J Statistika Vol. 15, No. 1, (2022), Hal. 152 - 157 www.unipasby.ac.id

Hotel Recommendation System with Content-Based Filtering Approach (Case Study: Hotel in Yogyakarta on Nusatrip Website)

Cheryl Ayu Melyani [(1)], Ayundyah Kesumawati [(2)], Raden Bagus Fajriya Hakim [(3)], Arum Handini Primandari [(4)]
Department of Statistics, Universitas Islam Indonesia
Jl.Kaliurang km 14,5, Yogyakarta 55584, Indonesia
e-mail: cherylayu25@gmail.com

ABSTRAK

Meningkatnya pandemi Covid-19 membuat aktivitas masyarakat menjadi terhambat seringkali memunculkan stress jika harus berada dirumah secara terus menerus. Hal ini menyebabkan meningkatnya tren staycation atau kegiatan berlibur di kota sendiri dengan menyewa sebuah hotel. Teknologi penyewaan hotel mulai dialihkan dengan adanya OTA (Online Travel Agent). Adanya berbagai macam hotel dengan berbagai macam fasilitas yang disuguhkan membuat masyarakat sering merasa kebingungan dalam memilih hotel yang akan ditempatinya. Untuk membantu mengatasi hal tersebut, peneliti membuat sebuah sistem rekomendasi untuk membantu calon penghuni hotel dalam memilih hotel sesuai dengan pilihannya. Selain itu juga dapat membantu perusahaan dalam meningkatkan pemesanan kamar hotel melalui website-nya. Dalam penelitian ini, peneliti membangun sebuah sistem rekomendasi hotel di Yogyakarta pada salah satu OTA di Indonesia menggunakan Content-Based Filtering Methods, pembobotan data teks menggunakan Term Frequency-Invers Document Frequency (TF-IDF) Methods, dan mengukur kemiripan dokumen menggunakan Cosine Similarity Methods. Data yang digunakan adalah deskripsi dari masimg-masing hotel. Berdasarkan hasil rekomendasi hotel Good Karma Yogyakarta sebagai contoh pengujian, didapatkan 10 hotel yang mirip yaitu Happy Buddha Yogyakarta – Hostel, Nextdoor Homestay, Hotel Puspita, OYO 426 Hotel Gading Resto, Omah Jegog Homestay, Prawirotaman Homestay, RedDoorz near Prawirotaman, Ayodhya Garden Hostel Yogyakarta by HOM, Bringin House Yogyakarta, dan House 24 Yogyakarta dengan nilai cosine similarity secara berturut-turut sebesar 0.956666513, 0.946570717, 0.917459394, 0.912534877, 0.886439718, 0.88221982, 0.881775275, 0.875845794, 0.872030219, dan 0.871514859.

Kata Kunci : Staycation, Hotel, Recommendation System, Content-Based Filtering

ABSTRACT

The increasing of Covid-19 pandemic has hampered people's activities, often causing stress if they are only stay at home continuously. This has led to an increasing trend of staycations or holiday activities in the city itself by renting a hotel. Hotel rental technology has begun to be transferred with the existence of OTA (Online Travel Agent). The existence of various kinds of hotels with various kinds of facilities that makes people feel confused in choosing which hotel to occupy. To help overcome this, the researchers create a recommendation system to help prospective hotel residents choose the hotel according to their choice. In addition, it can also assist companies in increasing hotel room reservations through its website. In this study, researchers build a hotel recommendation system in Yogyakarta at one of the OTAs in Indonesia using Content-Based Filtering Methods, weighting text data using Term Frequency-Inverse Document Frequency (TF-IDF) Methods and measuring document similarity using Cosine Similarity Methods. The data used is a description of each hotel. Based on the results of the Good Karma Yogyakarta hotel recommendations as a test example, 10 similar hotels were obtained, namely Happy Buddha Yogyakarta – Hostel, Nextdoor Homestay, Hotel Puspita, OYO 426 Hotel Gading Resto, Omah Jegog Homestay, Prawirotaman Homestay, RedDoorz near Prawirotaman, Ayodhya Garden Hostel Yogyakarta by HOM, Bringin House Yogyakarta, and House 24 Yogyakarta with cosine similarity values 0.956666513, 0.946570717, 0.917459394, 0.912534877, 0.886439718, 0.88221982, 0.881775275, 0.875845794, 0.872030219, and 0.871514859.

