Book data grouping in libraries using the k-means clustering method

Saut Parsaoran Tamba*, M Diarmansyah Batubara1, Windania Purba1, Maria Sihombing1, Victor Marudut Mulia Siregar2, Jaidup Banjarnahor3

1Department of Information System, Faculty of Technology and Computer Sciences, Universitas Prima Indonesia, Indonesia
2Faculty STMIK Politeknik Bisnis Indonesia, Indonesia
3Department of Informatics Management, AMIK Medan Business Polytechnic (MBP) Indonesia, Indonesia

saut_nabasa@yahoo.co.id

Abstract. Clustering is a process of sorting out a data set to become separate cluster groups and each has similarities, and aims to group the data into one cluster. This research aims to group the book information which is contained in Universitas Prima Indonesia, by using K-means clustering method. On this K-means clustering algorithm, the variables used as input are: NIM, Name, Book Title and Author. The output produced consists of 3 clusters, those are (C1) the most frequently borrowed book, (C2) book that is often borrowed, and (C3) book that is rarely borrowed. With the use of this K-means Clustering method, the final result obtained consists of member of cluster 1 as many as 19 members, member of cluster 2 as many as 22 members, and member of cluster 3 as many as 19 members. The information of grouping this book data can be used by the library. In the case of the selection of books that must be added to the library and to minimize the books that are rarely borrowed so as not to cause a buildup of books that are rarely borrowed, so there is a space for books to be added into the library.

1. Introduction

Data Mining is the process of discovering interesting patterns and knowledge from large amounts of data. Data mining is often treated as synonym for another popularly used term, Knowledge Discovery in Databases (KDD). The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically, by observing large data sets over a period of time, we can deduce previously-unknown and useful information concerning patterns, models, trends, and rules in the area of application.

Clustering is a process of grouping data objects into disjointed clusters so that the data in the same cluster are similar, but data belonging to different cluster differ. A cluster is a collection of data object that are similar to one another are in same cluster and dissimilar to the objects are in other clusters. Grouping is an important part of analyzing partitioned data dataset to the same subset of data points in each subset and different from data from other clusters. There are many proposed grouping techniques but K-means is one of the oldest and most numerous popular grouping techniques.

This problem also happened at the library at the Universitas Prima Indonesia (UNPRI). Along with developments, the number of collections of books in the UNPRI library is increasing. The library is now become a book warehouse, so some books are piled up and are not neatly arranged on
bookshelves because of limited space, besides the average number of visitors who come to the library every day to borrow books is increasing.

In the UNPRI library, the library admin always inputs data of books borrowed by students or lecturers and stored into the database. For the place where the data for borrowing books is still using Microsoft Excel. The amount of information books contained in the UNPRI library allows visitors to get all kinds of information. Among them is the book classification information. Most book information is not classified or classified according to the group so information about a particular book is difficult to find. From existing historical data, the author intends to process the data so that it can become a new information, namely for information about the most frequently borrowed books, often borrowed and books that are rarely borrowed [6].

In this research, the clustering technique used is using the k-means algorithm. The k-means method will be grouped to determine the most frequently borrowed books, often borrowed and books that are rarely borrowed from the UNPRI library [7-10]

2. Methods
The method used for grouping the most frequently borrowed books at the UNPRI library was carried out in several stages, carrying out the analysis needed by the system consisting of: the need for system input data to be built, namely book lending data in unprived libraries that could be used as samples for analysis. The next stage is done by processing or processing input data using clustering techniques using the K-means method, and the expected output is to produce 3 clusters namely the most frequently borrowed books (C1), frequently borrowed books (C2), and books that are rarely borrowed (C3). This research has been applied in unprivi libraries, by grouping books as many as 7826 books.

