IT Journal Research and Development (ITJRD)
Vol.8, No.2, March 2024, E-ISSN : 2528-4053 | P-ISSN : 2528-4061
DOI : 10.25299/itjrd.2024.13459

Knowledge Graph Analysis Using Graph Embedding Algorithms On English Wikipedia Pages

Yudistira Bagus Pratama[1], Haiyudi[2]
Department of Computer Science, Universitas Muhammadiyah Bangka Belitung[1]
Department of English Education, Universitas Muhammadiyah Bangka Belitung[2]
yudistira.bagus@unmuhbabel.ac.id[1], haiyudi@unmuhbabel.ac.id[2]

Article Info

Article history:

Received Jun 29, 2023
Revised Feb 15, 2024
Accepted Mar 27, 2024

Keyword:

Deep Learning
Data Mining
Community Detection
Wikipedia
Knowledge Graph

ABSTRACT

Analysis of social networks or online communities can be very difficult when working on large networks, as many measurements require expensive hardware. For example, identifying the community structure of a network is a very computationally expensive task. Embedded graph is a way to represent graphs with vectors, so that further analysis becomes easier. The purpose of this research is to analyze the knowledge graph from the wikipedia article data. This research aims to implement web scraping techniques on the wikipedia article search engine and display similar wikipedia pages and analyze them using a predetermined deep learning algorithm. Data collection in this research used scraping techniques to retrieve data from the unstructured wikipedia website and then processed it into structured data. The method used in this research is a standard cross-industry process for data mining by performing phases of data collection, data processing, proposed algorithms, testing and evaluation. The algorithm applied is deepwalk, kmeans, girvan newman. By doing this research, it is expected to provide knowledge about the deep learning approach for data representation of the wikipedia pages knowledge graph and can help users find similar wikipedia pages and enrich literacy on knowledge graph analysis.

Corresponding Author:

Yudistira Bagus Pratama
Department of Computer Science
Universitas Muhammadiyah Bangka Belitung
KH A Dahlan Street No.KM.4, Keramat, Rangkui, Pangkal Pinang City, Kepulauan Bangka Belitung 33134
Email: yudistira.bagus@unmuhbabel.ac.id

1. INTRODUCTION

Analysis of social networks or online communities can be very difficult when working in large networks, as many measurements require expensive hardware [1]. For example, identifying the community structure of a network is a very computationally expensive task [2]. Embedded graphs are a way to represent graphs with vectors, so that further analysis becomes easier [3].

Structured data as a graph in many places such as biology, chemistry,image, the system of decision-making, and social media networks [4]. Using these data in machine learning models proved difficult because of the nature of high-dimensional graph data [5]. Graph Neural network is a new technique in the knowledge graph, allowing to create a model of machine learning files

simultaneously the right tip for studying the structure of the chart data and adjust predictive models in it [6].

The neural networks learn graphical representation graph by making the insertion node graph in the lower-dimensional space [7]. Training to support this representation studied to end reflect the properties of the structural chart of interest for the problems encountered [8]. Representation of nodes iteratively embedding created by combining information from the environment each node [9]. Understanding complex networks can be greatly improved by the development of effective and efficient graph analytics. [10].

The purpose of this research is to test and find out how the results of the deep learning approach to represent knowledge graph data from wikipedia articles.

2. RESEARCH METHOD

2.1. Deep Learning

Deep learning is one of the fields of machine learning that utilizes artificial network to implement problems with large data sets. Deep learning techniques providing a very powerful architecture for supervised learning and unsupervised learning. In machine learning, there are techniques for using extraction features from training data and specialized learning algorithms for online community network classification and data representation. However, this method still has some drawbacks both in terms of speed and accuracy [11]. The application of deep neural networks can be seen in the existing machine learning algorithms so that now computers can learn with speed, accuracy, and on a large scale [12].

2.2. Graph Neural Network

A lot of data can be represented as a graph, the data are often used in fields such as biochemistry, social networking, recommendation systems, and even analysis computer program. Many applications are built to machine learning to make predictions using structured chart data. Not easy to combine information from structured graphical data as input to the machine learning models. The fact that the structured data dimensional non-Euclidean chart and is the main force to create a common and integrated way to use them as input to the model. Graph data is not uniform, with variable size and number of environmental variables. There are many approaches to transform the chart data is a feature that can be used for machine learning models using statistical summary graph, kernel function, or artificial engineering features. This approach is less flexible to adapt during the learning process [13].

