International Journal on Emerging Technologies **11**(3): 1164-1168(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Quality-Based Open Data Source Selection Using Ant Colony Optimization (ACO) Algorithm

Nor A.M. Sabri[1], Nurul A. Emran[2] and Noraswaliza Abdullah[3]

[1]Centre for Advanced Computing Technologies (CACT), Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia (UTeM), Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia.
[2]Associate Professor, Centre for Advanced Computing Technologies (CACT), Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia (UTeM), Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia.
[3]Senior Lecturer, Centre for Advanced Computing Technologies (CACT), Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia (UTeM), Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia.

(Corresponding author: Nurul A. Emran)
(Received 15 April 2020, Revised 18 June 2020, Accepted 20 June 2020)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Open data sources have been increasingly used by web data consumers due to its convenience in accessibility, variety of contents and free of charge attractions. Nevertheless, the decision to use the content of open data sources should also be based on quality. The challenge faced by open data community is on deciding the quality of open data sources. In this paper, we propose a model for open data source selection using quality features which are extracted using web data crawler. The model adopted a well-known meta-heuristic algorithm called as Ant Colony Optimization (ACO) which will perform an automated and seamless open data source selection based on the quality features. The implementation of this model will benefit search engines and open data consumers at large in selecting high quality open data sources.

Keywords: ant colony optimization, data quality, data source selection, open data.

Abbreviations: ACO, Ant Colony Optimization.

I. INTRODUCTION

Nowadays, open data sources are widely accepted by the community and being used in numerous fields. This current trend showed that open data sources are useful and helpful for software application development. Open data sources provide a free of charge service for everyone [1] without any consequences on copyrights or patents. Furthermore, the availability on retrieving the data freely have become the easiest way for everyone on gathering data in terms of time and cost. Subsequently, the services and applications based on open data are predicted to increase in the future[2]. The open data are legal to use as a secondary and legible mechanically [3]. Many governments nowadays are interested in sharing the data openly with the public sectors [1]. As a consequence, this situation offers new business openings to those who providing data, consuming the data and developing innovative services and applications using the data. In addition, many entrepreneurs mostly in IT companies agree on the importance of open data in achieving betterment of citizens in figuring out the current issues in a city [4].

However, there are some issues regarding on the open data that concern most of the researchers on the related fields. Firstly, the semantic of the open data schema. To understand the schema and the definition of data in the table is time consuming due to the lack of structures documentation on data dumps provided by open data source.

A more pressing issue is regarding the quality of open data sources. Observing the quality of data is important especially for open data as the data are used widely by everyone whenever these data are accessible. The importance of integrating the quality data also has been addressed from the context of cooperative data sources in healthcare domain in terms of completeness aspect [5]. The absence of information on the open data sources quality has left the public to choose their open data sources solely based on the content that might be low in quality. The efforts to filter open data at the data consumer's site require additional data processing and data cleaning. In fact, data cleaning would be the most complicated task, if we are dealing with the poor quality data [6].

Hence, there is a need to support the public data consumers in using open data sources through automated and seamless selection which considers the quality criteria of these data sources. This model, once it is implemented, it would benefit the search engines (such as Google) to rank open data sources based on quality features (in addition to the keyword-based search). In fact, the data quality concept is a well-known concept and defined as a measure of potential usefulness, clarity, correctness, and trustworthiness of data as well as datasets [7]. Even though there is no common standard of description for data quality[8][9]several quality dimensions have been mentioned in the literature such as accuracy, consistency, availability, completeness [9-11], conformance, credibility, process ability, relevance and timeliness [12-15]. The quality of data is determined by its fitness for reuse by data consumers [15].

Nevertheless, study on how meta-heuristics algorithm such as Ant colony optimization (ACO) can deal with open data sources selection is very limited. Thus, its benefits (and limitations) remain undiscovered. Hence, in this paper we are proposing a model for an open data

source selection using ACO, where the quality aspects are considered.
This limitation can also be seen in data source selection studies where methods other than meta-heuristics algorithms are used. For example, Deng (2017) applies the probability model in ranking deep web data sources [16]; Greedy algorithm has been used to assess the content of data sources in integrating big data [17]; IQIP model for web search engines evaluation [17]; the use of multi-agent in selecting digital source using the semantic web [18].
In the next section, the model of quality-based open data sources selection using ACO will be presented. Section III covers open data source scraping using open source web crawling application. Section IV presents an adapted ACO algorithm for open data source selection and finally, Section V concludes this paper.

