Jurnal ELTIKOM :
Jurnal Teknik Elektro, Teknologi Informasi dan Komputer

Vol. 5, No. 1, Juni 2021, hal. 25-31
ISSN 2598-3245 (Print), ISSN 2598-3288 (Online)
DOI: http://doi.org/10.31961/eltikom.v5i1.221
Tersedia online di http://eltikom.poliban.ac.id

FINDING THE MOST DESIRABLE CAR USING K-NEAREST NEIGHBOR FROM E-COMMERCE WEBSITES

Mohammad Farid Naufal, Yudistira Rahadian Wibisono
Teknik Informatika, Universitas Surabaya, Surabaya, Indonesia
e-mail: faridnaufal@staff.ubaya.ac.id, s160414068@student.ubaya.ac.id

Received: June 22, 2020 – Revised: October 25, 2020 – Accepted: October 26, 2020

ABSTRACT

The increasing number of cars that have been released to the market makes it more difficult for buyer to choose the choice of car that fits with their desired criteria such as transmission, number of kilometers, fuel type, and the year the car was made. The method that is suitable in determining the criteria desired by the community is the K-Nearest Neighbors (KNN). This method is used to find the lowest distance from each data in a car with the criteria desired by the buyer. Euclidean, Manhattan, and Minkowski distance are used for measuring the distance. For supporting the selection of cars, we need an automatic data col-lection method by using web crawling in which the system can retrieve car data from several ecommerce websites. With the construction of the car search system, the system can help the buyer in choosing a car and Euclidean distance has the best accuracy of 94.40%.

***Keywords**: E-Commerce, Euclidean Distance, K Nearest Neighbors, Manhattan Distance, Minkowski Distance.*

ABSTRAK

Meningkatnya penjualan mobil di pasaran membuat pembeli semakin sulit untuk memilih pilihan mobil yang sesuai dengan kriteria yang diinginkan seperti transmisi, jumlah kilometer, jenis bahan bakar, dan tahun pembuatan mobil. Metode yang cocok dalam menentukan kriteria yang diinginkan oleh masyarakat adalah K-Nearest Neighbors (KNN). Metode ini digunakan untuk mencari jarak terendah dari setiap data dalam mobil dengan kriteria yang diinginkan oleh pembeli. Jarak Euclidean, Manhattan, dan Minkowski digunakan untuk mengukur jarak. Untuk mendukung pemilihan mobil, kami memerlukan metode pengumpulan data otomatis dengan melakukan crawling web di mana sistem dapat mengambil data mobil dari beberapa situs web e-commerce. Dengan dibangunnya sistem pencarian mobil, sistem ini dapat membantu pembeli dalam memilih mobil dan jarak Euclidean memiliki akurasi terbaik yaitu 94,40%.

***Kata Kunci**: Euclidean Distance, K Nearest Neighbors, Manhattan Distance, Minkowski Distance, Web Crawling.*

I. INTRODUCTION

THE automotive world today continues to grow. This can be seen from the increasing purchasing power ratio of cars in the community. It is due to the increasing types of car accompanied by various price variants and promos offered by manufacturers. The more types of car on the market, it makes more difficult for buyer to make choices of cars according to the desired criteria. The many types of cars available in the market make prospective buyers set criteria for choosing cars [1].

The results of a survey states that the buyers take 2.9 months to determine desired car, while for a motorcycle states that the buyers take 3 weeks to determine the desired motorcycle [2]. Based on the problems, we need a system that can help speed up the search for cars based on the buyer desired criteria such as transmission, number of kilometers, fuel type, and the year the car was made.

K-Nearest Neighbor (KNN) is a method for comparing two objects based on the closest distance. KNN is one of top 10 data mining algorithms [3][4][5]. KNN has several advantages, including the calculation that is used simply and effectively in the calculation of large amounts of data [6]. KNN can be used for providing the user order of choice for buying cars based on their criteria. Because of there are several distance measurements in KNN, our research question is which one of the best distance measurements for finding the most desirable car for customer. This research will compare three distance measurements which are Euclidean, Manhattan, and Minkowski.

