Learning approaches can be used to automate recommendation systems and reduce human error. The algorithm used in this paper is Approximate Nearest Neighbor, a category of content-based filtering, which can solve the cold

start problem found in the previous report. With this benefit, we can give the recommendations to all products even the product didn't have any information related user-items relationships. Product titles, descriptions, price, and categories are used as input. While the Approximate Nearest Neighbor will predict recommended products that have a product vector that is adjacent to the input product [8].

2. LITERATURE REVIEW

2.1 Content-based Filtering

Content-based filtering is a recommendation system based on item description data, rather than predictions based on user associations. The user will receive a suggestion that has the same description data as what the user has viewed. With this system, the recommendations can be generated, even with a limited data collection, thus a user can get the recommendation which does not have any transaction [9].

2.2 Doc2Vec

Doc2Vec is a Word2Vec method that extends from the word level to the document level. Doc2vec's main objective is to analyze the neural network from the target word and the surrounding word to determine the optimal vector representation. In a similar way to Word2Vec, each word is represented by a d-dimensional continuous vector (d << | V |, that is a measure of vocabulary in the corpus). Moreover, the document is represented as a continuous vector in the same word vector space [10].

In the previous research conducted by Donghwa Kim and friends related to Doc2Vec, the research implemented three document representation methods, namely TF-IDF, LDA and Doc2Vec. In their own study of Doc2vec, the researchers used PV-DM and PV-DBOW and combined these results with other vectors to improve the performance of the classification technique. The vector combination is called as concatenate [10].

2.3 Approximate Nearest Neighbor

Approximate Nearest Neighbor is a strategy for reducing the cost of similarity queries by lowering the accuracy.

In the research conducted by Zhibin Pan and his friends, to overcome the problems caused by the large dimensions and the amount of data in the dataset, Approximate Nearest Neighbor (ANN) method was used with Product Quantization (PQ), which is a special form of Vector Quantization (VQ). The purpose of PQ is to divide the original database space into the Cartesian product from low-dimensional sub-spaces. On the one hand, to avoid a thorough search, the PQ method was combined with the inverted file index (IVF), which uses standard vector quantizer as the coarse quantizer to get a small size of candidate database vectors [11].

Basically, the IVF index will divide the data into clusters. Vector x is only compared with vectors that have the same cluster and adjacent clusters. There are two important parameters for the query process method, namely Ncells (the desired number of cells for output) and Nprobe (cells outside of Ncells which are compared to perform a search). The research shows IVF can effectively reduce costs and achieve substantial acceleration over the comprehensive searches [12].

In addition, a development made by Yu. A. Malkov increase the prediction speed of ANN algorithm using the combination of IVF index and HNSW (Hierarchical Navigable Small World). HNSW is a graph-based incremental ANN structure, which offers more efficient logarithmic complexity scaling. The main steps of HNSW are explicit selection of graph entry point nodes, split links with different scales, and utilize advanced heuristics to select neighbors. The HNSW algorithm can be considered as the development of a probabilistic skip list structure with a graph approach compared to a linked list [13].

2.4 Cosine Similarity

In mathematics, the cosine similarity analyzes similarity by measuring the cosine of the angle between two vectors [14]. In general, the implementation of cosine similarity is used in similarity texts. The output range of cosine similarity is 0-1. The higher the cosine similarity value means the higher similarity of two texts.

2.5 Evaluation Metrics

Evaluation metrics used in this research is recall@R for measuring the accuracy of ANN. For a query set with size Nq, we retrieve the first R positions for every query and include Nr nearest neighbors, then the recall@ R is the ratio of Nr to Nq. When R is long enough, several techniques may obtain extremely high recall@R, but most users are only interested in the top 100 or even less results for ANN search [11].

