DOI: 10.15849/IJASCA.220328.10

Int. J. Advance Soft Compu. Appl, Vol. 14, No. 1, March 2022
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Studying the Wikipedia Math Essential Pages using Graph Theory Metrics

Sajidah Shahadha Mahmood

Department of Public Relations, University of Al Iraqia, Baghdad, Iraq.
Email: sajidah.sh.mahmood@aliraqia.edu.iq

Abstract

COVID-19 pandemic enforced students in schools and universities all around the world to study using the online and blinded learning. In these learning models, students depend on the Internet for information searching of different scientific essentials to improve their skills and to overcome the gap of facing instructors. One of the most popular sources of information is Wikipedia. In this work, we attempt to study the relations of different math essential pages of Wikipedia to find the relation between these topics. A graph has been constructed for these pages. The graph theoretical metrics, such as, centrality, edge weights and clustering coefficient have been extracted of the constructed graph. The extracted values have been investigated to gain more insights of the math topics that should be studied first. The extracted results show that the in-degree property of the articles and the betweenness value of these articles are correlated. Moreover, there is no relation between the in /out-degree of the pages. Finally, the constructed graph has a small average shortest path and a high global cluster coefficient. This proves that the constructed graph follows the small world phenomenon.

***Keywords**: Graph metrics, Math essentials, Gephi, Small world phenomenon, Directed graph*

1 Introduction

COVID-19 pandemic has changed different aspects of our life. One of these aspects is the high education learning methods. A massive shifting has occurred in the way students interacted with instructors, classes and subjects [1]. In addition, the new learning method divides the learning responsibility between the students and the instructors in a way that the students are responsible of learning prerequisite material even if the instructor did not mention them [2, 3]. This makes the Internet

Received 16 December 2021; Accepted 28 February 2022

the main platform for searching for the information and materials. However, how to start searching for a subject? What is the first learning material to start with? What are the essential topics to understand before moving on? One website to start with is Wikipedia [4].

Wikipedia is an online free encyclopedia that is written in 309 different languages. According to Wikipedia statistics [5], it contains 29 billion words in 55 million articles in 309 languages. These statistics show how massive the number of articles that have been written in Wikipedia and how hard it is to find a starting point to study the essential topics or classes in any major. In addition, utilizing any search engine to find a specific information or topic requires a good background on the essentials of the topic. To tackle this issue, graph theory can be utilized to construct the relations between different essential topics in any major. Subsequently, the constructed graph can be leveraged to find a starting point in the essential topics.

A graph is defined a data structure that consists of nodes and links. Any two nodes are connected if a relation occurs between them [6]. Graphs have two main categories, direct and indirect. In the directed graph, moving from one node to another node must fallow the direction if the relation between the nodes. In this graph category, it is easy to find a starting point and to generate a tree from the graph with a root and different leafs. Graphs have many performance metrics that have many physical meanings. Theses metrics emerged in graph theory to create the network science [7]. Network science has been utilized to study social media relations [8], Internet autonomous system relations [9], scientific paper citation network [10] and the structure of wireless sensor networks [11].

In this work, network science and graph metrics have been leveraged to study the relation between different mathematical essentials. A directed graph has been created from math essential article pages that have been crawled from Wikipedia pages. Gephi [12, 13] graph analyzer has been utilized to calculate different graph metrics of the harvested Wikipedia articles. Betweenness, closeness, cluster coefficient, network diameter, nodes’ degrees and average paths have been calculated for the constructed graph.

The rest of this paper is organized as follows; section II overviews the related works that have been conducted to study Wikipedia. Section III introduces graph metrics and how to calculate them. Section V, overviews the conducted experiment. Section VI discusses the harvested results. We conclude this paper in section VII.