Idea behind graph neural network is for studying mapping graphs are embedded vertices or entire subgraphs, as in the vertices of a low-dimensional space vector. The aim is to increase this so that the geometric relationship in space reflected shown. This representational neural network learning task handles as a deep learning task itself [14].

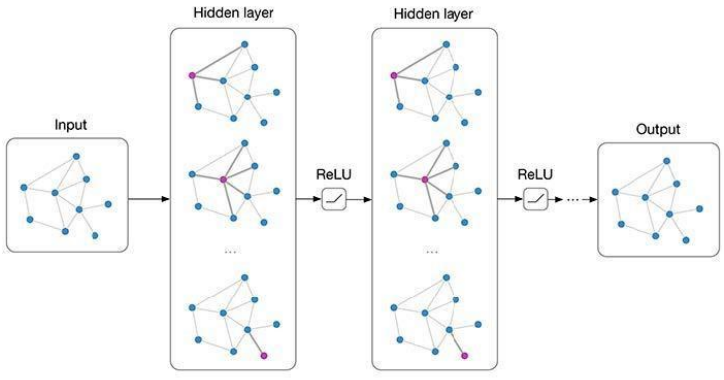

Figure. 1. Graph Neural Network

Figure 2. above about embedding a two-layer Graph neural network [15].

2.3. Deepwalk

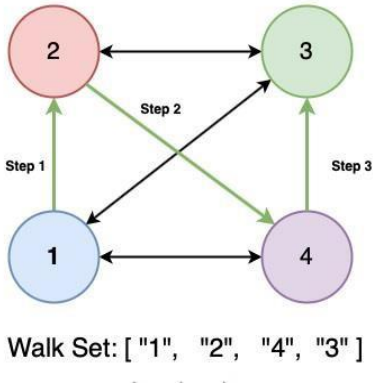

Figure. 2. Deepwalk

The deepwalk procedure is depicted in numerous phases in Figure 2. Do N "randomwalks" beginning from that node for each node. Consider each walk to be a series of node-id strings. Train the word2vec model on this string sequence using the skip-gram approach, given a list of each sequence.DeepWalk is a novel method for investigating the hidden representation of network nodes. This latent representation encodes the network's social ties [16].

Deepwalk is a neural network that acts on the target graph structure directly. This program use the randomwalk approach to gain insight into the network's local structure. The root of randomwalk is a graph G and a uniform sample of a random node. The sample is built up from the most recently visited neighborhood node until the maximum length (t) is achieved. DeepWalk does this by transforming randomwalks into sequences, which are subsequently utilized to train the skip-gram language model [16] .

2.4. Word2vec

As observed in Figure 3, w(t) is the specified target word. There is one hidden layer that computes the matrix weights and the input vector w(t). The hidden layer's computation results are sent to the output layer. The output layer computes the product of the hidden layer's output vector and the output layer's weight matrix. The softmax activation function is then used to determine the likelihood of words appearing in w(t) in a given context location [17].

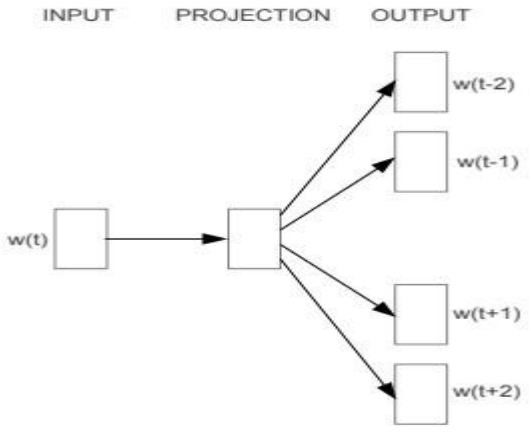

Figure. 3. Word2vec

2.5. Kmeans

One of the partitioning algorithms is the K-means technique, which is based on a preliminary estimate by determining the group centroid [18]. To create a cluster database, the K-means method employs an iterative procedure. It takes the desired beginning cluster number as input and returns the final centroid as output. The K-means clustering algorithm will select a random or random centroid as the starting point for the k pattern. A candidate at the random cluster centroid will alter the number of iterations required to reach the cluster centroid. So, in order to improve performance, the method might be developed by finding the centroid of the cluster as observed from the high initial data density [19].

When finished, the K-Means method will yield the centroid point, which is the algorithm's purpose. The K-means method will cluster data elements in a dataset based on their closest distance [20]. The distance with all data will be computed using the Euclidean Distance algorithm with the initial centroid value picked at random as the beginning center point. A cluster will be formed by data that is close to the centroid. This approach is repeated until there is no change in any of the groups [21].

