Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2024. Vol. 3. No. 2; e-ISSN : 2808-4519

Application for Recommending Tourist Attractions on The Island of Java with Content Based Filtering Using Cosine Similarity

Mutiara Sovina[1]*, Yusfrizal[2], Faisal Amir Harahap[3], Ivi Lazuly[4]

[1, 2, 3, 4]Universitas Potensi Utama

**mutiarasovina@gmail.com, yusfrizal80@gmail.com, faisalamirharahap5@gmail.com, ivilazuli2@gmail.com*

Abstract

Indonesia is a country with a high level of tourism, with natural beauty and historical places and other tourist destinations that continue to develop from year to year. Java Island is one of the islands visited by many tourists with various tourist attractions. Currently, many tourists like to travel, but during holidays tourists are confused about which tourist destination to visit. With advances in internet technology and the abundance of information in online media, it can make it easier for tourists to find information, but because there is so much information provided, it will make tourists confused about deciding and choosing a place. A tourist recommendation application is very necessary to provide good recommendation accuracy to make it easier for tourists to find tourist destinations according to the desired category. To get the best results, the Content Based Filtering method using Cosine Similarity used in this research will provide several tourist recommendations according to the level of similarity of various cities on the island of Java.

Keywords: *Tourism; Tourist Attraction; Content Based Filtering; Cosine Similarity*

1. Introduction

There are certainly many and varied tourists in Indonesia, both from nature lovers and historical and urban places, not only from Indonesia, many tourists from abroad also come to enjoy the beauty of Indonesia. Almost all provinces in Indonesia have tourist attractions that are usually visited by tourists, especially in this research recommendation, namely cities in Indonesia, totaling 100 tourist data obtained from accurate datasets. Tourism development in various cities and districts in Indonesia will certainly continue to increase every year. However, there are still many tourists who are still not familiar with all the tourist attractions in this country due to the lack of information available, therefore the need for recommendations for tourist attractions, apart from helping the local economy, introducing tourist attractions can also make Indonesia better known to other countries [1].

When going on a trip, of course there are many things that must be considered, such as choosing to find information about tourism in the area you are going to visit. This of course requires quite a lot of time and the information that is obtained can sometimes make tourists confused. Therefore, we need an application that is able to provide recommendations that can filter the amount of information available on social media, making it difficult for tourists to get the right information that suits their desired tastes. By looking at this problem, the idea was to build a recommendation application that would help tourists find tourist locations with various places that tourists want.

A recommendation application is an application that makes suggestions regarding certain items that can help system users make decisions [2]. Such extensive information can be narrowed down with this system; by analyzing data and information the results will be presented using recommendations system tactics or models, in this research using a Content Based Filtering approach. Which recommends suitable items for users when determining the distance between items, this method uses data from the items used as attributes, for recommendation results researchers use Cosine Similarity which will provide similar results from two vectors obtained from an existing dataset document [3].

This research uses the Cosine Similarity algorithm to classify according to the similarity of journal abstract texts. Cosine Similarity has been widely used to classify texts, such as classifying popular tweets, classifying exam questions, classifying exam answers, classifying student comments on learning evaluation systems and for classifying text documents [4].

2. Research Methodology

These are the stages carried out in this research:

1. Problem Analysis, after seeing the problems that exist in the world of tourism, with the lack of recommendation information on various places from every region in Indonesia, especially in this research, information on tourism on the island of Java was provided; therefore a tourist attraction recommendation application was built to make it easier for tourists to get information

about places. Tourism according to the desired criteria, there are several cities on the island of Java that will provide tourist recommendations according to the data in the dataset.

2. Data Collection, at this stage, data on tourist attractions is collected from the Kaggle website, totaling 100 data on tourist attractions in Jakarta, Yogyakarta, Bandung, Semarang and Surabaya with various categories, such as natural tourism, historical tourism, spiritual tourism and many more tourist attractions. Others will then be processed using a method to produce a recommendation application, the application will provide five recommendations for tourist attractions from each city with various categories in the dataset, using an implementation of the Python programming language which produces Cosine Similarity calculations.
3. Recommendation Calculation, using the Content Based Filtering method to provide recommendation results based on a comparison of criteria owned by tourist attractions, and the results of this method are tourist recommendations that have the highest level of similarity. Next, there is the Cosine Similarity method, which is used because it has very good accuracy and to find the level of similarity between data by calculating similarity vectors, this method can be called Cosine Similarity, and then the data obtained in the previous stage is processed using Cosine Similarity to get recommendation results.