2.6 CRISP-DM

CRISP-DM is one of the most popular frameworks for applying data mining projects. The

"Business Understanding" process is the first step of CRISP-DM that determine whether a machine learning modelling can affect business operations. The following step is "Data Understanding" which aims to determine hypotheses for hidden information about the objectives of a machine learning modeling project that is formed based on experience and qualified assumptions [15]. The third process is called "Data Preparation", the process aims to collect relevant data and prepare the data for the next process, namely modeling [16]. Modelling is a workflow built to find the desired parameter settings for the selected algorithm and to run machine learning modeling tasks [17]. The next phase is "Evaluation", in which the model is trained, then tested against a test dataset and evaluated against the underlying business objectives. If the findings of this evaluation are positive, this machine learning model will be implemented on a larger scale or production scale, or also known as the "Deployment" process [17].

3. METHODOLOGY

In this study, the framework used is CRISP-DM including data understanding, data preparation, modelling, and evaluation.

3.1 Data Understanding

In this stage, the data is collected, namely the product detail info on the product detail page. Web scraping was used to get the data from numerous products in Tokopedia E-commerce. By using web scrapping, data is retrieved through the text in the html on the product detail page. The parameters retrieved from product details vary widely, such as product titles, price, locations, categories and etc. Next phase is filtering parameters by checking whether these parameters are suitable for data preparation process and the final parameters are product title, category, product description and price. The categories chosen in this research are men's fashion, women's fashion, books, electronic devices, and health. In total the amount of collected training data is 50.000.

3.2 Data Preparation

The data preparation aims to get clean data in vector form. Data requires preprocessing before its features can be extracted. Doc2Vec, One-hot Encoding, and normalizing are the techniques utilized in this procedure. The product title will be cleaned with a stopword and stemming procedure before being entered into the Doc2Vec algorithm.

3.3 Modelling

The next process is a main process in machine learning, called modelling. The outcome of data preparation, which includes product title, category, and price, is the input for this modelling. A vector with a length of 128 dimensions is the result of data preparation. The result is in the form of a model, which is saved in a file with the extension of .pkl(pickle), joblib dan numpy.

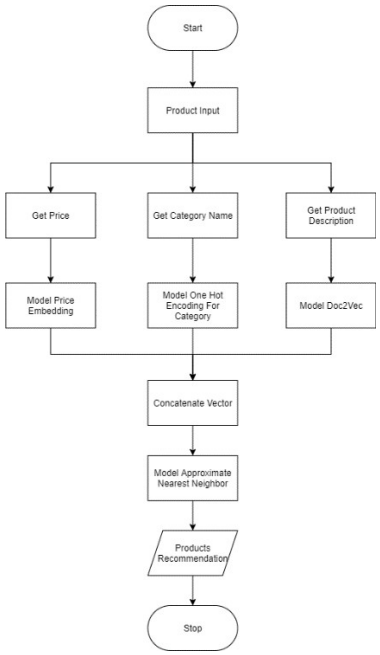

Figure 1. Flowchart from Data Preparation until Predicting Products

The Doc2Vec technique that was used is PV-DBOW to convert text into vectors, and the Approximate Nearest Neighbor algorithm was utilized to forecast neighboring products. The programming language is Python, and the libraries used are FAISS, Scikit-learn, and gensim. Index types such as IVFlat, IVF, and combination of IVF and HNSW are utilized in the Approximate Nearest Neighbor modelling process.

3.4 Evaluation

After the modelling process, some parameters will be evaluated against the Approximate Nearest Neighbor model, such as

- Evaluate the balanced Nprobe value based on recall@R value
- Evaluating each index using recall@R with a maximum k value of 100
- Evaluate the prediction time of each index

Journal of Theoretical and Applied Information Technology
30th June 2022. Vol.100. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. RESULT AND DISCUSSION

4.1 Data Understanding

The total data obtained from the web scrapping is 50.000 product data spread across 5 categories. The output of this stage is the data in csv file and the retrieved features are listed below.