II. BACKGROUND

A. *Previous Research*

Chen et al. [7] proposes user price preferences in boosting recommendation in unexplored categories. Experiment results show fusing price preference improves both top item recommendation and overall item ranking in all settings. Li et al. [8] proposes Matrix Factorization recommender system by considering user-product rating. Janjarasuk et al. [9] proposes product recommendation for recording studio by using genetic algorithm. The method provides reasonable solutions only for a case study in recording studio for power unit selection. [10] uses sentiment analysis, topic modelling, and distributed representations to provide for product recommendations in flipkart e-commerces. Zhao et al. [11] build product recommender system by using user purchase history and their demographic information in their microblog. Prasetyo et al. [12] uses K Means clustering for finding the cheapest product in e-commerce. It uses products from bukalapak, Lazada, and blibli as dataset. Cho et al. [13] uses web usage mining and decision tree induction. The methodology of web usage mining is data mining methods which are decision tree, association rule mining, and product taxonomy.

From those previous researches, there is no product recommender system especially for used car. Used car is different with other product recommendation from previous researches, because only one used car can be sold then product reviews could not be used as considered attribute for used car. Used car also has unique quantitative attributes such as mileage and years of car manufacturing which can easily be considered by using KNN for product recommendations. Figure 1 shows the kind of dataset from previous research.

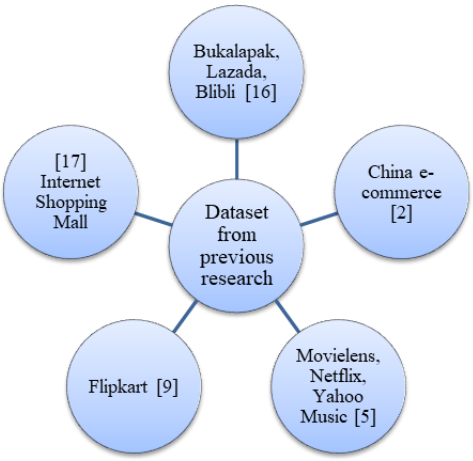
Bukalapak, Lazada, Blibli [16]
[17] Internet Shopping Mall
China e-commerce [2]
Dataset from previous research
Flipkart [9]
Movielens, Netflix, Yahoo Music [5]

Figure 1. Some kind of dataset from previous research

B. *K-Nearest Neighbor (KNN)*

KNN uses distance measurement to determine the level of similarity between two data objects. The common distance measurements are Euclidean, Manhattan, and Minkowski Distance [14]. KNN is widely used in many problems like image processing [15] [16] [17] and text classification [18] [19] [20]. KNN method is simple, works based on calculations to determine the shortest distance between desired car by buyer with obtained car data from ecommerce websites. From the results of the similarity level that is closest to the desired car will be used to display car to prospective buyers. Equation 1, 2, and 3 show the distance measurement of Euclidean, Manhattan, and Minkowski respectively.

$$dE(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \tag{1}$$

$$dMan(x,y) = \sum_{i=1}^{n} | xi - yi | \quad (2)$$

$$dMin(x,y) = \left(\sum_{i=1}^{n} | xi - yi |^3 \right)^{\frac{1}{3}} \quad (3)$$

Where *n* is number of criteria. *x* (*x*1,*x*2,…,*xn*) are the first object point in *n*-spaces. *y* (*y*1,*y*2,…,*yn*) are the first object point in *n*-spaces.

C. Normalization

Normalization is the process of scaling an attribute by using a linear transformation on the original data. Normalization balancing the value to fall within a certain range [21]. Normalization is used to create several attributes to balance the weights on some unbalanced attribute values, the way it works is using scaling of values from several categories which helps the numeric to be made with more valued values. This process makes calculation of the nearest KNN does not weigh on one of the category attributes and to make the weight value of the criteria more balanced to calculate the distance. Equation 4 shows the normalization function.

$$Ni = \frac{Vi - Min\ i}{Max\ i - Min\ i} \quad (4)$$

Where *Ni* is Normalization of *Vi*. *Vi* is value of attribute *i* to be normalized. *Maxi* is maximum value of attribute *i*. *Mini* is minimum value of attribute *i*.

D. Web Crawling

Web crawling is a technique for retrieving data information from a website automatically without having to copy data from the destination website manually. The purpose of web crawling is to retrieve information data based on a particular search topic from a website that aims to collect data, which is then entered into the website pages that are created. We did web crawling from March until July 2019. Web crawling has several steps in processing data retrieval.