II. QUALITY-BASED OPEN DATA SOURCES SELECTION USING ACO

As shown in Fig. 1, there are three stages in the model namely Scraping, Quality measurement and Selection. The input of the model is a set of pre-defined open data sources. Data sources for this model are identified based on the URLs of open data web sites. These URLs are used to perform scraping using web crawlers. Scraping process is initiated once the input is received.

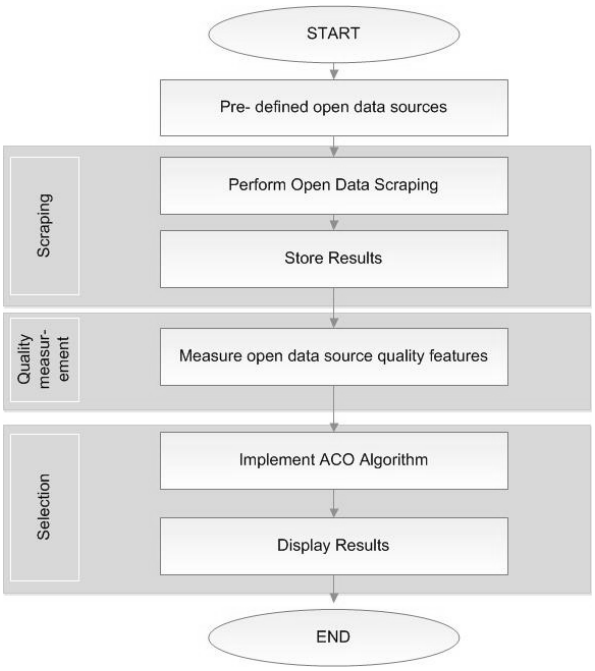

Fig. 1. The Proposed Model for Quality-based Open Data Source Selection using ACO.

The results obtained from the scraping process are analyzed to measure the quality of the pre-defined open data sources based on a set of quality dimensions (for examples accuracy, completeness and consistency). The results of the multidimensional quality measurements become the input for ACO. In the selection stage, ACO will perform selection and suggest the optimal results as well as the ranking of the open data sources.
The rank of the data sources will be based on the quality measures from all the open data web sites under the measure.
In the next section, details on the scraping phase will be provided.

III. SCRAPING OPEN DATA SOURCES USING WEB CRAWLER

In order to extract the information from the open data web sites, scraping technique is used. In particular, as most open data sources are on the web, these data sources can be accessed through a web scraper (with some programs and pre-written libraries) [19]. In the meantime, web scraping is heavily used for web indexing. Web indexing is supported by bot or web crawler and known as universal technique by most search engines. Web scraping transforms the unstructured data into structured data format. Then the data will be stored and analyzed in a central local database or spreadsheets.
Web crawler uses one or more starting references or a seed list to initialize a queue for frontier navigation. The full content associated with each reference can be retrieved based on the list. To find the desired web pages, the crawlers will employ search engines to exhaustively crawl the web to find and index web pages to serve user queries efficiently [20]. All the web page contents are recorded including the related web links and requirements [21]. The process of gathering the requirements usually involves multiple stakeholders and tools [22].
In the proposed model, the scraping process is performed using a web scraping technique based on a web crawler engine called as Scrapy. Scrapy is an open-source application framework for crawling web sites. It is also known for structured data extraction for a large range of useful applications (for example data mining, information processing or historical archival). Scrapy is claimed to be useful in synchronous request processing and scheduling. This means a new request can be accepted and processed without having to wait for the previous request to be completed. Moreover, if any requests have failed, other requests will not be affected. Therefore, fast crawls can be achieved.
Scrapy is used for extracting data using an APIs or as a web crawler for general-purpose [23]. Scrapy is also known as a python-based framework tool developed for web extracting data. Scrapy contains some predefined libraries to enable us to perform data extraction from online which makes work becomes easier [24]. The reason of using Scrapy is due to its ability to understand broken HTML. Scrapy also has a vibrant community which can help users to learn and use it easily. In addition, Scrapy is maintained by the community and is a well-organized code. Besides, Scrapy grows in features and stability fixes [23]. There are strong reasons to use Scrapy as it is built upon years of experience in extracting massive amounts of data efficiently and in a robust manner. There are several methods available in extracting data in addition to Scrapy as shown in Table 1. Based on the table we can see that Scrapy is a popular tool for web data extraction in various applications i.e. ventilation, power generation, food drying [3-5].

