International Journal of Technology 10(8): 1567-1575
ISSN 2086-9614

PROFILING ACADEMIC LIBRARY PATRONS USING K-MEANS AND X-MEANS CLUSTERING

Aisyah Larasati[1,4*], Apif Miftahul Hajji[2,4], Anik Nur Handayani [3], Nabila Azzahra[1], Muhammad Farhan[1], Puji Rahmawati[1]

[1]*Department of Industrial Engineering, Universitas Negeri Malang, Jl. Semarang No.5, Malang 65145, Indonesia*
[2]*Department of Civil Engineering, Universitas Negeri Malang, Jl. Semarang No.5, Malang 65145, Indonesia*
[3]*Department of Electrical Engineering, Universitas Negeri Malang, Jl. Semarang No.5, Malang 65145, Indonesia*
[4]*PUI-PT Disruptive Learning Innovation (DLI) Universitas Negeri Malang, Jl. Semarang No.5, Malang 65145, Indonesia*

(Received: November 2018 / Revised: January 2019 / Accepted: September 2019)

ABSTRACT

Information technology is now used very often, especially by individuals born between 1982 and 2002 (the Millennial generation). The academic library, which from its beginnings has been a storehouse for information through collections, is becoming less attractive for Millennials because of the influence of information technology. This study aimed to use k-means and x-means clustering algorithms to identify the characteristics of academic library patrons, particularly Millennial patrons. K-means is a well-known algorithm due to its simplicity, while x-means is a relatively new algorithm for performing clustering and provides the capability to determine an optimal number of clusters, the number of cluster that minimizes differences within each cluster and maximizes differences between clusters. In this study, data were collected using questionnaires, both in online and offline forms. A total of 935 responses were collected. The results show that k-means performs better than x-means since it results in a lower Davies-Bouldin index value. However, x-means provides better descriptions of the patrons' behavior on each cluster. Both k-means and x-means clustering methods create five clusters based on the behavior of academic library patrons. One of the clusters resulting from k-means and x-means also confirms that not all patrons come to the academic library for the book collection; they come because of invitations from friends or to use internet services.

Keywords: Academic library; Clustering; K-means; X-means

1. INTRODUCTION

The Millennial generation (also known as Generation Y) is the group of individuals born between 1982 and 2002 (Kotz, 2016). Members of this generation have unique characteristics. For example, many Millennials do not wear watches because their cell phones display the time. Furthermore, rather than using physical photograph albums, Millennials store their photographs on Facebook, Instagram, and other social media platforms. Millennials enjoy using technology. Indeed, they are the first generation to have become dependent on technology (Smith & Nichols, 2015). They live in an age when they can instantly access whatever information they want, for

*Corresponding author's email: aisyah.larasati.ft@um.ac.id, Tel. +62-341-551312
Permalink/DOI: https://doi.org/10.14716/ijtech.v10i8.3440

example, academic data or information through their smartphones (Maiers, 2017) as part of their use of existing technology. According to Maiers (2017), social networks have become one of the strongest motivators for Millennials, rather than interactions in the real world.

In this age, information technology is one of the most frequently used tools in life, including for academic libraries because most have chosen to integrate technology into their information content (Walton, 2014). Many academic library patrons surveyed for the study reported that obtaining information from the internet is easier than having to search in a library. It is also easier because not everything patrons need is available in libraries. Moreover, technology is one of the most commonly used communication tools, and the majority of users who use it to communicate are Millennials (Maiers, 2017).

Learning methods must continually adapt to engage and educate this generation (Nicholas, 2008). Millennials tend to have a different learning method than previous generations. Millennials prefer to have skills and creativity in arts, games, video lectures, field trips, and other activities that do not depend only on books and theories. They tend to work beyond required working hours and have less social time (DeVaney, 2015). Moreover, Millennials are fluent in the uses of technology or perhaps even dependent on it (Nicholas, 2008).

To keep pace with and adjust to evolutions in Millennials' learning methods, academic libraries must be able to convert some of their traditional services into digital services. Academic libraries are transitioning from a collections-based model to a broader services-based model (Gleason, 2018). Library services, most of which are printed books, must be converted into digital services augmented by free internet services and other offerings to accommodate Millennials' needs. Fulfilling users' needs may increase customer satisfaction and affect an institution's success. User satisfaction may be achieved by identifying service quality attributes and their effects on user satisfaction (Zuna et al., 2016).