Keywords : Staycation, Hotel, Recommendation System, Content-Based Filtering

1. INTRODUCTION

The COVID-19 pandemic, which has lasted for approximately 2 years has an impact on various sectors. Activities that people do are hampered. On the one hand, people do not want to be infected with the Corona Virus and decide to carry out all activities in their respective homes. On the other hand, they realized that if they stay at home all time, it will create stress. Based on Google Trends data, the graph of the staycation trend tends to rise from the beginning of 2020 until now. According to HowStuffWorks, a staycation is a combination of taking a vacation and staying at home (Layton, 2009). A survey conducted by Wego.co.id regarding the staycation trend, shows that Yogyakarta is one of the most popular destinations as a destination city for staycation. This is shown by the number of hotel searches in May-August 2020 on their platform (Ahsan, 2020).

In this digital era, the development of information technology has grown rapidly where various innovations in offering a product or service have emerged. One of the innovations that has evolved from conventional to modern is in the tourism sector (Sandi Wachyuni & Wiweka, 2020). The existence of an Online Travel Agent (OTA) is able to replace hotel booking activities where previously book a hotel directly, now it can be done online using website-based OTA (Hendriyati, 2019). There are various kinds of star and non-star hotels in Yogyakarta, people who want to do a staycation will be confused in choosing a hotel according to the criteria and facilities of interest. OTA provides a variety of hotel options that can be tailored to the needs and desires of the user. All information regarding the location, facilities, and advantages of the hotel which includes explanations about rooms, restaurants, swimming pools, parking areas, and others are explained in the description column on an OTA website. In this study, researchers create a hotel recommendation system in Yogyakarta at one of the OTAs in Indonesia which is Nusatrip.com using Content-Based Filtering Methods which is recommendation method based on content or features of the items, weighting text data using Term Frequency-Inverse Document Frequency (TF-IDF) , and measuring document similarity using Cosine Similarity to make it easier for someone in choosing a hotel. One of studies about recommendation system is "Research Supervisor Recommendation System Based on Topic Conformity". the purpose of this research is to make it easier for students to determine their supervisor according to the research topic. This study applies word weighting using TF_IDF method and calculating the similarity between two objects using Cosine Similarity Method. Based on the results of the analysis, the accuracy value of comparison of the recommendations from the system with the actual data is 75% .

2. METHOD AND MATERIALS

2.1. Data and Data Source

This study used secondary data that obtained form Nusatrip.com website on November 16, 2021. The variables used are name, description, and hotel star data in Yogyakarta.

2.2. Recommendation System

Recommendation system is a software and a technique that suggests an item of interest to users and intended to support its use in various decision-making processes (Larasati & Februariyanti, 2021). For example, in the decision-making process such as what items will users buy, what music will be heard, what hotels will be ordered, what news will be read, and much more. In how its works, the recommendation system will not only recommend items that are most liked by user but also provide recommendations personally to each user which will provide item recommendations according to user (Mondi et al., 2019). The recommendation system is one of the most powerful and popular information discovery tools on a website to overcome information overload and can help them to make good choices (Ricci et al., 2015).

2.3. Content-Based Filtering

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 (Raghuwanshi & Pateriya, 2019). 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. 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 (Purnaramadhan, 2021). The advantages and disadvantages of the Content-Based Filtering are (Aamir & Bhusry, 2015):

Advantages:

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:

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.

2.4. Text Preprocessing

The first step that must be passed before the analysis process is prepare the data to be analyzed. A text usually has a messy data structure, so it cannot be processed immediately. At this stage, the text will be cleaned of unnecessary parts until its prepared into a structured text and ready to be processed further.

1. Case Folding: the process of converting capital letter to lowercase. It its intended that the same word but containing capital letters is not detected differently from words but containing capital letters is not detected differently from words that do not contain capital letters.
2. Remove Punctuation: The stage for remove punctuation characters.
3. Remove Whitespace: Process of removing excess whitespace in text or documents.
4. Remove Stopword: Step to remove words that have no meaning or meaningless.
5. Lemmatization: A stage to find the basic form of a word (lemma) according to morphological analysis and dictionaries (Supriyati & Iqbal, 2018).

Tokenizing: Process to separating words in a text or document based on hyphens (-) and spaces.

