Tourist Places Recommender System Using Cosine Similarity and Singular Value Decomposition Methods

Theriana Ayu Waskitaning Tyas, Z K Abdurahman Baizal’, Ramanti Dharayani

School of Computing, Informatics Study Program, Telkom University, Bandung, Indonesia

Abstract—Tourism in the city of Bandung has various potentials in the field of culture, regional specialties, buildings, and other tourist attractions. On the Tripadvisor page there are many reviews from users who have visited tourist attractions in the city of Bandung. In this case, user reviews are an important element for analysis. The analysis process is carried out using rule-based sentiment analysis. In conducting the review analysis, we use vaderSentiment to weight the positive and negative values. Positive values are subtracted from negative values to get a compound value and converted to a rating value. The rating value obtained is then processed using the Cosine Similarity and Singular Value Decomposition methods to obtain recommendations for tourist attractions in the city of Bandung. For this method, we use the Root Mean Square Error method as a measure of the level of accuracy between the predicted values. The results of the measurement of the level of accuracy produce a value of 3,489 in the Cosine Similarity method, while the Singular Value Decomposition method gets a value of 1,231. The value in the Singular Value Decomposition method is smaller than the Cosine Similarity method with a difference of 2,258 values.

Keywords: Tourism; Recommendation; Cosine Similarity; Singular Value Decomposition; Rating; Review

1. INTRODUCTION

Tourism is a trip to visit tourist sites for recreation or certain purposes. Tourism data shows a significant increase from the Indonesian tourism sector. Indonesian tourism data can show the increasing number of tourists visiting tourist attractions every year [1]. In 2019, the Indonesian Central Statistics Agency showed that the number of tourists visiting tourist attractions for entertainment was 47.61%. The number of tourists in 2019 reached 282 million, both domestic and foreign tourists. Not only Indonesian tourism is experiencing development, but tourism in West Java has developed in various tourism sectors.

West Java has 496 tourist attractions in the form of cultural tourism, nature tourism, building tourism, culinary tourism, and special interest tourism. According to the Department of Tourism and Culture, West Java has great natural potential, with 303 natural attractions [2]. As the capital of West Java, the city of Bandung has a high physical and cultural appeal. The culture the city of Bandung is diverse such as traditional musical instruments, regional languages, regional dances, and special foods. A physical point of view, the city of Bandung has old Dutch buildings and many are still well preserved. On September 25, 2013 by UNESCO determined that the city of Bandung became one of the cities as a world tourist destination [3]. It can be seen from the development of the city of Bandung in terms of transportation, buildings, tourist attractions, and culinary. This makes the development of the city of Bandung which is quite rapid, making many tourist destinations on offer and new tourist attractions continue to emerge from nature to culinary tourism. This development makes more and more tourists visit every year. The growth of visiting tourists can be seen in Figure 1.

![Wisatawan Domestik](image)

**Figure 1.** Bandung City Domestic Tourism Chart

The development of tourism is inversely proportional to the available information about the available tourist attractions. In these conditions, users will be faced with a choice of tourist attractions that have been around for a long time and have only popularity. Many new tourist attractions are not yet known by tourists, due to lack of information and the results of other user reviews. TripAdvisor is a travel site that has many tourist destinations in
various regions. The tourist attractions listed on the TripAdvisor site are quite diverse, such as indoor, outdoor attractions, historical buildings, special foods and much more. Users who have visited certain tourist attractions, will provide a review on the TripAdvisor site. User reviews on the site become an important marketing value, the quality of the reviews shows the level of user satisfaction [4].

From the problems above, it is necessary to have a system that can provide recommendations to users in choosing tourist attractions. In this case, we will do two things, namely converting user reviews into rating values by calculating polarity values and generating recommendations for tourist attractions using the Cosine Similarity and Surprise Factorization Matrix in the Singular Value Decomposition [5],[6].

Matrix Factorization is an approach that is often used in the process of recommender systems. Matrix Factorization process by filling in empty ratings that utilize the characteristics of matrix factorization which is positioned as preprocessing. The next stage is to look for similarities with Cosine Similarity, selecting the same neighbor values and predicting the number of weights. Based on the results of weighting that has a small value, it cannot be used as an appropriate recommendation, and items with a low rating are not recommended properly [7],[8].

Surprise is a Python matrix factor library that supports the use of the Singular Value Decomposition method, which allows us to visualize similarities between items in the data set being used. Supports model evaluation results such as cross-validation iterators and scikit-learn metrics, model selection process and automated hyper-parameter search, namely GridSearch. The use of Surprise to support the Singular Value Decomposition method is the result of fast execution in the exploration of recommendation results [9]–[11].