2 Related Works

Studying and analyzing Wikipedia articles and pages has attracted researcher over the past decade. Many methods and techniques have been leveraged for the mining process of useful information. For example in [14], the authors attempted to study the popular chemical substances using the number of views of each page mentioned these substances. The Wikipedia chemical structure explorer (WCSE) developed by Novartis along with Python script allowed the researcher to count all the chemical materials mentioned in all pages with the number of views. In [15], the author conducted a quantitative study or a statistical study of the top ten languages used in Wikipedia articles and their evolution over the time. In [16], the authors attempted to study the citation quality of Wikipedia pages to real scientific papers. More than 115 million citations and 800K papers have been investigated. The authors found that most paper were uncited or untrusted. In [17], the author studied Wiki-trends to gain more insight of a special disease. The author compared the Wiki-trend of the disease with Google trends. In [18], the author studied and investigated the history of diseases over time in Wikipedia articles. Their purpose was to show the evolution of information of a disease over the time. In [19], authors studied the media coverage of COVID-19 news utilizing the page views number of Wikipedia articles. The author found that the number of views of pages in Wikipedia reflects the media coverage news. In [20], the quality of different articles written in different languages has been investigated. The authors have shown that the quality of any article depends on the language and the popularity of the article.

In the field of network science and graph metric studies, Wikipedia articles, authors and their relations have been studied heavily in the past decades. In [21], the author leveraged graph theory to study the authors' behavior of the English articles in Wikipedia. A social graph has been constructed. The author found that articles in Wikipedia are classified into two main categories; narrow focus and broad articles. In [22], authors generated to graphs of Wikipedia data. The first graph is category graph and the second graph is the articles graph. Different graph metrics have been investigated to gain more insight of the structure of the articles and categories of Wikipedia. In [23], another graph has been constructed to investigate the quality of Wikipedia articles. The authors attempted to answer the question 'why the quality of articles differ?' using social metrics. In [24, 25] a revision history graph has been constructed and different graph metrics have been implemented on it.

Our work in this study differs from all the other works in three main folds. First, we have utilized narrow focused articles of Wikipedia, mathematical essentials, in this study. These essentials have been studied for the coronation with the number of views and different deep learning models in [26]. Second, the constructed graph emphasized on understanding the physical meaning of different graph metrics. Finally, this work can be re-conducted for different essentials' of any major.

3 Graph Theoretical Background

In this section, different graph metrics leveraged in this work are overviewed. We start by defining the graph itself. Subsequently, the simple metrics, such as, edge degree, diameter and average path will be shown. Finally, the centrality metrics are introduced. Table 1 shows the definition of all variable used in these metrics.

Table 1: Variables' Definitions

Variable	Definition

$\mathbf{n}$	Total number of nodes in the graph
$\mathbf{d}(\mathbf{N},\mathbf{y})$	Shortest path between node N and node y
$\boldsymbol{\partial}_{\mathbf{st}}(\mathbf{N})$	Shortest path between node S and node T that pass through node N
$\mathbf{m}$	Number of links

3.1Graph

A graph G = (V, E) is a data structure that consists of nodes or vertices (V) and relations, links, vectors or edges (E). A graph is known as undirected graph if the relation or the link between two nodes can be passed from both sides. Moreover, it is known as directed if these links are vectors or an arrow with one node at the arrow start and the other node at its head. The graph size depends on the number of nodes in the graph and the graph order is the number of links between these nodes in the graph.

3.2Node degree

The number of links that start or end at the node is known as the node degree. If the graph is undirected, two degrees are shown, in-degree and out-degree. Node degree can define the popularity of any node in the graph.

3.3 Graph diameter

It is a distance measurement tool. It is known as the longest shortest path calculated between the nodes in the graph. To calculate it, all the shortest paths between all nodes in the graph should be calculated. Subsequently, the longest value is the graph diameter. In highly connect graphs; the diameter has a small value.