As mentioned above, Millennials often look for academic references on the internet rather than physically searching in a library. In one study, 79.5% of college students reported that they are experts at using the internet to search for information efficiently and effectively, but only 56.4% said that they are skilled in using the college library (Lippincott, 2012). Thus, academic libraries must understand the characteristics of Millennials in order to create an environment that is attractive to them; for example, in providing books Millennials need and promoting such services, libraries can attract the attention of patrons to persuade them to continue using books (Lippincott, 2012). Each customer may have a different perspective on the attributes that affect his preferences since customer preferences can be influenced by the completeness of the product/service attributes and the transaction process (Suzianti et al., 2015).

At the present time, academic libraries are providing number of services that Millennials would find attractive and may not able to find in online sources such as data management, information about digital scholarship, copyright management, citation management, open educational resources, and others (Dempsey & Malpas, 2018).

Millennials students exhibit a number of common characteristics: They are more focused on achievement, they prefer to question everything and use all means available to get information, and they use technology not only to find information on the internet but also for typing notes in class (Freeman et al., 2014). Currently, data integration and analysis are still rarely used to support decision-making, although many academic libraries have applied technology to obtain various reader information. Tremendous amounts of collected data remain to be analyzed in a simple analysis such as correlation (Wang et al., 2011). Thus, in the present study, in order to obtain information about the characteristics of students who use the library often, the most suitable method was clustering because it can be used to identify unique distributions or patterns in data and discover groups of data (Halkidi et al., 2001).

Table 1 Changes in the function of the library

Terms	Collection-Based Library	Services-Based Library
Library	Explained as library collection, reference	Explained as users' needs, such as lecturer and student research
Organization	The system used is a bureaucracy that prioritizes the production of the facilities offered by the library.	The system used is enterprising, which is focused on changing its goals.
Ability	Process and subject	Focused on learning, research, skills, etc.
Systems	Back office	Shared system with workflow systems (scholarship information and e-books)
Space	Focused on collection of books	Focused on service or user experience
Collection	Based on consumption from users	The facilities already exist and between one facility and another are collective.

Source: (Dempsey & Malpas, 2018)

Clustering methods are unsupervised classification methods aimed at facilitating the discovery process by combining a set of objects to create a collection of data subjects that have homogenous groups (Bader et al., 2006; Padmaja et al., 2008). The cluster members in one group have maximum similarities but minimum similarities with other cluster group members. Clustering is different from classification. Clustering is the segmenting of data into a group, while classification segments some data by assigning it into groups (Chen & Chen, 2006). The quality of clustering data depends on how high the intra-class similarities are and how low the inter-class similarities are. A common measure of cluster accuracy is the Euclidean distance. Computational time may also be used as a measure of cluster performance (Aparna & Mydhili, 2016).

By using a data mining method such as clustering, it is possible to discover different behaviors of patrons and possibly use those behaviors to determine whether a library's service and collection match Millennials' learning methods. Two methods used to conduct the data integration are k-means and x-means clustering. K-means clustering is a data mining algorithm that divides n objects into k clusters so that the members of one cluster have high similar characteristics while the members of different clusters are dissimilar (Ahmar et al., 2018). X-means clustering is an extension of k-means clustering that refines the clustering by continuously splitting the cluster until the selection criterion is reached.

The aim of the present study was to profile the behavior of academic library patrons, particularly patrons who are categorized as members of the Millennial generation, by comparing the clusters resulting from the k-means and x-means clustering methods.

2. METHODOLOGY

This study aims to profile the behavior of academic library patrons, particularly patrons who are categorized as members of the Millennial generation, by applying the k-means and x-means clustering methods. The main benefit of the k-means algorithm is the high speed of computational processes when the k (number of clusters) is small, even for large variables (Dubey et al., 2018).

This study used k-means and x-means algorithms as clustering methods. Data were collected using closed-ended questions. A total of 935 responses were collected (online and offline form) from eight faculties and 32 departments. The offline survey was conducted by distributing the questionnaire directly to library patrons during the operational hours of the library in May 2019. The online survey was also available during May 2019. For the purposes of the reliability and validity tests, the study collected 50 responses using a self-administered survey. The reliability

test was conducted using the Pearson product-moment correlation and the validity test was based on the construct validity. Since the correlation product-moment for each item of the questionnaire > 0.3 and the Cronbach's alpha for each construct > 0.7, the instrument was valid and reliable for collecting data.