The purpose of this study is to convert user reviews into rating values, then develop a system to provide the best recommendations for users and that the computing era has developed with systems that assist in data processing. A system for recommending tourist attractions based on user reviews on the TripAdvisor site. Identifying the problem above, we want to compare two different methods to show how user reviews can be converted as rating values to generate recommendations for users. The comparison process between the Cosine Similarity and Singular Value Decomposition methods by looking at the accuracy obtained from the two methods using the Root Mean Square Error (RMSE). Root mean square error has been used as a standard statistical metric to measure model performance in meteorological research studies to measure model error [16].

2. RESEARCH METHODOLOGY

2.1 System Design

In this study, we process user reviews into rating prediction values at the pre-processing stage. The rating prediction value is used to recommend tourist attractions using the Cosine Similarity and Singular Value Decomposition methods. These results were tested using the Root mean square error and comparison was made. An overview of the system is shown in figure 2.

![System Design Diagram](image)

**Figure 2.** System Design

2.2 Dataset

In this study, we use data from TripAdvisor which contains 2872 data. The data includes details of tourist attractions in the city of Bandung (tour id, tourist place name, address, category), reviewer id, reviewer name, review title, and review comments given by users on certain tourist attractions. The data we get is then processed.
by predicting the rating and taking the column that will be used in the next process. The following is the form of the
dataset shown in table 1.

**Table 1. TripAdvisor Dataset**

| id_wisata | id_user | name   | address                | category       | Name_reviewer | Date_reviewer | title            | review                                           |
|-----------|---------|--------|------------------------|----------------|---------------|---------------|------------------|-------------------------------------------------|
| 1         | 1118914 | Alam   | Jl. Kol. Masturi KM 14  | restaurant     | ariadian      | 1512019       | great little escape | the venue was great, i feel so far from the city. the food is very good!... |
|           |         | Wisata | No 157-Cimahi          |                |               |               |                  |                                                 |
| 4         | 612     | Sari Ater Hot Spring | Jalan Raya Ciater Subang | Hot Springs & Geysers | Fatima I     | 16122018 | Half Day Pamper Session | We as a group of friends decided to spend... |
|           |         |        |                        |                |               |               |                  |                                                 |
| 7         | 1379    | Cimahi Waterfall | Jl. Raya Kolonel Masturi, Bandung | Waterfalls | Mary Grace B | 1652018       | Amazing view | Going down is easy but going up is quite a... |

2.3 Rating Prediction Using Vader

In this study, we use Vader (Valence Aware Dictionary and Sentiment Reasoner) in the rating prediction process. Vader is a rule-based sentiment analysis tool adapted from sentiment results on social media[12]. The rating prediction takes information from review_tempatwisata.csv, using the information in column id_wisata, id_user, and review.

2.4 Cosine Similarity

Cosine Similarity is a method used to calculate the similarity value between data and then compare to see the similarity level [13],[14]. This can be done by looking for the similarity value of the data to produce recommendations. The recommendation process is used to measure how similar the data are based on the similarity of different criteria. This method of measuring similarity has several advantages, namely the normalization of long data. This minimizes the effects caused because the data used has quite a lot of data. The following is the algorithm used, as follows:

\[
\cos(\theta) = \frac{A \cdot B}{|A||B|} \quad (1)
\]

Based on the above algorithm, \(A \cdot B\) is a dot vector multiplication of A and B by calculating with \(\sum A \cdot B\), \(|A|\) is the length of vector A, calculated with \(\sqrt{\sum A^2}\), \(|B|\) is the length of vector B, calculated with \(\sqrt{\sum B^2}\).

2.5 Singular Value Decomposition

Singular Value Decomposition is factoring a matrix by dividing it into two matrices and becoming a diagonal matrix containing the factorization value [10]. The parameters that underlie the Singular Value Decomposition model include item factors using rank values by predicting the relationship between items [15]. The Singular Value Decomposition method is applied to a wide variety of application systems including dimension reduction, computer vision, signal processing, etc. The following is the algorithm used, as follows:

\[
R = U \sum V^T \quad (2)
\]

Based on the above algorithm, we can describe that R is the ranking matrix of the users, U is the user feature matrix, V is the user feature matrix and Sigma is the diagonal of the single value matrix. In U and V are orthogonal matrices with user features is how relevant to each tourist spot.

