ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse344942020.pdf
https://doi.org/10.30534/ijatcse/2020/344942020

Open Government Data (OGD) portals selection using Ant Colony Optimization (ACO) Algorithm

Nor A. M. Sabri[1], Nurul A. Emran[2], Norharyati Harum[3]
[1,2,3]Centre for Advanced Computing Technologies (CACT), Fakulti Teknologi Maklumat Dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM), Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia.
noramalinamohdsabri@gmail.com[1], nurulakmar@utem.edu.my[2], norharyati@utem.edu.my[3]

ABSTRACT

The article describes a new method of open data source selection using the Ant Colony Optimization (ACO) algorithm. The work is driven by open government data (OGD) that has been increasingly used by data consumers which motivates a proliferation of software applications for the civilians. As the number of government open data portals increases, the decision on which portal to use often relies on the portal with the most relevant content. Nevertheless, even though the relevant portals are known, data consumers are still left with the question of whether the sources in hand subscribe to open data quality standards. This is because the quality of the data source will affect decision making. Manual checking is feasible, but it is not an attractive option as the number of data sources grows. Thus, in this article, we present the results of implementing open data portals selection with quality consideration. The enhancement of the ACO algorithm is made through quality filters. OGD policies from 30 Open Government Data (OGD) Portals are retrieved to establish the set of data quality requirements that are used in the filtering process. The model considers three selection parameters to assess 40 random countries. The parameters are quality requirements coverage in the OGD policy, the readiness aspect of the OGD portal, and the content of government data. The implementation of this model gives an insight into how the ACO algorithm can be used to deal with selection problem that involves multi-sources and multi-selection criteria. The results of this paper contribute to improving the quality of data provided by OGD providers for open data consumers.

Key words : Ant colony optimization, data quality, data source selection, open government

1. INTRODUCTION

We can witness the adoption of open data initiatives by governments worldwide since the advocation of open government concept. For example, the establishment of the organization such as the Organisation for Economic Co-operation and Development (OECD)[1] shows worldwide countries' commitment to open data. As a result, government data can be accessed easily through open data portals and many software applications can be developed at a rapid pace. Open data has been regarded as valuable assets as more businesses have embraced the idea of open data to foster innovation and revenue generation [1],[25].

With many open data portals made available, the decision on which portal to use often relies on the portal with the most relevant content. Content-based selection nowadays can be fulfilled by most search engines. Nevertheless, even though the relevant portals are known, data consumers are still left with the question of whether the sources in hand subscribe to open data quality standards [26]. In this case, the selection decision is no longer depends on the content of the open data portals alone. The concern on the quality of data that a particular open data portal provides is based on the effect that open data will impose on the end product of the data usage, such as in analysis or in drawing important decisions. While manual checking on open data portal's commitment to ensuring data quality is feasible (i.e through data policy) this is no longer an attractive option especially in the context where the number of data portals to consider is large. Thus, deciding which open data portal to use is a selection problem that calls for ways to search for relevant and high-quality portals at an acceptable time limit.

Optimization algorithms have a long track record in addressing selection problems. The principle of optimization that aims to maximize output and to minimize the computational time while sustaining the right leverage between quality and performance gained makes it a promising way to deal with the selection problem [2], [3]. The optimisation algorithm is used to improve the performance of the selection process. The selection process becomes a hard problem due to the huge size of data. When the performance increased and computational cost is decreased, the stability of the selection process is reached [4]. To achieve optimal performance, an optimization algorithm should be able to

[1] https://www.oecd.org/about/v

reduce the maintenance cost and also the sum of total running time [5].

There are two variants of optimization algorithms namely heuristic and metaheuristic. The metaheuristic algorithms such as Ant Colony Optimization (ACO), Genetic algorithm (GA) and P Swarm Optimization (PSO) are broadly applied in the optimization field and are commonly agreed to show better performance as compared to heuristic algorithms. Metaheuristics algorithms frequently used in solving complex optimization problems and the one that habitually imitating a few successful characteristics in nature are known as nature-inspired [6]. Metaheuristics algorithms also offer discovering search space to find optimal or near best solutions [7].