The data material to be used in the data grouping of this book consists of borrowing books, which consists of several attributes needed as a basis for conducting the clustering process are Name, Book Title and Author as seen in Table 1.

| NO | Name | Book Title                | Author            | Number of Books |
|----|------|---------------------------|-------------------|-----------------|
| 1  | P1   | Statistik kesehatan       | Riwidikdo, Handoko| 8               |
| 2  | P2   | Statistik kesehatan       | Riwidikdo, Handoko| 8               |
| 3  | P3   | Teknik Komplilasi         | Utdirartatmo, Firrar| 2               |
| 4  | P4   | Metode Numerik            | Munir, Rinaldi    | 2               |
| 5  | P5   | Pengolahan Citra Digital  | Putra, Darma      | 2               |
| 6  | P6   | Pengolahan Citra Digital  | Putra, Darma      | 2               |
| 7  | P7   | Analisis kritis atas laporan keuangan | Harahap, Sofyan Syafri | 9               |
| 8  | P8   | Analisis kritis atas laporan keuangan | Harahap, Sofyan Syafri | 9               |
| 9  | P9   | Asas-asas hukum pidana    | Hamzah, Andi      | 19              |
| 10 | P10  | Linguistik umum           | Chae, Abdul       | 20              |
| 11 | P11  | English syntax: an introduction | Kim, Jong-Bok; Sells Peter | 13              |
| 12 | P12  | Asas-asas hukum pidana    | Moeljatno         | 19              |
| 13 | P13  | Asas-asas hukum pidana    | Moeljatno         | 19              |
| 14 | P14  | Asas-asas hukum pidana    | Hamzah, Andi      | 19              |
| 15 | P15  | Psikologi pendidikan      | Dalyono, M        | 15              |
| 16 | P16  | Prinsip dasar ilmu gizi   | Almatsier, Sunita | 5               |
| 17 | P17  | Pengantar Akuntansi       | Sujarweni, V Wiratna| 10             |
| 18 | P18  | Understanding and using english grammar | Azar, Betty Schrampfer | 20             |
| Page | Title                                      | Author(s)                        | Page |
|------|-------------------------------------------|----------------------------------|------|
| 19   | Psikologi Kepribadian                     | Alwisol                          | 13   |
| 20   | Psikologi Kepribadian                     | Suryabrata, Sumadi               | 13   |
| 21   | Teknik Kompilasi                          | Utdirartatmo, Firrar             | 2    |
| 22   | Teknik Kompilasi                          | Utdirartatmo, Firrar             | 2    |
| 23   | Pengolahan Citra Digital                  | Putra, Darma                     | 2    |
| 24   | Teknik Kompilasi                          | Slamet, Sumantri; Heru Suhartanto | 2    |
| 25   | Akuntansi sektor publik                   | Mardiasmo                        | 8    |
| 26   | Pengantar Akuntansi                       | Warren, Carl S; Reeve, James M Fess, Philip E | 10 |
| 27   | Linguistik umum                           | Chaer, Abdul                      | 20   |
| 28   | English syntax: an introduction           | Kim, Jong-Bok; Sells Peter       | 13   |
| 29   | Understanding and using english grammar   | Azar, Betty Schrampfer           | 20   |
| 30   | Understanding and using english grammar   | Azar, Betty Schrampfer           | 20   |
| 31   | Understanding and using english grammar   | Azar, Betty Schrampfer           | 20   |
| 32   | Asas-asas hukum pidana                    | Hamzah, Andi                      | 19   |
| 33   | Kitab Undang-Undang Hukum Perdata         | Subekti, R.; R. Tjitrosudibio    | 8    |
| 34   | Psikologi pendidikan                      | Sugihartono; dkk                 | 15   |
| 35   | Psikologi pendidikan                      | Santrock, John W                 | 15   |
| 36   | Psikologi pendidikan                      | Santrock, John W                 | 15   |
| 37   | Psikologi Kepribadian                     | Alwisol                          | 13   |
| 38   | Kenakalan Remaja                          | Kartono, Kartini                 | 4    |
| 39   | Kenakalan Remaja                          | Sudarsono                        | 4    |
| 40   | Metode Numerik                            | Munir, Rinaldi                   | 2    |
| 41   | Metode Numerik                            | Setiawan, Agus                   | 2    |
| 42   | Akuntansi sektor publik                   | Renyowijoyo, Muindro             | 8    |
| 43   | Akuntansi sektor publik                   | Mahmudi                          | 8    |
| 44   | Akuntansi sektor publik                   | Nordiawan, Deddi                 | 8    |
| 45   | Linguistik umum                           | Chaer, Abdul                      | 20   |
| 46   | Linguistik umum                           | Chaer, Abdul                      | 20   |
| 47   | Kitab Undang-Undang Hukum Perdata         | Subekti, R.; R. Tjitrosudibio    | 8    |
| 48   | Teori dan hukum konstitusi               | Thaib, H. Dahlan; Jazim Hamidi; Nı’matul Huda | 3    |
| 49   | Teori dan hukum konstitusi               | Thaib, H. Dahlan; Jazim Hamidi; Nı’matul Huda | 3    |
| 50   | Psikologi pendidikan                      | Sugihartono; dkk                 | 15   |
| 51   | Kitab Undang-Undang Hukum Perdata         | Subekti, R.; R. Tjitrosudibio    | 8    |
| 52   | Ilmu kesehatan masyarakat: Prinsip-prinsip dasar | Notoatmodjo, Soekidjo          | 8    |
| 53   | Pengantar Akuntansi                       | Yadiati, Winwin; ilham Wahyudi   | 10   |
| 54   | Pengantar Akuntansi                       | Yadiati, Winwin; ilham Wahyudi   | 10   |
After the input data is collected, the next step is the transformation of data. At this stage the data change process is carried out, which aims to make the data can be processed using the K-means Clustering algorithm. This algorithm can only be processed with numbers. Data variables borrowing books consisting of names, titles, and authors must be converted into numbers because the data is not in the form of numbers. Data transformation can be done with the following steps: sort the data based on the frequency of occurrence; initialize the data starting from the highest data with a value of 1, then the next data 2, 3 and so on.