- Product Title: This feature is an identifier on the product that distinguishes it from other products. The data type obtained from this product name is a string.
- Product Description: Product description is a more detailed explanation of the products offered in the store. The data type for the product description is string.
- Category Name: This column is part of the classification system entered by the shop owner. With the category, users can search for an item without having to search any keywords. In the scrapping process, the data type used in the category name is string. In this work, we used five categories, which are education books, health, electronic, men's fashion, and women's fashion.
- Price: Contains the nominal that must be paid by the user to get the item. The data type for the obtained price is integer.
- URL: It stands for Uniform Resource Locator, which is the internet address for a certain web page or file. In the scrapping process, this field is utilized for the second step.

4.2 Data Preprocessing

The product's data is transformed into the vector as the final result of data processing (figure 2). This vector will be used as a data input for modelling phase.

```
[[ 0.00242383 -0.18854553  0.25608432 -0.03566263  0.23377299  0.10460666
   0.06399751 -0.10496061  0.25636855 -0.02256762 -0.10681847 -0.03928288
  -0.00428684 -0.12762788  0.05694487  0.19325864  0.02636613 -0.23805663
  -0.01192074 -0.1432479   0.15262653 -0.41101792 -0.045863   -0.12191287
  -0.13252652 -0.05893086 -0.00663318 -0.0178797   0.15144216  0.05852488
  -0.16822775 -0.01795682 -0.03745513 -0.00962356 -0.0312625   0.15032497
   0.1097229   0.13192704  0.05514287 -0.21123746 -0.01380176 -0.20322008
  -0.2205382   0.20771448 -0.13093153 -0.08893307 -0.20503081 -0.08664417
   0.04279458  0.15826018  0.16624385  0.11724231 -0.01171989  0.09261835
   0.23261613  0.02567687  0.00502252  0.03574204 -0.033773   -0.01369787
   0.27657712  0.12590416 -0.10990014 -0.12975471 -0.29270303 -0.14085296
  -0.14888045 -0.35573497  0.24423039 -0.36948124  0.03531016  0.04413002
  -0.22725686 -0.0064602  -0.08095612  0.07601814 -0.02303726  0.01610803
  -0.39405292  0.02841392 -0.10575467  0.02277299 -0.05300098  0.21779336
  -0.021326    0.15673007  0.11109421  0.05313873 -0.24924436  0.07290609
  -0.19897963 -0.25104031  0.11793074 -0.06602298 -0.16442478 -0.10477549
   0.22768225 -0.17779233  0.04124251  0.14002891  0.08851343 -0.20346895
  -0.0416136  -0.21948104 -0.27755976 -0.27880085 -0.06469191  0.06366134
   0.01947488  0.089955   -0.06799984 -0.02025569  0.13911539 -0.14082949
  -0.15398635 -0.2653673  -0.19193542 -0.03920398  1.          0.
   0.          0.          0.78456592  0.          0.          0.        ]
```

Figure 2. Output Example to Predict the Vector from Input

4.3 Modelling

The Approximate Nearest Neighbor training procedure contains many tunings that may be modified according to the demands and aims. The following are the details.

- Index Type. The Flat index is the first and most often used as an index. Without any extra pre-processing, such as data compression, all data will be directly inserted into the index. This results in a high level of accuracy, but a slightly longer prediction time. The IVF Flat (Inverted File) index and HNSW (Hierarchical Navigable Small World) also utilized in this research.
- Nprobe. Nprobe expands the number of buckets that may be searched. As a result, the computational load increases and the search performance decreases, but the search precision increases. The situation may differ per data set with different distributions.

The output obtained in this process is the input product along with the recommended product which is written in the csv file. An example of the output can be seen in figure 3.

4.4 Evaluation

As previously explained, the algorithm used in this study is the Approximate Nearest Neighbor algorithm. In the validation and tuning process, there are several variables utilized to get the maximum results, which will be evaluated though Nprobe on the accuracy of the testing data, the level of precision, and the recall on each index (to check the speed of the index to predict the product).