In this article, we explore the capability of a metaheuristic algorithm namely Ant Colony Optimization (ACO) in dealing with selection problems specifically for open data portals where the quality aspect of the portals is taken into consideration.

In the next section, a review of related work on metaheuristic algorithms and source selection will be presented. Section 3 covers the description of the proposed model and implementation setup. Section 4 covers the results and finally, Section 5 concludes this article.

2. RELATED WORKS

Nowadays, the trend of using global search algorithms or metaheuristic algorithms such as the Genetic algorithm, P swarm optimization algorithm, and Ant colony optimization (ACO) algorithm is increasing from other search algorithms. This is because searching the high-quality solutions within a realistic time can be improved by the global search algorithms [8].

Difficult combinatorial problems such as quadratic assignment, a traveling salesman, and scheduling have been solved by the ACO successfully. Chen et al., (2010) approved that heuristic information is not able to guide search to the optimal minimal subset but ants can obtain the finest feature combinations due to their capability to traverse the graph [9]. ACO is commonly used in the field of a traveling salesman to search the shortest path. Nevertheless, the algorithm is not only valid in finding the shortest path but also in other scopes such as selection and scheduling. ACO has become more acceptable and is broadly used due to its ability in searching the optimal result with a good performance. However, the role of ACO in selection problems particularly in data sources selection has received less attention.
Even though the application of metaheuristic algorithms in solving data sources selection is not widely explored, researchers found several reasons that will make the algorithms useful [10], [11]. Firstly, the increasing amount of data sources under consideration makes the selection process more difficult and complex. Thus, this problem is difficult to handle specifically by analytical methods. Secondly, evolutionary algorithms have a strong capability in solving optimization and search problems. Moreover, as the algorithm can be useful in finding the near-optimal sources from all existing sources, the data sources selection problem is considered as one of the searches and optimization problems. The third reason is searching for the selection of data sources based on the highest relevance in a huge space is important by exploring and exploiting each region around the search space. This condition is suitable for evolutionary algorithms to handle.

The best example of the role of metaheuristic algorithm that we can highlight is the contribution of GA in two research works. Kumar, Singh, and Kumar (2015) propose search engine selection on the relevance between the user query and search engine [12]. Lebib, Mellah, and Drias (2017) deal with the problem of obtaining the best selection in distributed information retrieval for large space of multi-sources [11]).
GA is widely used in scientific communities because of its usefulness and simplicity of implementation. This algorithm's robustness and efficiency have been tested in outperforming the analytical methods particularly involving huge datasets. Besides, the algorithm can search good quality solutions with the operators such as crossover and mutation where the highest relevance of sources can be gained by exploring and exploit each region of search space [11].

In research work by Kumar, Singh, and Kumar (2015) the selection criteria are made on search engines to recognize the useful content for a user query so that the appropriate search engine is selected [12]. The selection process is driven by the problems of retrieving irrelevant query results. They suggested that, if the data is ranked, relevant information can be easily determined.

In a multi-sources' environment, users are required to search relevant sources to fulfil their information needs. The efficiency of search results is usually affected by the existence of irrelevant documents and undesired information. Source selection efficiency is crucial especially in the case where the number of relevant sources is so small relative to the number of the available sources.

The contribution of metaheuristic algorithms that can be seen in finding relevant sources based on their contents and their strength in dealing with a huge number of sources makes them an appropriate choice. Moreover, the problems of data sources selection are similar to problems of search and optimization problems where the aim is to find the

near-optimal solutions or sources from the huge number of available sources [10].
Nevertheless, as there is a growing concern on finding relevant sources based on the source's quality, further exploration of metaheuristic algorithms is needed due to its limited coverage.

