Cek Similariti Game Popularity Level During Covid-19 Pandemic Using Agglomerative Hierarchical Clustering

by Informatika Teknik

Submission date: 10-Apr-2023 02:39PM (UTC+0700)
Submission ID: 2060386854
File name: ovid-19_Pandemic_Using_Agglomerative_Hierarchical_Clustering.pdf (796.73K)
Word count: 3347
Character count: 17153

Game Popularity Level During Covid-19 Pandemic Using Agglomerative Hierarchical Clustering

Wildan Budiawan Zulfikar
Department of Informatics
UIN Sunan Gunung Djati
Bandung, Indonesia
wildan.b@uinsgd.ac.id

Agung Wahana
Department of Informatics
UIN Sunan Gunung Djati
Bandung, Indonesia
wahana.agung@uinsgd.ac.id

Richcy Dian Sukma
Department of Mathematics
UIN Sunan Gunung Djati
Bandung, Indonesia
richcyveliz@gmail.com

Diena Rauda Ramdania
Department of Informatics
UIN Sunan Gunung Djati
Bandung, Indonesia
diena.rauda@uinsgd.ac.id

Dian Saadilah Maylawati
Department of Informatics
UIN Sunan Gunung Djati
Bandung, Indonesia
diansm@uinsgd.ac.id

***Abstract*— During the COVID-19 pandemic, various activities of people outside the home were disrupted and made people move more indoors. For some companies take advantage of this pandemic period as their advantage, especially digital game industry companies. Various games have been released and promoted, these games are published on various game platforms. Currently, Steam is one of the biggest gaming platforms. On this platform, there are a lot of games offered by game developers and provide game pages that are currently popular. However, the website does not provide the popularity level of the currently popular games. This causes ambiguity in determining which games have high, medium, or low popularity. This study tries to create a machine learning model to cluster these games into groups using Agglomerative Hierarchical Clusterin. The distance measure used is euclidean, cosine and manhattan/cityblock and uses single, average, complete and ward linkage. Based on the evaluation results, the best cluster results are the silhouette value of 0.639 and the calinski-harabasz value of 90.192.**

Keywords—C4.5, ID3, Machine Learning

I. INTRODUCTION

Nowadays, the development of the technology industry has given birth to various types of games. The Investopedia article said that the gaming industry's profits in 2020 were US$155 billion. By 2025, this industry will generate revenues of more than USD 260 billion. The end of 2019 the Covid-19 pandemic has caused various disruptions to the normal lifestyle of the community. However, most of the gaming industry experienced a very high increase in orders. Adults and children stay at home and need outdoor entertainment. One of the chosen entertainment is games [1] [2].

Currently, more than half of players (55%) say they are playing more games during the pandemic, and most players (90%) say they will continue to play after the country opens, according to a survey of 4,000 adults conducted by a market research firm. Ipsos in February for the Entertainment Software Association [4]. For players, who are in lockdown or quarantine during the pandemic, video games are a great source of stress relief and distraction. Video games also serve as an escape and rest for children, said 71% of parents surveyed. More than half of parents (59%) say their children play educational games and two-thirds of parents (66%) say video games make the transition to distance learning easier for their children [3][4].

Game is an activity that is carried out to realize a certain goal or situation, is limited by only using a method that is allowed by a certain rule, and the function of making these rules is to allow these activities to occur. One of the best gaming PC platforms right now is Steam. Steam was developed by the Valve company which provides more than 30,000 games that can be easily accessed by PC game players. Steam also provides a tool called Steamworks that can help game developers to release their games online [7]. On the other hand, there are several software developers who provide analytics to measure the popularity of Steam and its web-based games such as SteamDB and SteamCharts. Basically, both websites store all application data and package history from Steam and then provide that info online [8] [9].

Furthermore, PC game players and game developers will use the Steam service to find games they want to play or can be used as ideas for developing a PC game. In addition, the SteamDB site can also be used as a more complete and better insight about the Steam platform and all the data that is in its database [10]. The site provides game data information in tabular form such as Most Played Games, Trending Games, Popular Releases, and Hot Releases. In the Most Played Games table, all types of data are presented in the form of ratings, then each data has more detailed information such as ID, developer, publisher, Release Date, and so on.

