Bulletin of Electrical Engineering and Informatics
Vol. 9, No. 6, December 2020, pp. 2492~2498
ISSN: 2302-9285, DOI: 10.11591/eei.v9i6.2018

Marketplace affiliates potential analysis using cosine similarity and vision-based page segmentation

Wildan Budiawan Zulfikar[1], Mohamad Irfan[2], Muhammad Ghufron[3], Jumadi[4], Esa Firmansyah[5]
[1,3]Department of Informatics, UIN Sunan Gunung Djati, Indonesia
[2]Department of ICT, Asia E University, Malaysia
[4]School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Indonesia
[5]Department of Informatics, STMIK Sumedang, Indonesia

Article Info

Article history:

Received Aug 15, 2019
Revised Jan 28, 2020
Accepted Mar 1, 2020

Keywords:

Cosine similarity
Marketplace affiliates
Page segmentation
Vision
Web scraping

ABSTRACT

One success factor of an online affiliate is determined by the quality of the content source. Therefore, affiliate marketplaces need to do an objective assessment to retrieve content data that will be used to choose the right product in the appropriate product filter. Usually, the selection is not made using a good and measured system so that the selection of product content is only based on parts that are not in accordance with what is seen or subjective. However, if analyzed using a good and measurable system will produce an objective product content and can have a positive impact on users because the selection is based on factual data. The purpose of this research is to analyze the potential of the affiliate marketplace by combining cosine similarity with vision-based page segmentation. This is a new breakthrough made for optimization to get the best content in accordance with the required criteria. This work will produce a number of product recommendations that are appropriate for publication and then made use of for comparison that matches the required criteria. At the limited evaluation stage, the performance of the proposed model obtained satisfactory results, in which 5 queries tested were all as expected.

Corresponding Author:

Wildan Budiawan Zulfikar,
Department of Informatics,
UIN Sunan Gunung Djati,
105th A.H. Nasution Street, Bandung, 40614, Indonesia.
Email: wildan.b@uinsgd.ac.id

1. INTRODUCTION

Nowadays, information technology has created new types and business opportunities where more and more business transactions are being made online. Therefore, everyone might easily carry out buying and selling transactions [1-3]. Many companies try to offer a variety of products using this media [4, 5]. One of the benefits of the existence of the internet is as a media promotion of a product. A product that is online via the internet can bring huge benefits to entrepreneurs because the product is known throughout the world [4, 6].

Web scraping is the process of extracting information from a website. Web scraping is an alternative way that chose because the required data is not always available in the API, another source like shared database or data warehouse, or even they do not provide the API at all [7-12]. This research has used product attribute data obtained from several marketplace affiliates using web scraping techniques. It used one of the web scraping methods, vision-based page segmentation. Vision-based page Segmentation is an algorithm for website page metadata. Based on previous research, this method of extracting tag tree data can detect content

structures quickly [13, 14]. It transforms the deep web into a visual tree [13, 15]. The result is divided into several segments and can be processed using DOM parser before it can finally be processed and modeled [13].

In addition, the proposed model applies Cosine Similarity and TF-IDF. Cosine-Similarity is one algorithm that functions to compare similarities between documents. In this case, what is compared is a query with a training document [16-18]. In calculating cosine similarity, first, do a scalar multiplication between the query and the document then add up, then do the multiplication between the length of the document and the length of the query that has been squared, after that the square root count is calculated [16, 19-23]. Furthermore, the results of the scalar multiplication are divided by the results of the multiplication of the length of the document and query.

2. RESEARCH METHOD

In this article, it will be explained that the existing attribute data is sourced from some marketplace data. Data sources are taken directly from the original website. Online marketplace data taken is product data that is still active in the product category list. Detailed marketplace affiliate data used in this work is described in Table 1.

Table 1. List of marketplace affiliate

Marketplace	URL	Role
Tokopedia	https://www.tokopedia.com	Main marketplace
Bukalapak	https://www.bukalapak.com	2nd marketplace
Blanja	https://www.blanja.com	3rd marketplace
Lazada	https://www.lazada.co.id	4th marketplace

In this work, the main marketplace is Tokopedia. Then, one product will compare to another marketplaces. The use of this method is divided into two processes namely the first process will be scraping product data based on all selected web marketplace data. This method uses the id category and name category attributes of each product. When the process of web scraping, product data will be divided into several categories that will be done using the cosine similarity and vision-based page segmentation methods. After the data is formed in the form of HTML dom, the system will determine one of the data used to do the process to display the data. The category becomes one of the data used as a reference for this data filtering process because it shows the level of each product data based on the category and is appropriate in retrieving accurate data and filters in the price and rating order. Product availability is the second factor because it supports product competency.