In table 2, it can be seen the results of data initialization for book title variables and authors.

| No | Book Title       | Author         | Number of books |
|----|------------------|----------------|-----------------|
| 1  | 2                | 2              | 8               |
| 2  | 2                | 2              | 8               |
| 3  | 3                | 6              | 2               |
| 4  | 4                | 6              | 2               |
| 5  | 5                | 6              | 2               |
| 6  | 6                | 10             | 9               |
| 7  | 6                | 10             | 9               |
| 8  | 6                | 10             | 9               |
| 9  | 1                | 7              | 19              |
| 10 | 8                | 3              | 20              |
| 11 | 9                | 11             | 13              |
| 12 | 1                | 12             | 19              |
| 13 | 1                | 12             | 19              |
| 14 | 1                | 7              | 19              |
| 15 | 7                | 18             | 15              |
| 16 | 10               | 1              | 5               |
| 17 | 11               | 19             | 10              |
| 18 | 12               | 4              | 20              |
| 19 | 14               | 13             | 13              |
| 20 | 14               | 20             | 13              |

After the transformation process the next step is to process data by applying the k-means clustering algorithm. The first step that must be done is to determine the number of clusters to be formed, in this study the clusters that will be formed are as many as 3 clusters. Then determine the initial center point of each cluster, which is determined randomly. The center point of each cluster based on the data is as follows:

For the initial center cluster (C1) the book title = 2; author = 65; book number = 51, for center cluster (C2) book title = 25; author = 16; book number = 11, for center cluster (C3) book title = 40; author = 35, number of books = 20.
3. Result and Discussion
After the central point is obtained, then the distance from each book data will be calculated to the center point of the cluster, until the last data is the 60th data. The results of these calculations can be seen in Table 3.