2.6. Girvan Newman

By deleting the edges of the original network, the Girvan-Newman method discovers progressive communities. Communities are the remaining network's linked components. The Girvan-Newman algorithm focuses on edges that are likely to be "between" communities rather than attempting to construct a metric that informs us where the edges are most central to the community. The vertex betweenness of a node indicates its importance in the network. The vertex betweenness of each node is defined as the proportion of the shortest path between the pair of nodes that cross it. This is important for modifying a network model where the transfer of commodities between the beginning and terminating sites is known, assuming the transfer takes the shortest possible path [22].

2.7. Javascript

JavaScript is an object-based scripting language that allows users to control many user interactions in an HTML document. Where such objects may include windows, frames, URLs, documents, forms, buttons, or other [23].

2.8. Wikipedia

Wikipedia is a free and open networked multilingual encyclopedia project [24]. Since its official launch on January 15, 2001, the English Wikipedia has experienced a tremendous growth in the number of articles. Overall, Wikipedia in all languages reached 1 million articles in September 2004 and then continued to climb to more than 1 million articles. 55 million articles in 2021 [25].

2.9. Method

This research uses the Cross Industry Standard Process for Data Mining (CRISP-DM) method. CRISP-DM is the most representative method for planning overall data extraction, experimental design and evaluation. Since this research has exploratory and experimental objectives, the first and last steps of CRISP-DM, understanding business and deployment, were not implemented. Only the data understanding, data preparation, modeling and evaluation. Web scraping is the process of retrieving information from existing websites. Web scraping applies indexing by browsing HTML documents from a website from which information will be retrieved [26]. Wikipedia Website English Pages was the object for this research

3. RESULTS AND ANALYSIS

Figure 4 below showing a web scraping flowchart starts with opening wikipedia with a web browser, then searching for the keywords of the page want to search for. Set a search filter to only internal links to related wikipedia articles. Then with the code script entered, if the data is successfully collected it will be saved in tsv format and done.

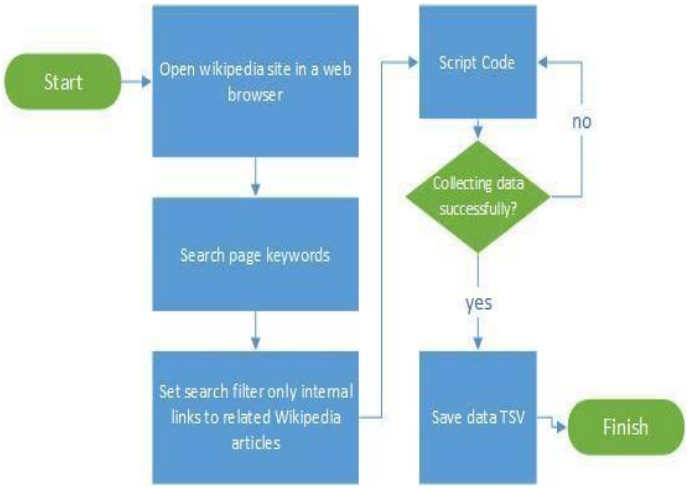

Figure. 4. Flowchart Web Scraping

The result of this research is a web application that implements web scraping techniques on the wikipedia article search engine to display similar wikipedia pages using the javascript programming language and the search results are stored in a table using a local database from a web browser. Then the results of the web scraping are analyzed using a deep learning approach using the python programming language to evaluate whether the results of web scraping are in accordance with the expected results.