The respondents were undergraduate students between semesters 1 and 14. The k-means algorithm was used because it is simple and can be used for a wide variety of data types. It is also quite efficient, even though multiple runs are often performed (Tan et al., 2005). The k-means algorithm is known to converge to a local minimum of the distortion measure (that is, the average squared distance from points to their class centroids). It is also known to be too slow for practical databases. K-means is fully deterministic, based on the starting centers. Improper initial centers may have a great impact on both performance and distortion (Pelleg & Moore, 2000). Another algorithm that is used is x-means, a new algorithm that quickly estimates the k (the number of clusters) (Pelleg & Moore, 2000). However, x-means is not very sensitive to the number of clusters when the value of R changes. Overall, it can be said that the x-means process runs twice as fast as the k-means process (Pelleg & Moore, 2000). Thus, this study compared k-means and x-means clustering results to profile the academic library patrons.

The research methodology of this study consisted of three processes. The process flowchart of this research is shown in Figure 1.

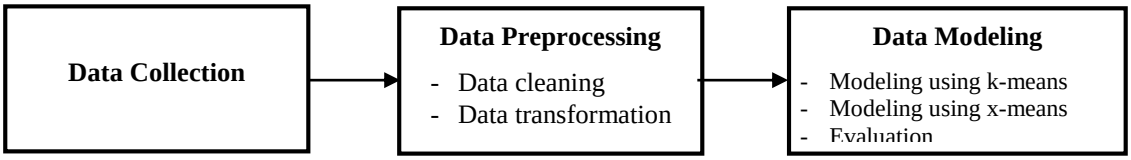

Figure 1 Flowchart of the research

The first procedure entailed data collection. Data were collected using online and offline surveys. This study used an accidental sampling technique. The indicators used in the questionnaire are shown in Table 2.

The survey respondents were students at Universitas Negeri Malang. The usable rate of the responses was 96.14%. From a total of 935 samples, data from 898 could be processed. The second step was to clean the data to eliminate missing values and duplicated data. The number of classes for each question was also changed to five classes so that the class of each item was uniform. This made the responses easier to enter into the model. Transformed items can be seen in Table 3.

Table 2 Questionnaire indicators

Indicator	Questionnaire items
Respondent Profile	4 profile questions
Motivation	Items 1, 2, 3, 4, 5, 6
Types of services	Items 7, 8, 9
User behavior	Items 10, 11, 12, 13, 14, 15
Book collection	Items 16, 17, 18, 19
Service quality	Items 20, 21, 22, 23, 24, 25, 26, 27

In the last step, the processed data were then formed to build a cluster model using the k-means and x-means algorithms. The value on the Davies-Bouldin index was used as a parameter to assess the cluster model performance; smaller was better. The Davies-Bouldin index performs better than the Dunn index for finding the best cluster based on the internal cluster validity (Kryszczuk & Hurley, 2010).

Table 3 Data transformation

Item number	Attribute	Number of classes	Transformation number of classes
7	Frequently used services	9	5
8	Preferred service innovation	9	5
21	Library service that needs to improve	9	5

3. RESULTS

The k-means cluster algorithm was run for 10 cycles. The numerical measure was the Euclidian distance and it applies 100 steps for optimization. The k-means cluster algorithm resulted in five clusters. Cluster 1 was the most populated cluster, while the least populated cluster was cluster 0. The Davies-Bouldin index value was 4,831 and the average cluster distance was 102,829. The clusters generated and the characteristics of each cluster produced from the k-means algorithm are shown in Figure 2 and Figure 3.