3.4Graph density

It measures the number of links that shown from the possible number of links. It depends on the number of nodes and the number of links in the graph. Eq.1 shows how to calculate this value.

$$d(G) = \frac{m}{n(n-1)} \qquad (1)$$

3.5 Centrality Metrics

Different metrics can be used to measure the node centrality in the graph. The most popular metrics are; betweenness, closeness, eccentricity and cluster coefficient.

- Betweenness: is defined as the number of shortest paths between any pair of nodes that passes through the node. Eq.2 is utilized to measure the betweenness of any node
$$b(N) = \sum_{s \neq N \neq t} \frac{\partial_{st}(N)}{\partial_{st}} \qquad (2)$$
- Closeness: Is defined as the sum of all shortest paths between the node and all other nodes in the graph. If the node is central, it will have a shortest path to all nodes. Eq.3 shows how to calculate this metric
$$C(N) = \frac{n}{\sum_y d(N,y)} \qquad (3)$$
- Cluster Coefficient: This metric has is divided into two categories a local and a global one. The local cluster coefficient is calculated for each node. However, the global cluster coefficient is calculated for the graph itself. The local cluster coefficient is defined as the number of triangles that the node participates in over the number that should occurs. In other words, this metric measure how much the neighbors of a node know each other without passing through the node. Eq.4 shows how to calculate the local cluster coefficient. The global cluster coefficient is the average of the local cluster coefficient of all nodes as in Eq.5. It worth mentioning that a graph with a high global cluster coefficient follows small world phenomenon.

$$C^i = \frac{\#\ of\ triangles\ that\ contain\ node\ i}{k_i(k_i - 1)} \qquad (4)$$
$$C_{global} = \frac{1}{n} \sum c^i \qquad (5)$$

3.6 Average shortest Path

A graph is a small world if it has a small average shortest path. The path should be less than the natural log of the number of nodes. This is shown in Eq.6.

$$Avergae\ Shortest\ Path \leq \ln(n) \qquad (6)$$

3.7 Eccentricity

It is defined as the maximum distance in edges from a node to all other nodes in the network. For nodes with small edge degree, the eccentricity value should be high.

4 Experiment

Math essentials that can be found in Wikipedia pages have been crawled with the relation between them. We have utilized the list of these essential from [12] as the seed list for our crawler. The crawler has been written in Python. 1068 different pages have been crawled with 27078 links between them. Table II shows general statistics of the harvested data. All the pages are connected. In other words, the pages can create one connected graph. Finally, the pages have been given ID numbers and a connected list between these nodes have been created and fed to Gephi to generate one directed graph. Figure 1 shows the directed graph that has been constructed by Gephi. We can observe from the figure that the constructed graph has many nodes with few connections and few nodes with massive number of connections. This graph has been leveraged for graph metrics calculations that shown in the result section.

Table 2: Statistics of the Harvested Data

Variable	Value
Number of pages	1068
Number of references in all pages for other Wikipedia pages	27078
Number of pages in the seed list	5 articles
Type of constructed graph	Direct
Number of graphs “connected components”	1

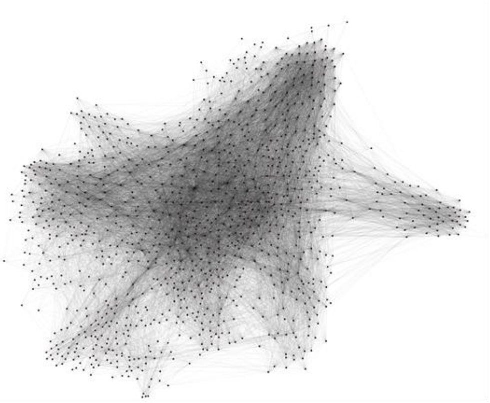

Fig. 1. The Directed Constructed Graph in Gephi

5 Results

Table III shows a general result statistics that have been calculated in Gephi. We can observe from the table that the average shortest path value is approximately 2.8

which is lower than ln(1068). Moreover we can observe that the global cluster coefficient has a high value. These two number show that the directed constructed graph of the math essentials follows the small world phenomenon. Finally, the graph has a small density value. This is obvious with the small number of connections that have been harvested.