IV. AN ADAPTED ACO ALGORITHM FOR OPEN DATA SOURCE SELECTION

In the proposed model, the quality criteria are considered as "features" of interest that will determine the optimal solution in ACO. Feature selection is defined as the process in choosing a subset of features from the original set of features which is forming patterns in a given dataset [25]. Feature selection is one of the most fundamental problems especially in the field of machine learning [26]. This problem usually deals with high dimensional space of features [27]. Primarily, there are three aspects that become the aims of feature selection which are the reduced cost of extracting features, improvement of the classification accuracy and performance reliability [27]. Feature selection is important in order to reduce the problem size and search space for learning algorithms. Feature selection has become a method to treat abundant amount of noise, irrelevant or misleading features for its ability in handling the inconsistent and imprecise information in real world problems [14].

Table 1: Web Data Extraction Methods.

Author	Scope	Method
Chen *et al.,* (2016) [28]	Collect traffic information from various open data website to learn and predict the patterns of traffic conditions	Present a deep learning approach with a stacked LSTM model as a particular type of Recurrent Neural Network (RNN) to predict traffic condition
Farah and Correal, (2013) Farah *et al*., (2014) [29, 30]	Analyze the information in GitHub in order to facilitate proper selection of software components based on repository mining technique	Develop a tool to support the analysis and selection of components and software applications in GitHub using Archnalyzer
Chang *et al*., (2015) [21]	Collect sale data from different online malls by using web crawlers by applying a data mining technique	Develop a Web-Crawler based sale management system to solve the inventory problem of the online malls and the distribution problem of physical malls
Bonifacio *et al*., (2015) [31]	Present a software tool to fetch, download and consolidate climate data on multiple web pages to easily access in a bulk format	Deconstruct URL and modifies the date parameters to download large volume of data, remove individual file headers, merge data file into one output file. The tool is coded as Microsoft Excel Macro
Rao *et al*., (2015) [24]	Analyse commodity price data available on various e-commerce sites	Introduce data scraping technique to collect data using Scrapy
Shi and Lin, (2016) [23]	Monitor news web pages using web crawler and stores the updated news in database	Implement incremental Python web crawler using Scrapy to crawls news web pages and then remove repetition web links using Bloom filter
Wang, (2012) [32]	Analyze online marketing transaction in e-commerce	Apply Scrapy crawl architecture to crawl Taobao shares platform to analyze relationship between sellers and buyers
Zhiqiang, (2015) [33]	Introduce the features of Tibetan language news in the field of search engine	Modify Scrapy and Solar to enhance user experience in using Tibetan search engine

Table 2: Methods Used in Feature Selection

Author	Scope	Result	Method/ Tool
Cadenas *et al*., (2013) [34]	Select selection to handle crisp and low quality data	• Classification accuracy • number of features selected • high dimensional datasets and low quality datasets	Fuzzy Random Forest
Schiezaro and Pedrini, (2013) [35]	Investigate, implement and analyze feature selection method to classification of different data sets.	• Reduce number of features • Classification accuracy	Artificial Bee Colony
Tabakhi et al., (2014) [36]	Identify unsupervised feature selection to find an optimal feature subset without using learning algorithms.	• Low computational Complexity	ACO
Bae et al., 2010) [37]	Modify Particle swarm optimization (PSO) called Intelligent Dynamic Swarm (IDS) for feature selection	• Search capability • Reduce features	PSO
Moradi and Rostami, (2015) [25]	Propose novel feature selection method based on graph clustering approach and Ant colony optimization for classification problems	• Classification accuracy • Number of features • Execution time	ACO
Xue *et al*., (2014) [38]	Develop a novel feature selection approach based on a new approach of PSO	• Number of features • Computational time	PSO
Inbarani *et al*., (2013) [39]	Supervise new feature selection methods based on a hybridization of PSO for disease diagnosis	• Classification accuracy • Number of features • Computational time	Hybrid PSO
Saraç *et al*., (2014) [40]	Select best feature in web pages	• Improve runtime • accuracy	ACO

Moreover, in machine learning, feature selection is an active research especially in the area of eliminating features with little or no predictive information and redundant features [34].