Behind the presentation of data that is quite complete, the weakness of the web does not provide a class stating how high the popularity of the Steam game is. In addition to the Most Played Games table, all game data is updated every 5 to 10 minutes so it requires an analysis that determines the class of Steam game popularity levels during the Covid-19 pandemic. This also occurs on the Steam platform itself and causes the popularity of a game to be ambiguous and causes confusion in determining the level of Steam game popularity.

Clustering refers to grouping records, observations, or cases into similar object classes. A cluster is a collection of

978-1-6654-5122-2/22/$31.00 ©2022 IEEE

records that are similar to each other and different from records in other clusters. Clustering differs from classification in that there is no target variable for clustering. The clustering task does not attempt to classify, estimate, or predict the value of the target variable. In contrast, clustering algorithms seek to group the entire data set into relatively homogeneous subgroups or clusters, where the similarity of the data within the cluster is maximized, and the similarity to the data outside this cluster is minimized [5], [6].

The hierarchical clustering method works by grouping data objects into hierarchies or "trees" of clusters. Representing data objects in a hierarchical form is useful for summarizing and visualizing data. Hierarchical Clustering method is divided into 2 categories, namely Agglomerative Hierarchical Clustering and Divisive Hierarchical Clustering [7]–[9].

The agglomerative clustering method uses a bottom-up strategy. It usually starts by letting each object form its own cluster and iteratively merges the clusters into larger and larger clusters, until all objects are in one cluster or certain termination conditions are met [10], [11]. The single cluster becomes the root of the hierarchy, then for the merging step, this method finds the two clusters that are closest to each other (according to some measure of similarity), and combines them to form one cluster. Since two clusters are combined per iteration, where each cluster contains at least one object, the agglomerative method requires at most n iterations. It is widely implemented in classifying diverse data [12]–[15].

The purpose of this research is to try to determine the popularity level of games on Steam using the Agglomerative Hierarchical Clustering algorithm during the covid-19 pandemic.

II. METHODOLOGY

The purpose of this research is to add a label or detail variable for the popularity level of each Steam game in the hope that it can help players or game developers determine whether the game has high, medium, or low popularity. In Data Mining, this goal can be done using the clustering method. Of the several clustering techniques offered in data mining, this study will use the Agglomerative Hierarchical Clustering technique.

The data involved in this research is sourced from the SteamDB web with the web address https://steamdb.info/. Then use the list of games from the Most Played Games table. The following data is provided for each Steam game that is in the Most Played Games table.

At this stage, the data is loaded and corrected before entering modeling. The first step at this stage is to retrieve data from the Most Played Games table on the SteamDB web. The data collection technique is done conventionally, namely by copying and pasting it into Microsoft Excel. The following data will be taken and used in this study:

1. Data Owner Estimation: Owner estimation data by PlayTracker.
2. Playtime Estimation by SteamSpy data: Average total playtime count data.
3. Data Store Data: Positive reviews count data in percent.
4. Monthly Breakdown Table: Peak data from December 2019 to July 2021.

Avg peak player data is calculated using (1):

$$app = \frac{peak\ Dec19 + \cdots + peak\ Jul21}{20} \quad (1)$$

In this case, app is the variable name of avg peak players which is the average of the peak player count each month from December 2019 to July 2021.

Owner estimation, average total playtime, positive reviews are formulated using (2):

$$qgs = oe \times atp \times \left(\frac{pr}{100}\right) \quad (2)$$

In this case, *qgs* is the variable name of the quality game score which is the calculation of the owner estimation (*oe*) times the average total playtime (*atp*) and multiplied by the percentage of positive reviews (*pr*). The calcution result of peak player and quality game score detailly describe on table 1.