4.4.1 Evaluation The Effect of Nprobe on Recall@100

There are six indices utilized in the graph illustrated in Figure 4. IVF1024,Flat, IVF1024 HNSW32,Flat, IVF2048,Flat, IVF2048 HNSW32,Flat, IVF3000,Flat, IVF3000 HNSW32,Flat, IVF3000 HNSW32,Flat, IVF3000 HNSW32,Flat, IVF3000 HNSW32,Flat, IVF3 IVF and HNSW were chosen because they fit the distance metric, which is cosine similarity. The result shows the higher the Nprobe value, the higher the recall@R test results. However, with the higher Nprobe value, the time needed to predict recommendations will be longer. Therefore, the purpose of this section is to determine the optimized Nprobe. The optimal Nprobe is realized

in the area of 6-7 percent, as can be seen from the graph. Because the results of recall <6% still indicate a considerable rise in recall value, meanwhile, an Nprobe value of >7% suggests that the recall value tends to stay the same or grow only with an unsignificant value.

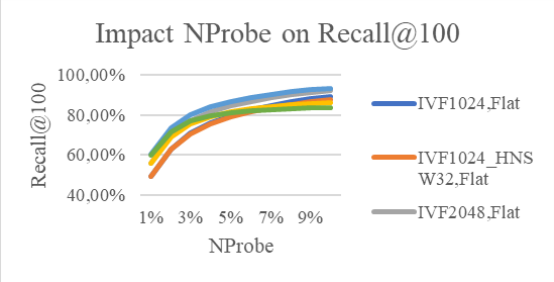

Figure 4. Impact of NProbe on Recall

4.4.2 Recall@R

Based on the result from the effect of NProbe, this research used 6% as a NProbe input. In figure 5, all the indexes tested have a fairly high recall rate where the recall value is above 82% when k is 100. However, compared to all the indexes that have been created, the IDMap, Flat index has the highest recall rate, compared to other indices with recall@1, recall@10, recall@50 and recall@100 reach more than 98%, followed by index IVF3000,Flat, IVF2048,Flat, IVF1024,Flat, IVF2048_HNSW32,Flat, IVF1024_HNSW32,Flat, and IVF3000_HNSW32,Flat.

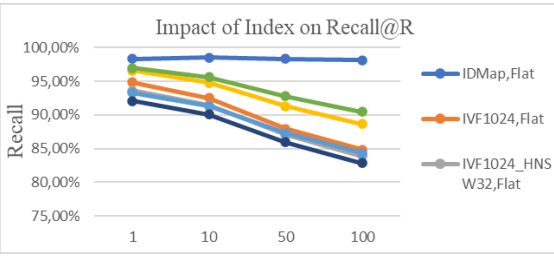

Figure 5. Impact of Index on Recall

4.4.3 Prediction Latency

The next assessment is a comparison of the prediction speed of each index created by the Approximate Nearest Neighbor algorithm. These results, as shown in table 1, contrast with the prior recall and accuracy values that were evaluated. The actual result is the sorted original Nearest Neighbor algorithm result based on the closest distance. When compared with other indices, the prediction time of the actual result is significantly different compared to IVF1024_HNSW32,Flat which has the fastest average prediction time, while the actual result is 1211 times slower.

Table 1: The Impact of Index on Prediction Speed in Millisecond

Index Name	Max(ms)	Min(ms)	Avg(ms)
Actual result	247,89490	166,26800	173,3857
IDMap,Flat	14,91213	1,191854	1,320133
IVF1024,Flat	10,44345	0,125647	0,163683
IVF1024_HNSW32,Flat	11,97028	0,106812	0,143052
IVF2048,Flat	11,83796	0,175476	0,238814
IVF2048_HNSW32,Flat	6,761198	0,100613	0,158663
IVF3000,Flat	9,018421	0,235081	0,27671
IVF3000_HNSW32,Flat	8,912325	0,095129	0,196271

4.5 Result

Five product inputs are used in this section within five different categories such as education book, electronic, health, men's fashion, and women's fashion. The product inputs are science book, HP cartridge, cataflam medicine, cotton t-shirt, and negligee following five categories respectively. Mostly, the results show similar products with the product input.