2.1. Recommendation System

A recommendation system is an application that functions to predict an item that is of interest to the user, for example recommendations for films, music, books, news and so on [5]. There are two types of methods applied to recommendation systems, namely Collaborative Filtering and Content Based Filtering. Collaborative Filtering is a recommendation system algorithm where recommendations are given based on consideration of data from other users. Meanwhile, Content Based Filtering provides recommendations by exploring the contents of user profiles, product descriptions or matters related to forming user choices for an item. This research uses the Content Based Filtering method to form items that appear in the recommendations given to users [6].

2.2. Content Based Filtering

The content-based filtering method forms a user profile based on the attributes that make up an item. The content-based filtering method algorithm is explained in the following stages [7]:

1. An item of goods is separated based on a vector of its constituent components.
2. Users will give a like or dislike rating to the item.
3. The system will form a user profile based on the vector weights of the components that make up an item. Creating user profiles can use the TF-IDF (term frequency-inverse document frequency) algorithm. TF is the number of terms in a document.

The system will carry out an assessment based on an analysis of the similarity of the user profile with the component vectors that form the item. If the item will be liked by the user then the item will be recommended to the user. The main drawback of this method is that it is unable to recommend types of items that are new or have never been seen to a user. This is because this method is based on items that have been rated by the user [8].

The content-based filtering method forms a user profile based on the attributes that make up an item. The content-based filtering method algorithm is explained in the following stages [9]:

1. An item of goods is separated based on a vector of its constituent components.
2. Users will give a like or dislike rating to the item.
3. The system will form a user profile based on the vector weights of the components that make up an item. Creating user profiles can use the TF-IDF (term frequency-inverse document frequency) algorithm. TF is the number of terms in a document.

The system will carry out an assessment based on an analysis of the similarity of the user profile with the component vectors that form the item. If the item will be liked by the user then the item will be recommended to the user [10]. The main drawback of this method is that it is unable to recommend types of items that are new or have never been seen to a user. This is because this method is based on items that have been rated by the user.

The Content-Based Filtering method is a recommendation system method based on content or features of the items then compared with the items that user liked before. The recommendation system using Content-Based Filtering method is only based on the item that user is looking for or likes own and not involve other users in making the recommendations [7]. Thus, if the user changes, the technique with Content-Based Filtering is still possible to adjust the recommendation or suggestion of the appropriate item in a short time. The advantages and disadvantages of the Content-Based Filtering are:

Advantages [11]:

1. The Content-Based Filtering method only requires the content of the item and the user profile itself for recommendations.
2. The Content-Based Filtering method can explain the features of the item on which the recommendation is based to the user.
3. New item can be recommended to users even though they don't have ratings from other users because they are based on the content of the item.

Disadvantages [11]:

1. If the content on an item does not include complete and sufficient information to accurately distinguish it from other items, the recommendations will be less precise.
2. Serendipity constraints (unexpected events), where the system with this method will be difficult to provide recommendations or suggestions that are not unexpected items that are selected only based on content.

2.3. Cosine Similarity

The adjusted cosine similarity equation is used to calculate the similarity value between items. This similarity calculation is a modification of the vector-based similarity calculation which takes into account the fact that each user has a different rating scheme.

Sometimes users give a high rating to item a, on the other hand users give a very low rating to item b. Therefore, for each rating, it is reduced by the average rating given by the user [12].
Cosine similarity is a method for measuring the similarity between two n-dimensional vectors which is usually used in the information search field to compare two texts or documents. If two texts or documents are increasingly similar, the value of cosine similarity will be closer to 1, whereas if the value of cosine similarity is close to 0 then the two texts or documents are increasingly dissimilar [13].

2.4. Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) weighting method is a method used to gibe weight to the relationship of a term in a document by combining to concept for weight calculation, namely Term Frequency (TF) which is the frequency of occurrence of words in the document and Inverse Document Frequency (IDF) which is the inverse frequency of document containing words [14].

3. Results and Discussion

In these results and discussion, researchers obtain a dataset, content based, cosine similarity which will be presented in table form and an explanation of the results that have been obtained.