TABLE 1. AVG PEAK PLAYER AND QUALITY GAME SCORE

Game	Avg peak players	Quality game score
Call of Duty Black Ops III	4.927,1	202,74
Fallout New Vegas	7.245,4	57,42
Geometry Dash	8.844	172,63
American Truck Simulator	9.605	223,66
Plants vs. Zombies GotY	5.328	100,23

Next, Agglomerative Hierarchical Clustering is implemented. The first step is to make each data into different clusters. Then calculate the distance matrix of each cluster. Next, combine the 2 closest clusters into 1 cluster. Then, update the distance matrix value with the desired linkage technique. Next, repeat steps 3 and 4 until all data becomes 1 cluster. To simplify the results of the cluster can be represented in the form of a dendrogram according to the cluster formed and the distance value. The next step is to make each data into different clusters as described in table 2.

TABLE 2. CLUSTER OF ALL SAMPLE GAME

Game	Avg peak players	Quality game score	Cluster
Call of Duty Black Ops III	4.927,1	202,74	0
Fallout New Vegas	7.245,4	57,42	1
Geometry Dash	8.844	172,63	2
American Truck Simulator	9.605	223,66	3
Plants vs. Zombies GotY	5.328	100,23	4

After that, calculate each distance in each cluster that is formed using the Euclidean formula (3). So that the values represented in the matrix below can be obtained.

$$d = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]} \quad (3)$$

	0	1	2	3	4
0	0	2323	3917	4678	401
1	2323	0	1603	2365	1918
2	3917	1603	0	763	3517
3	4678	2365	763	0	4279
4	401	1918	3517	4279	0

Based on the matrix that has been formed, the next step is to combine the two closest clusters. In the first calculation, the closest distance value is in cluster 0 with 4 and the value is 401. After merging the clusters, the distance matrix value will be updated with the single linkage grouping technique.

	(0,4)	1	2	3
(0,4)	0	1918	3517	4279
1	1918	0	1603	2365
2	3517	1603	0	763
3	4279	2365	763	0

The second process, the closest clusters formed next are clusters 2 and 3 with a distance matrix value of 763. Next, recalculate the distance matrix value for the second calculation stage.

	(0,4)	(2,3)	1
(0,4)	0	3517	1918
(2,3)	3517	0	1603
1	1918	1603	0

In the third process, the closest clusters formed are clusters (2,3) and 1 with a distance matrix value of 1,603. After that, the new distance matrix value is also determined.

	(2,3,1)	(0,4)
(2,3,1)	0	1918
(0,4)	1918	0

The calculation process for updating the distance matrix value with a single linkage has been completed, because the last cluster formed has become 1. The final step of this description is to make a dendrogram according to the cluster formed and the distance value that has been calculated above.

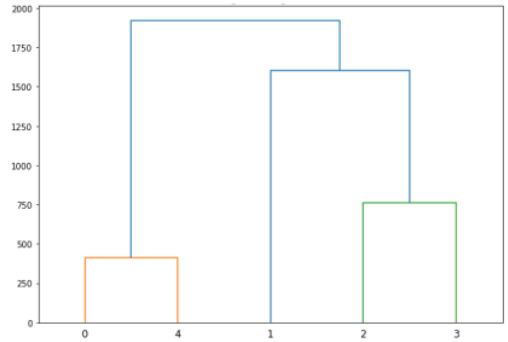


FIG 1. DENDROGRAM SAMPLE 5 GAMES WITH EUCLIDEAN DISTANCE AND SINGLE LINKAGE

III. RESULT AND DISCUSSION

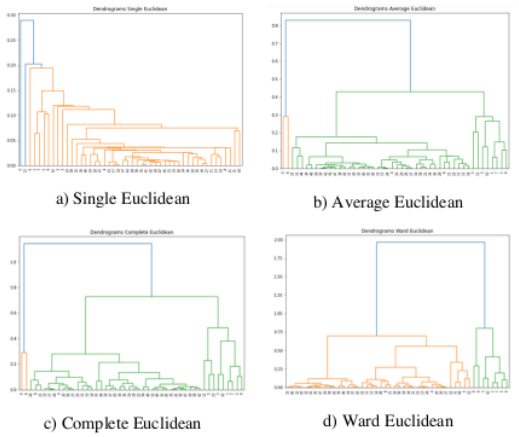
In this study, of the 50 game data used as samples, there were 2 data that included data outliers after normalizing the data. This is because the Z value of the game data produces a value of more than 3. In the table below it can be seen that the Counter-Strike Global Offensive and Dota 2 game data has a Z value of more than 3. Therefore, the data is not included.