Beyond the scope of metaheuristic algorithms, there are existing works that deal with source selection. For example, Deng (2017) proposes the use of a probability model to rank deep web data sources where rely on source correlation and the content of documents [13]. Lin, Wang, Li, and Gao (2019) propose the use of a greedy algorithm where recall and precision measures are used to evaluate the content (contribution) of data sources for big data integration [14].
Awareness in considering quality in source selection can be seen as early as 2005 when Knight and Burn (2005) propose a model called IQIP to incorporate quality criteria in web search engines [15]. Later in 2009, there is a proposal for deep web source selection using quality criteria and three quality dimensions namely completeness, consistency, and size of data source are used to measure the sources [16]. The importance of data completeness also has been highlighted in special purpose database [27]. Neumaier, Umbrich, and Polleres (2016) propose quality assessment to support open data portals selection. In their work, the assessment is based on the source's metadata where quality metrics are developed to assess open data portals [17]. In 2018, a source selection model called SOURCERY which is based on user preference and the quality of the source content has been proposed [18][19]. In this work, users are offered more flexibility in determining specific information that is relevant to their needs. The quality dimensions that are considered are correctness, relevance, usefulness, consistency, conciseness, and interpretability.

In the light of existing work on source selection, it appears that an extension is needed in understanding the capability of metaheuristic algorithms in quality-based data source selection. Existing works on quality-oriented source selection provide room for improvement by considering a greater amount of quality dimensions against a greater amount of data sources.

Thus, in this article, we present an investigation of a metaheuristic algorithm namely ACO in addressing the limitation of the current work. ACO has been recognized as an efficient algorithm in solving selection problems such as materialized views and feature selection [8][20][21]. However, the question of how ACO can be used in quality-based data source selection is an open problem despite its reputation in improving runtime, accuracy, and computational complexity. In the next section, the implementation of ACO in open data source selection will be presented.

3. QUALITY-BASED OPEN DATA SOURCES SELECTION USING ACO

In conducting open data portal selection using ACO, we implemented the steps as shown in Figure 1. There are four main stages involved in selecting the OGD portals namely Data Extraction, Quality Measurements, Filtering, and Selection. 40 Open Government Data (OGD) Portals which represent 40 countries are randomly selected during the implementation of ACO.

Three selection criteria of the portals are considered namely: 1) the coverage of data quality requirements, 2) the readiness aspect of the government in providing the content of open data, and 3) the content of the government data available in the portals. The required data are prepared during the data extraction stage to fulfill the selection criteria. To measure the coverage of data quality requirements, we first extract data quality requirements stated in OGD policies provided by the portals. OGD Quality requirements are also known as OGD principles in OGD policy. OGD policy is endorsed by a government that agrees to commit open data to guide the public and OGD contributors in fulfilling the requirements of open data shared in the government's web portals. Knowing the requirements stated in the policy is helpful in understanding governments' expectations in realizing open government initiatives (Mexico Open Government, 2016). There are initially 36 raw OGD quality requirements extracted from OGD policies of 31 portals (of 25 countries). After filtering synonyms and related terms in similarity analysis, the total number quality criteria is reduced to 15, which are Public, Machine readability, Timely, Accessibility, Completeness, Reusable, License-free, Standard formats, Primary, Permanence, Non-discriminatory, Manageable, Trusted, Protected and Non-proprietary (A similar extraction process can be referred from our early work in [22]).

Next, the readiness scores data set of open data providers are retrieved from Open Data Barometer (ODB) organization where the organization measures policy readiness as one of the criteria to evaluate the country's commitment to adopting the International Open Data Charter Principles[2] or the G20 Anti-Corruption Open Data Principles (Open Data Barometer Organization, 2017). The readiness scores will become the input for the filtering stage together with the data quality coverage scores (that will be measured in the quality measurement stage).

[2] List of open data principles in Open Data Charter: https://opendatacharter.net/principles/

In the data extraction stage, we also extract the Global Open Data Index scores data set provided by the Open Knowledge Network that measures 15 categories[3] of government data that indicate the "openness" of a particular government. The score shows the extent of the content of the government data available in the portals. The categories are Government Budget, National Statistics, Procurement, National Laws, Administrative Boundaries, Draft Legislation, Air Quality, National Maps, Weather Forecast, Company Register, Election Results, Locations, Water Quality, Government Spending, and Land Ownership. This data set will be used during the selection stage. In the measurement stage, data cleaning against the raw OGD quality requirements as mentioned earlier is performed before data quality requirements coverage scores are computed. A similar coverage measure (as used in [24]) is adapted to measure the coverage of quality requirements of OGD portals as follows:

$$p_i = |QR_i| \cap |UQR| / |UQR| \quad (1)$$

where,

- P is a set of OGD portals under measure
- p_i is the ith portal in P,
- QR_i is the set of quality requirements of p_i
- UQR is the union of quality requirement sets of P