| No | Book Title | Author | Number of Books | C1  | C2  | C3  | Nearest Cluster Distance |
|----|------------|--------|----------------|-----|-----|-----|--------------------------|
|    |            |        |                | C1  | C2  | C3  | C1 | C2 | C3  |
| 1  | 2          | 2      | 8              | 0.0000 | 7,8102 | 19,3391 | 1 | 0 | 0  |
| 2  | 2          | 2      | 8              | 0.0000 | 7,8102 | 19,3391 | 1 | 0 | 0  |
| 3  | 3          | 5      | 2              | 6,7823 | 2,2361 | 18,0000 | 0 | 1 | 0  |
| 4  | 4          | 9      | 2              | 9,4340 | 3,1623 | 14,5945 | 0 | 1 | 0  |
| 5  | 5          | 6      | 2              | 7,8102 | 0,0000 | 16,4012 | 0 | 1 | 0  |
| 6  | 5          | 6      | 2              | 7,8102 | 0,0000 | 16,4012 | 0 | 1 | 0  |
| 7  | 6          | 10     | 9              | 9,0000 | 8,1240 | 10,3441 | 0 | 1 | 0  |
| 8  | 6          | 10     | 9              | 9,0000 | 8,1240 | 10,3441 | 0 | 1 | 0  |
| 9  | 1          | 7      | 19             | 12,1244 | 17,4929 | 18,0278 | 1 | 0 | 0  |
| 10 | 8          | 3      | 20             | 13,4536 | 18,4932 | 19,1050 | 1 | 0 | 0  |
| 11 | 9          | 11     | 13             | 12,4499 | 12,7279 | 8,7750  | 0 | 0 | 1  |
| 12 | 1          | 12     | 19             | 14,8997 | 18,4662 | 15,1565 | 1 | 0 | 0  |
| 13 | 1          | 12     | 19             | 14,8997 | 18,4662 | 15,1565 | 1 | 0 | 0  |
| 14 | 1          | 7      | 19             | 12,1244 | 17,4929 | 18,0278 | 1 | 0 | 0  |
| 15 | 7          | 18     | 15             | 18,1659 | 17,8045 | 6,4807  | 0 | 0 | 1  |
| 16 | 10         | 1      | 5              | 8,6023 | 7,6811 | 18,7083 | 0 | 0 | 1  |
| 17 | 11         | 19     | 10             | 19,3391 | 16,4012 | 0,0000  | 0 | 0 | 1  |
| 18 | 12         | 4      | 20             | 15,7480 | 19,4165 | 18,0555 | 1 | 0 | 0  |
| 19 | 14         | 13     | 13             | 17,0294 | 15,8430 | 7,3485  | 0 | 0 | 1  |
| 20 | 14         | 20     | 13             | 22,2036 | 19,9499 | 4,3589  | 0 | 0 | 1  |

After all data is in the closest cluster, then recalculate the new cluster center based on the average member in the cluster.

Re-generation of new centroid is calculated by the following formula $C = \frac{\sum m}{n}$

Description: C is the data centroid, m is a member of data that belongs to a particular centroid, n is the amount of data that is a member of a particular centroid.

Based on the minimum values generated in Table 3 above in determining the centroid value, the grouping results are obtained as shown in Table 3.

| Clusters | Cluster Name | Cluster Member |
|----------|--------------|----------------|
|          | Cluster 1    | 1, 2, 9, 10, 12, 13, 14, 18, 27, 29, 30, 31, 32, 45, 46, 57, 58 |
|          | Cluster 2    | 3, 4, 5, 6, 7, 8, 16, 21, 22, 23, 33, 40, 47, 51, 55, 56, 59, 60 |
|          | Cluster 3    | 11, 15, 17, 19, 20, 24, 25, 26, 28, 34, 35, 36, 37, 38, 39, 41, 42, 43, 44, 48, 49, 50, 52, 53, 54 |

After the calculation process above, the new centroid will be obtained with the following values. For the initial center cluster (C1) the book title = 5.4706; author = 4.7647; the number of books =
16.8824, for the center cluster (C2) the title of the book = 7.7222; author = 5.5556; book number = 4.6111, for center cluster (C3) book title = 11.2800; author = 19.3600, number of books = 9.5200

After the new center point is obtained, then do the second iteration with the same calculation, namely between the data with the new cluster center.