TABLE 3. NORMALIZED AVG PEAK PLAYER AND QUALITY GAME SCORE

Game	*Avg peak players (Z-score normalized)*	*Quality game score (Z-score normalized*
Counter-Strike Global Offensive	5,646085	3,880056
Dota 2	3,744850	5,684950
Grand Theft Auto V	0,702258	0,036587
Team Fortress 2	0,225908	0,082790
…..	…..	…..
The Isle	-0,370069	-0,307107

A. Euclidean Distance

At the evaluation stage, calculations are carried out using single, average, complete, and ward Euclidean. The results are represented in the form of a dendrogram which is described in the figure below.



a) Single Euclidean

b) Average Euclidean

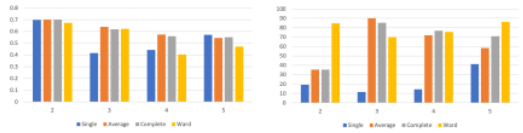
c) Complete Euclidean

d) Ward Euclidean

FIG 2. DENDROGRAM OF EUCLIDEAN

Figure 4 describes the histogram of Silhouette values for Euclidean Distance. It can be seen that in the number of clusters 2 the average and complete calculations have a silhouette value that is close to 1 and the highest of all calculations with a value of 0.699716. This makes the number of clusters 2 with linkage average and complete a candidate for determining the best number of clusters through this value. The following below is a detailed table for each silhouette value calculation for Euclidean Distance.

Based on Calinski-Harabasz for euclidean distance, the number of clusters 3 calculations average has the highest calinski-harabadz value of all calculations with a value of 90,192. This makes the number of clusters 3 with the linkage average a candidate for determining the best number of clusters through this value. Below is a detailed table for each Calinski-Harabasz value calculation for the Euclidean distance.

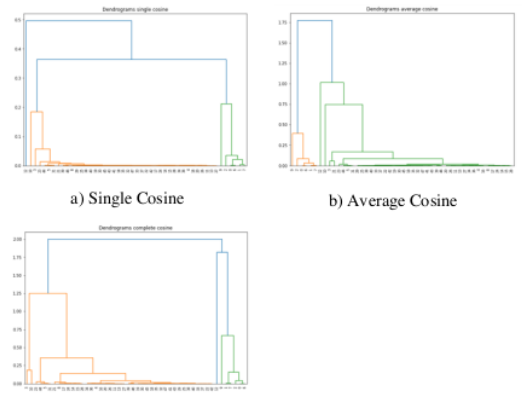


a) Silhouette b) Calinski-Harabasz

FIG 3. SILHOUETTE AND CALINSKI-HARABASZ FOR EUCLIDEAN

B. *Cosine Distance*

At the evaluation stage, calculations are carried out using single, average, complete *cosine distance*. The results are represented in the form of a dendrogram which is described in the figure below.



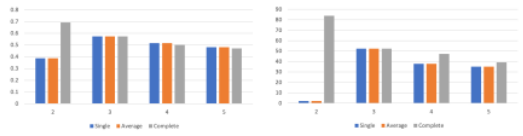
a) Single Cosine b) Average Cosine

c) Complete Cosine

FIG 4. DENDROGRAM OF COSINE

With the number of clusters 2, the complete calculation has a silhouette value close to 1 and the highest of all calculations with a value of 0.692135. This makes the number of cluster 2 with complete linkage a candidate for determining the best number of clusters through this value. Below is a detailed table for each calculation of the silhouette value for the cosine distance.