2.5. 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 (Nurjannah & Fitri Astuti, 2013). The equation used to calculate TF-IDF:

$$W_{ij} = TF_{ij} \times \left(\ln\left(\frac{D}{df_j}\right) + 1\right) \quad (1)$$

with:
W_{ij} : Weight of term j to document i
TF_{ij} : Frequency of occurrence of term j in document i
IDF_j : Inverse document frequency of term j
D : Total documents
df_j : Total documents from term j

2.6. Cosine Similarity

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 (Jannach et al., 2011). 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. The equation used to calculate cosine similarity:

$$Cos \propto = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (2)$$

With:
A : Vector A
B : Vector B
A_i : Term i in document A
B_i : Term I in document B
$A \cdot B$: Dot product between vector A and vector B
$|A|$: Vector length A
$|B|$: Vector length B
$|A||B|$: Cross product between $|A|$ and $|B|$

3. RESULT AND DISCUSSION

3.1. Descriptive Analysis

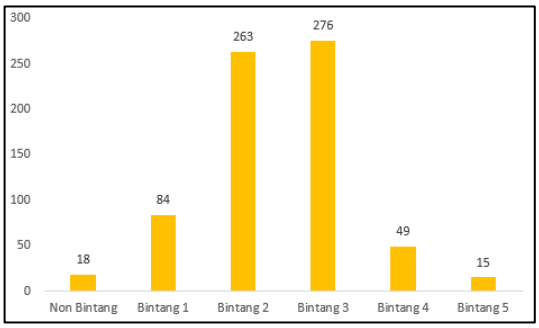

Figure 1. Number of Hotels in Yogyakarta by Star

The graph shows the number of hotels in Yogyakarta on Nusatrip.com website classified by star. From the graph, the largest distribution of hotels is in two-star and three-star hotels. One of the factors that the hotel has a high of frequency is adapting to the pattern of consumer tendencies. The two-star and three-star hotels has quite complete facilities and the price is not too expensive. This is in line with the opinion expressed by Co-Founder and Chief Marketing Office of Tiket.com in Republika.co.id article where consumers tend to choose accommodation with low process, get discounts, and promotions. This sees the tendency of consumers to choose hotel accommodation (Asrianti & Azizah, 2020).

3.2. Preprocessing Data

The first stage in the analysis process is data preprocessing. Data preprocessing carried out in this research is Case Folding, Remove Punctuation, Stopword Removal, Lemmatization, and Tokenizing text. This is an example of the data preprocessing stage using 132nd document that present in Table 1.

Table 1. Original Data

Hotel Descriptions
This property located near Malioboro Street. This smoke-free hostel features a 24-hour front desk, tour/ticket assistance, and a garden. WiFi in public areas is free. Rooms offer free WiFi and showers. Take in the views from a terrace and a garden and make use of amenities such as complimentary wireless Internet access. Nearby Attractions Batik Plentong 0.5 km / 0.3 mi Lana Gallery 1 km / 0.6 mi Masjid Jogokariyan 1.3 km.

This is the result of a document that has gone through the data preprocessing stage document that present in Table 2.

Table 2. Data after preprocessing

Hotel Descriptions
property located near malioboro street smoke free hostel feature 24 hour front desk tourticket assistance garden wifi public area free room offer free wifi shower view terrace garden use amenity complimentary wireless internet access… nearby attraction batik plentong 0 5 km 0 3 mi lana gallery 1 km 0 6 mi masjid jogokariyan 1 3 km ……

After that, the data that has been processed is separated for each word namely tokenizing.

Table 3. Word tokenizing

Hotel Descriptions
'property', 'located', 'near', 'malioboro', 'street', 'smoke', 'free', 'hostel', 'feature', '24', 'hour', 'front', 'desk', 'tourticket', 'assistance', 'garden', 'wifi', 'public', 'area', 'free', 'room', 'offer', 'free', 'wifi', 'shower', 'view', 'terrace', 'garden', 'use', 'amenity', 'complimentary', 'wireless', 'internet', 'access',…. , 'nearby', 'attraction', 'batik', 'plentong', '0', '5', 'km', '0', '3', 'mi', 'lana', 'gallery', '1', 'km', '0', '6', 'mi', 'masjid', 'jogokariyan', '1', '3', 'km', …..