Table 3: Results Statistics

Variable	Value
Diameter	12
Average Degree	25.5
Average Path Length	2.794
Graph Density	0.024
Global Clustering Coefficient	0.214

Figure 2 shows the distribution of nodes' out-degree property. As mentioned, the constructed graph is a directed graph. This gives each node two types of degrees; in and out. We can observe from the figure that small number of nodes have a very high out-degree and the rest of the nodes have a small out-degree number. This follows the small world phenomenon were rich get richer. Table 3 shows the top five pages with the highest value of out-degree. We can observer from the table that these pages are general math essentials. Moreover, these pages are the five nodes in the seed list of our crawler.

Table 4: Out-Degree top five pages

Page	Out-degree value
Mathematics	532
Real Numbers	290
Function (mathematics)	270
Complex Numbers	214
Integer	201

Figure 3 shows the betweenness values of the graph's nodes. As in the out-degree figure, the betweenness values have few nodes with massive betweenness values and the other node with small value. This means that there are few pages or articles that a lot of shortest paths follow from them. These articles are important articles and need to be reviewed. Table 4 shows the top six articles with the highest betweenness values. We can observe from the table that these articles have no similarities with the articles in table 3. However, table 5 shows the top 5 pages with the highest in-degree values. We can observe the similarity between the pages in the betweenness table and the in-degree table. We also can observe from the table

5 that the values of in-degree property are much smaller than the values of out-degree table.

Figure 4 shows the relation between the betweenness value of the node and its in-degree value. We can observe from the figure that with higher betweenness values the in-degree value is also high. This has been reflected in the table of the top 5 pages in the in-degree property and the betweenness. Figure 5 shows the relation between the in-degree and the out-degree properties. We can observe that the figure looks like a random figure with no correlation between these values.

Table 5: Betweenness top Six pages

Page	Out-degree value
Bijection	12472.16
Logarithm	12176.67
Domain of a Function	9896.27
Exponentiation	8617.249
Directed graph	8163.047
Big O notation	7894.13

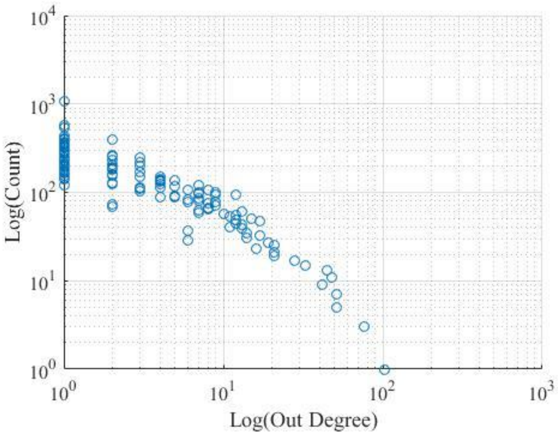

Fig. 2. Out-degree Distribution

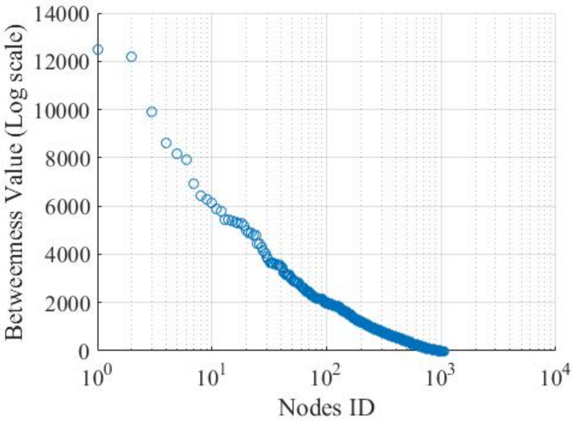

Fig. 3. betweenness values of the graph’s nodes

Table 6: The top 5 pages with the highest in-degree value

Page	Out-degree value
Logarithm	101
Domain of a Function	99
Bijection	89
Laplace transform	89
Factorial	82

