Indonesian Journal of Electrical Engineering and Computer Science
Vol. 22, No. 2, May 2021, pp. 780~786
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v22.i2.pp780-786

Network analysis of YouTube videos based on keyword search with graph centrality approach

Edi Surya Negara, Ria Andryani, Riyan Amanda
Data Science Interdisciplinary Research Center, Information System Departement, Universitas Bina Darma, Indonesia

Article Info

Article history:

Received Oct 18, 2020
Revised Jan 6, 2021
Accepted Feb 25, 2021

Keywords:

Degree centrality
Graph
Scrapping
Social network analytics
Youtube

ABSTRACT

YouTube is a social media that has billions of users, with this can be used as a promotional media, trends, business, and so forth. This study aims to analyze the correlation between YouTube videos by utilizing hashtags on video using graph theory. Data collection in this study uses scraping techniques taken from the YouTube website in the form of links, titles, keywords, and hashtags. The method used in this research is social network analysis, the measurements used in this study are degree centrality and betweenness centrality. The results of this study indicate that the most popular hashtags with the keyword search for "viruses" are #KidflixPT, #Portugues, and #Mondo with degree centrality values equal to 0.071875. and the correlation between the most closely related videos about #Coronavirus with a value of betweenness centrality of 0.082626.

Corresponding Author:

Edi Surya Negara
Data Science Interdisciplinary Research Center
Information System Department
Universitas Bina Darma, Indonesia
Email: e.s.negara@binadarma.ac.id

1. INTRODUCTION

Social media is a technology that is currently closely related to human life, it is inconceivable that one day the world will lose all of its technology. Social media is used for various purposes such as promotions, business, hospitality, and so on. Social media that is popular right now is YouTube which is a web-based video sharing social media. Many people use YouTube to share videos on various topics such as news, activities (vlogs), tutorials, and so on. With billions of users, YouTube has a very large data set. Therefore, by analyzing the data set can find new correlations [1], [2]. Many methods for analyzing social media data sets, like text analytics, topic modeling and including social network analysis (SNA) [3]-[5].

SNA is a methodology that emerged based on graph theory [6]. In addition, SNA has been successfully applied to many research domains [6]-[8]. YouTube is a free platform from Google that has billions of users and people usually watch hundreds of millions of hours on YouTube and generate billions of viewers [9], [10]. This can be used by content creators to find out the right trends, behavior, and #hashtags in promoting their videos.

In this study, the authors analyze YouTube data to find out the correlation between videos by utilizing the #hashtag in the video using graph theory. In addition to knowing the correlation between videos, in the future, it is expected to find out the #hashtags that are most popular related to certain video topics. Data retrieved using the scrapping technique with the python programming language. The attributes used in this study are links, titles, keywords, and # hashtags. After the data is obtained, researchers can use it for calculations, explanations, and simulations [11], [12]. The measures used in this study are the degree centrality and betweenness centrality.

This study proposes the use of the hashtag modeling approach to the network on the YouTube site to find the relationship between the hashtag and determine the most popular hashtag on the YouTube site with the network analysis approach. This approach is a new approach to determine the relationship between videos and determine the most popular videos using keywords and hashtags.

2. LITERATURE REVIEW

2.1. Graph

A graph is a mathematical approach which is the main approach in SNA. Graph theory is derived from a mathematical investigation carried out by Euler and provides a method for studying all types of networks [13], [14]. Informally, a graph is a group of objects called nodes connected by edges. Usually, graphs are described as a set of points connected by lines [15].

There are two ways to describe relationships in a graph, namely directed graph and bonded-tie graph. A directed graph is a collection of nodes that are connected by a direct line, while a bonded-tie graph is a collection of nodes that are connected to a directionless line. A simple graph example can be seen in Figures 1 and 2.

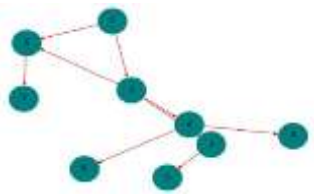

Figure 1. Simple graph

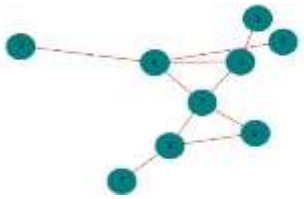

Figure 2. Directed graph

2.2. Social network analytics

Social network analysis (SNA) can be defined as a study of human relations through graph theory [16]. Through graph theory, SNA can examine the structure of social relations within a particular group to reveal informal relationships between individuals. It was developed to understand the interaction of actors in the system with 2 focus in certain social contexts [17]. The typical task of SNA involves identifying the most influential, prestigious or central actors, using statistical measures [18]. There are several variations in SNA, one of which is centrality [19], [20]. There are four ways to measure centrality, namely by degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality. In this study, only degree centrality and betweenness centrality will be used.