In performing the selection process, the proposed model applies metaheuristic algorithm-which is known as Ant Colony Optimization (ACO), where it is able to perform a selection on features. ACO is claimed to be great at decreasing the computational time due to the distributed problem-solving nature and its ability to be implemented in parallel. Moreover, ACO also has local search capabilities. Stochastic component of ACO explores the search space efficiently where the problem of being trapped in local minimum can be avoided [36]. In ACO, the better solutions are found by ants while updating pheromones. The pheromone used to decrease exploration ability in the algorithm as additional information. Moreover, due to the great level of self-organization and the ability to perform complex tasks, ACO is regarded as an intelligent entity. It also inspires many researchers to develop a new clarification for problem optimization in computer science [41, 42].

In addition to ACO, Table 2 shows the methods used for feature selection such as Particle Swarm Optimization (PSO) and Artificial Bee Colony. From these works, the accuracy of selection and selection speed are used to evaluate most of the selection methods. In order to understand how ACO works with quality features, we have adapted ACO accordingly (from ACO pseudo-code presented in [36]) as shown in Fig. 2.

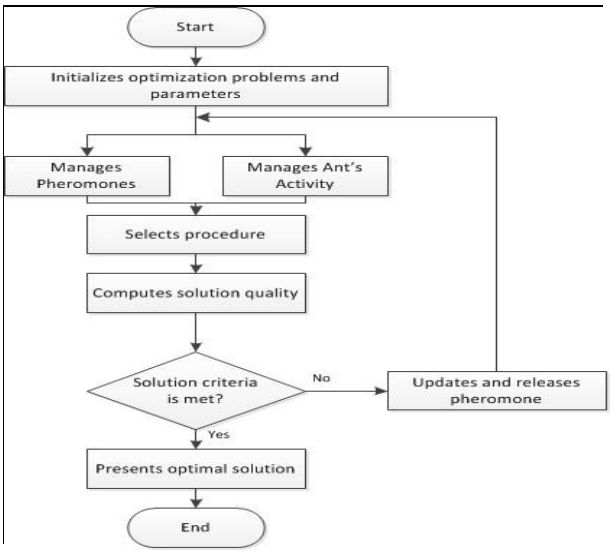

Fig. 2. Flow of ACO in Feature Selection.

Fig. 2 shows the flow of the ACO in feature selection based on three steps:

Step 1: Initialize the optimization problems and parameters of ACO including the numbers of ants, the maximum number of iterations, manage pheromone by initiating its level and the ant activity.

Step 2: Construct solution procedure, select node (data sources) randomly, select the best features (quality dimensions) and deselect low features. All nodes are visited by each ant to compute solution quality results.

Step 3: Meet the solution criteria by selecting the best feature subset. If the solution is met, the optimal solution results are gained, and the process is ended. Else, the pheromone is updated and released again. The global best subset is defined by the highest accuracy among all local best solutions.

V. CONCLUSION

As conclusion, this paper has presented the proposal of a quality-based open data source selection model using a well-known metaheuristic algorithm called Ant Colony Optimization (ACO). The model adopts an open-source web data crawling (Scrapy) to collect quality features of open data sources that are available on the web. ACO, which is known for its ability to deal with highly multidimensional features and robustness has been selected to perform the selection.

VI. FUTURE SCOPE

To understand the benefits (and limitations) of this model, its practicality must be evaluated and tested in our future work

ACKNOWLEDGEMENTS

The authors would like to thank the Centrefor Advanced Computing Technologies (CACT), Fakulti Teknologi Maklumat Dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM) for supporting this research.

Conflict of Interest. No.

REFERENCES

[1]. A. Immonen, M. Palviainen, and E. Ovaska (2014). Requirements of an Open Data Based Business Ecosystem. Access, IEEE, *2*, 88–103.