4.5.1 Education Book

The first category represents education book products. In this example, a science book is the product input. After data processing and modelling using Approximate Nearest Neighbor, the result shows all recommendations are within the same category i.e., education book. The result can be found in figure 6.

4.5.2 Health

The second category used in this paper is health which uses cataflam medicine as a product input. Cataflam is an anti-inflammatory drug to reduce pain and inflammation. The recommendations are within the same health category with five anti-inflammatory products, similar to product input, while there is an anti-allergic medicine called cetirizine. Limited datasets can cause these results.

4.5.3 Electronic

The third category uses a cartridge for HP printer as the product input. The result represents the same category i.e., electronic with the same product i.e., HP cartridge. The 3 recommendations have the same version with the product input (HP932XL) and other recommendations shows

different versions of HP cartridges. The process of data pre-processing filters numerical data and only use text data as the product input. The result can be seen in figure 8.

4.5.4 Men's Fashion

The fourth category uses cotton t-shirts as a product input and the results show the same cotton t-shirt products.

4.5.5 Women's Fashion

The last category utilizes pajamas as a product input, and the findings reveal that all the suggestions are for women's pajamas. It indicates that the advice is effective in the women's fashion area.

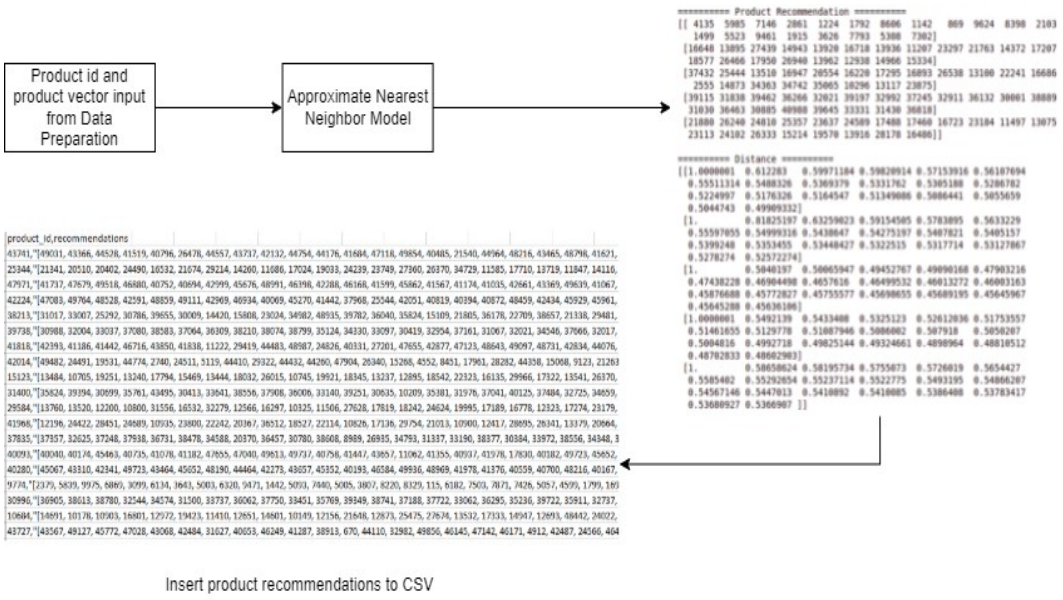

Figure 3. Output Example to Predict Recommendations from Product Vector

Figure 6. Product Input and Recommendations for Education Book Category

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 7. Product Input and Recommendations for Health Category