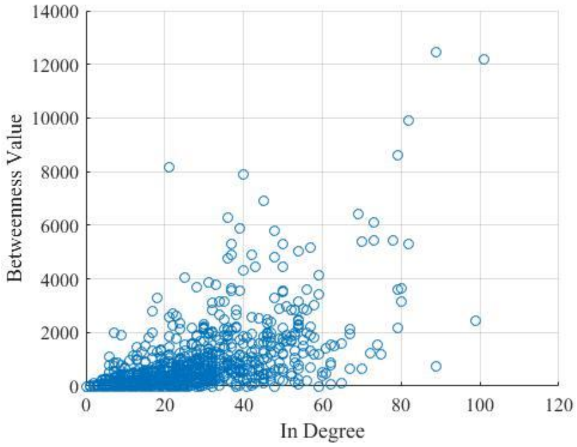

Fig. 4. Betweenness and In-degree

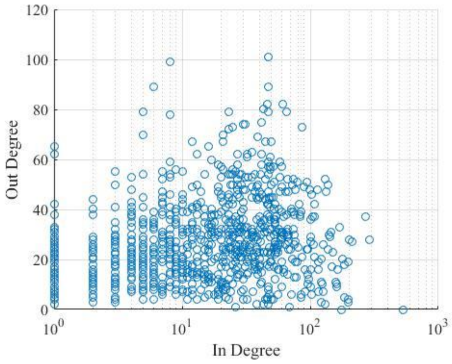

Fig. 5. In-degree and Out-degree

Figure 6 shows the CDF of closeness values calculated for the harvest articles. We can observe that 10% of the articles have zero closeness which means that their paths to all other nodes are massive. These 10% of the nodes are the leaves of the graph. In addition we observed that 5% of the nodes have 1 closeness which means that they are in the central of the graph. Another observation from the figure is than more than 70% of the nodes have a low closeness value.

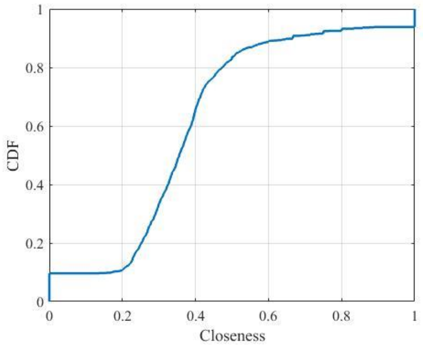

Fig. 6. CDF of Closeness

Figure 7 shows CDF of local cluster coefficient of the articles. We can observe from the figure that the highest cluster coefficient of the nodes is 0.5. Moreover, we can observe that 90% of the nodes have a value between 0.1 and 0.4. This means that for each node at least 10% of its neighbors know each other. For the nodes with 0.4 values, 40% of its neighbors know each other. This result is not high which mean that more references should be added to these articles.

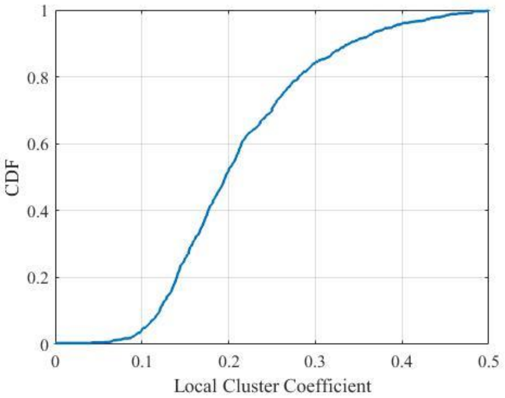

Fig. 7. CDF of Local cluster coefficient

Finally, figure 8 shows the histogram of the eccentricity values of the nodes in the constructed graph. We can observe that more than 30% of the nodes have eccentricity value of 5. This means that starting from these articles, a reader have to read 5 different articles to reach all the articles in math essentials. Moreover, for few nodes, the eccentricity reaches 12 hops. These means that this articles are very advance articles and hard to be visited easily by readers