2.5 Comparison Using RMSE

Root Mean Square Error (RMSE) is a method of measuring the level of error by calculating the difference in the value of the predicted model, the form of statistical metrics is also used to measure the performance of the model [16],[17]. Where the calculation results are getting smaller (closer to 0) the RMSE value, the prediction results will be more accurate and better. The following is the algorithm used, as follows:

\[
RMSE = \sqrt{\frac{\sum(X-Y)^2}{n}} \quad (3)
\]
Based on the above algorithm, the values of X and Y are the data to be compared. While the value of n is the amount of existing data.

3. RESULT AND DISCUSSION

3.1 Data Pre-Processing

The first step in our research is to process user review data into rating values. This data processing process aims to change the form of review data information into rating data. The resulting information is then processed using the Cosine Similarity and Singular Value Decomposition methods. The steps in the pre-processing process are eliminating stop words and conducting analysis to get the rating value.

3.1.1 Remove Stopword

Stopword is the process of removing unnecessary characters for analysis. Deletion of words that are considered inappropriate or appear frequently, such as: (.), (,), (!), (?), (;), ($), (%), (:), (_). The process of removing stopwords involves entering a list of stopwords into the clean_data array, storing the input data into variables and splitting the string variables using the string split function into the clean_data array. The results of eliminating Stopword data are shown in table 2.

Table 2. Data Pre-Processing

| Index | Positive Value | Negative Value |
|-------|----------------|----------------|
| 0     | 36.6           | 0.0            |
| 1     | 10.6           | 6.7            |
| 2     | 19.9           | 0.0            |
| 3     | 7.0            | 0.0            |
| 4     | 11.7           | 0.0            |

The resulting data is then processed into rating predictions. SentimentIntensityAnalyzer from vaderSentiment to provide sentiment score as output by dividing three values, namely positive values, negative values and compound values. This value is processed using python library and only compound scores are extracted. After extraction of the combined scores for each user opinion, based on the sum of the positive, negative and neutral scores calculated which are given in table 3.

Table 3. Value Data Pre-Processing

| Index | Positive Value | Negative Value |
|-------|----------------|----------------|
| 0     | 36.6           | 0.0            |
| 1     | 10.6           | 6.7            |
| 2     | 19.9           | 0.0            |
| 3     | 7.0            | 0.0            |
| 4     | 11.7           | 0.0            |

3.2 Rating Result

The previous step is Data Pre-Processing, to convert the results of user reviews into positive and negative values using the Python library (shown in table 2). The determination of the rating value is based on the positive value minus the negative value, so the higher the positive value, the less negative value. The rating value is then converted into a scale from 0 to 5 as shown in Table 4.
The process of assigning a rating value on a scale of 0 to 5 uses the sector division guidelines that have been determined in the algorithm. The rating value that has been generated is then stored in a new data document with the required tables such as id_user, id_wisata and rating. Table selection is based on how important the information in the table is, the data will be reused in the next process.

### 3.3 Cosine Similarity

The similarity value between items is calculated using the calculation of the cosine angle value of two variables or vectors. By comparing each vector and given a predicted rank based on the cosine similarity for each category. In the process of calculating the matrix value, the data is divided into two, namely training data and testing data. The distribution of data is 80% training data and 20% testing data. The larger amount of training data used can affect the value of the output data testing.

#### 3.3.1 Matrix Factorization – Cosine Similarity

The Matrix Factorization technique aims to test the training data to calculate the matrix values that have adjacent similarity values [7]. The matrix value based on factorization is concluded to get a ranking pattern [8]. The algorithm can be seen in table 5.

| Table 5. Algorithm Cosine Similarity |
|--------------------------------------|
| Algorithm 1 Matrix Factorization – Cosine Similarity |
| 1. csm(A,B): |
| 2. value=np.dot(A,B) |
| 3. p1=np.sqrt(np.sum(A**2,axis=1))[:,np.newaxis] |
| 4. p2=np.sqrt(np.sum(B**2,axis=1))[:,np.newaxis,:] |
| 5. return value/(p1*p2)print(csm(A,B)) |

The above algorithm is to calculate the matrix value. The generalization of the cosine similarity is compared with the values in the data matrix B (cosine similarity matrix A vs. B) and compares with itself (cosine similarity matrix using A vs. A). After the Matrix Factorization has been carried out, it is continued with calculations using the Cosine Similarity method to find a set of items that are similar to their neighbors. Vector values cannot have a negative rating and items that do not have a rating are normalized as zero rank. The ranking is shown in Table 6.