[2]. M. Saha and M. Singh (2017). Sustainable Urbanisation: An integrated approach towards future India. Int. J. Emerg. Technol., *8*(1), 84–90.

[3]. T. Maki, K. Takahashi, T. Wakahara, A. Yamaguchi, Y. Ichifuji, and N. Sonehara (2016). A new multiple label propagation algorithm for linked open data. *Proc. - 2016 10th Int. Conf. Innov. Mob. Internet Serv. Ubiquitous Comput. IMIS 2016*, 202–208.

[4]. H. Dong, S. Member, and G. Singh (2017). Open Data-Set of Seven Canadian Cities, 529–543.

[5]. F. N. M. Leza and N. A. Emran (2014). Data accessibility model using QR code for lifetime healthcare records. *World Appl. Sci. J.*, *30*(30), 395–402.

[6]. Q. Li and B. Li (2011). Mining open source software data using regular expressions. In CCIS2011 - Proceedings: *2011 IEEE International Conference on Cloud Computing and Intelligence Systems*, 550–554.

[7]. Robinson, R. L. M., Lynch, I., Peijnenburg, W., Rumble, J., Klaessig, F., Marquardt, C., & Karcher, S. (2016). How should the completeness and quality of curated nanomaterial data be evaluated? *Nanoscale*, *8*(19), 9919-9943.

[8]. Frank, M., & Walker, J. (2016). User Centred Methods for Measuring the Value of Open Data. *The Journal of Community Informatics*, *12*(2), 47-68.

[9]. N. A. Emran, S. Embury, and P. Missier (2008). Model-driven component generation for families of completeness. In *6th International Workshop on Quality in Databases and Management of Uncertain Data, Very Large Databases (VLDB)*.