The results of the calculation of each data to each cluster center point for the second iteration can be seen in table 5.

| No | Book Title | Author | Number of Books | C1  | C2  | C3  | Nearest Cluster Distance |
|----|------------|--------|----------------|-----|-----|-----|--------------------------|
| 1  |            |        |                | 2   | 2   | 8   | 9.9290 7.5412 19.7433    |
| 2  |            |        |                | 2   | 2   | 8   | 9.9290 7.5412 19.7433    |
| 3  |            |        |                | 3   | 5   | 2   | 15.0879 5.4246 18.2022   |
| 4  |            |        |                | 4   | 9   | 2   | 15.5430 5.7041 14.7268   |
| 5  |            |        |                | 5   | 6   | 2   | 14.9409 3.7981 16.5674   |
| 6  |            |        |                | 5   | 6   | 2   | 14.9409 3.7981 16.5674   |
| 7  |            |        |                | 6   | 10  | 9   | 9.4773 6.4793 10.7591    |
| 8  |            |        |                | 6   | 10  | 9   | 9.4773 6.4793 10.7591    |
| 9  |            |        |                | 1   | 7   | 19  | 5.4284 15.9473 18.6633   |
| 10 |            |        |                | 8   | 3   | 20  | 4.3854 15.6021 19.7038   |
| 11 |            |        |                | 9   | 11  | 13  | 8.1491 10.0821 9.3380    |
| 12 |            |        |                | 1   | 12  | 19  | 8.7647 17.1394 15.8025   |
| 13 |            |        |                | 1   | 12  | 19  | 8.7647 17.1394 15.8025   |
| 14 |            |        |                | 1   | 7   | 19  | 5.4284 15.9473 18.6633   |
| 15 |            |        |                | 7   | 18  | 15  | 13.4557 16.2270 7.0851   |
| 16 |            |        |                | 10  | 1   | 5   | 13.2619 5.1081 18.9515   |
| 17 |            |        |                | 11  | 19  | 10  | 16.7507 14.8505 0.6621   |
| 18 |            |        |                | 12  | 4   | 20  | 7.2758 16.0480 18.6086   |
| 19 |            |        |                | 14  | 13  | 13  | 12.4757 12.8532 7.7433   |
| 20 |            |        |                | 14  | 20  | 13  | 17.8868 17.8445 4.4630   |

The next step is calculating the distance closest to the cluster. Compare members from each iteration cluster 1 and iteration 2, if the cluster member position is still changing, then proceed to iteration 3. If the cluster member position does not change, then the iteration is stopped. Calculations to determine the new cluster center point from the new data is done by summing the values of all cluster members then divided by the total number of cluster members.

Based on the results of the grouping of all data using the k-means clustering method, the final grouping results up to the 5th iteration, where the center point no longer changes and no data moves between clusters. The cluster results formed from each of these data can be seen in table 6.

| Name Cluster | Cluster member                  |
|--------------|--------------------------------|
| Cluster 1    | 9, 10, 11, 12, 13, 14, 18, 27, 28, 29, 30, 31, 32, 34, 35, 36 45, 46, 50 |
| Cluster 2    | 1, 2, 3, 4, 5, 6, 7, 8, 16, 21, 22, 23, 33, 40, 47, 51, 55, 56, 57, 58, 59, 60 |
| Cluster 3    | 15, 17, 19, 20, 24, 25, 26, 37, 38, 39, 41, 42, 43, 44, 48, 49, 52, 53, 54 |

Based on the results of the clusters formed in table 10, it can be seen that the characteristics of cluster 1 members are the most frequently borrowed groups of 19 members, cluster2 is a group of books that are often borrowed as many as 22 members, and cluster3 is a group of books that are rarely
borrowed as many as 19 members. From the loan data above it can be concluded that cluster 2 has more members compared to cluster1 and cluster 3. So in this case, this information can be used by the Library in terms of selecting books that must be added in the library and to minimize books that rarely borrowed so that books are rarely borrowed, so there is a place for books to be added to the library.