Meanwhile, with the number of clusters 2 in calinski-carabasz the complete calculation has the highest value of all calculations with a value of 83.77064. This makes the number of cluster 2 with complete linkage a candidate for determining the best number of clusters through this value.

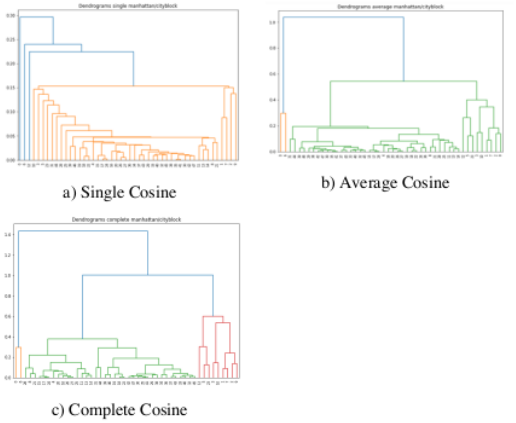


a) Silhouette b) Calinski-Harabasz

FIG 5. SILHOUETTE AND CALINSKI-HARABASZ FOR EUCLIDEAN

C. *Manhattan/cityblock Distance*

At the evaluation stage, calculations are carried out using single, average, complete *Manhattan/Citiblock distance*. The results are represented in the form of a dendrogram which is described in the figure below.



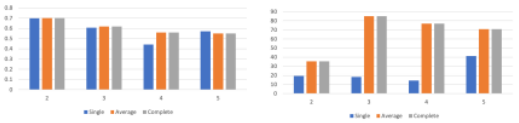
a) Single Cosine b) Average Cosine

c) Complete Cosine

FIG 6. DENDROGRAM OF MANHATTAN/CITYBLOCK

In the number of clusters 2, the average and complete calculations have a silhouette value close to 1 and the highest of all calculations with a value of 0.699716. This makes the number of clusters 2 with linkage average and complete a candidate for determining the best number of clusters through this value. Below is a detailed table for each silhouette value calculation for Manhattan/cityblock distance.

Meanwhile, the clinski-harabasz with the number of clusters 3 has the highest calinksi-harabasz value of all calculations with a value of 85,13218. This makes the number of clusters 3 with linkage average and complete a candidate for determining the best number of clusters through this value.



a) Silhouette b) Calinski-Harabasz

FIG 7. SILHOUETTE AND CALINSKI-HARABASZ FOR EUCLIDEAN

IV. CONCLUSION

Based on the evaluation of the distance matrix and linkage criteria, the best cluster evaluation value lies in the Euclidean distance using the linkage average with the number of clusters is 3. The result of the calculation of the silhouette cluster value is 0.639. The silhouette value is quite good because it is close to 1 and is the highest of all distance matrix and linkage criteria calculations. In addition, the value of the calinski-harabasz cluster is 90.192.

The calinski-harabasz value is the best because it has the highest value of all distance matrix calculations and linkage criteria. The final game data formed from the best distance and linkage calculations are 38 games including low categorization, 8 games including medium categorization, and 2 games including high categorization. These low, medium and high categories are variables that state the level of popularity of each Steam game that has been researched.

The game cluster data that has been formed from this research can be implemented in the Steam web or Steam analysis web as a detailed variable that determines the level of popularity of each Steam game. The results of this study can be used by Steam users as a comparison between Steam games. In addition, it can also be used by game analysts as a variable that determines the level of popularity of Steam games.

REFERENCES

[1] S. Ahn, J. Kang, and S. Park, "What makes the difference between popular games and unpopular games? Analysis of online game reviews from steam platform using word2vec and bass model," *ICIC Express Lett.*, vol. 11, no. 12, pp. 1729–1737, 2017, doi: 10.24507/icicel.11.12.1729.

[2] A. Beattie, "How the Video Game Industry Is Changing." .

[3] "Just how popular were video games were during COVID-19? | World Economic Forum." .

[4] M. Snider, "Video games 2021: COVID-19 pandemic led to more game-playing Americans," 2021. .