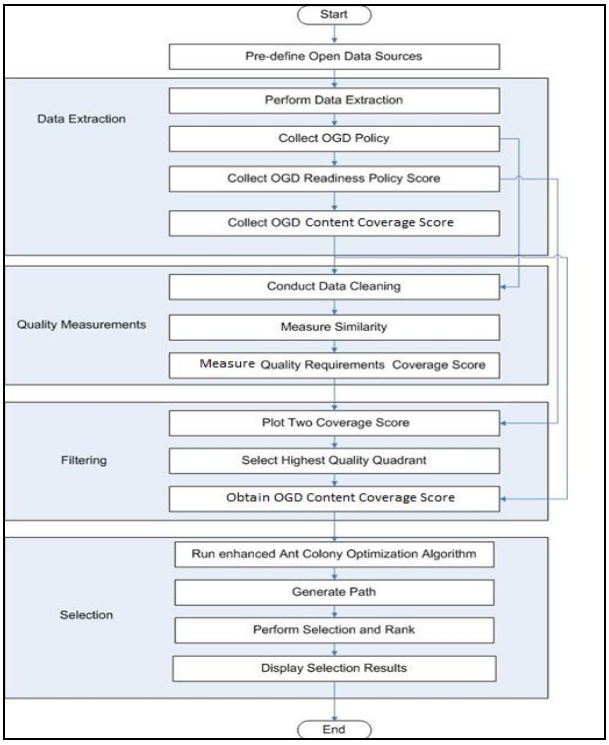

Figure. 1: Implementation Steps of OGD Portals Selection

Next, in the filtering stage, the OGD portals with the high score of both quality requirement coverage and readiness score will be retained in for the next process. The results are plotted into a quadrant graph where scores from 50% to 100% represent high scorers, while scores below 50% are considered low scores. As one might notice, the number of data portals under consideration will become smaller, which will give an advantage for the algorithm to run faster with a small workload.

Finally, in the selection stage, the ACO algorithm will get the content coverage scores for 15 categories and the set of filtered OGD portals as the inputs for further process. The proposed algorithm is enhanced through the adaptation of a directed graph that will reduce the time complexity of ant in traversing the graph. The original ACO algorithm allows the ant to construct a solution randomly, where all nodes are visited. However, the enhanced ACO algorithm in this study has been improved by allowing the ant traversing the directed graph so that the ant does not need to visit all of the nodes to construct the solution. The enhanced algorithm will follow the directed graph based on the value of quality attributes and multi-layer arrangements of nodes. In this way, the ant is designed to visit all nodes layer by layer and not randomly visited all nodes without any direction. This characteristic allows the algorithm to reduce the computing time during the selection process. Figure 2 illustrates the enhancement made on ACO for OGD portals selection.

The flow can be summarized into several steps. It starts with the initialization of the optimization problem and parameters. These include the number of ants, the maximum value of iteration, pheromone level initialization value, and the ant's activity. Then, the solution procedure will be executed which involved the selection of nodes (in this case the OGD portals based on the directed content coverage).

In this step, the algorithm will select the node with the highest score as the local solution result. Selected nodes that are visited by each ant will be used to compute the global solution results. In our case, the global solution result is the path of OGD portals visited most. We use a frequency measure (in percentage) to measure the number of visits for each portal as follows:

$$f(p_i) = fp_i / n \times 100, \quad (2)$$

where p_i is the ith portal in the set of filtered OGD portals, fp_i is the number of visits for pi and n is the total number of the content category.

Once the solution criteria are met, the optimal solution is yielded, and this will end the process. Otherwise, the pheromone will be updated and being released. The best global score is defined by the highest frequency score that will determine the ranking of the portals. We use MATLAB software to aid in interpreting the ACO algorithm path.

[3] https://index.okfn.org/dataset/

In the next section, the results of implementing OGD portals selection based on the method described will be presented.