3.3. Term Weighting using TF-IDF

The first step to calculate term weighting is calsulate the vealue of Term Frequency (TF). This is a table of term frequency:

Table 4. TF Table

Doc	laundry	nearby	⋯	wifi
1	1	0	⋯	0
2	1	0	⋯	0
3	0	1	⋯	1
4	**1**	2	⋯	2
5	0	0	⋯	0
⋮	⋮	⋮	⋯	⋮
705	0	0	⋯	0
df	**164**	461	⋯	456

The example used in this calculation is the word “laundry” in 4th document. A value of 0 in TF

table means that in a document, a certain word does not appear in the document, while a value of 1 means that in a document, a certain word appears 1 time in the document. The TF value for the word "laundry" in 4th document is 1, so it means that the word "laundry" appears 1 time in the 4yh document. Next, look for the df value which is the document frequency value. For example, in word "laundry", there are 164 documents containing that word. D is total document, which is 705 documents.

$$W_{ij} = TF_{ij} \times \left(\ln\left(\frac{D}{df_j}\right) + 1\right)$$

$$W_{4;laundry} = 1 \times \left(\ln\left(\frac{705}{164}\right) + 1\right)$$

$$W_{4;laundry} = 1 \times (1{,}458331 + 1) = 2{,}458831$$

Based on the result, the weight of the word "laundry" in 4th document is 2.458831.

3.4. Cosine Similarity

Previously, the value of word weighting has been calculated using TF-IDF, then calculate the document similarity. After getting the similarity value of a document with other documents, it will be sorted from the largest and the top 10 documents with the highest similarity value will be taken as recommendations. For example, the 132nd and 170th documents as a sample for manual calculations.

$$Cos \propto = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

$$Cos \propto = \frac{1638{,}29225}{\sqrt{1708{,}006} \times \sqrt{1812{,}564}} = 0{,}956666513$$

From the results of cosine similarity calculation, the 132nd document has the highest similarity value to the 170th document by 96% and it followed by nine other documents. According to the data obtained, the 132nd document is a document from Good Karma Yogyakarta-Hostel. Thus, 10 sequences hotel recommendations that are similar with Good Karma Yogyakarta Hostel from the recommendation system are shown in Table 5. The table contains the document number, hotel's name, and cosine similarity value of the hotel document to other hotel documents. Then 10 order of hotel documents which have largest cosine similarity value are displayed in Table 5.

Table 5. Recommendation Result

No	Doc	Hotel's Name	Cosim Value
1	170	Happy Buddha Yogyakarta - Hostel	0,9566
2	303	Nextdoor Homestay	0,9465
3	210	Hotel Puspita	0,9174
4	360	OYO 426 Hotel Gading Resto	0,9125
5	315	Omah Jegog Homestay	0,8864
6	422	Prawirotaman Homestay	0,8822
7	501	RedDoorz near Prawirotaman	0,8817
8	45	Ayodhya Garden Hostel Yogyakarta by HOM	0,8758
9	63	Bringin House Yogyakarta	0,8720
10	223	House 24 Yogyakarta	0,8715

When the user clicks on Good Karma Yogyakarta-Hostel, the system will provide 10 other hotel recommendations that have the highest similarity to the hotel that user has chosen. The similarity is measured based od the similarity of the description of each hotel.

3.5. Application of the Recommendation System on Website

Deploying a recommendation system that has been create into a simple website using Flask Python and Heroku to make it easier for users to find hotels that are similar with the previous selected hotel. The appearance pf the recommendation system application website presented in Figure 2 and Figure 3.