[10]. Emran, N. A. (2015). Data completeness measures. In *Pattern Analysis, Intelligent Security and the Internet of Things* (pp. 117-130). Springer, Cham.
[11]. Xie, H., Li, L., & Xuan, P. (2019). An Effective Source Selection Algorithm for Filling Missing Tuples. In *2019 IEEE International Conference on Power Data Science (ICPDS)* (pp. 91-95). IEEE.
[12]. Herrera-Viedma, E., Pasi, G., Lopez-Herrera, A. G., & Porcel, C. (2006). Evaluating the information quality of web sites: A methodology based on fuzzy computing with words. *Journal of the American Society for Information Science and Technology*, *57*(4), 538-549.
[13]. D. UK, (2013). The Six Primary Dimensions for Data Quality Assessment.
[14]. J. Ma, Q. Wang, C. Dong, and H. Li (2017). Data Descriptor: The research infrastructure of Chinese foundations, a database for Chinese civil society studies, 1–7.
[15]. M. Dekkers, N. Loutas, M. De Keyzer, and S. Goedertier (2014). Presentation metadata Open Data & Metadata Quality. *2014 Eur. Comm.,* 1–37.
[16]. S. Deng (2017). Non-Cooperative Deep Web Data Source Selection Based on Subject and Probability Model. *J. Softw., 28*(12), 3241–3256.
[17]. Y. Lin, H. Wang, J. Li, and H. Gao (2019). Data source selection for information integration in big data era. *Inf. Sci. (Ny)., 479,* 197–213
[18]. J. de Mooij, C. Kurtan, J. Baas, and M. Dastani (2020). A multiagent framework for querying distributed digital collections. In ICAART 2020 - Proceedings of the *12th International Conference on Agents and Artificial Intelligence.*
[19]. D. K. Mahto and L. Singh (2016). A Dive into Web Scraper World, 689–693.
[20] S. K. S. (2014). A Service Crawler Framework for Similarity Based Web Service Discovery.
[21]. Y. J. Chang, F. Y. Leu, S. C. Chen, and H.L. Wong (2015). Applying Web Crawlers to Develop a Sale Management System for Online Malls. *2015 9th Int. Conf. Innov. Mob. Internet Serv. Ubiquitous Comput.,* 408–413.
[22]. Anuar, U., Ahmad, S., & Emran, N. A. (2015). A simplified systematic literature review: Improving Software Requirements Specification quality with boilerplates. In *2015 9th Malaysian Software Engineering Conference (MySEC)* (pp. 99-105). IEEE.
[23]. Z. Shi and W. Lin, (2016). The Implementation of Crawling News Page Based On Incremental Web Crawler, 348–351.
[24]. M. K. Rao, R. Lagisetty, M. S. V. K. Maniraj, K. N. S. Dattu, and B. S. Ganga, (2015). Commodity Price *Data Analysis Using Web Scraping, 4*(4), 146–150.
[25]. P. Moradi and M. Rostami (2015). Knowledge-Based Systems Integration of graph clustering with ant colony optimization for feature selection. *Knowledge-Based Syst., 84*, 144–161.
[26]. Y. Chen, D. Miao, and R. Wang (2010). A rough set approach to feature selection based on ant colony optimization. *Pattern Recognit. Lett.*, *31*(3), 226–233.
[27]. X. Zhao, D. Li, B. Yang, C. Ma, Y. Zhu, and H. Chen (2014). Feature selection based on improved ant colony optimization for online detection of foreign fiber in cotton. *Appl. Soft Comput. J.*, *24*, 585–596.
[28]. Y. Chen, Y. Lv, Z. Li, and F. Wang (2016). Long Short-Term Memory Model for Traffic Congestion Prediction with Online Open Data. *Int. Conf. Intell. Transp. Syst.*
[29]. G. Farah and D. Correal (2013). Analysis of intercrossed open-source software repositories data in GitHub. *2013 8th Comput. Colomb. Conf. 8CCC 2013*.
[30]. G. Farah, J. S. Tejada, and D. Correal (2014). OpenHub : A scalable architecture for the analysis of software quality attributes. *MSR 2014 Proc. 11th Work. Conf. Min. Softw. Repos.,* 420–423.
[31]. C. Bonifacio, T. E. Barchyn, C. H. Hugenholtz, and S. W. Kienzle (2015). Computers & Geosciences CCDST : A free Canadian climate data scraping tool. *Comput. Geosci.*, *75*, 13–16.
[32]. J. Wang (2012). Scrapy-based Crawling and User-behavior Characteristics Analysis on Taobao, 44–52.
[33]. H. Zhiqiang (2015). Research on Tibetan News Sites ' Web Crawler and Search Engine, no. *Lemcs,* 607–611.
[34]. J. M. Cadenas, M. C. Garrido, and R. Martínez, (2013). Expert Systems with Applications Feature subset selection Filter – Wrapper based on low quality data, *40,* 6241–6252.
[35]. M. Schiezaro and H. Pedrini, (2013). Data feature selection based on Artificial Bee Colony algorithm. *EURASIP J. Image Video Process*, 1–8.
[36]. S. Tabakhi, P. Moradi, and F. Akhlaghian, (2014). Engineering Applications of Artificial Intelligence: An unsupervised feature selection algorithm based on ant colony optimization. *Eng. Appl. Artif. Intell., 32*, 112–123.
[37]. C. Bae, W. Yeh, Y. Ying, and S. Liu (2010). Expert Systems with Applications Feature selection with Intelligent Dynamic Swarm and Rough Set. *Expert Syst. Appl., 37*(10), 7026–7032.
[38]. B. Xue, M. Zhang, and W. N. Browne (2014). Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms. *Appl. Soft Comput. J., 18*, 261–276.
[39]. H. H. Inbarani, A. Taher, and G. Jothi (2013). Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. *Comput. Methods Programs Biomed., 113*(1), 175–185.
[40]. E. Saraç, S. Ay, and G. Özel, (2014). An Ant Colony Optimization Based Feature Selection for Web Page Classification. *Sci. world J.*, 1–16.
[41]. Abd-Alsabour, N. (2014). A review on evolutionary feature selection. In *2014 European Modelling Symposium* (pp. 20-26). IEEE.
[42]. Sabri, N. A. M., & Emran, N. A. (2018). Review of Materialized Views Selection Algorithm for Cyber Manufacturing. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *10*(2-8), 15-19.

How to cite this article: Sabri, N. A. M., Emran, N. A. and Abdullah, N. (2020). Quality-Based Open Data Source Selection Using Ant Colony Optimization (ACO) Algorithm. *International Journal on Emerging Technologies*, *11*(3): 1164–1168.