Figure 8. Product Input and Recommendations for Health Category

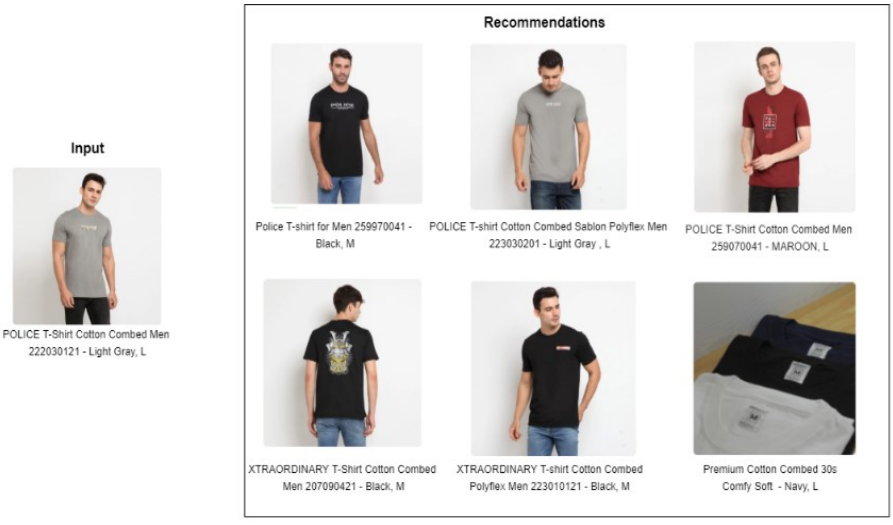

Figure 9. Product Input and Recommendations for Men's Fashion Category

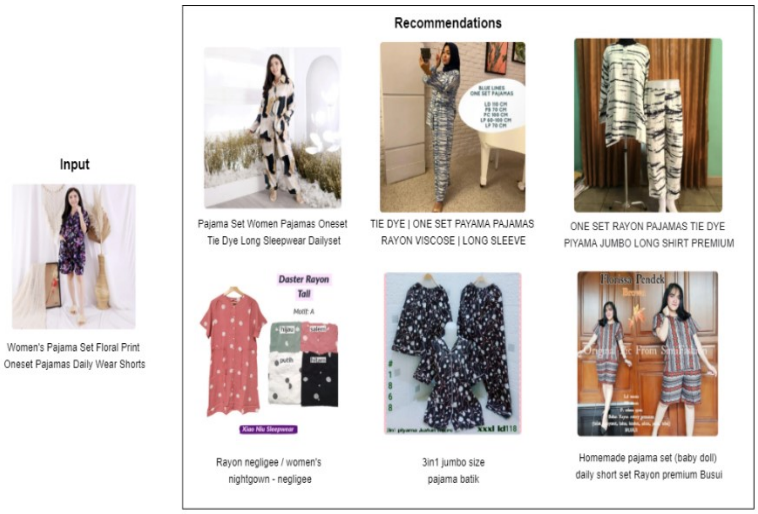

Figure 10. Product Input and Recommendations for Women's Fashion Category

5. CONCLUSION

The recommendation system is one of the important features in E-Commerce to increase user traffic. One of the suitable algorithms is the Approximate Nearest Neighbor, through which the users will get recommendations similar to the input product or the product being viewed by the user. ANN can give recommendations even the product is new or did not have any interaction with user, means the algorithm can be utilized to solve the cold start issue.

The results shows the use of FAISS tools with the Approximate Nearest Neighbor algorithm plays an important role in getting recommendations faster than KNN by only reducing a small level of performance, where Flat has a fairly high recall value of 98.13%, by using the IDMap,Flat index or if it is lowered slightly by using the IVF3000 index,Flat which has a recall value of 90.46% but has a very high prediction speed, which is below 0.3 milliseconds on the average. In addition, the results also show the optimized NProbe value is 6-7%, because increasing the Nprobe value will increase the recall results, but the prediction speed will also increase. While the NProbe value above 7% indicates an insignificant increase in the recall value. The index is determined mostly by the requirements, particularly the amount of recall and forecast speed. In this study, the index with the maximum recall@R are IDMap,Flat. While the index that has the fastest prediction speed is IVF1024_HNSW32,Flat.