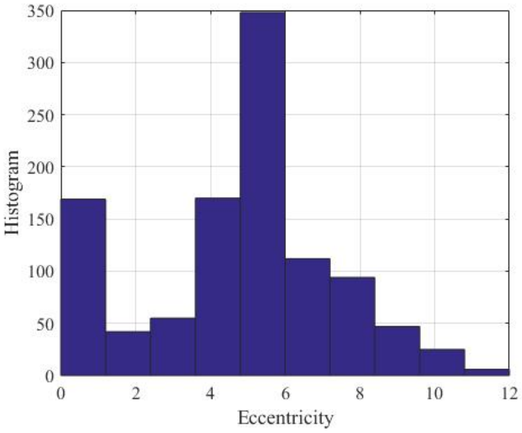

Fig. 8. Eccentricity Histogram

6 Conclusion

The shifting to online and blinded learning motivated students to study different topics and articles from the Internet. A good point to start learning is Wikipedia. However, where to start is the question. In this work, we have crawled the math essentials from Wikipedia pages to find the relation between these pages and to create a starting point for the learners. A directed graph has been generated and different graph metrics have been studied. Our results show that betweenness and in-degree parameters can be used as a good start for studying the essentials. Moreover, the eccentricity ranked the articles according to their maximum hop count to other articles in the field. Finally, we found that all the essentials generated one connected graph without any disconnected components. This means, a crawler with only one seed article could crawl all of these articles.

References

[1] Alheyasat, O. (2016). Investigation and analysis of research gate user's activities using neural networks. *Int. Arab J. Inf. Technol*, *13*(2), 320-325.

[2] Alkhalil, S., Manasrah, A., & Masoud, M. (2021). Let's Learn with a Jigsaw! Implementing a Unique Collaborative Online Learning in an Engineering Course. *2021 International Conference on Information Technology (ICIT) on* (pp. 396-399), doi: 10.1109/ICIT52682.2021.9491692.

[3] Alshahrani, A. (2021). Readiness of Higher Education Institutions for e-learning: A Case Study of Saudi Universities During the COVID-19 Pandemic. *International Journal of Advances in Soft Computing & Its Applications, 13*(1), 149-161.

[4] Barabási, A. Network (2013). Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. *The Royal Society Publishing 371*(1987), 20120375.

[5] Cao, Y., Mehta. H., Norcross, A., Taniguchi, M., & Jonathan, L. (February 2020) Analysis of Wikipedia pageviews to identify popular chemicals. In Reporters, Markers, Dyes, Nanoparticles, and Molecular Probes for Biomedical

Applications. *International Society for Optics and Photonics XII on* (pp. 11256-112560).

[6] Gozzi, N., Tizzani, M., Starnini, M., Ciulla. F., Paolotti, D., Panisson, A., & Perra, N. (2020). Collective response to media coverage of the COVID-19 pandemic on Reddit and Wikipedia: mixed-methods analysis. *Journal of medical Internet research*, *22*(10).

[7] Iba, T., Nemoto, K., Peters, B., & Gloor, P. (2010). Analyzing the creative editing behavior of Wikipedia editors: Through dynamic social network analysis. *Procedia-Social and Behavioral Sciences*, *2*(4), 6441-6456.

[8] Jannoud, I & Masoud, M. (2015). Advanced Computer and Communication Engineering Technology. *On understanding centrality in directed citation graph* (pp. 43-51) Springer Cham.

[9] Johnson, E. (2019). Gephi-Network analysis and visualization.