3.1. Datasets

The collection of datasets obtained was sourced from the Kaggle.com website, where the website provides a wide variety of datasets that are accurate and in accordance with what is needed. In this research the author chose a dataset of tourism recommendations in Indonesia, especially on the island of Java, from the data above it has 100 data of which there are 50% natural tourism and 50% historical tourism so that it reaches 100 data from various cities on the island of Java such as Jakarta, Yogyakarta, Bandung, Surabaya and Semarang. After taking the dataset, the author will implement it in the form of a recommendation application using the Content Based Filtering and Cosine Similarity methods.

Table 1: Example Datasets

Pack	City	Place_Tourism1	Place_Tourism2	Place_Tourism3	Place_Tourism4	Place_Tourism5
1	Jakarta	Pasar Tanah Abang	Taman Ayodya	Museum Tekstil	-	-
2	Jakarta	Pasar Tanah Abang	Pasar Taman Puring	Pasar Petak Sembilan	-	-
3	Jakarta	Perpustakaan Nasional	Monas	Masjid Istiqlal	-	-
4	Jakarta	Pulau Tidumg	Pulau Bidadari	Pulau Pari	Pulau Pramuka	Pulau Pelangi
…	Bandung	Gunung Tangkupan Perahu	Gunung Papandayan	Gunung Manglayang	Curug Dago	Curug Batu Templek
100	Surabaya	Taman Buah Surabaya	Hutan Bambu Keputih	Taman Barunawati	Kebun Bibit Wonorejo	Taman Mundu

From some of the data in the dataset, there are some that are missing, for example in the city of Jakarta where there is no tourist list in place_tourism 4 and 5, while for the cities of Bandung and Surabaya it appears to be completely filled in place_tourism 1 to 5.

3.2. Recommendation Calculation Results

In the calculation results using an implementation using the Python programming language which will provide tourist recommendations according to the package that the user will input, the author will try 5 tests by inputting different packages. Each test will produce 5 tourist recommendations in the dataset which will be sorted based on the highest similarity.

Table 2: Recommendation Results

Package_tourism	Package	Place_Tourism 1
10	88	Rumah Batik
	53	Rainbow Garden
	9	Taman Impian Jaya Ancol
	52	Alun-Alun Kota Bandung
	72	Monumen Palagan Ambarawa
Package_tourism	**Package**	**Place_Tourism 2**
10	88	Jembatan Merah
	53	Kota Mini
	9	Kota Tua
	52	Taman Balai Kota Bandung
	72	Benteng Pendem
Package_tourism	**Package**	**Place_Tourism 3**
10	82	Klenteng Sanggar Agung
	94	Museum De Javasche Bank
	0	Museum Tekstil
	71	Masjid Agung Ungaran
	70	Benteng Pendem
Package_tourism	**Package**	**Place_Tourism 4**
10	76	La Kana Chapel
	28	Embung Tambakboyo
	68	Wisata Eling Bening
	65	Brown Canyon

	73	-
Package_tourism	**Package**	**Place_Tourism 5**
10	72	Semarang Chinatown
	71	-
	70	Hutan Pinus Kayon
	69	-
	68	Gua Maria Kerep Ambarawa
Package_tourism	**Package**	**Place_Tourism 1**
92	99	Taman Mundu
	18	Taman Spathodea
	19	Taman Lapangan Banteng
	97	Taman Flora Bratang Surabaya
	98	Taman Air Mancur Menari Kenjeran
Package_tourism	**Package**	**Place_Tourism 2**
92	16	Museum Layang-layang
	9	Museum Fatahillah
	85	Museum Mpu Tantular
	98	Museum Mpu Tantular
	54	Museum Pos Indonesia
Package_tourism	**Package**	**Place_Tourism 3**
92	85	Balai Kota Surabaya
	98	Surabaya Museum (Gedung Siola)
	0	Museum Tekstil
	50	Kota Mini
	45	Kota Mini
Package_tourism	**Package**	**Place_Tourism 4**
92	64	Masjid Agung Ungaran
	0	Taman Ayodia
	71	Masjid Kapal Semarang
	70	Monumen Plagan Ambarawa
	69	Pura Giri Natha
Package_tourism	**Package**	**Place_Tourism 5**
92	92	Museum TNI AL Loka Jala Crana
	15	Museum Joang 45
	56	Museum Gedung Sate
	28	Gembira Loka Zoo
	4	Museum Satria Mandala

The results show that there are five tourist recommendations according to the level of similarity of various cities on the island of Java. These results were obtained from Cosine Similarity calculations using the Python programming language, as in the table above there are several tourist attractions that were not found according to the contents of the existing dataset. . The five tourist attractions are taken from several cities on the island of Java such as Jakarta, Yogyakarta, Bandung, Semarang and Surabaya. With various categories ranging from natural tourism, historical places to other destinations. The implementation of the Cosine Similarity method for tourist recommendations has accurate accuracy based on testing using Google Collab tools, influenced by the level of similarity in tourist categories and similarity of places. A system that applies a classification process can be an alternative to simplify and speed up analysis in making decisions about recommending tourist attractions according to what users or tourists want.