2.3. Degree centrality

Degree Centrality is based on the idea that important nodes are the nodes that have the largest number of ties to other nodes in a graph [21]. Degree centrality is measured by the total number of direct connections with other nodes [21], estimating the number of interactions made by a node [21]. To calculate the degree centrality value of the nodes in the graph can be done using the formula in (1) [22].

$$C\boldsymbol{D}(\boldsymbol{ni}) = \boldsymbol{d}(\boldsymbol{ni}) \quad (1)$$

Where d (ni) is the number of interactions possessed by this node with other nodes in the graph.

2.4. Betweenness centrality

Betweenness centrality is used to measure a node which is the connecting role in the graph. If a node is the only path traveled by other nodes, such as communication, connections, transportation or transactions, then the node has an important role and could have a high centrality betweenness [23]. Nodes that have the highest betweenness centrality are nodes that have the role of being the best link between nodes in the graph [23]. To calculate the value of betweenness centrality can be done with the formula in (2) [22].

$$CB(ni) = \sum gjk(ni) \,/\, gjk \quad (2)$$

Where gjk (ni) is the number of shortest paths from node j to node k that passes through node i. And gjk is the number of shortest paths between 2 (two) nodes in a graph.

3. RESEARCH METHOD

In general, this research goes through several stages, namely data collection, data preprocessing, and graph visualization. The flow of the research process can be seen in Figure 3.

Figure 3. Process flow research stages

3.1. Data collection

In this study, the data was taken from the YouTube website. Data is taken by a web scrapping technique using the python programming language. Web scrapping is a form of data mining that aims to mine information from websites that are different from unstructured and turn them into structured so that they can be understood [24], [25]. Data attributes in this study are link, title, keyword, #hashtag. Data was taken on February 2, 2020, with the keyword "virus" search, and the results of the video search were filtered based on video only and relevance. The HTML anchor tag taken to get the data in this study is "id = video-title", then copied as Xpath which will later be used in a python application to pull data. The results of data retrieval can be seen in Figure 5.

3.2. Data preprocessing

At this stage the data will be cleared by deleting rows with no values, transforming the links attribute to numeric, removing spaces at the beginning and end, deleting all punctuation, changing all letters into lower, and repeating " FOR "to add a new line according to the number of #hashtags owned by the node. An example of repetition on the hashtags attributes as shown in Figure 4, links A has 3 (three) hashtags namely # 1, # 2, # 3, and links B has 2 (two) hashtags namely # 2, # 4. Look at Table 1. After repeating the hashtag attribute, the table will look like Table 2.

Out[3]:

	links	title	keywords	hashtags
0	C1xWXTFZAwA	Coronavirus: Ten passengers on cruise ship tes...	virus	NaN
1	GbJJqgX5u0Y	China reports 73 new deaths from virus outbreak	virus	#China #death #virus
2	SLKlGe99IZU	Vlog: how a community health center in Wuhan i...	virus	NaN
3	ias2lSdvq	NaN	virus	NaN
4	HJ2Ec7pgQIQ	NaN	virus	NaN
...	...	...	...	...
410	qgiML4b8zOM	¿Quien esta detrás del virus de china? aumenta...	virus	NaN
411	TqhCjCvNQR	NaN	virus	NaN
412	XCrOde-JYs0	NaN	virus	NaN
413	QPSq3kSjfg	Qayamat - City Under Threat (HD) - Hindi Movie...	virus	NaN
414	5Et76r7HI	CORONA VIRUS OUTBREAK, SYMPTOMS, TRANSMISSION ...	virus	#CoronaVirus #CoronaVirusSymptoms #CoronaVirus...

415 rows × 4 columns

Figure 4. Raw data

Table 1. Examples of raw data

links	Hashtags
A	#1, #2, #3
B	#2, #4

Table 2. Examples of final data

Links	Hashtags
A	#1
A	#2
A	#3
B	#2
B	#4

After the preprocessing and looping process is finished, then the graph is visualized.