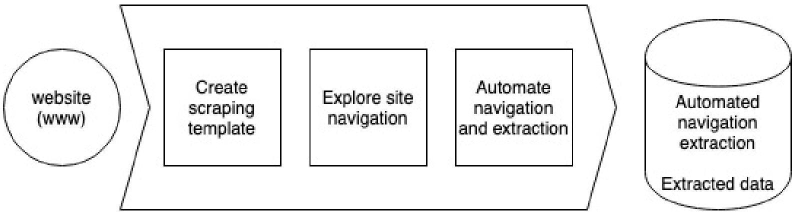

Figure 2. Web Crawling Process

The process of retrieving data by web crawling can be seen in Figure 2. Here are the steps in the web crawling process.

1) *Create Template Scraping.*
   Retrieval of information data from a website is generally taken from websites that use markup languages such as HTML and XHTM. In the process of creating scraping template the program can read the html or xhtml tags that you want to retrieve data from.
2) *Explore Site Navigation.*
   The program can get information based on the location of the html tag contained on the website.
3) *Automatic Navigation and Extraction.*
   After steps 1 and 2 have obtained the required information, a web crawling program is created so that it can be used automatically in retrieving information on the website.

4) *Extracted Data and Package History*.

Information that has been collected through web crawling will be stored in a database that has been provided.

III. Methodology

The detail of this study approaches will be described in this section. There are 4 steps in this research to find the most desirable car from ecommerce websites as in Figure 3.

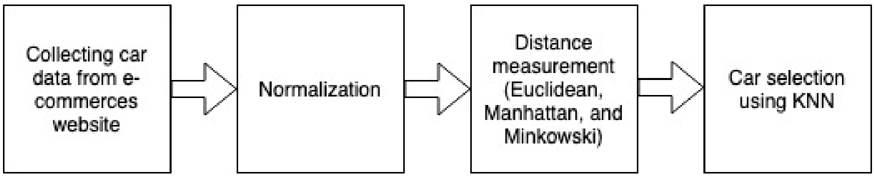
Collecting car data from e-commerces website
Normalization
Distance measurement (Euclidean, Manhattan, and Minkowski)
Car selection using KNN

Figure 3. Methodology

A. Collecting Dataset

Web crawling method is used to retrieve car data from several e-commerce websites that have been promoted by car sellers. The e-commerce websites are olx.co.id, seva.id, carmudi.co.id, and id.priceprice.com in 2019. Figure 4 shows the car product list from those e-commerce websites.

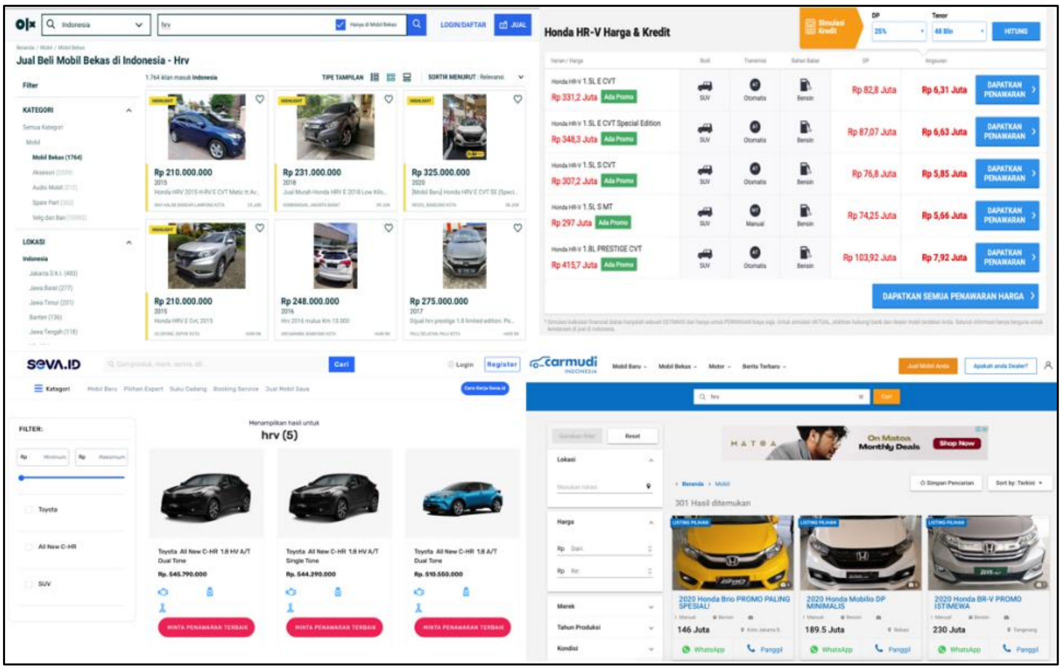
Jual Beli Mobil Bekas di Indonesia - Hrv
Honda HR-V Harga & Kredit
SEVA.ID
carmudi
hrv (5)