2.1. Cosine similarity

The following is a simulation or example of data used in the process. Category data can be seen in Table 2. Conditions are adjusted to each category which in this case is limited to 6 categories. The product attributes that will be analyzed in the work in detail can be seen in Table 3.

Table 2. Product categories

Categories	Code
Fashion	Cat_001
Health	Cat_002
Beauty	Cat_003
Smartphone and tablet	Cat_004
Laptop	Cat_005
Computer	Cat_006

Table 3. Product attributes

Attributes	Code
Name	prod_name
ID	prod_id
SKU	prod_sku
Link	prod_link
Figure	prod_fig
Price	prod_price
Category code	prod_cat_id
Category description	prod_cat_name
Advertiser	prod_ads

Table 4 is the query data that will be calculated using TF-IDF based on a specific query. This work involves six queries and four affiliate marketplaces as explained in Table 4. Figure 1 is a visualization of Table 4 to make it easier to read valid data and the same or almost the same then the table is converted into a graph diagram. Pictures from the graph diagram of the query and matching with each of the place list lists can be seen in the Figure 1.

Tabel 4. TF-IDF

Query	Marketplace 1 (D1)	Marketplace 2 (D2)	Marketplace 3 (D3)	Marketplace 4 (D4)
Galaxy s7	1	1	0	0
Samsung	1	1	1	1
iPhone X 128Gb Black	1	0	0	0
Galaxy+S7	1	1	1	1
MacBook Air	1	0	1	1
Keyboard Razer	1	0	1	0

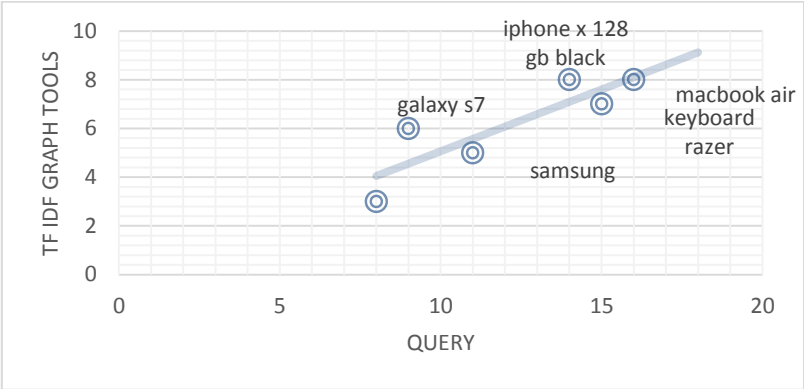

Figure 1. TF-IDF data on graph

The first stage of vision-based page segmentation is to determine the initial weight of each query manually. For example, the first weight is filled by Samsung's query and the second group is weighted by Galaxy. Then obtained: Centroid 1=0.3 and Centroid 2=0.3 as explained in Table 5 and the visualization explained in Figure 2.

Table 5. TF-IDF data and weighting

Code	D1	D2	Description
Oppo	0	0	
Lenovo	0	0	
Samsung	0.3	0	Weight 1
Asus	0	0	
Galaxy	0	0.3	Weight 2
Iphone	0	0	

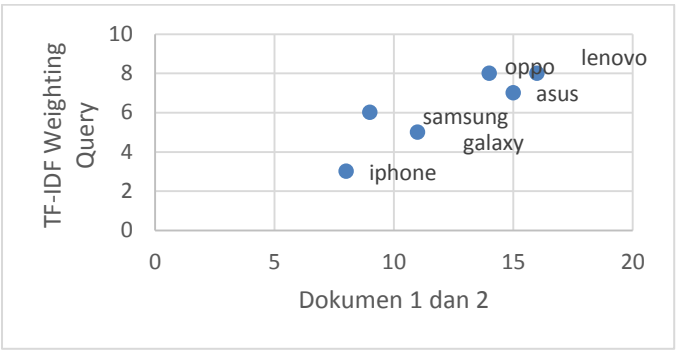

Figure 2. TF-IDF data in a graph diagram with weighted queries

Next calculate the distance of each data with each weight using (1) [24, 25]:

$$idf = Log10(\frac{N}{df1}) \quad (1)$$

The next work to calculate the weight by comparing query data 1 with each query taken that has weight. The query data weighting can be seen in Table 6. Then, look for the average of each weight value to be used as a new query weight namely:

W1 New: (0.3, 0.3)
W2 New: (0.0, 0.0)

This step will continue to be repeated until the conditions are met. The desired condition is that there is no change in the weighting of the data source which means there is no difference between the data query and the query in the previous iteration. Then the second iteration will be performed using a new weighting.