3.3. Graph visualization

At this stage, the data that has gone through the preprocessing stage will be visualized using graphs [26], with the size discussed in the previous sections, namely degree centrality and betweenness centrality. At this stage, you will also find the most influential node in the graph. In this study, the pandas package is used for data analysis and the NetworkX package is for network analysis [27], [28].

4. RESULTS AND DISCUSSION

4.1. Preprocessing

At this stage the data preprocessing will go through several stages, namely deleting rows that have nan values (no values), repeating the hashtags attribute, case folding, and transformation (label encoder) on the links attribute. The encoder label is the process of transforming the values in the links attribute to numeric with values from 0 to n where n is a different value. This is done to make graph visualization better. For this preprocessing step, see Figure 5 which shows the data after pereprocessing.

Out[13]:

	links	titles	keywords	hashtags
0	56	china reports new deaths from virus outbreak	virus	#China
1	56	china reports new deaths from virus outbreak	virus	#death
2	56	china reports new deaths from virus outbreak	virus	#virus
3	48	coronavirus whistleblower doctor is online her...	virus	#CNN
4	48	coronavirus whistleblower doctor is online her...	virus	#CNNI
...	...	...	...	...
406	63	chaos erupts in china m under quarantine as...	virus	#Quarantine
407	63	chaos erupts in china m under quarantine as...	virus	#Epidemic
408	18	corona virus outbreak symptoms transmission ...	virus	#CoronaVirus
409	18	corona virus outbreak symptoms transmission ...	virus	#CoronaVirusSymptoms
410	18	corona virus outbreak symptoms transmission ...	virus	#CoronaVirusTransmission

411 rows × 4 columns

Figure 5. Data after preprocessing

4.2. Graph data visualization

As explained in previous sections that graph is a collection of objects called nodes connected by edges. From the dataset that has been processed in this study, the number of nodes is 321 and the number of edges is 411 with an average degree of 2.5607. The graph form of the dataset that has been processed in this study can be seen in Figure 6. Based on the objective, which is to find out the correlation between videos and the most popular #hashtags based on video topics, the measures to be used are degree centrality and betweenness centrality.

Figure 6. Graph visualization form preprocessing data

4.2.1. Degree centrality

Degree centrality can be used as a measure to find the popularity of a node. The results of the calculation of degree centrality can be seen in Figure 7. From Figure 7, it can be seen that the highest degree of centrality is found in the hashtags #KidflixPT, #Portugues, and #Mondo. It can be assumed that the three (3) hashtags are the ones that have the highest popularity compared to other hashtags. 3 (three) hashtags are also hashtags that have the most relationship (edge) compared to other hashtags. The form of graph degree centrality can be seen in Figure 10.

In Figure 8, the size of the node is based on the degree of degree centrality. The greater the value of the degree centrality, the greater the size of the node. Likewise, with the color of the node, the color of the node represents the degree centrality value of a node.