Figure 4. The car product list from e-commerce websites

B. Normalization

During the calculation process, the first step taken is to normalize each criterion of car to balance the value. Equation 4 is used for normalizes the criterions. The criterions are price, the year the car was made, and car mileage in Kilometers. Table 1 shows the example of cars criteria that will be normalized. Table 2 shows the normalized cars attribute.

Table 1
The Car Category Values Before Normalization

Car	Price (Rp)	Years	Mileage (KM)
1	98.000.000	2011	150
2	100.000.000	2013	190
3	103.000.000	2015	170

TABLE 2
THE CAR CATEGORY VALUES AFTER NORMALIZATION

Car	Price (Rp)	Years	Mileage (KM)
1	0	0	0
2	0.4	0.5	1
3	1	1	0.5

For example, car 2 price is 100.000.000, the minimum price of the car from the list is 98.000.000 and maximum price of the car from the list is 103.000.000. By using Equation 4 then the normalized value of price for car 2 is:

$$Nprice\ (Car\ 2) = \frac{100.000.000 - 98.000.000}{103.000.000 - 98.000.000} = 0.4$$

The next step is normalizing the buyer desired car criteria. Table 3 shows the buyer desired car before and after normalization.

TABLE 3
THE BUYER DESIRED CAR BEFORE AND AFTER NORMALIZATION

Criteria	Before Normalization	After Normalization
Price (Rp)	100.000.000	0.4
Years	2014	0.75
Mileage (KM)	150	0

C. Distance Measurement

Each car data will be calculated its distance by using Euclidean, Manhattan, and Minkowski distance. After obtaining the distance value, the system will display the car data based on the distance value. The distance calculation result can be seen in Table 4.

TABLE 4
THE DISTANCE CALCULATION RESULT

Car	Euclidean	Manhattan	Minkowski
1	0.85	0.786155007	1.15
2	1.030776406	1.00518144	1.25
3	0.820060973	0.709148619	1.35

D. Car Selection Using K Nearest Neighbors

The system uses the KNN method to get a choice of cars with a degree of distance from the criteria owned by the buyer with the criteria available on the ecommerce websites. The system will display a K-list of choices based on the smallest distance. Table 5 shows the recommendations by using 3 nearest neighbors for each distance. The best distance approaches will be explained in Result section.

TABLE 5
RECOMMENDATIONS BY USING 3 NEAREST NEIGHBORS

Car	Recommendation
Euclidean	Car 3
Manhattan	Car 2
Minkowski	Car 3

E. Evaluation

In determining the accuracy, it is done by making a calculation table by providing the top of 20 cars (K = 20) provided by the system. Respondents will make their order of choice from 1 to 20. The choice will be matched with the order list of cars provided by the system. If the best choice of buyer is first on the list of distance ranking, then the accuracy is 100%, then if the best choice of buyer is second on the list then the accuracy is 95%, then if the best choice of buyer is 20th of the list then the accuracy is 0% and so on. The accuracy and average accuracy formula can be seen in Equation 5 and 6.

$$Accuracy = 100 - \frac{100}{k}(x - 1) \tag{5}$$

$$Accuracy\ Average = \frac{(\sum_{i=1}^{n} Yi)}{n} \tag{6}$$

Where *k* is the number of nearest neighbors. *x* is order of respondent’s choice in distance calculation. *n* is the number of respondents. *Yi* is accuracy value of respondent *i*.