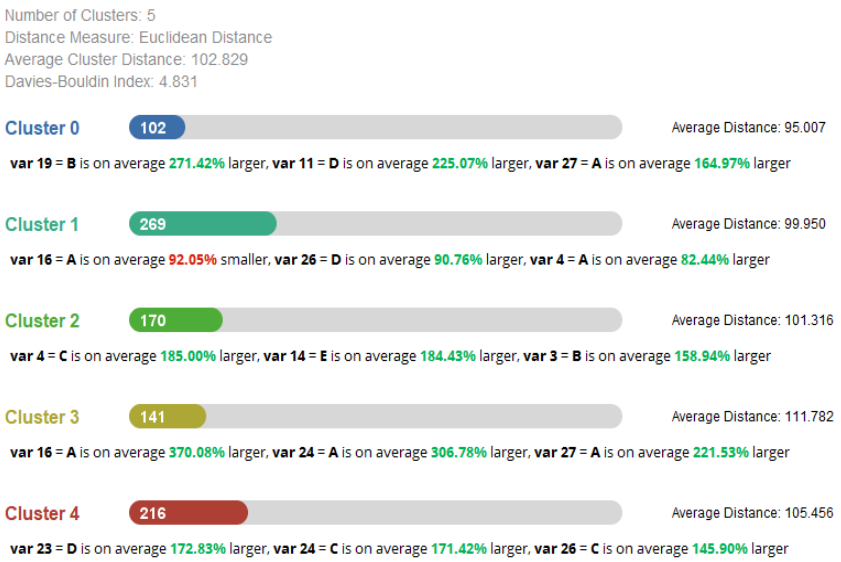

Figure 2 K-means algorithm cluster distribution

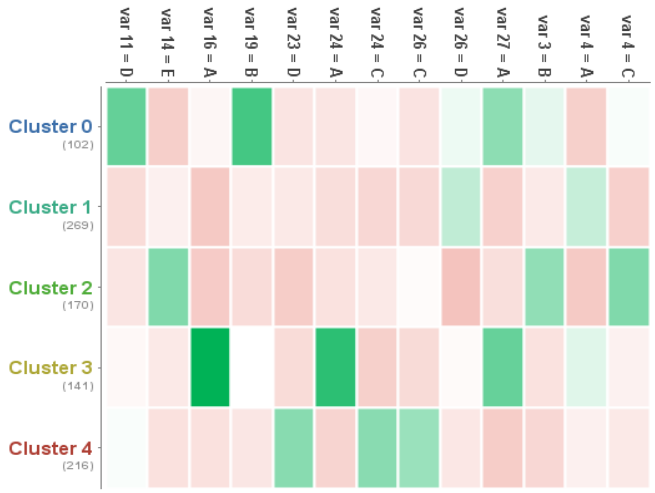

Figure 3 K-means algorithm heat map

The x-means algorithm was run for 10 cycles. The numerical measure was cosine similarity and it applies 100 steps for optimization. The x-means cluster algorithm resulted in five clusters. The total number of responses in each cluster ranged from 159 to 213. The Davies-Bouldin index value was 4,882 and the average cluster distance was 102,940. The clusters generated and the characteristics of each cluster produced from the x-means algorithm are shown in Figure 4 and Figure 5.