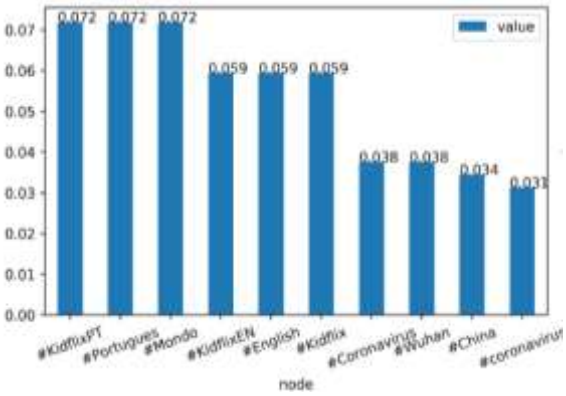

Figure 7. Results of degree centrality calculation

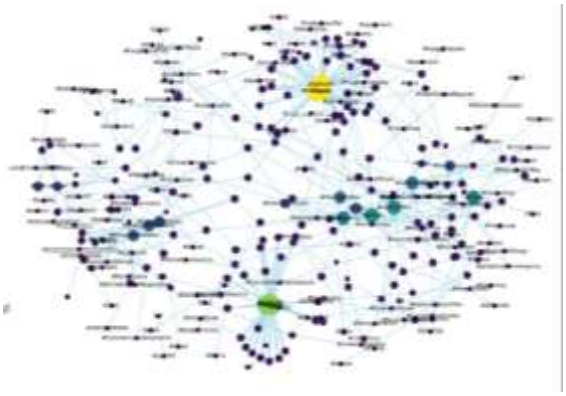

Figure 8. Graph degree centrality

4.2.2. Betweenness centrality

To see the relationship between videos can be used as a measure of betweenness centrality. The results of the calculation of betweenness centrality can be seen in Figure 9. From Figure 9, it can be seen that the highest value of betweenness centrality is found in the #Coronavirus hashtag. It can be assumed that the hashtag which has the best connecting role is #Coronavirus. And it can be assumed that the correlation between videos with the keyword "virus" is the most about "Corona Virus". The form of graph betweenness centrality can be seen in Figure 12.

From Figure 10, it can be seen that the relationship between the video with the keyword "virus" is not only about "Corona Virus" but also has to do with "Wuhan", "China". It can be assumed that the most "Corona Virus" outbreaks occurred in or from the "Wuhan" country of "China". This can be proven by filtering graphs based on related words in the video title as relations between nodes. The results can be seen in Figures 11 and 12.

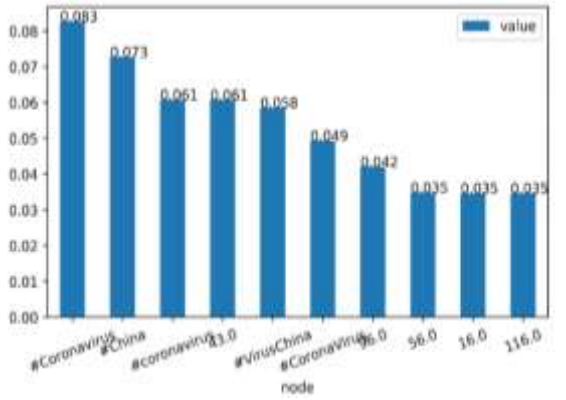

Figure 9. The results of the calculation of betweenness centrality

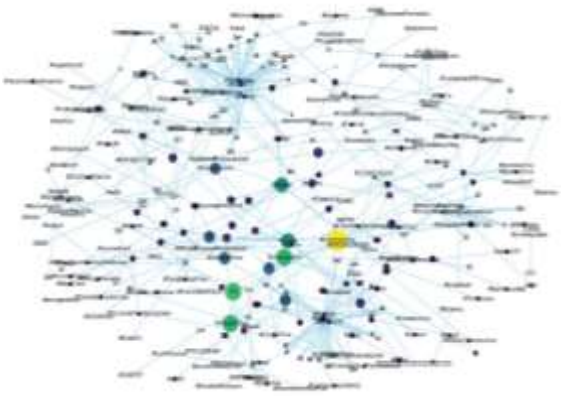

Figure 10. Graph betweenness centrality

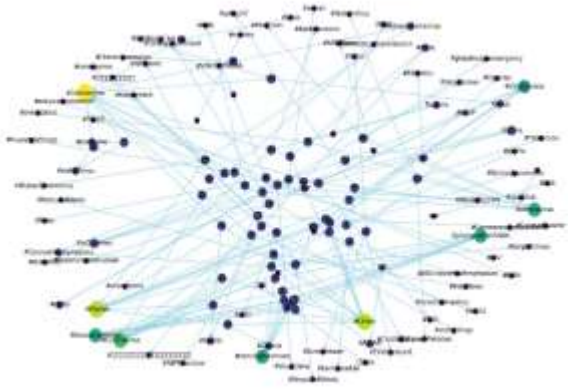

Figure 11. Graph of video title containing the word "corona"

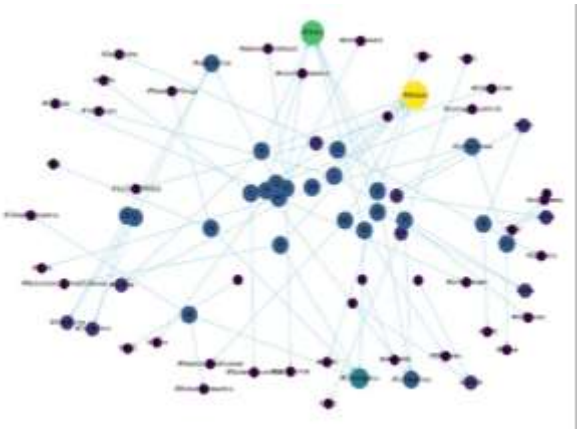

Figure 12. Graph of video title containing the word "China and Wuhan"