| Table 6. Vector Value |
|-----------------------|
| id_user | id_wisata | 14 | 16 | 45 | 61 | 118 | 119 |
| 19 | 0.0 | 0.0 | 2.0 | 1.0 | 0.0 | 0.0 |
| 20 | 4.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 21 | 0.0 | 3.0 | 0.0 | 0.0 | 0.0 | 4.0 |
| 22 | 0.0 | 0.0 | 0.0 | 0.0 | 3.0 | 0.0 |

#### 3.3.2 Output – Cosine Similarity

The results of recommendations using the Cosine Similarity method are a list of recommendations tourist attractions and predictive values addressed to users based on the dataset we currently have. The sample results can be seen in table 7.

| Table 7. Output Cosine Similarity |
|----------------------------------|
| id_wisata | rating |
| 10 | 5 |
| 21 | 5 |
| 8 | 5 |
| 16 | 5 |
3.4 Singular Value Decomposition

The first stage in this method is dividing the data into two. The distribution of data is 80% training data and 20% testing data. The greater the amount of training data used can affect the value of the output data testing. The process in the Singular Value Decomposition method, the process that needs to be done is Tuning hyperparameters as well as modeling using existing train data and testing the designed model to the test data. The test results will be compared with the Cosine Similarity method. The data is processed using optimal parameters in the Singular Value Decomposition method called Grid Search. To get a certain grid vulnerability value and optimal parameter by traversing all points on the grid by dividing the parameters used, this method becomes a solution when the distance is small and the global value when the optimal interval is quite wide[18].

3.4.1 Surprise–Singular Value Decomposition

After doing data splitting, then the data is processed to be carried out for modeling. Python surprise library has provided Singular Value Decomposition function. Surprise is useful in data exploration processes that require fast time to generate predictions. The algorithm can be seen in table 8.

| Table 8. Algorithm Singular Value Decomposition |
|-----------------------------------------------|
| Algorithm 2 Surprise - Singular Value Decomposition |
| 1. algo_svd = SVD() |
| 2. prediction_matrixF = algo_svd.fit(trainset).test(testset) |
| 3. # Prediksi |
| 4. prediction_matrixF |

3.4.2 Output–Singular Value Decomposition

The results of the recommendations using the Singular Value Decomposition method are a list of recommended tourist attractions and predictive values addressed to users based on the dataset used. The sample results can be seen in table 9.

| Table 9. Output Singular Value Decomposition |
|---------------------------------------------|
| id_wisata | value |
|----------|-------|
| 23       | 1164  |
| 21       | 1125  |
| 20       | 993   |
| 19       | 850   |

3.5 Comparison using Root Mean Square Error

The evaluation process of this study uses the Root Mean Square Error (RMSE) method. RMSE as an error metric to show a good numerical prediction by comparing the prediction errors of different models or configuring the model for certain variables. The algorithm can be seen in table 10.

| Table 10. Algorithm Root Mean Square Error |
|-------------------------------------------|
| Algorithm 3 Root Mean Square Error         |
| 1. # Calculating RMSE using rmse()       |
| 2. result = rmse(actual, predicted)      |
| 3. # Printing the value                  |
| 4. print(result)                          |

The results of the evaluation were compared to see the value of the Singular Value Decomposition method and the Cosine Similarity method. From the value generated by the RMSE method, the Cosine Similarity method gets a score of 3.489 and the Singular Value Decomposition method gets a score of 1.231. Seen in figure 3.

Figure 3. Comparison Root Mean Square Error
4. CONCLUSION

The system for providing recommendations for tourist attractions by taking Bandung city tourism as a dataset in this study. Based on the results of this study, the tests carried out using the RMSE method on the Cosine Similarity method using Matrix Factorization and the Singular Value Decomposition method using the Python Surprise library. The evaluation value generated by the Cosine Similarity method is 3.489 and the SVD method is 1.231. In the results of the recommendation system for tourist attractions, the evaluation value of the Cosine Similarity method is greater than that of the Singular Value Decomposition method. This shows that the level of accuracy produced by the Singular Value Decomposition method is better than the Cosine Similarity method. The smaller the evaluation results of the RMSE method, the better the accuracy value obtained. Suggestions for further research are to be able to use data from user reviews that use Indonesian, and to expand stopword removal. The use of parameters that can produce more precise accuracy values, data that can be processed using faster methods.