IV. Result and Discussion

A. Result

At this stage the respondent is needed for determining the accuracy. Then it will be known the distance measurement which has the best accuracy. Testing process is conducted by testing the system directly to 42 respondents who looking a car in e-commerce websites. Each respondent will be given 20 choices of cars provided by the system. Order of respondent’s choice in each distance calculation can be seen in Table 6. The accuracy of each distance for every respondent can be seen in Table 7.

TABLE 6
ORDER OF RESPONDENT’S CHOICE IN EACH DISTANCE CALCULATION

Respondent	Order of choice		
	Euclidean	Manhattan	Minkowski
1	1	2	1
2	1	1	1
3	4	4	4
4	1	1	1
…	…	…	…

TABLE 7
THE ACCURACY OF EACH DISTANCE FOR EVERY RESPONDENT

Respondent	Accuracy (%)		
	Euclidean	Manhattan	Minkowski
1	100	95	100
2	100	100	100
3	85	85	85
4	100	100	100
…	…	…	…

TABLE 8
AVERAGE ACCURACY FOR EACH DISTANCE

Rank	Distance	Average Accuracy (%)
1	Euclidean	94.40
2	Manhattan	93.80
3	Minkowski	92.29

B. Discussion

The average accuracy of each distance can be seen in Table 8and Figure 5. Based on the results that has been calculated in Table 7, it can be concluded that the Euclidean has a better accuracy which is 94.40% than the Manhattan and Minkowski which are 93.80% and 92.29% respectively.

Another research [16] uses K Means clustering for finding the cheapest product in e-commerce. It can be concluded that the precision and recall in this research are 74.7% and 73.8% respectively. Although this research uses different dataset, but it only considering the price of the product. If we compared [16] with this research, our method also considering some criteria of the car product which is have a lot of consideration.

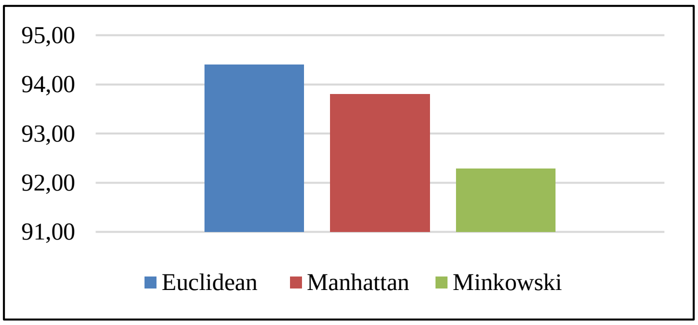

Figure 5. Average accuracy of each distance

V. Conclusion

From the results, it can be concluded that the KNN with Euclidean, Manhattan, and Minkowski as distance measurement can help the buyers in choosing cars originating from more than one ecommerce websites. The system also can assist the buyers in choosing cars based on some different criterions. This can be seen from average percentage of accuracy of the three distances measurement. Euclidean distance has the best accuracy compared to the other two distances.

For further development the method can be combined with the option to add weight values from each criterion then the buyers can choose their preferred criteria. More options for car sources will be added so that the list of cars is displayed more and can minimize the existence of car ads that are no longer active. Clustering process also can be added because it can make searching process faster.

References

[1] “Automotive Revolution & Perspective Towards 2030,” 2016. doi: 10.1365/s40112-016-1117-8.

[2] A. Awalinah, S. Arifin, M. Saf. Sistem Pendukung Keputusan Pembelian Mobil dengan Membandingkan Metode Analytic Hierachy Process dan Fuzzy Associative Memory. Jurnal Teknologi dan Sistem Informasi, pp. 89-100. 2017

[3] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, “Efficient kNN classification with different numbers of nearest neighbors,” IEEE Trans. Neural Networks Learn. Syst., vol. 29, no. 5, pp. 1774–1785, 2018, doi: 10.1109/TNNLS.2017.2673241.

[4] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, “Learning k for kNN Classification,” ACM Trans. Intell. Syst. Technol., vol. 8, no. 3, 2017, doi: 10.1145/2990508.

[5] X. Wu et al., Top 10 algorithms in data mining, vol. 14, no. 1. 2008.