3.1. Crawler Application

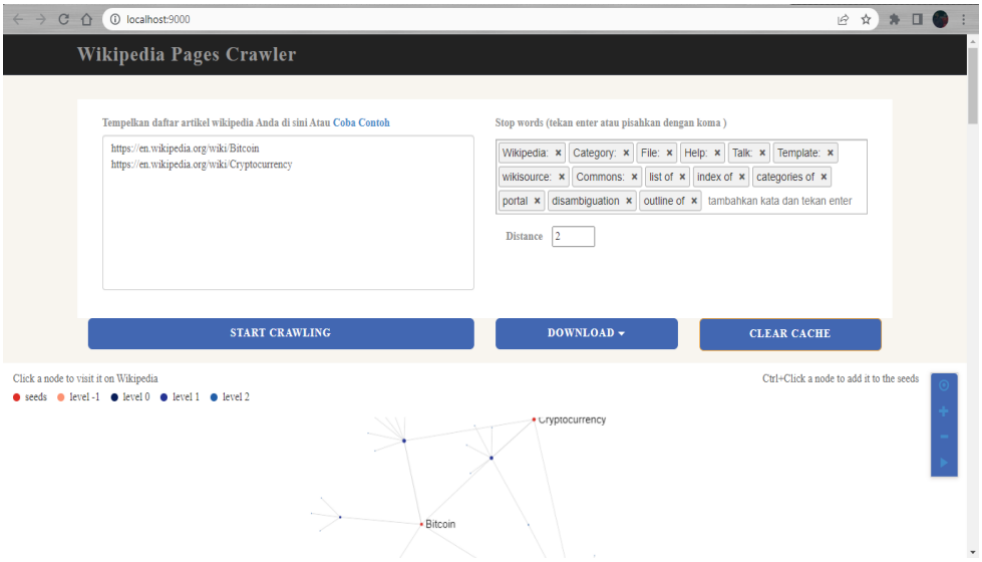

Figure. 5. Crawler Application

When this scraping web application is run, the web application will display a page that contains the functions found in web scraping. The value of "distance" determines the number of iterations. With a distance of 1, will get the original page and the page in the "see also" section. increase in value, the application will perform the same operation on every page that is fetched. With the "stop words" column it is possible to specify which pages should be discarded. The application will search for every "stop word" in the article title, if there is a match then the article will be discarded. Results are cached in local browser storage for 24 hours, allowing to quickly recreate previous scraps or restart canceled scraps. Click the "clear cache" button to reset it. "Source" contains

the analyzed article. "Target" contains the article reference page. "Level" is the distance from the original node.

3.2. Preprocessing Analysis

At this stage the data preprocessing will go through several stages, namely: case folding and stopword removal. This process aims to transform the data for the better so that the data does not have a lot of noise that can affect the level of accuracy in classification. The following flowchart for preprocessing data can be seen in figure 6.

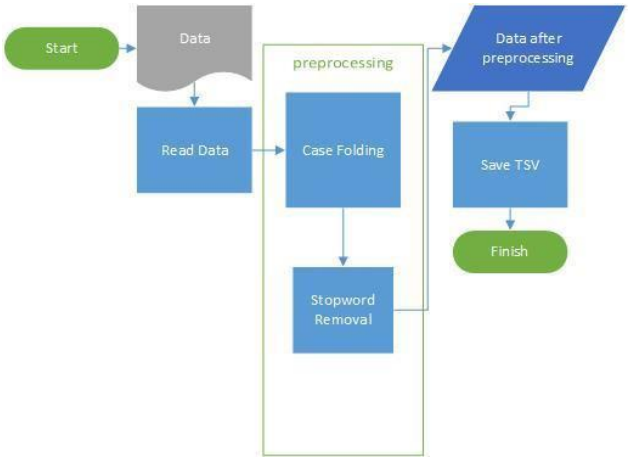

Figure. 6. Flowchart Preprocesing

Read Data In this process then the data stored as a document will be read first and converted into a pandas dataframe for preprocessing. Case Folding This process is done to filter data by removing numbers in sentences, spaces at the beginning and end, and changing all sentences to lower case. Stopword removal In this stage, the words that are not used in the classification process will be removed, such as words: at, list of, with and others. This process will be implemented using the NPM library of javascript. Save TSV The data that has gone through the entire process will be saved in the form of a document in TSV format. Then The data in TSV format will be analyzed.

In the case folding process, the data will be transformed into lowercase, deleting other than letters in the sentence. The stopword removal process is carried out on the source and target attributes. The results of the preprocessing can be seen in Figure 7.

Out[123]:

	source	target	depth
0	bitcoin	alternative currency	1
1	bitcoin	base58	1
2	bitcoin	crypto-anarchism	1
3	bitcoin	virtual currency law in the united states	1
4	cryptocurrency	2018 crypto crash	1
5	cryptocurrency	blockchain-based remittances company	1
6	cryptocurrency	crypto-anarchism	1
7	cryptocurrency	cryptocurrency bubble	1
8	cryptocurrency	cryptocurrency exchange	1
9	cryptocurrency	cryptographic protocol	1

Figure. 7. Data Preprocesing

3.3. Kmeans Clustering

Before doing the kmeans analysis. First determine the number of clusters to be used. There are various methods to determine the optimal number of clusters, that is the average silhouette and elbow methods.