4. Conclusion

Based on the results of the design, discussion and testing of the system implementation carried out. So it can be concluded as follows:

1. The tourist recommendation application using the Content Based method using Cosine Similarity calculations is running well.
2. The application can group tourist recommendations based on the selected tourism package with the result that there are 5 places that the application will recommend.
3. Users can choose whatever package they want with a range of 1-100.
4. In this application there are 100 tourist attractions in various cities on the island of Java, such as Jakarta, Yogyakarta, Bandung, Semarang and Surabaya.
5. With the Content Based method and Cosine Similarity calculations, the results provided will be accurate with a high and appropriate level of similarity to the criteria.

Acknowledgement

Thank you to all those who have helped complete this research.

References

[1] R. Dodds, A. Ali, and K. Galaski, “Mobilizing knowledge: Determining key elements for success and pitfalls in developing community-based tourism,” *Curr. Issues Tour.*, vol. 21, no. 13, pp. 1547–1568, 2018.

[2] S. L. Kolasinski *et al.*, "2019 American College of Rheumatology/Arthritis Foundation guideline for the management of osteoarthritis of the hand, hip, and knee," *Arthritis Rheumatol.*, vol. 72, no. 2, pp. 220–233, 2020.
[3] E. Bolturk and C. Kahraman, "A novel interval-valued neutrosophic AHP with cosine similarity measure," *Soft Comput.*, vol. 22, pp. 4941–4958, 2018.
[4] D. Gunawan, C. A. Sembiring, and M. A. Budiman, "The implementation of cosine similarity to calculate text relevance between two documents," in *Journal of physics: conference series*, 2018, vol. 978, p. 12120.
[5] M. Nilashi, O. Ibrahim, and K. Bagherifard, "A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques," *Expert Syst. Appl.*, vol. 92, pp. 507–520, 2018.
[6] G. Geetha, M. Safa, C. Fancy, and D. Saranya, "A hybrid approach using collaborative filtering and content based filtering for recommender system," in *Journal of Physics: Conference Series*, 2018, vol. 1000, p. 12101.
[7] S. H. Nallamala, U. R. Bajjuri, S. Anandarao, D. D. Prasad, and P. Mishra, "A Brief Analysis of Collaborative and Content Based Filtering Algorithms used in Recommender Systems," in *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 981, no. 2, p. 22008.
[8] D. Liu, X. Chen, and D. Peng, "Some cosine similarity measures and distance measures between q-rung orthopair fuzzy sets," *Int. J. Intell. Syst.*, vol. 34, no. 7, pp. 1572–1587, 2019.
[9] R. Glauber and A. Loula, "Collaborative filtering vs. content-based filtering: differences and similarities," *arXiv Prepr. arXiv1912.08932*, 2019.
[10] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, 2020.
[11] L. Fu and X. Ma, "An improved recommendation method based on content filtering and collaborative filtering," *Complexity*, vol. 2021, pp. 1–11, 2021.
[12] J. Huetle-Figueroa, F. Perez-Tellez, and D. Pinto, "Measuring semantic similarity of documents with weighted cosine and fuzzy logic," *J. Intell. Fuzzy Syst.*, vol. 39, no. 2, pp. 2263–2278, 2020.
[13] I. Indriyanto and I. D. Sumitra, "Measuring the level of plagiarism of thesis using vector space model and cosine similarity methods," in *IOP Conference Series: Materials Science and Engineering*, 2019, vol. 662, no. 2, p. 22111.
[14] S. Sintia, S. Defit, and G. W. Nurcahyo, "Product Codefication Accuracy With Cosine Similarity And Weighted Term Frequency And Inverse Document Frequency (TF-IDF)," *J. Appl. Eng. Technol. Sci.*, vol. 2, no. 2, pp. 62–69, 2021.