In this experiment, the algorithm will be completed in the third iteration. The final results are presented in the form of a Cartesian diagram to make it easier to see the closeness of the data between the weighting and each data as explained on Table 6. Table 6 explains the list of Queries included in the category. Products that are in the first weighting are Q3, Q5, Q6 and in the second weighting are Q1, Q2, Q4.

Table 6. Clustering results

W1	W2
Q3	Q1
Q5	Q2
Q6	Q4

2.2. Vision-based page segmentation

Query retrieved adjusted to the query that has been selected. The higher the weighting value of the selected query the higher the query used and conversely the lower the weighting of the query the lower weighting of the query is used. Next, calculate the normalization of data according to the vision-based page segmentation formula then multiplied by the weights that have been determined at the initialization stage namely (2) for profit and (3) for costs:

$$r_{ij} = \left\{ \frac{x_{ij}}{\underset{i}{Max}\, x_{ij}} \right\} \quad (2)$$

$$r_{ij} = \left\{ \frac{\underset{i}{Min}\, x_{ij}}{x_{ij}} \right\} \quad (3)$$

3. RESULTS AND DISCUSSION

After conducting the previous weighting phase, product data search will be performed based on the Vision-based Page Segmentation algorithm. Based on some data that has been classified, only one product data will be taken that matches the previous TF-IDF process and to be compared using this algorithm. The query groups to be selected are based on the query to be searched. If the query is appropriate, then the appropriate group will be taken, and while the position is not appropriate, page 404 or product page will not be selected. In this case, the first query is taken that is Samsung Galaxy. The following is the use of the vision-based page segmentation algorithm as described in the previous chapter:

a. First step is determining new product data. Figure 3 describe the default position of product detail including name, dimension, price, and any related data of product.

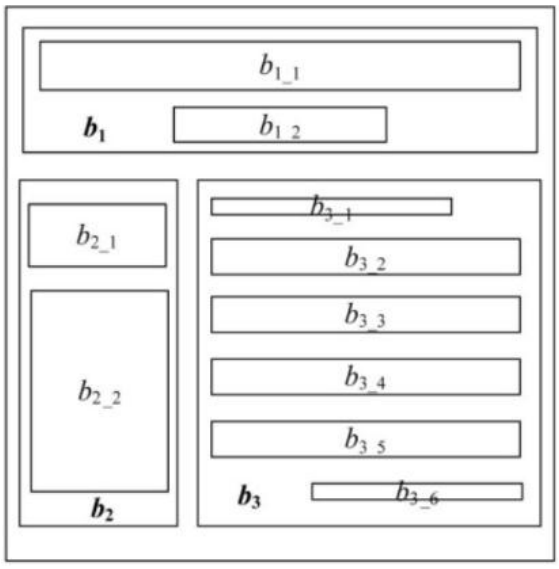

Figure 3. Scheme design vision-based page segmentation product scraping data

b. Then, each product attributes parse by its category and subcategory as describe on Figure 4.

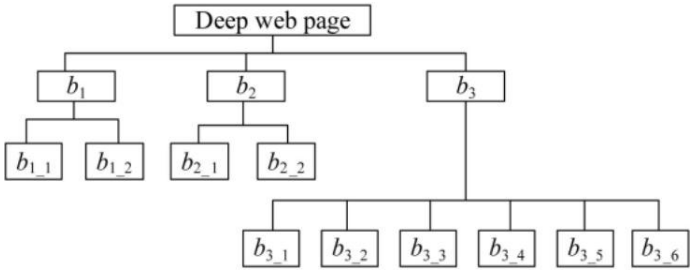

Figure 4. Web process data page product data scraping

c. Extract data that will be searched using (4).

$$TSV = w(1) * r1R \tag{4}$$

Where w(1) is the query data that is input with the total weight that will be searched from the weight value of R and r is the number of segments related to the query that already has a value.

d. Validation of data that has been processed and normalize data according to (5).

$$TSV = \left\{ \frac{\underset{i}{Min} x_{ij}}{x_{ij}} \right\} \tag{5}$$

The results of data normalization using are explained in Table 7.