5. CONCLUSION

This research was successfully conducted and answered the problem formulation and objectives of this study with a dataset obtained from YouTube based on the keyword search "virus". The closest correlation between videos with the #Coronavirus hashtag with a value of betweenness centrality of 0.082626, and the most popular hashtag is the hashtag #KidflixPT, #Portugues, and #Mondo with a degree of centrality equal of 0.071875. The suggestion that the author can convey for further research is to try to predict the hashtag of the video taken because not all videos have a hashtag.

REFERENCES

[1] Shaila S.G, Prasanna MSM, and K. Mohit, "Classification of YouTube Data based on Sentiment Analysis," *International Journal of Engineering Research in Computer Science and Engineering (IJERCSE*), vol. 5, no. 6, 2018.

[2] Khan, A.U.R., Khan, M. and Khan, M.B., "Naïve Multi-label classification of YouTube comments using comparative opinion mining," *Procedia Computer Science*, vol. 82, pp. 57-64, 2016, doi: 10.1016/j.procs.2016.04.009.

[3] M. Tsvetovat and A. Kouznetsov, "Social network analysis for startups," *O'Reilly Media Inc*, 2011.

[4] T. Sutabri, A. Suryatno, D. Setiadi and E. S. Negara, "Improving Naïve Bayes in Sentiment Analysis For Hotel Industry in Indonesia," *2018 Third International Conference on Informatics and Computing (ICIC)*, Palembang, Indonesia, pp. 1-6, 2018, doi: 10.1109/IAC.2018.8780444.

[5] E. S. Negara, D. Triadi and R. Andryani, "Topic Modelling Twitter Data with Latent Dirichlet Allocation Method," *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, Batam Island, Indonesia, pp. 386-390, 2019, doi: 10.1109/ICECOS47637.2019.8984523.

[6] Tacheva, Z. and Simpson, N, "Social network analysis in humanitarian logistics research," *Journal of Humanitarian Logistics and Supply Chain Management*, vol. 9, no. 3, pp. 492-514, 2019, doi: 10.1108/JHLSCM-06-2018-0047.

[7] C.-S. Wang, I.-H. Ting, and Y.-C. Li, "Taiwan Academic Network Discussion via Social Networks Analysis Perspective," *in 2011 International Conference on Advances in Social Networks Analysis and Mining*, Kaohsiung City, Taiwan, pp. 685-689, 2011. doi: 10.1109/ASONAM.2011.99.

[8] E. S. Negara, D. Kerami, I. M. Wiryana, and T. B. M. Kusuma, "Researchgate Data Analysis To Measure The Strength Of Indonesian Research," *FJEC*, vol. 17, no. 5, pp. 1177-1183, 2017, doi: 10.17654/EC017051177.

[9] Li, T., Lin, L., Choi, M., Fu, K., Gong, S. and Wang, J., "Youtube av 50k: an annotated corpus for comments in autonomous vehicles," *In 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, IEEE, pp. 1-5, 2018.

[10] Chen, Y.L., Chang, C.L. and Yeh, C.S., "Emotion classification of YouTube videos," *Decision Support Systems*, vol. 101, pp.40-50, 2017, doi: 10.1016/j.dss.2017.05.014.

[11] J. Zhang and Y. Luo, "Degree Centrality, Betweenness Centrality, and Closeness Centrality in Social Network," *in Proceedings of the 2017 2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM2017)*, Bankog, Thailand, 2017, doi: 10.2991/msam-17.2017.68.

[12] Yoganarasimhan, H., "Impact of social network structure on content propagation: A study using YouTube data," *Quantitative Marketing and Economics*, vol. 10, no. 1, pp. 111-150, 2012, doi: 10.1007/s11129-011-9105-4.

[13] J. Scott, "Social network analysis: developments, advances, and prospects," *SOCNET*, vol. 1, no. 1, pp. 21-26, 2011, doi: 10.1007/s13278-010-0012-6.