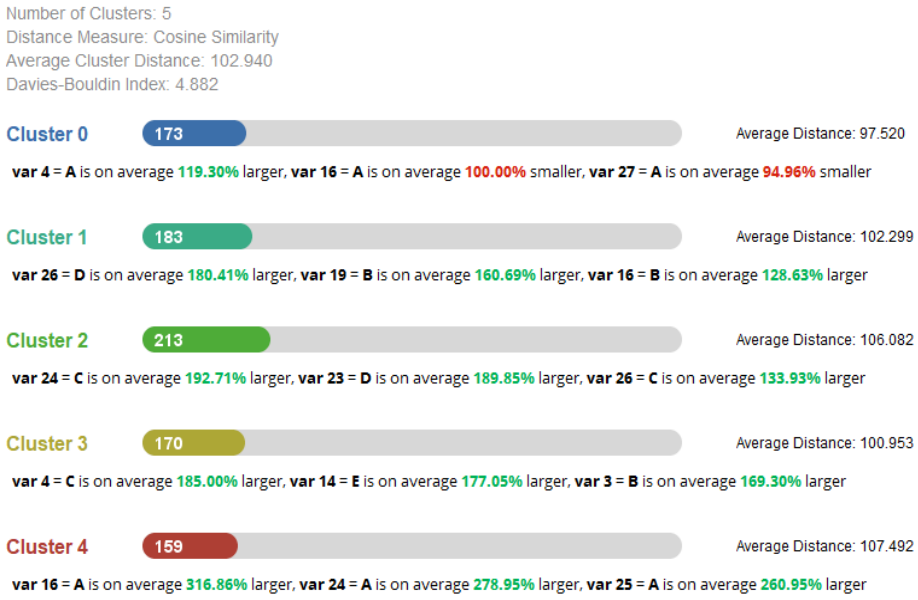

Figure 4 X-means algorithm cluster distribution

Figure 5 X-means algorithm heat map

4. DISCUSSION

The results of the cluster distribution and heat map, as shown in Figure 2 and Figure 3, provide information about the characteristics of each cluster produced in the k-means algorithm. Students grouped in **cluster 0** generally borrow books at other universities (var19 = b), visit the library between 16.00-19.00 (var11 = c), and assess the services between libraries as very easy to use (var27 = a). Students grouped in **cluster 1** generally think that the bag storage space is less secure (var26 = d) and always go to the library when they do not have a book (var4 = a). Students grouped in **cluster 2** generally never borrow books from the library (var14 = e), go to the library because they are invited by friends (var3 = b), and usually search for books on the

internet (var4 = c). Students grouped in **cluster 3** generally have their needs fulfilled by the book collection (var16 = a), think that the reading room is comfortable (var24 = a), say that the services between libraries are very easy to use (var27 = a), and always go to the library if they do not have a book (var4 = a). Students grouped in **cluster 4** generally think that OPAC (Online Public Access Catalog) services are less convenient (var23 = d), the reading room in the library is quite comfortable (var24 = c), and the storage of bags in the library is quite safe (var26 = c).

The results of the cluster distribution and the heat map of the X-means clustering are shown in Figure 4 and Figure 5, explaining the characteristics of each cluster produced in the x-means algorithm. Students grouped in **cluster 0** generally go to the library when they do not have a book (var4 = a) and judge that the storage room for bags is not safe (var26 = d). Students grouped in **cluster 1** generally think that the storage room for bags is less secure (var26 = d), borrow books at other universities (var19 = b), use the book collection according to their needs (var16 = b), say that the inter-library services are very easy to use (var27 = a), and read books in the library because they were invited by friends (var3 = b). Students grouped in **cluster 2** generally think that OPAC services are less easy to use (var23 = d), say that the reading room in the library is quite comfortable (var24 = c), and think that the storage of bags in the library is quite safe (var26 = c). Students grouped in **cluster 3** generally never borrow books from the library (var14 = e), go to the library because they are invited by friends (var3 = b), and usually search for books on the internet (var4 = c). Students grouped in **cluster 4** generally have their needs fulfilled by the book collections (var16 = a), think that the reading room is comfortable (var24 = a), assess the library officers as very responsive (var25 = a), say that the inter-library services are very easy to use (var27 = a), and always go to the library if they do not have a book (var4 = a).

From a comparison of the results of the k-means and x-means clustering, it can be seen that the results of the k-means algorithm are better than those of the x-means algorithm. The Davies-Bouldin index value for the k-means was 4,831, while the value for the x-means was 4,882. This result indicates that k-means clustering has a better performance than x-means clustering in profiling the academic library patrons. However, according to Halkidi et al., 2001, the k-means clustering algorithm gives the best results for their clusters when the data are partitioned into several clusters. Although x-means has a higher Davies-Bouldin index value, the algorithm supports better descriptions or characteristics distribution to each cluster.

5. CONCLUSION

Based on the Davies-Bouldin index parameter, the k-means produced a value of 4.831 and the x-means produced a value of 4.882. Thus, this study demonstrates that k-means performs better at clustering academic library patrons' behavior than the x-means since the value of the Davies-Bouldin index is smaller than that of the x-means. However, although the x-means has a higher Davies-Bouldin index value, it is better able to provide detailed information about the characteristics of the respondents in each cluster.

This study has a limitation in the number of iterations used to compare the k-means and x-means clustering. Further research is needed to deepen the analysis of the research findings.

6. ACKNOWLEDGEMENT

The authors would like to acknowledge Universitas Negeri Malang (UM) and PUI-PT Disruptive Learning Innovation (DLI) Universitas Negeri Malang for their funding of this research through an Islamic Development Bank (IsDB)-UM Research Grant No. 26.3.34/UN32.14.1/LT/2019.