Using the average silhouette scoring, after the calculation is done, the silhouette score is 0.74. The following is a Silhouette analysis performed on the above plot with the aim of selecting the optimal value for n_cluster. As can be seen at figure 8. N_cluster value as 3, 4 and 5 seems not optimal to data provided for the following reason: The presence of clusters with a score silhouette below the average fluctuation width in a plot size silhouette. The value for n_cluster as 2 looks optimal. The silhouette score for each cluster is above the average silhouette score. Also, the size fluctuations are almost similar. the thickness is more uniform than others. Then, the optimal number of clusters that can be selected is 2.

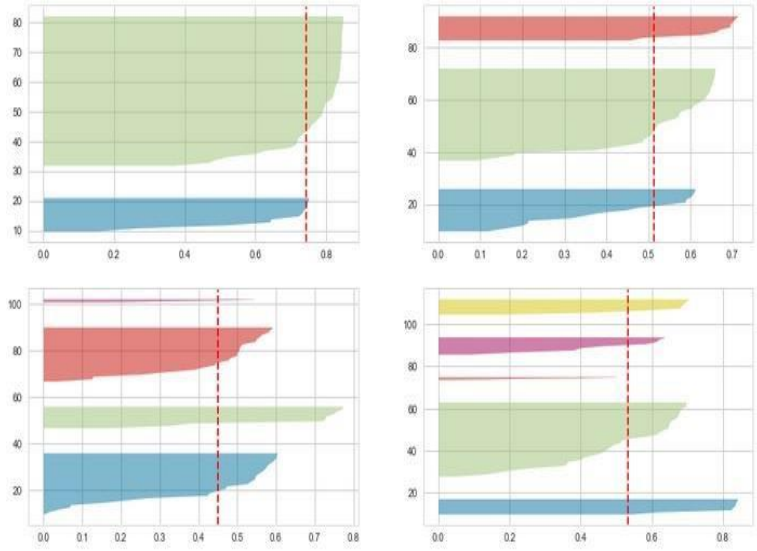

Figure. 8. Silhouette Scoring

Using elbow method by running k-means grouping for the k cluster range and for each value, the sum of the squared distances from each point to the point is calculated. Set center (distortion), when the distortion is plotted and the plot looks like an arm then the "elbow" (the inflection point on the curve) is the best k value. It can be seen that "elbow" is the optimal number 2 for this research. Thus K-Means can be run using n_cluster number 2.

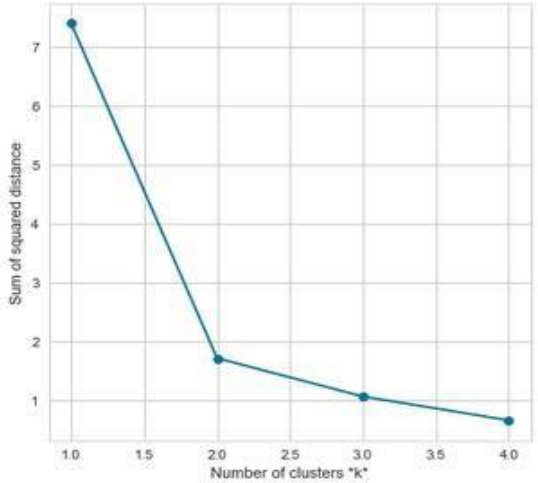

Figure. 9. Elbow Method

Using kmeans clustering, after obtaining the optimal number of clusters through the calculation process on the average silhouette scoring & elbow method algorithm. The Kmeans algorithm is applied whose calculation results can be seen in figure 10.

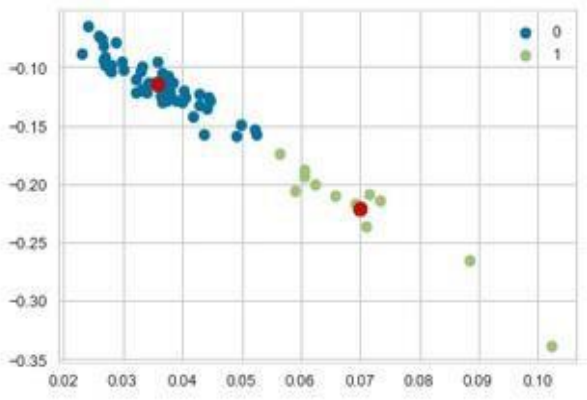

Figure. 10. Kmeans Cluster

The graph above depicts a visualization of the data supplied by the clusters to which it belongs. Cluster 1 is shown in blue on the plot, whereas Cluster 2 is shown in green. k-means clustering attempts to arrange comparable things into clusters by identifying similarities between items and grouping them into groups. The prior siloet scoring and elbow technique methods are utilized to determine the appropriate number of clusters. Each cluster has its own centeroid, which is shown in red. The two clusters obtained correspond to the quantity of data entered during the data crawling procedure.