The limitations of this research are data limitations caused by web scrapping data collection and some indexes cannot be used because of distance cosine similarity limitation. For future research, the machine learning algorithm Approximate Nearest Neighbors can be used in other attributes such as image similarity search or use of text similarity using other distances such as Euclidean distance and Inner Product.

REFERENCES

[1] A. Berisha-Shaqiri and M. Berisha-Namani, "Information Technology and the Digital Economy," *Mediterr. J. Soc. Sci.*, 2015, doi: 10.5901/mjss.2015.v6n6p78.

[2] Tim APJII, "Penetrasi Internet Indonesia 2018," 2019.

[3] T. Mauritsius, S. Alatas, F. Binsar, R. Jayadi, and N. Legowo, "Promo abuse modeling in e-commerce using machine learning approach," 2020, doi: 10.1109/ICOT51877.2020.9468744.

[4] Monavia Ayu Rizaty, "Transaksi E-Commerce Indonesia Diproyeksikan Capai Rp 403 Triliun pada 2021," *Databoks*, 2021.

[5] R. Franedya, "Tokopedia Cs Terlalu Dominan, 20 e-Commerce Gulung Tikar," *CNBC Indonesia*, 2019. https://www.cnbcindonesia.com/tech/20190513133753-37-72085/tokopedia-cs-terlalu-

dominan-20-e-commerce-gulung-tikar.

[6] B. Morgan, "How Amazon Has Reorganized Around Artificial Intelligence And Machine Learning," 2018. https://www.forbes.com/sites/blakemorgan/2018/07/16/how-amazon-has-re-organized-around-artificial-intelligence-and-machine-learning/?sh=75d836f27361.

[7] K. Wang, T. Zhang, T. Xue, Y. Lu, and S. G. Na, "E-commerce personalized recommendation analysis by deeply-learned clustering," *J. Vis. Commun. Image Represent.*, 2020, doi: 10.1016/j.jvcir.2019.102735.

[8] S. An *et al.*, "Quarter-Point Product Quantization for approximate nearest neighbor search," *Pattern Recognit. Lett.*, 2019, doi: 10.1016/j.patrec.2019.04.017.

[9] C. S. D. Prasetya, "Sistem Rekomendasi Pada E-Commerce Menggunakan K-Nearest Neighbor," *J. Teknol. Inf. dan Ilmu Komput.*, 2017, doi: 10.25126/jtiik.201743392.

[10] D. Kim, D. Seo, S. Cho, and P. Kang, "Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec," *Inf. Sci. (Ny).*, 2019, doi: 10.1016/j.ins.2018.10.006.

[11] Z. Pan, L. Wang, Y. Wang, and Y. Liu, "Product quantization with dual codebooks for approximate nearest neighbor search," *Neurocomputing*, 2020, doi: 10.1016/j.neucom.2020.03.016.

[12] D. Danopoulos, C. Kachris, and D. Soudris, "Approximate similarity search with FAISS framework using FPGAs on the cloud," 2019, doi: 10.1007/978-3-030-27562-4_27.

[13] Y. A. Malkov, "Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.

[14] Z. Geng, Y. Li, Y. Han, and Q. Zhu, "A novel self-organizing cosine similarity learning network: An application to production prediction of petrochemical systems," *Energy*, vol. 142, 2018, doi: 10.1016/j.energy.2017.10.017.

[15] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, "DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model," 2019, doi: 10.1016/j.procir.2019.02.106.

[16] T. Mauritsius, A. S. Braza, and Fransisca, "Bank marketing data mining using CRISP-DM approach," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 5, 2019, doi: 10.30534/ijatcse/2019/71852019.

[17] C. Schröer, F. Kruse, J. Marx, F. Kruse, and J. Marx, "A Systematic Literature Review on Applying Process Model on Applying CRISP-DM Process Model," *Procedia Comput. Sci.*, vol. 181, no. 2019, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.