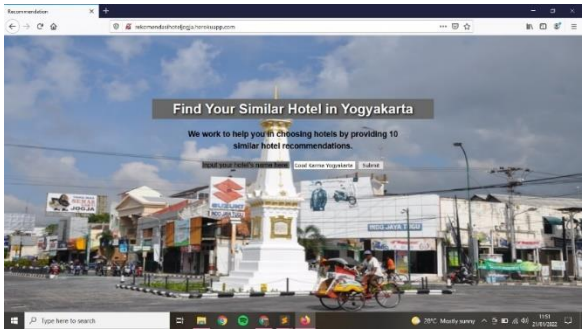

Figure 2. Search Page

Figure 3. Recommendation Result Page

Accessible via https://rekomendasihoteljogja.herokuapp.com/

4. Conclusion

Based on the result and discussion of this study, it can be concluded that document similarity calculation using cosine similarity in this study as an example of using the 132nd document. There were 10 documents with the highest similarity values which are 170th, 303rd, 210th, 360th, 315th, 422nd, 510th, 45th, 63rd, and 223rd documents. These documents have a cosine similarity value of 0.956666513, 0.946570717, 0.917459394, 0.912534877, 0.886439718, 0.88221982, 0.881775275, 0.875845794, 0.872030219, and 0.871514859. The result of hotel recommendations in Yogyakarta as an example using the 132nd document, namely Good Karma Yogyakarta-Hostel with Content-Based Filtering generated in this study based on the similarity of hotel descriptions are Happy Buddha Yogyakarta – Hostel, Nextdoor Homestay, Hotel Puspita, OYO 426 Hotel Gading Resto, Omah Jegog Homestay, Prawirotaman Homestay, RedDoorz near Prawirotaman, Ayodhya Garden Hostel Yogyakarta by HOM, Bringin House Yogyakarta, and House 24 Yogyakarta.

REFERENCES

Aamir, M., & Bhusry, M. (2015). Recommendation System: State of the Art Approach. *International Journal of Computer Applications*, *120*(12), 25–32.

Ahsan. (2020.) Staycation Kala Pandemi, Bandung dan Yogyakarta Jadi Incaran. Diambil kembali dari https://travel.wego.com/berita/staycationkalA-pandemi-bandung-dan-yogyakarta-jadi-incaran/

Asrianti, S, &Azizah, N. (2020). Pandemi Corona Ubah Kecenderungan Wisatawan Pilih Akomodasi. Diambil kembali pada https://www.republika.co.id/berita/qc6j66463/pandemicoronaubah-kecenderungan-wisatawan-pilih-akomodasi

Hendriyati, L. (2019). Online travel agent. *Jurnal Media Wisata*, *17*(1), 1–10.

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2011). *Recommender systems: an introduction*. Cambridge University Press.

Larasati, F. B. A., & Februariyanti, H. (2021). Sistem Rekomendasi Product Emina Cosmetics Dengan Menggunakan Metode Content - Based Filtering. *Jurnal Manajemen Informatika Dan Sistem Informasi*, *4*(1), 45.

Layton, Julia. 2009. What is a Staycation. Diambil kembali pada https://money.howstuffworks.com/personal-finance/budgeting/staycation.htm

Mondi, R. H., Wijayanto, A., & Winarno. (2019). Recommendation System With Content-Based Filtering Method for Culinary Tourism in Mangan Application. *Itsmart*, *8*(2), 65–72.

Nurjannah, M., & Fitri Astuti, I. (2013). Penerapan Algoritma Term Frequency-Inverse Document Frequency (Tf-Idf) Untuk Text Mining. Mahasiswa S1 Program Studi Ilmu Komputer FMIPA Universitas Mulawarman Dosen Program Studi Ilmu Komputer FMIPA Universitas Mulawarman. *Jurnal Informatika Mulawarman*, *8*(3), 110–113.

Purnaramadhan, R. (2021). Recommendation System Model Untuk Merekomendasikan Produk Pada Website Menggunakan Metode Content-Based. *Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Islam Indonesia*.

Raghuwanshi, S. K., & Pateriya, R. K. (2019). Recommendations System : Techniques, Challenges, Application, and Evaluation. *Advances in Intelligent Systems and Computing*, *2*(January), 107–119.

Ricci, F., Rokch, L., & Shapira, B. (2015). *Recommender Systems Handbook* (F. Ricci, L. Rokach, & B. Shapira (eds.); Second Edi). Springer Science+Business Media.

Rismanto, R., Syulistyo, A. R., & Agusta, B. P. C. (2020). Research supervisor recommendation system based on topic conformity. *International Journal of Modern Education and Computer Science*, *12*(1), 26–34

Sandi Wachyuni, S., & Wiweka, K. (2020). Kepuasan Wisatawan Dalam Penggunaan E-Commerce Agoda Dalam Pemesanan Hotel. *Journal of Tourism Destination and Attraction*, *8*(1), 275.

Supriyati, E., & Iqbal, M. (2018). Pengukuran Similarity Tema Pada Juz 30 Al Qur'an Menggunakan Teks Klasifikasi. *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, *9*(1), 361–370.