3.4. Deepwalk analysis

After applying the deepwalk algorithm to the data knowledge graph, then enter some keywords from the wikipedia page for example to be tested. As a result, similar wikipedia entities are grouped together. For example, "crypto-anarchism", "digital gold currency", "2018 crypto crash", and "cryptocurrency and crime" are all pages that are directly related to cryptocurrency pages. While other pages that do not reference each other will have a range of distance from each other.

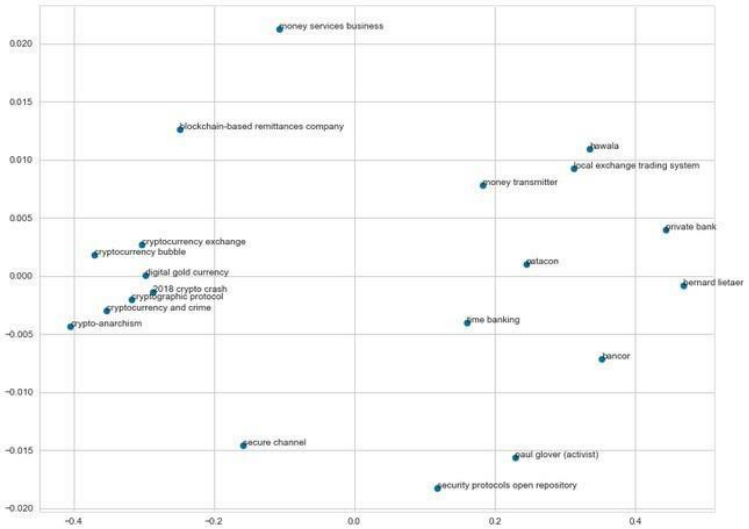

Figure. 11. Deepwalk

3.5. Community Detection

By doing community detection girvan newman on the crawled data obtained 2 communities. From this data, a community structure is then made that shows the interconnected nodes in the graph by targeting the 2 main node colors that show the community. Similar wikipedia pages will be linked to each other and grouped with the same color, which means the crawler application has performed its function properly, that is displaying similar wikipedia pages.

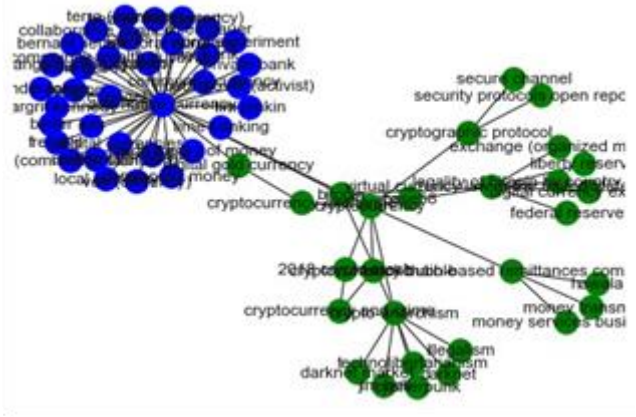

Figure. 12. Graph Community

From the figure above, it can be seen that the wikipedia page community graph is formed by successfully linking wikipedia pages that reference each other.

3.6. Betweenness Centrality

The degree to which a node resides on the path between other nodes is measured by betweenness centrality. Betweenness centrality in this study specifically analyzes the extent to which a wikipedia page is on the shortest path connecting other wikipedia pages in the network. The greater the betweenness centrality rating, the more dependent a page is on other pages.

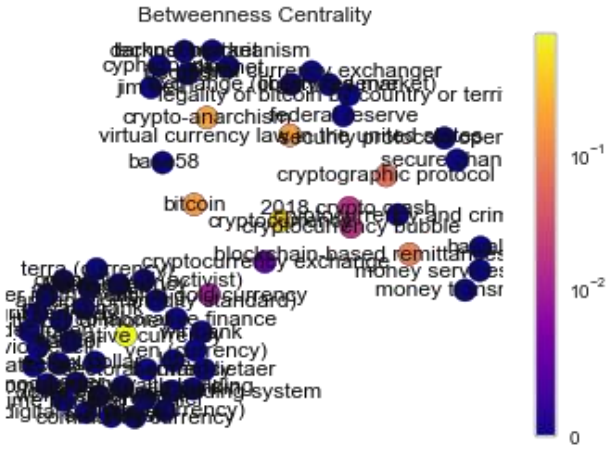

Figure. 13. Graph Betweenness Centrality

From the figure above, it can be seen that the graph betweenness centrality of the wikipedia page is formed by successfully sorting the wikipedia page that has the highest influence based on the color of the heatmap.