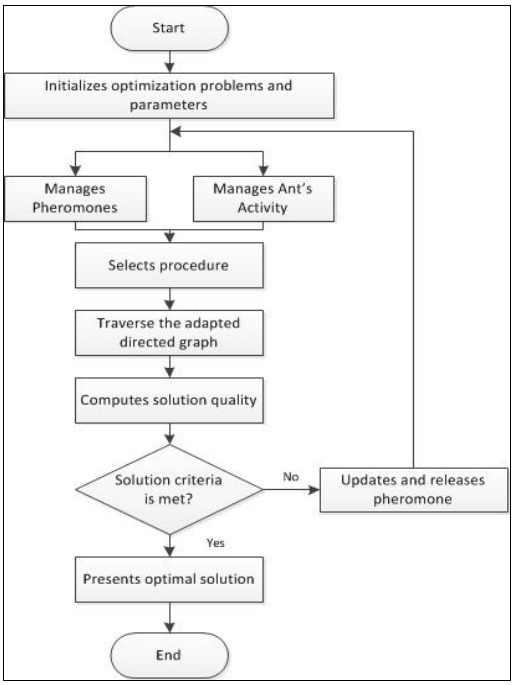

Figure. 2: Flow of the Enhanced Ant Colony Optimization Algorithm [23]

4. RESULTS AND DISCUSSION

The result of the filtering stage is as shown in Figure 3. The scope of the results is 40 countries that have been randomly selected in the assessment. Out of 40 portals, 8 portals exhibit high scores in both policy readiness and quality requirement coverage (in quadrant Q1). These are Canada, Netherlands, Australia, Italy, Albania, Germany, India, and Malaysia.

Most of the OGD portals belong to Q2 quadrants. In this quadrant, these portals are highly prepared in their open data policy but, but lack of quality requirement coverage in their policy. Five OGD portals with missing policy readiness scores are Oman, Afghanistan, Taiwan, Iran and Bhutan. This is because these countries are not yet included in ODB assessment. Nevertheless, Oman shows a high score for quality requirements coverage with 73.33% (see Q4). Our concern is for countries in the Q3 quadrant that show low scores in both criteria. Details of the results for all 40 countries is available in the Appendix, where countries that belong to Q1 are highlighted.

Different filtering results can be yielded depending on the threshold values set for the quadrants. As in this article, we set scores from 50% to 100% as high scores, while scores below 50% are as low scores.

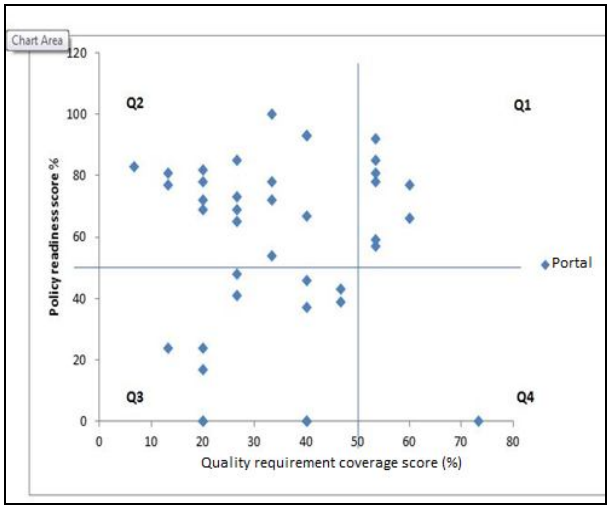

Figure. 3: Quadrant Graph from Filtering Stage

Figure 4 illustrates the generated path during the selection stage once ACO algorithm is implemented. This stage starts with a source node (Node 1) and ends with a destination node (Node 122). Each node represents a portal's content coverage score, while each layer represents a content category. The colored nodes are the node with the highest content coverage score in the layer (category) they belong to.

To illustrate further, Figure 5 shows the selection path generated by MATLAB software. For example, the local solution for the first layer which is for the Air Quality category is node 62 that belongs to India's OGD portal. Using ACO algorithm, the local solution is retrieved layer by layer before the final global solution is met.