[14] Majeed, A. and Rauf, I., "Graph Theory: A Comprehensive Survey about Graph Theory Applications in Computer Science and Social Networks," *Inventions*, vol. 5, no. 1, p. 10, 2020, doi: 10.3390/inventions5010010.
[15] F. Riaz and K. M. Ali, "Applications of Graph Theory in Computer Science," *in 2011 Third International Conference on Computational Intelligence, Communication Systems and Networks*, Bali, Indonesia, pp. 142-145, 2011, doi: 10.1109/CICSyN.2011.40.
[16] D. F. Brianna, E. Surya Negara, and Y. N. Kunang, "Network Centralization Analysis Approach in the Spread of Hoax News on Social Media," *in 2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, Batam Island, Indonesia, pp. 303-308, 2019, doi: 10.1109/ICECOS47637.2019.8984526.
[17] M. Cordeiro, R. P. Sarmento, P. Brazdil, and J. Gama, "Evolving Networks and Social Network Analysis Methods and Techniques," *in Social Media and Journalism - Trends, Connections, Implications, J. Višňovský and J. Radošinská, Eds. InTech*, vol. 101, no. 2, 2018, doi: 10.5772/intechopen.79041.
[18] L. Tang and H. Liu, "Graph Mining Applications to Social Network Analysis," *in Managing and Mining Graph Data, C. C. Aggarwal and H. Wang, Eds. Boston, MA: Springer* US, vol. 40, pp. 487-513. 2010, doi: 10.1007/978-1-4419-6045-0_16.
[19] V. Latora and M. Marchiori, "A measure of centrality based on network efficiency," *New J. Phys*., vol. 9, no. 6, pp. 188-188, Jun. 2007, doi: 10.1088/1367-2630/9/6/188.
[20] Lin, C.C., Huang, W., Liu, W.Y. and Wu, S.F., "A novel centrality-based method for visual analytics of small-world networks," *Journal of Visualization*, vol. 22, no. 5, pp. 973-990, 2019, doi: 10.1007/s12650-019-00582-5.
[21] Salavati, C., Abdollahpouri, A. and Manbari, Z., "Ranking nodes in complex networks based on local structure and improving closeness centrality," *Neurocomputing*, vol. 336, pp. 36-45, 2019, doi: 10.1016/j.neucom.2018.04.086.
[22] Knoke, D. and Yang, S., "Social network analysis," *Sage Publications*, vol. 154, 2019.
[23] Fan, C., Zeng, L., Ding, Y., Chen, M., Sun, Y. and Liu, Z., "Learning to identify high betweenness centrality nodes from scratch: A novel graph neural network approach," *In Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 559-568, 2019, doi: 10.1145/3357384.3357979.
[24] S. Anand V, P. Kedar G, and G. Shweta A, "An Overview On Web Scraping Techniques And Tools," *IJFRCSCE*, vol. 4, no. 4, pp. 363-367, 2018.
[25] Thomas, D.M. and Mathur, S., "Data analysis by web scraping using python," *In 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, pp. 450-454, 2019, doi: 10.1109/ICECA.2019.8822022.
[26] Linhares, C.D., *et al*., "A scalable node ordering strategy based on community structure for enhanced temporal network visualization," *Computers & Graphics*, vol. 84, pp. 185-198, 2019, doi: 10.1016/j.cag.2019.08.006.
[27] Jimenez-Marquez, J.L., Gonzalez-Carrasco, I., Lopez-Cuadrado, J.L. and Ruiz-Mezcua, B., "Towards a big data framework for analyzing social media content," *International Journal of Information Management*, vol. 44, pp. 1-12, 2019, doi: 10.1016/j.ijinfomgt.2018.09.003.
[28] de Siracusa Pedro, C., MR Jr, G.L. and Artur, Z., "New perspectives on analysing data from biological collections based on social network analytics," *Scientific Reports (Nature Publisher Group)*, vol. 10, no. 1, pp. 1-10, 2020, doi: 10.1038/s41598-020-60134-y.

BIOGRAPHIES OF AUTHORS

Edi Surya Negara has obtained his bachelor's and master of informatics from Universitas Bina Darma and Doctor of Information Technology from Gunadarma University. He has 10 years of teaching and research experience. He published 9 research papers at the international level.

Ria Andryani has obtained her bachelor's and master of informatics from Universitas Bina Darma. She has 12 years of teaching experience. She published 4 research papers at the international level.

Riyan Amanda has obtained his bachelor's and now he is a master student of informatics from Universitas Bina Darma. He published 1 research paper at the international level.