Table 1. Betweenness Centrality

Rank	Wikipedia Page	Centrality Measure
1	alternative currency	0.8130618720253833
2	cryptocurrency	0.4456196016217168
3	crypto-anarchism	0.18975850520007054
4	bitcoin	0.16596157236030318
5	virtual currency law in the united states	0.16014454433280453
6	blockchain-based remittances company	0.09518773135906927
7	cryptographic protocol	0.06398730830248546

Table 1 presents data on betweenness centrality for various Wikipedia pages related to alternative currencies, cryptocurrencies, and related concepts. Betweenness centrality is a measure used in network analysis to quantify the importance of a node within a network based on its position in connecting other nodes. Alternative currency has the highest betweenness centrality, indicating that it plays a significant role in connecting other pages within the network. Cryptocurrency follows as the second most central node, albeit with a lower betweenness centrality compared to alternative currency. Crypto-anarchism and bitcoin also have notable betweenness centrality measures, indicating their importance in connecting various concepts within the network. Virtual currency law in the United States, blockchain-based remittances company, and cryptographic protocol have lower betweenness centrality measures compared to the top-ranked pages but still play significant roles in connecting certain parts of the network. In summary, the data in Table 1 provide insights into the structural importance of different Wikipedia pages within the network of alternative currencies, cryptocurrencies, and related topics, as measured by their betweenness centrality.

4. CONCLUSION

This research was successfully carried out using a deep learning approach as an evaluation process. Application crawler successfully displays similar wikipedia pages. Analysis using the kmeans algorithm can group the training data into 2 clusters according to the input of training data. The analysis using the girvan newman algorithm finds 2 communities in the training data and describes them in graph form according to the input of training data. Analysis using the deepwalk algorithm produces a group of wikipedia pages in the form of words in low dimensions according to the input of training data. Analysis using the betweenness centrality can find most referenced page from wikipedia dataset. This research can be used as a reference for further research and try to use other algorithms to enrich research literacy related to the knowledge graph.

REFERENCES

[1] S. Peng *et al.*, "A survey on deep learning for textual emotion analysis in social networks," *Digit. Commun. Networks*, vol. 8, no. 5, pp. 745–762, Oct. 2022, doi: 10.1016/J.DCAN.2021.10.003.

[2] F. Karimi, S. Lotfi, and H. Izadkhah, "Multiplex community detection in complex networks using an evolutionary approach," *Expert Syst. Appl.*, vol. 146, p. 113184, May 2020, doi: 10.1016/j.eswa.2020.113184.

[3] X. Li, G. Xu, L. Jiao, Y. Zhou, and W. Yu, "Multi-layer network community detection model based on attributes and social interaction intensity," *Comput. Electr. Eng.*, vol. 77, pp. 300–313, Jul. 2019, doi: 10.1016/j.compeleceng.2019.06.010.

[4] J. Zhou *et al.*, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, Jan. 2020, doi: 10.1016/J.AIOPEN.2021.01.001.

[5] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Trans. neural networks Learn. Syst.*, vol. 32, no. 1, pp. 4–24, 2020.