[6] S. Mutrofin, A. Mu'alif, R,V,H. Ginardi, and C Fatichah. Optimasi Teknik Klasifikasi Modified k Nearest Neighbor Menggunakan Algoritma Genetika. Jurnal Gamma, pp:2. 2019

[7] J. Chen, Q. Jin, S. Zhao, S. Bao, L. Zhang, Z. Su, Y. Yu. “Boosting Recommendation in Unexplored Categories by User Price Preference,” ACM Trans. Inf. Syst, pp. 12:1-12:27. 2019

[8] Li, H., Chan, T.N., Yiu, M.L., Mamoulis, N. “FEXIPRO: Fast and Exact Inner Product Retrieval in Recommender Systems,” in Proceedings of the 2017 ACM International Conference on Management of Data, pp 835-850. 2017

[9] U. Janjarasuk, and S. Puengrusme, “Product Recommendation based on Genetic Algorithm,” in Proceedings of the 14th International Conference on Applied Sciences, and Technology (ICEAST), pp. 1-4. 2019

[10] M. Chelliah and S. Sarkar. “Product Recommendations Enhanced with Reviews,” in Proceedings of the Eleventh ACM Conference on Recommender Systems, pp 398-399. 2017

[11] X, W. Zhao, Y. Guo, Y. He, H. Jiang, Y. Wu, and X. Li “We Know What You Want to Buy: A Demographic-based System for Product Recommendation on Microblogs,” in Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 1935-1944. 2014

[12] V. R. Prasetyo, "Searching Cheapest Product on Three Different E-Commerce Using K-Means Algorithm," 2018 International Seminar on Intelligent Technology and Its Applications (ISITIA), Bali, Indonesia, 2018, pp. 239-244, doi: 10.1109/ISITIA.2018.8711043.

[13] Y. H. Cho, J. K. Kim, and S. H. Kim, “A personalized recommender system based on web usage mining and decision tree induction,” Expert Syst. Appl., vol. 23, no. 3, pp. 329–342, 2002, doi: https://doi.org/10.1016/S0957-4174(02)00052-0.

[14] P. Tan, M, Steinbach and V. Kumar. Introduction to Data Mining first edition, Addison-Wesley Longman Publishing Co., Inc. 2005

[15] T. Jinhui, H, Richang, Y. Shuicheng, C. Tat-Seng, Q Guo-Jun, J. Ramesh. “Image annotation by kNN-sparse graph-based label propagation over noisily tagged web images” in ACM Transactions on Intelligent Systems and Technology, vol 3. 2011

[16] J. Ding, H. D. Cheng, M. Xian, Y. Zhang, and F. Xu, “Local-weighted Citation-kNN algorithm for breast ultrasound image classification,” Optik (Stuttg)., vol. 126, no. 24, pp. 5188–5193, 2015, doi: https://doi.org/10.1016/j.ijleo.2015.09.231.

[17] K. S. Angel Viji and D. Hevin Rajesh, “An Efficient Technique to Segment the Tumor and Abnormality Detection in the Brain MRI Images Using KNN Classifier,” Mater. Today Proc., vol. 24, pp. 1944–1954, 2020, doi: https://doi.org/10.1016/j.matpr.2020.03.622.

[18] B. Trstenjak, S. Mikac, and D. Donko, “KNN with TF-IDF based Framework for Text Categorization,” Procedia Eng., vol. 69, pp. 1356–1364, 2014, doi: https://doi.org/10.1016/j.proeng.2014.03.129.

[19] Z. Chen, L. J. Zhou, X. Da Li, J. N. Zhang, and W. J. Huo, “The Lao Text Classification Method Based on KNN,” Procedia Comput. Sci., vol. 166, pp. 523–528, 2020, doi: https://doi.org/10.1016/j.procs.2020.02.053.

[20] S. Tan, “An effective refinement strategy for KNN text classifier,” Expert Syst. Appl., vol. 30, no. 2, pp. 290–298, 2006, doi: https://doi.org/10.1016/j.eswa.2005.07.019.

[21] M. Han, K. P. Jian. Data Mining Concepts and Techniques Third Edition, Morgan Kaufmann Publishers is an imprint of Elsevier. 225 Wyman Street, Waltham, MA 02451, USA, ISBN 978-0- 12-381479-1. 2012