[5] L. Rokach and O. Maimon, "Clustering Methods," *Data Min. Knowl. Discov. Handb.*, pp. 321–352, May 2005, doi: 10.1007/0-387-25465-X_15.

[6] L. Rutkowski, "Data clustering methods," *Comput. Intell.*, pp. 349–369, 2008, doi: 10.1007/978-3-540-76288-1_8.

[7] L. R. Emmendorfer, "An empirical evaluation of two novel linkage criteria for hierarchical agglomerative clustering," *Proc. - 2019 Brazilian Conf. Intell. Syst. BRACIS 2019*, pp. 622–626, Oct. 2019, doi: 10.1109/BRACIS.2019.00114.

[8] R. J. Gil-García, J. M. Badía-Contelles, and A. Pons-Porrata, "A general framework for agglomerative hierarchical clustering algorithms," *Proc. - Int. Conf. Pattern Recognit.*, vol. 2, pp. 569–572, 2006, doi: 10.1109/ICPR.2006.69.

[9] J. W. Sangma, M. Sarkar, V. Pal, A. Agrawal, and Yogita, "Hierarchical clustering for multiple nominal data streams with evolving behaviour," *Complex Intell. Syst. 2022 82*, vol. 8, no. 2, pp. 1737–1761, Jan. 2022, doi: 10.1007/S40747-021-00634-0.

[10] V. Van Hai, H. L. L. Le Nhung, and R. Jasek, "Toward Applying Agglomerative Hierarchical Clustering in Improving the Software Development Effort Estimation," pp. 353–371, 2022, doi: 10.1007/978-3-031-09070-7_30.

[11] M. L. Zepeda-Mendoza and O. Resendis-Antonio, "Hierarchical Agglomerative Clustering," *Encycl. Syst. Biol.*, pp. 886–887, 2013, doi: 10.1007/978-1-4419-9863-7_1371.

[12] T. D. Nguyen and C. K. Kwoh, "Efficient agglomerative hierarchical clustering for biological sequence analysis," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, vol. 2016-January, Jan. 2016, doi: 10.1109/TENCON.2015.7373194.

[13] A. Nugraha, M. Arista Harum Perdana, H. Agus Santoso, J. Zeniarja, A. Luthfiarta, and A. Pertiwi, "Determining the Senior High School Major Using Agglomerative Hierarchial Clustering Algorithm," *Proc. - 2018 Int. Semin. Appl. Technol. Inf. Commun. Creat. Technol. Hum. Life, iSemantic 2018*, pp. 225–228, Nov. 2018, doi: 10.1109/ISEMANTIC.2018.8549834.

[14] R. Liu, J. Zhang, P. Song, F. Shao, and G. Liu, "An agglomerative hierarchical clustering based high-resolution remote sensing image segmentation algorithm," *Proc. - Int. Conf. Comput. Sci. Softw. Eng. CSSE 2008*, vol. 4, pp. 403–406, 2008, doi: 10.1109/CSSE.2008.1017.

[15] Smarika, N. Mattas, P. Kalra, and D. Mehrotra, "Agglomerative hierarchical Clustering technique for partitioning patent dataset," *2015 4th Int. Conf. Reliab. Infocom Technol. Optim. Trends Futur. Dir. ICRITO 2015*, Dec. 2015, doi: 10.1109/ICRITO.2015.7359281.

Cek Similariti Game Popularity Level During Covid-19 Pandemic Using Agglomerative Hierarchical Clustering

ORIGINALITY REPORT

17%	17%	%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	digilib.uinsgd.ac.id Internet Source	3%
2	www.therecordherald.com Internet Source	3%
3	hanj.cs.illinois.edu Internet Source	3%
4	docplayer.net Internet Source	3%
5	www.atlantis-press.com Internet Source	2%
6	etd.aau.edu.et Internet Source	1%
7	searchworks.stanford.edu Internet Source	<1%
8	www.ijisae.org Internet Source	<1%
9	digitalcommons.usu.edu Internet Source	<1%

10	link.springer.com Internet Source	<1%
11	ejournal.nusamandiri.ac.id Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		