[6] I. Chami, S. Abu-El-Haija, B. Perozzi, C. Ré, and K. Murphy, "Machine Learning on Graphs: A Model and Comprehensive Taxonomy," *J. Mach. Learn. Res.*, vol. 23, 2022.
[7] D. Matsunaga, T. Suzumura, and T. Takahashi, "Exploring Graph Neural Networks for Stock Market Predictions with Rolling Window Analysis," Sep. 2019, Accessed: Jun. 26, 2023. [Online]. Available: http://arxiv.org/abs/1909.10660
[8] B. P. Adedeji and G. Kabir, "A feedforward deep neural network for predicting the state-of-charge of lithium-ion battery in electric vehicles," *Decis. Anal. J.*, vol. 8, p. 100255, Sep. 2023, doi: 10.1016/J.DAJOUR.2023.100255.
[9] S. Alaie and S. J. Al'Aref, "Application of deep neural networks for inferring pressure in polymeric acoustic transponders/sensors," *Mach. Learn. with Appl.*, p. 100477, Jun. 2023, doi: 10.1016/J.MLWA.2023.100477.
[10] M. Xu, "Understanding Graph Embedding Methods and Their Applications," *SIAM Rev.*, vol. 63, no. 4, pp. 825–853, 2021, doi: 10.1137/20M1386062.
[11] R. FIRMANSYAH, "IMPLEMENTASI DEEP LEARNING MENGGUNAKAN CONVOLUTIONAL NEURAL NETWORK UNTUK KLASIFIKASI BUNGA," *Pap. Knowl. . Towar. a Media Hist. Doc.*, vol. 3, no. 2, p. 6, 2021.
[12] S. R. Dewi, "Deep Learning Object Detection Pada Video," *Deep Learn. Object Detect. Pada Video Menggunakan Tensorflow Dan Convolutional Neural Netw.*, pp. 1–60, 2018, [Online]. Available: https://dspace.uii.ac.id/bitstream/handle/123456789/7762/14611242_Syarifah Rosita Dewi_Statistika.pdf?sequence=1
[13] P. Rodríguez, E. M. Thesis, and V. Arias, "Graph Neural Networks and its applications Master in Innovation and Research in Informatics," 2019.
[14] X. Li, L. Sun, M. Ling, and Y. Peng, "A survey of graph neural network based recommendation in social networks," *Neurocomputing*, vol. 549, p. 126441, Sep. 2023, doi: 10.1016/J.NEUCOM.2023.126441.
[15] P. Shao, J. He, G. Li, D. Zhang, and J. Tao, "Hierarchical graph attention network for temporal knowledge graph reasoning," *Neurocomputing*, vol. 550, p. 126390, Sep. 2023, doi: 10.1016/J.NEUCOM.2023.126390.
[16] J. Wang, K. Yue, L. Duan, Z. Qi, and S. Qiao, "An efficient approach for multiple probabilistic inferences with Deepwalk based Bayesian network embedding," *Knowledge-Based Syst.*, vol. 239, p. 107996, Mar. 2022, doi: 10.1016/J.KNOSYS.2021.107996.
[17] J. J. Zhu and Z. J. Ren, "The evolution of research in resources, conservation & recycling revealed by Word2vec-enhanced data mining," *Resour. Conserv. Recycl.*, vol. 190, p. 106876, Mar. 2023, doi: 10.1016/J.RESCONREC.2023.106876.
[18] M. Ay, L. Özbakır, S. Kulluk, B. Gülmez, G. Öztürk, and S. Özer, "FC-Kmeans: Fixed-centered K-means algorithm," *Expert Syst. Appl.*, vol. 211, p. 118656, Jan. 2023, doi: 10.1016/J.ESWA.2022.118656.
[19] M. Z. Islam, V. Estivill-Castro, M. A. Rahman, and T. Bossomaier, "Combining K-MEANS and a genetic algorithm through a novel arrangement of genetic operators for high quality clustering," *Expert Syst. Appl.*, vol. 91, pp. 402–417, Jan. 2018, doi: 10.1016/j.eswa.2017.09.005.
[20] F. D. Bortoloti, E. de Oliveira, and P. M. Ciarelli, "Supervised kernel density estimation K-means," *Expert Syst. Appl.*, vol. 168, Apr. 2021, doi: 10.1016/j.eswa.2020.114350.
[21] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (Ny).*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.
[22] W. Huang, L. Li, H. Liu, R. Zhang, and M. Xu, "Defense resource allocation in road dangerous goods transportation network: A Self-Contained Girvan-Newman Algorithm and Mean Variance Model combined approach," *Reliab. Eng. Syst. Saf.*, vol. 215, p. 107899, Nov. 2021, doi: 10.1016/J.RESS.2021.107899.
[23] Z. Jiang, H. Zhong, and N. Meng, "Investigating and recommending co-changed entities for JavaScript programs," *J. Syst. Softw.*, vol. 180, p. 111027, Oct. 2021, doi: 10.1016/J.JSS.2021.111027.
[24] T. Azar, "Wikipedia: One of the last, best internet spaces for teaching digital literacy, public

writing, and research skills in first year composition," *Comput. Compos.*, vol. 68, p. 102774, Jun. 2023, doi: 10.1016/J.COMPCOM.2023.102774.

[25] Wikipedia Milestones. Accessed 20 May 2023. Accessed from https://meta.wikimedia.org/wiki/Wikipedia_milestones

[26] A. A. Maulana, A. Susanto, and D. P. Kusumaningrum, "Rancang Bangun Web Scraping Pada Marketplace di Indonesia," *JOINS (Journal Inf. Syst.*, vol. 4, no. 1, pp. 41–53, 2019, doi: 10.33633/joins.v4i1.2544.