The final global solution results are as shown in Table 1. The results show the top three portals based on their frequency scores. The top portal is Australia with a 33.33% frequency score. This shows that Australia is the best OGD portal in terms of its policy readiness, quality requirement coverage, and also the content of government data. Four portals namely Italy, India, Netherlands, and Canada share the second rank with 13.33%, followed by Albania and Germany at the 3rd place. In implementing ACO algorithm for OGD portal selection, one can customize not only the filtering criteria and selection criteria, but also the threshold for the quadrants.
For example, more portals can be put under consideration for the selection stage by setting a lower threshold value for the quadrants. Different results will be yielded with such customization. Another possible customization is on the filtering criteria that are used to rule out the portals that fail to meet the condition.

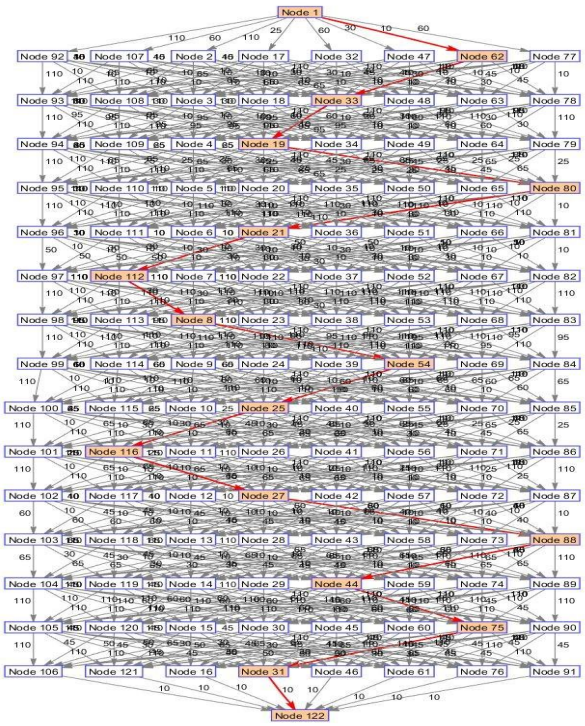

Figure. 4: The Flow of ACO Algorithm in Selection Stage

Table 1: The Global Solution Results

Ranking	OGD Portals	Frequency score (%)
1st	Australia	33.33
2nd	Italy, India, Netherlands, Canada	13.33
3rd	Albania, Germany	6.67

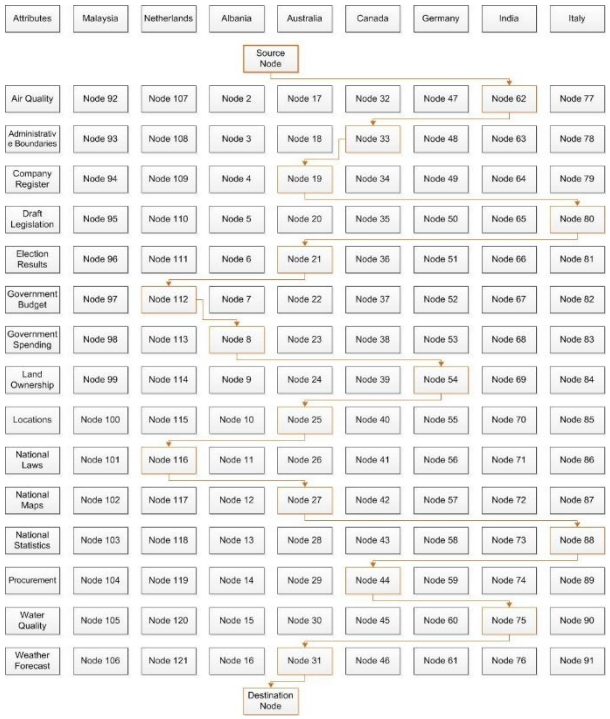

Figure 5: Selection Results Illustration in MATLAB

The results also suggest that the countries from the European region are within the top three countries that exhibit high scores in the assessment. In practice, the results can be used by open data consumers in deciding which OGD portal to use in retrieving the government data sets. As for the data provider, OGD portals can use the results to improve the coverage and quality of the data they provide to the public.