4. Conclusion
The conclusion that can be drawn from this research is With K-means Clustering Method can be used to view the most borrowed books, frequent books and books that are rarely borrowed from the library, so that the information can be used by library management to determine the books to add to the library, and to sort books that are rarely borrowed in order to minimize the availability of places.

The results of the clustering process can be used to determine the location of the book according to the most frequently borrowed books, which are often borrowed, and which are rarely borrowed so that students can easily find the desired book. The results of this clustering process can also be used by the campus to request books to be donated to the library.

Acknowledgments
This paper is funded by Universitas Prima Indonesia.

References
[1] Shruti Aggarwal, Parminder Singh 2014 Comparative Study of Various Enhanced K-Means Clustering Algorithms 5 179 - 187.
[2] Unnati R. Raval, Chaita Jani 2016 Implementing & improvisation of k-means clustering algorithm 5 191 - 203.
[3] Asmita Yadav and Sandeep Kumar Singh 2016 An improved k-means clustering algorithm 5 88 – 103.
[4] Shiv Ram Dubey, Pushkar Dixit, Nishant Singh, Jay Prakash Gupta 2013 Infected fruit part detection using k-means clustering segmentation technique 2 65 – 72.
[5] K. Selvakumar, L. Sai Ramesh, A. Kannan 2015 Enhanced k-means clustering algorithm for evolving user groups 8 80192.
[6] W. Purba, S. Tamba, and J. Saragih 2018 The effect of mining data k-means clustering toward students profile model drop out potential 1 1-6.
[7] Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur 2012 Efficient k-means clustering algorithm using ranking method in data mining 1 85 - 91.
[8] Oyelade, O. J, Oladipupo, O. O, Obagbowa, I. C 2010 Application of k-means clustering algorithm for prediction of students’ academic performance 7 292 - 295.
[9] Shruti Kapil, Meenu Chawla, Mohd Dilshad Ansari 2016 On k-means data clustering Algorithm with genetic algorithm 4 202 - 206.
[10] V. M. M. Siregar and H. Sugara 2018 Implementation of artificial neural network to assessment the lecturer's performance 420 12112.
[11] K. Aparna1 and Mydhili K. Nair 2016 Incorporating stability and error-based constraints for a novel partitional clustering algorithm 4 691-700.
[12] Juanying Xie, Shuai Jiang, Weixin Xie, Xinbo Gao 2011 An efficient global K-means clustering algorithm 6 271 - 279.
[13] Zhongxiang Fan, Yan Sun, Hong Luo 2017 Clustering of College Students Based on Improved K-means Algorithm 28 195 - 203.
[14] Bhavesh Patel, Chetan Gondaliya 2017 Student Performance Analysis Using Data Mining Technique 5 64 - 71.
[15] Ni Nyoman Emang Smrti 2013 Otomatisasi klasifikasi buku perpustakaan dengan menggabungkan metode K-NN dengan k-medoids 4 201 - 214.
[16] Widya Safira Azis1 dan Dedy Atmaja2 2016 Pengelompokan minat baca mahasiswa menggunakan metode k-means 8 201 - 214.
[17] Ferlyna K Wardhani, Erma Suryani, dan Ahmad Mukhlason 2012 Penerapan metode GA-K-means untuk pengelompokan pengguna pada bapersip provinsi jawa timur 1 545 - 550.