REFERENCES

[1] BPS-Statistics Indonesia, "Domestic Tourism Statistics 2019", BPS RI / BPS-Statistics Indonesia, 2019.
[2] Utami, D. D., Sinaga, E. K., Desiria, M. K., Febrian, N., Prayitno, R. A., Department, T., Tinggi, S., & Bandung, P. (2020). Potential of Smart Tourism Destination in Bandung City. TEST Engineering and Management.
[3] Fajri, K., & Riyanto E.S. N. (2016). STRATEGI PENGEMBANGAN DESTINASI PARIWISATA KOTA BANDUNG DALAM MENINGKATKAN TINGKAT KUNJUNGAN WISATAWAN ASAL MALAYSIA. Tourism Scientific Journal. https://doi.org/10.32659/tsj.v1i2.9
[4] Xie, K. L., Chen, C., & Wu, S. (2016). Online Consumer Review Factors Affecting Offline Hotel Popularity: Evidence from Tripadvisor. Journal of Travel and Tourism Marketing. https://doi.org/10.1080/10548408.2015.1050538
[5] Vagliano, I., Monti, D., & Morisio, M. (2017). SemRevRec: A recommender system based on user reviews and linked data. CEUR Workshop Proceedings.
[6] Al-Ghuribi, S. M., & Mohd Noah, S. A. (2019). Multi-Criteria Review-Based Recommender System-The State of the Art. In IEEE Access. https://doi.org/10.1109/ACCESS.2019.2954861
[7] Zhao, Q., Zhang, Y., Ma, J., & Duan, Q. (2019). Factored Item Similarity and Bayesian Personalized Ranking for Recommendation with Implicit Feedback. Arabian Journal for Science and Engineering. https://doi.org/10.1007/s13369-018-3358-0
[8] Ardimansyah, M. I., Huda, A. F., & Baizal, Z. K. A. (2017). Preprocessing matrix factorization for solving data sparsity on memory-based collaborative filtering. Proceeding - 2017 3rd International Conference on Science in Information Technology: Theory and Application of IT for Education, Industry and Society in Big Data Era, ICSITech 2017. https://doi.org/10.1109/ICSTech.2017.8257168
[9] Hug, N. (2020). Surprise: A Python library for recommender systems. Journal of Open Source Software. https://doi.org/10.21105/joss.02174
[10] Gutflaish, E., Kantorovich, A., Salzano, S., Biller, O., & Sofer, O. (2019). Temporal anomaly detection: Calibrating the surprise. 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, AAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019. https://doi.org/10.1609/aaai.v33i01.33013755
[11] Sageri Fikri Madramdan, ZK Abdurahman Baizal and Rita Rismala.(2020). Lodging Recommendations Using the SparkML Engine ALS and Surprise SVD. Jurnal Media Informatika Budidarma. https://doi.org/10.30865/mib.v25i2.2257.
[12] Kumar, A., Seth, S., Gupta, S., & Shubham. (2020). Sentiment-enhanced content-based system for online recommendations and rating prediction. International Journal of Gaming and Computer-Mediated Simulations. https://doi.org/10.4018/IGCMS.202004101
[13] Joy, J., & Renumol, V. G. (2020). Comparison of generic similarity measures in E-learning content recommender system in cold-start condition. 2020 IEEE Bombay Section Signature Conference, IBSSC 2020. https://doi.org/10.1109/IBSSC51096.2020.9332162
[14] Protasiewicz, J., Pedrycz, W., Kozłowski, M., Dadas, S., Stanislawek, T., Kopacz, A., & Gałęzewska, M. (2016). A recommender system of reviewers and experts in reviewing problems. Knowledge-Based Systems. https://doi.org/10.1016/j.knosys.2016.05.041
[15] Rodpsysh, K. V., Mirabedini, S. J., & Banirastam, T. (2021). Employing singular value decomposition and similarity criteria for alleviating cold start and sparse data in context-aware recommender systems. Electronic Commerce Research. https://doi.org/10.1007/s10660-021-00848-7
[16] Rahmawati, S., Nurjanah, D., & Rismala, R. (2018). Analisis dan Implementasi pendekatan Hybrid untuk Sistem Rekomendasi Pekerjaan dengan Metode Knowledge Based dan Collaborative Filtering. Indonesian Journal on Computing (Indo-JC). https://doi.org/10.21108/indojc.2018.3.2.210
[17] Ćalasan, M., Abdel Aleem, S. H. E., & Zobaa, A. F. (2020). On the root mean square error (RMSE) calculation for parameter estimation of photovoltaic models: A novel exact analytical solution based on Lambert W function. Energy Conversion and Management. https://doi.org/10.1016/j.enconman.2020.112716
[18] Wang, X., Zhang, F., Kung, H. te, Johnson, V. C., & Latif, A. (2020). Extracting soil salinization information with a fractional-order filtering algorithm and grid-search support vector machine (GS-SVM) model. International Journal of Remote Sensing. https://doi.org/10.1080/01431161.2019.1654142