5. CONCLUSION

In conclusion, this article presented the results of implementing OGD portals selection by considering the quality aspects of the portals. We described how the selection problem in this domain can be addressed by incorporating quality criteria within a well-known meta-heuristic algorithm called Ant Colony Optimization (ACO). The results of the analysis made against OGD policies from 30 Open Government Data (OGD) Portals help us to establish the set of data quality requirements that are used in the filtering process. The selection model implemented in this study has demonstrated how one can customize and set selection criteria beyond the three parameters used within the scope of this article. The feature of the ACO algorithm that allows the generation of selection paths based on the parameters offers one to deal with multi-sources and multi-criteria inputs. The results that suggest the position of the countries (and regions) in the selection ranking contribute towards initiating the improvement for OGD portals that will benefit open data consumers globally.

APPENDIX

No	OGD Portals	Coverage of Quality Requirements	Readiness Scores
1	Oman	73.33	-
2	Afghanistan	40	-
3	Taiwan	40	-
4	Bhutan	20	-
5	Iran	20	-
6	Great Britain	33.33	100
7	United States	40	93
8	New Zealand	40	93
9	Canada	53.33	92
10	Netherlands	53.33	85
11	British Columbia	26.67	85
12	Mexico	6.67	83
13	Singapore	20	82
14	Australia	53.33	81
15	Norway	13.33	81
16	Italy	53.33	78
17	Ireland	33.33	78
18	Austria	20	78
19	India	60	77
20	Japan	13.33	77
21	Czech	26.67	73
22	Belgium	33.33	72
23	Sweden	20	72
24	Philippines	26.67	69
25	Brazil	20	69
26	Russia	40	67
27	Malaysia	60	66
28	Finland	26.67	65
29	Albania	53.33	59
30	Germany	53.33	57
31	Denmark	33.33	54
32	Chile	26.67	48
33	Ukraine	40	46
34	Bangladesh	46.67	43
35	Indonesia	26.67	41
36	Argentina	46.67	39
37	Saints Lucia	40	37
38	Tunisia	20	24
39	South Africa	13.33	24
40	Latvia	20	17

ACKNOWLEDGEMENT

We would like to thank the Centre for Advanced Computing Technologies (CACT), Fakulti Teknologi Maklumat Dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM) for supporting this research, the Open Data Barometer (ODB) and Open Knowledge Network for providing the data sets.

REFERENCES

1. F. Zeleti and A. Ojo, "**Capability Matrix for Open Data**," in *15th Working Conference on Virtual Enterprises (PROVE)*, 2016.
2. S. Zagorac and R. Pears, "**Web materialization formulation: Modelling feasible solutions**," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014.
3. D. De, D. Banerjee, S. Mukherjee, and J. G. Dastidar, "**A Simplistic Mechanism for Query Cost Optimization**," *Int. J. Adv. Comput. Res.*, vol. 5, no. 19, pp. 205–211, 2015.
4. Alka and A. Gosain, "**A Comparative Study of Materialised View Selection in Data Warehouse Environment**," in *2013 5th International Conference and Computational Intelligence and Communication Networks*, 2013, pp. 455–459. https://doi.org/10.1109/CICN.2013.100
5. R. P. P. Karde and R. V. M. Thakare, "**Materialized View Selection Approach Using Tree Based Methodology**," *Int. J. Eng. Sci. Technol.*, vol. 2, no. 10, pp. 5473–5483, 2010.
6. X. S. Yang, ***Nature Inspired Metaheuristics Algorithms***. 2010.
7. I. Khennak, H. Drias, and S. Kechid, "**A New Modeling of Query Expansion Using an Effective Bat-Inspired Optimization Algorithm**," *IFAC-PapersOnLine*, vol. 49, no. 12, pp. 1791–1796, Jan. 2016.
8. X. Zhao, D. Li, B. Yang, C. Ma, Y. Zhu, and H. Chen, "**Feature selection based on improved ant colony optimization for online detection of foreign fiber in cotton**," *Appl. Soft Comput. J.*, vol. 24, pp. 585–596, 2014. https://doi.org/10.1016/j.asoc.2014.07.024
9. Y. Chen, D. Miao, and R. Wang, "**A rough set approach to feature selection based on ant colony optimization**," *Pattern Recognit. Lett.*, vol. 31, no. 3, pp. 226–233, 2010.
10. P. Bhatnagar and N. K. Pareek, "**A Combined Matching Function based Evolutionary Approach for development of Adaptive Information Retrieval System**," vol. 2, no. 6, 2012.
11. F. Z. Lebib, H. Mellah, and H. Drias, "**Enhancing information source selection using a genetic algorithm and social tagging**," *Int. J. Inf. Manage.*, vol. 37, no. 6, pp. 741–749, 2017.
12. R. Kumar, S. K. Singh, and V. Kumar, "**A Heuristic Approach for Search Engine Selection in Meta-search Engine**," in *International Conference on Computing, Communication and Automation, ICCCA 2015*, 2015, no. July 2015, pp. 865–869. https://doi.org/10.1109/CCAA.2015.7148496
13. S. Deng, "**Non-Cooperative Deep Web Data Source Selection Based on Subject and Probability Model**," *J. Softw.*, vol. 28, no. 12, pp. 3241–3256, 2017.