[10] Lagunes-García, G., Rodríguez-González, A., Lucía Prieto-Santamaría, García del Valle E., Massimiliano Z., & Menasalvas-Ruiz, E. (2020). How Wikipedia disease information evolve over time? An analysis of disease-based articles changes. *Information Processing & Management*, *57*(3).

[11] Lewoniewski, W., Węcel, K., & Abramowicz, W. (2019). Multilingual ranking of Wikipedia articles with quality and popularity assessment in different topics. *Computers*, *8*(3), 60.

[12] Liu, J., & Ram, S. (2018). Using big data and network analysis to understand Wikipedia article quality. *Data & Knowledge Engineering, 115,* 80-93.

[13] Manasrah, A., Masoud, M., & Jaradat, Y. (2021, July). Short Videos, or Long Videos? A Study on the Ideal Video Length in Online Learning. *In 2021 International Conference on Information Technology (ICIT),* (pp. 366-370). IEEE.

[14] Masoud, M., Jaradat, Y. & Ahmad A. (2017). Machine learning approach for categorizing internet autonomous systems' links. *ICGHIT, Hanzhou, China Google Scholar*.

[15] Masoud, M., Jaradat, Y., Jannoud, I. & Al Sibahee, M. (2019). A hybrid clustering routing protocol based on machine learning and graph theory for energy conservation and hole detection in wireless sensor network. *International Journal of Distributed Sensor Networks 15*(6), 1-19, 1550147719858231.

[16] Masoud, M., Jaradat, Y., Jannoud, I., & Hong,H. (2017). The Impact of 16-bit and 32-bit ASNs Coexistence on the Accuracy of Internet AS Graph. *Journal of Network and Systems Management, 25*(2).

[17] Mathieu, B., Heymann. S, & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *In Third international AAAI conference on weblogs and social media. 3*(1) (pp. 361-362).

[18] Nicholson, J., Uppala, A., Sieber, M., Grabitz, P., Mordaunt, M. & Rife, S. (2021). Measuring the quality of scientific references in Wikipedia: an analysis of more than 115M citations to over 800,000 scientific articles. *The FEBS Journal 288*(14), 4242-4248.

[19] Raman, N., Sauerberg, N., Fisher, J. & Narayan, S. (2020). Classifying Wikipedia article quality with revision history networks. *In Proceedings of the 16th International Symposium on Open Collaboration,* (pp. 1-7).

[20] Raman, N., Sauerberg, N., Partida, A., & Fisher, J. (2020). Revisionist History: Predicting Wikipedia Article Quality With Edit Histories.

[21] Rozemberczki, B., Scherer, P., He, Y., Panagopoulos, G., Riedel, A., Astefanoaei, M. & Kiss O., et al. (2021). Pytorch geometric temporal: Spatiotemporal signal processing with neural machine learning models. *In Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, (pp. 4564-4573).

[22] Sciascia, S., & Radin, M. (2017). What can Google and Wikipedia can tell us about a disease? Big Data trends analysis in Systemic Lupus Erythematosus. *International journal of medical informatics 107*, 65-69.

[23] Soto, O., & Felipe, J. (2012). *Wikipedia: A quantitative analysis*.

[24] Sternberger, A.,Wyatt, S. (2021). Wikipedia in the science classroom. *CourseSource.*

[25] Wikipedia statistics, “en.wikipedia.org/wiki/ Wikipedia: Size_comparisons”

[26] Zesch, T., & Gurevych, I. (2007). Analysis of the Wikipedia category graph for NLP applications. *In Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing.* 1-8.

Authors’ information

Sajidah Shahadha Mahmood is an assistant lecturer at the Department of Public Relations, University of Al Iraqia, Baghdad, Iraq. She was born in Baghdad, Iraq in 1977. She received her B.SC. Degree in Control and Systems Engineering\ Control Engineering in 2002 from University of Technology in Baghdad, Iraq and she received her M.SC. Degree in Control and Systems Engineering\ Computer Engineering in 2020.