Table 7. The result of data normalization

No.	Kode	Div.Id	Div.Attr	Div.Class	Cache	Alpha result
1	Q1	0.875	1	0.8	1	0.5
2	Q2	1	1	0.8	0.954545455	1
3	Q4	0.9375	0.875	1	0.909090909	1

e. Normalization results are multiplied by the weights and summed to find out the final result of the preference value with (6) and final preference result explained in Table 8.

$$V_I = \sum_{j=1}^{n} w_j r_{ij} \tag{6}$$

Table 8. Final preference results

No.	Kode	Div.Id	Div.Attr	Div.Class	Cache	Alpha result	Result
W		0.3	0.2	0.2	0.15	0.15	
1	Q1	0.875	1	0.8	1	0.5	0.8475
2	Q2	1	1	0.8	0.954545455	1	0.953181818
3	Q4	0.9375	0.875	1	0.909090909	1	0.942613636

Based on Table 8, it can be concluded that the most recommended Query data is the Query data with the Q2 code. Query data Q2 gets 0.95 results and is only 0.01 points different from Q4.

4. CONCLUSION

If evaluated from performance, the proposed model gets the appropriate results. 5 queries tested everything as expected. The cosine similarity algorithm successfully improvised the vision-based page segmentation algorithm and was able to adjust product 1 to other products that were eligible to be selected by searching product data processed by TF-IDF. Further work, we suggest comparing this model with different methods.

REFERENCES

[1] Y. U. Chandra, S. Karya, and M. Hendrawaty, "Decision support systems for customer to buy products with an integration of reviews and comments from marketplace e-commerce sites in Indonesia: A proposed model," *International Journal Advamced Science Engineering Information Technology.*, vol. 9, no. 4, pp. 1171-1176, 2019.

[2] I. O. Sfenrianto, A. Christiano, and M. P. Mulani, "Impact of e-service on customer loyalty in marketplace in Indonesia," *Journal of Theoretical and Applied Information Technology*, vol. 96, no. 20, pp. 6795-6805, 2018.

[3] G. J. A. Santoso and T. A. Napitupulu, "Factors affecting seller loyalty in business e-marketplace : A case of Indonesia," *Journal of Theoretical and Applied Information Technology*, vol. 96, no. 1, pp. 162-171,·Jan. 2018.

[4] M. A. Fauzan, A. S. Nisafani, and A. Wibisono, "Seller reputation impact on sales performance in public e-marketplace Bukalapak," *TELKOMNIKA Telecommunication Computing Electronic and Control*, vol. 17, no. 4, pp. 1810-1817, Aug. 2019.

[5] H. Yoganarasimhan, "The value of reputation in an online freelance marketplace," *Marketing Science*, vol. 32, no. 6, pp. 860-891, Nov. 2013.

[6] M. N. Alraja and M. A. Said Kashoob, "Transformation to electronic purchasing: an empirical investigation," *TELKOMNIKA Telecommunication Computing Electronic and Control*, vol. 17, no. 3, pp. 1209-1219, June 2019.

[7] S. K. Malik and S. Rizvi, "Information extraction using web usage mining, web scrapping and semantic annotation," *2011 International Conference on Computational Intelligence and Communication Networks*, Gwalior, pp. 465-469, 2011.

[8] K. Sriraghav, S. Jayanthi, N. Vidya, and V. S. Felix Enigo, "ScrAnViz-A tool to scrap, analyze and visualize unstructured-data using attribute-based opinion mining algorithm," *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*, Vellore, pp. 1-5, 2017.

[9] R. Murali, "An intelligent web spider for online e-commerce data extraction," *2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT)*, Bangalore, India, pp. 332-339, 2018.

[10] S. Mehak, R. Zafar, S. Aslam, and S. M. Bhatti, "Exploiting filtering approach with web scrapping for smart online shopping: Penny wise: A wise tool for online shopping," *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, Sukkur, Pakistan, pp. 1-5, 2019.

[11] Đ. Petrović and I. Stanišević, "Web scrapping and storing data in a database, a case study of the used cars market," *2017 25th Telecommunication Forum (TELFOR)*, Belgrade, pp. 1-4, 2017.

[12] J. G. Thomsen, E. Ernst, C. Brabrand, and M. Schwartzbach, "WebSelF: A web scraping framework," *International Conference on Web Engineering 2012, Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, vol. 7387, pp. 347-361, 2012.