14. Y. Lin, H. Wang, J. Li, and H. Gao, "**Data source selection for information integration in big data era**," *Inf. Sci. (Ny).*, vol. 479, pp. 197–213, 2019.
15. S.-A. Knight and J. Burn, "**Developing a Framework for Assessing Information Quality on the World Wide Web**," *Inf. Sci.J .*, vol. 8(5), pp. 159–172, 2005.
16. X. F. Xian, P. P. Zhao, W. Fang, J. Xin, and Z. M. Cui, "**Quality-based data source selection for web-scale deep web data integration**," in *Proceedings of the 2009 International Conference on Machine Learning and Cybernetics*, 2009, vol. 1, no. 75, pp. 427–432.
17. S. Neumaier, J. Umbrich, and A. Polleres, "**Automated quality assessment of metadata across open data portals**," *J. Data Inf. Qual.*, vol. 8, no. 1, pp. 1–29, Oct. 2016.
18. E. Abel *et al.*, "**User driven multi-criteria source selection**," *Inf. Sci. (Ny).*, vol. 430–431, pp. 179–199, Mar. 2018.
19. E. Abel *et al.*, "**Sourcery: User driven multi-criteria source selection**," in *International Conference on Information and Knowledge Management, Proceedings*, 2018, no. 1, pp. 1947–1950.
20. R. Goswami, D. K. Bhattacharyya, and M. Dutta, "**Materialized view selection using evolutionary algorithm for speeding up big data query processing**," *J. Intell. Inf. Syst.*, Mar. 2017.
21. N. A. M. Sabri, N. A. Emran, and N. Khushairi, "**Materialized Views Selection Algorithm for Cyber Manufacturing**," vol. X, no. X, pp. 1–5, 2018.
22. N. A. M. Sabri, N. A. Emran, and N. Harum, "**Government open data portals: A measurement of data veracity coverage**," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 12, pp. 1975–1983, 2019.
23. N. A. M. Sabri, N. A. Emran, and N. Abdullah, "**Quality-Based Open Data Source Selection Using Ant Colony Optimization (ACO) Algorithm**," *Int. J. Emerg. Technol.*, vol. 11, no. 3, pp. 1164–1168, 2020.
24. N. A. Emran, "**Data completeness measures**," in *Advances in Intelligent Systems and Computing*, vol. 355, 2015, pp. 117–130.
25. D. W. Jacob, M. F. M. Fudzee, M. A. Salamat, and N. H. A. Rahman, "**Analyzing the barrier to open government data (OGD) in Indonesia**," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 1.3S1, pp. 136–139, 2019. https://doi.org/10.30534/ijatcse/2019/2681.32019
26. P. Kadu and A. V. Zadgaonkar, "**Knowledge extraction from text document using open information extraction technique**," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 2, pp. 2280–2283, 2020. https://doi.org/10.30534/ijatcse/2020/208922020
27. N. A. Emran, S. Embury, and P. Missier, "**Measuring Population-Based Completeness for Single Nucleotide Polymorphism (SNP) Databases**," in *Advanced Approaches to Intelligent Information and Database Systems*, J. Sobecki, V. Boonjing, and S. Chittayasothorn, Eds. Cham: Springer International Publishing, 2014, pp. 173–182.