7. REFERENCES

Ahmar, A.S., Napitupulu, D., Rahim, R., Hidayat, R., Sonatha, Y., Azmi, M., 2018. Using K-Means Clustering to Cluster Provinces in Indonesia. *Journal of Physics: Conference Series*, Volume 1028, pp. 1–6

Aparna, K., Mydhili, K.N., 2016. Incorporating Stability and Error-based Constraints for a Novel Partitional Clustering Algorithm. *International Journal of Technology*, Volume 7(4), pp. 691–700

Bader, S., Urfer, W., Baumbach, J.I., 2006. Reduction of Ion Mobility Spectrometry Data by Clustering Characteristic Peak Structures. *Journal of Chemometrics: A Journal of the Chemometrics Society*, Volume 20(3–4), pp. 128–135

Chen, A.-P., Chen, C.-C., 2006. A New Efficient Approach for Data Clustering in Electronic Library using Ant Colony Clustering Algorithm. *The Electronic Library*, Volume 24(4), pp. 548–559

Dempsey, L., Malpas, C., 2018. *Academic Library Futures in a Diversified University System*. Higher Education in the Era of the Fourth Industrial Revolution. Springer, pp. 65–89

DeVaney, S.A., 2015. Understanding the Millennial Generation. *Journal of Financial Service Professionals*, Volume 69(6), pp. 11–14

Dubey, A.K., Gupta, U., Jain, S., 2018. Comparative Study of K-means and Fuzzy C-means Algorithms on the Breast Cancer Data. *International Journal on Advanced Science, Engineering and Information Technology*, Volume 8(1), pp. 18–29

Freeman, S., Eddy, S.L., McDonough, M., Smith, M.K., Okoroafor, N., Jordt, H., Wenderoth, M. P., 2014. Active Learning Increases Student Performance in Science, Engineering, and Mathematics. *Proceedings of the National Academy of Sciences*, Volume 111(23), pp. 8410–8415

Gleason, N.W., 2018. *Higher Education in the Era of the Fourth Industrial Revolution*. Springer

Halkidi, M., Batistakis, Y., Vazirgiannis, M., 2001. On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, Volume 17(2–3), pp. 107–145

Kotz, P.E., 2016. Reaching the Millennial Generation in the Classroom. *Universal Journal of Educational Research*, Volume 4(5), pp. 1163–1166

Kryszczuk, K., Hurley, P., 2010. Estimation of the Number of Clusters using Multiple Clustering Validity Indices. *In*: International Workshop on Multiple Classifier Systems, Springer, pp. 114–123

Lippincott, J.K., 2012. Information Commons: Meeting Millennials' Needs. *Journal of Library Administration*, Volume 52(6–7), pp. 538–548

Maiers, M., 2017. Our Future in the Hands of Millennials. *The Journal of the Canadian Chiropractic Association*, Volume 61(3), pp. 212–217

Nicholas, A., 2008. *Preferred Learning Methods of the Millennial Generation*. Faculty and Staff - Articles & Papers. 18. Available Online at https://digitalcommons.salve.edu/fac_staff_pub/18

Padmaja, P., Vikkurty, S., Siddiqui, N.I., Dasari, P., Ambica, B., Rao, V.V., Rudraraju, V.J.P.R., 2008. Characteristic Evaluation of Diabetes Data using Clustering Techniques. *IJCSNS*, Volume 8(11), pp. 244–251

Pelleg, D., Moore, A.W., 2000. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. *In*: Proceedings of the 17th International Conference on Machine Learning, Volume 1, pp. 727–734

Smith, T.J., Nichols, T., 2015. Understanding the Millennial Generation. *The Journal of Business Diversity*, Volume 15(1), pp. 39–47

Suzianti, A., Faradilla, N.D.P., Anjani, S., 2015. Customer Preference Analysis on Fashion Online Shops using the Kano Model and Conjoint Analysis. *International Journal of Technology*, Volume 6(5), pp. 881–885

Tan, P.-N., Steinbach, M., Kumar, V., 2005. Chapter 8: Cluster Analysis: Basic Concepts and Algorithms. *In*: Introduction to Data Mining. Available Online at: https://doi.org/10.1016/0022-4405(81)90007-8

Walton, E.W., 2014. Why Undergraduate Students Choose to use E-books. *Journal of Librarianship and Information Science*, Volume 46(4), pp. 263–270

Wang, R., Tang, Y., Liu, G., Li, Y., 2011. K-means Clustering Algorithm Application in University Libraries. *In*: IEEE 10th International Conference on Cognitive Informatics and Cognitive Computing (ICCI-CC'11), IEEE, pp. 419–422

Zuna, H.T., Hadiwardoyo, S.P., Rahadian, H., 2016. Developing a Model of Toll Road Service Quality using an Artificial Neural Network Approach. *International Journal of Technology*, Volume 7(4), pp. 562–570