[13] M. Cormer, R. Mann, K. Moffatt, and R. Cohen, "Towards an improved vision-based web page segmentation algorithm," *2017 14th Conference on Computer and Robot Vision (CRV)*, Edmonton, AB, pp. 345-352, 2017.

[14] A. Bhardwaj and V. Mangat, "A novel approach for content extraction from web pages," *2014 Recent Advances in Engineering and Computational Sciences (RAECS)*, Chandigarh, pp. 1-4, 2014.

[15] P. Ko, S. Kang, and H. Kumar, "Web page dependent vision based segementation for web sites," *Seventh IEEE/ACIS International Conference on Computer and Information Science (ICIS 2008)*, Portland, OR, pp. 690-694, 2008.

[16] D. Xue and Y. Wang, "Applying cosine similarity to discount evidence," *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, Hangzhou, pp. 516-519, 2017.

[17] X. Wang, Z. Xu, X. Xia, and C. Mao, "Computing user similarity by combining SimRank++ and cosine similarities to improve collaborative filtering," *2017 14th Web Information Systems and Applications Conference (WISA)*, Liuzhou, pp. 205-210, 2017.

[18] P. P. Gokul, B. K. Akhil, and K. K. M. Shiva, "Sentence similarity detection in Malayalam language using cosine similarity," *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, Bangalore, pp. 221-225, 2017.

[19] M. Alodadi and V. P. Janeja, "Similarity in patient support forums using TF-IDF and cosine similarity metrics," *2015 International Conference on Healthcare Informatics*, Dallas, TX, pp. 521-522, 2015.

[20] Hilary I. Okagbue, Sheila A. Bishop, Pelumi E. Oguntunde, Patience I. Adamu, Abiodun A. Opanuga, and Elvir M. Akhmetshin, "Modified CiteScore metric for reducing the effect of self-citations," *TELKOMNIKA Telecommunication Computing Electronic and Control*, vol. 17, no. 6, pp. 3044-3049, Dec. 2019.

[21] A. Hamdy and M. Elsayed, "Towards more accurate automatic recommendation of software design patterns," *Journal of Theoretical & Applied Information Technology*, vol. 96, no. 15, pp. 5069-5079, 2018.

[22] D. Jayasri and D. D. Manimegalai, "An efficient cross ontology-based similarity measure for bio-document retrieval system," *Journal of Theoretical & Applied Information Technology*, vol. 54, no. 2, pp. 245-258, 2013.

[23] Y. Kawada, "Cosine similarity and the Borda rule," *Social Choice and Welfare*, vol. 51, no. 1, pp. 1-11, Jun. 2018.

[24] W. Uther, "TF–IDF," in C. Sammut and G. I. Webb (Ed.), *Encyclopedia of Machine Learning*, Boston, MA: Springer US, pp. 986-987, 2011.

[25] Li-Ping Jing, Hou-Kuan Huang, and Hong-Bo Shi, "Improved feature selection approach TFIDF in text mining," *Proceedings. International Conference on Machine Learning and Cybernetics*, Beijing, vol. 2, pp. 944-946, 2002.

BIOGRAPHIES OF AUTHORS

Wildan Budiawan Zulfikar received the B.Eng degree from UIN Sunan Gunung Djati, Indonesia, and M.Cs from STMIK LIKMI, Indonesia. He currently a lecturer at UIN Sunan Gunung Djati, Indonesia. His research area is in Information System and Data Mining.

Mohamad Irfan received the B.Eng degree from UIN Sunan Gunung Djati, Indonesia, and M.Cs from STMIK LIKMI, Indonesia. He currently a Ph.D student of Asia E University, Malaysia. His research focused on Information System.

Muhammad Ghufron received the B.Eng degree from UIN Sunan Gunung Djati, Indonesia. He currently a research assistant of Informatics Department. His research interest is Business Information System.

Jumadi received the B.Eng degree from Dharma Negara Business and Infromatics School, Indonesia and M.Cs from Universitas Gajah Mada, Indonesia. He currently a Ph.D student of School of Electrical Engineering and Informatics, Institute Teknologi Bandung, Indonesia. His research focussed on Semantics Information Retrieval.

Esa Firmansyah received the B.Eng degree from STMIK PMBI Bandung, and M.Kom from STTIBI Jakarta, Indonesia. He currently a Ph.D student of Asia E University, Malaysia. His research focussed on Information Technology & Information System.