A Hybrid Recommender Model based on Information Retrieval for Mexican Tourism Text in Rest-Mex 2022

Victor Giovanni Morales Murillo[1], David Eduardo Pinto Avendaño[1] and Franco Rojas López[2]

[1]*Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, Puebla, México.*
[2]*Universidad Politécnica Metropolitana de Puebla, Ingeniería en Sistemas Computacionales, Puebla, México*

Abstract

Nowadays, the tourism is a principal economic sector for the world due to the exportations are improved, the jobs number is enhanced and the economic is developed. In México, the tourism represents 8.7% of GDP and generates 4.5 million direct jobs, however this economic sector has been affected by COVID-19 pandemic. For these reasons, a hybrid recommender model based on information retrieval is presented in this research to tackle the recommendation systems task of Rest-Mex 2022. A vector space model with tf-idf weighting scheme and cosine similarity is implemented. Besides, a hybrid recommender model is generated applying the recommendation techniques item-item collaborative filtering, content-based filtering and switching hybrid approach. Finally, our proposal won the second and third place in the competition.

Keywords

Hybrid recommender system, information retrieval, natural language processing, Mexican tourism.

1. Introduction

The World Tourism Organization (UNWTO) defines tourism concept as "a social, cultural and economic phenomenon which entails the movement of people to countries or places outside their usual environment for personal or business/professional purposes. These people are called visitors (which may be either tourists or excursionists; residents or non-residents) and tourism has to do with their activities, some of which involve tourism expenditure." [1] Nowadays, the business volumes of oil exports, food products and automobiles have been surpassed by the tourism. For this reason, it is a main economic sector for the world because the exportations are improved, the jobs number is enhanced and the economic is developed by this activity [2]. In México, the tourism represents 8.7% of the national gross domestic product (GDP) and it generates 4.5 million direct jobs. However, this economic sector has been affected by COVID-19 pandemic, which spread out in Mexico in March 2020. Therefore, the quality and safety of touristic products and services must be improved to restore Mexican tourism [3].

Thus, the event called Recommendation System, Sentiment Analysis and Covid Semaphore Prediction for Mexican Tourist Texts (Rest-Mex) [4] releases three task on natural language processing (NLP), these tasks are (1) the recommendation system task, (2) the sentiment analysis task and (3) the epidemiological semaphore task. Besides, Mexican tourism can be fortified by new NLP mechanisms generated in the Rest-Mex tasks. The recommendation system task is tackled by this paper, this task is defined as follows: "Given a TripAdvisor tourist and a Mexican tourist place, the goal is to automatically obtain the degree of satisfaction (between 1 and 5) that the tourist will have when visiting that place." The motivation task is that few recommendation systems for tourist sites are based on a user's profile's affinity compared to each place's description. The data collections to train these systems are from users and places in English-speaking countries. Considering the importance of Ibero-American

IberLEF 2022, September 2022, A Coruña, Spain.
EMAIL: vg055@hotmail.com (V.G. Morales Murillo); david.pintoavendano@viep.com.mx (D.E. Pinto Avendaño) ORCID: 0000-0002-6786-9232 (V.G. Morales Murillo); 0000-0002-8516-5925 (D.E. Pinto Avendaño); 0000-0002-2907-1334 (F. Rojas López)


countries in tourism, it is vitally important to generate Spanish resources that allow the generation of systems that help develop intelligent systems in tourism.

Recommender systems (RS) are a software tool that suggests several products called items as commercial products, web sites, friends, jobs, films, songs, touristic places, hotels, restaurants and other items, these suggestions are based on users' preferences [5]. Nowadays, RS represent a high economic, social and technological impact at international level due the main companies as Google, Facebook, Twitter, LinkedIn, Netflix, Amazon, Microsoft, Yahoo!, eBay, Pandora, Spotify and others more have been used these systems in their leading services [6]. Further, these companies are economically beneficiated by RS because information overload problem in e-commerce is tackled, users' experience is improved and users' decision making is helped by these systems [7]. On the other hand, RS performance can be optimized with hybrid approaches that combined two or more recommender algorithms to complement their disadvantages with advantages of other algorithms. For this reason, big money amounts are inverted to optimize algorithms and to develop new research on hybrid approach [8].

Information retrieval (IR) looks for text documents in large documents collections based on an information need [9]. The IR techniques have been utilized widely by a recommendation technique called content-based filtering (CBF) that recommends items similarities [10], where a user profile is generated by attributes and textual descriptions of user's items rated, then items similarities to user profile are recommended, also this technique uses items text metadata as main information source [11]. Although, RI techniques have been used traditionally by CBF, these techniques can be used by other recommendation technique called collaborative filtering (CF), which uses an user-item rating matrix as main information source to look for users similarities (user-user CF) or items similarities (item-item CF) called near neighbors, moreover the users' ratings are usually numeric values between 1 to 5 and the users' ratings predictions are realized with the nearest neighbors identified [12].

In this work a hybrid recommender model based on information retrieval is proposed for the recommendation task of Rest-Mex 2022, where item-item CF and CBF recommendation techniques are applied with information retrieval techniques. The paper structure is as follows. The related works are introduced in section 2, the model proposed is presented in section 3, the discussion of experimental results is showed in section 4 and the conclusions are introduced in section 5.

2. Previous works on recommendation systems task of Rest-Mex 2021

The recommendation task in Rest-Mex 2021 edition [3] is to predict the satisfaction degree of a tourist will have when visiting a place of Nayarit, México. The dataset has 2,263 instances with 2,011 tourist and 18 touristic places from Nayarit and its information is gotten of TripAdvisor. The dataset is divided in 70/30, this means that, training dataset contents 1,587 instances and test dataset contents 681 instances. Besides, mean average error (MAE) metric is utilized for the competition evaluation. Besides, two teams participate in this competition edition, these teams are Alumni-MCE 2GEN and Labsemco-UAEM. The Table 1 describes the results of Rest-Mex 2021, the best results with 0.31 and 0.32 of MAE are presented by Alumni-MCE 2GEN team and the Labsemco-UAEM team presents 1.65 of MAE.

Table 1
Results of Rest-Mex 2021 [3].

Team	MAE
Alumni-MCE 2GENRun1	0.31
Alumni-MCE 2GENRun2	0.32
Baseline	0.73
Labsemco-UAEM	1.65

The Alumni-MCE 2GEN team presents a research called *An Embeddings Based Recommendation System for Mexican Tourism. Submission to the REST-MEX Shared Task at IberLEF 2021* [13] that proposes two variants of the same model, where the most relevant change is a vector word

representation. The information dataset is preprocessed to clean and get completely Spanish information, then a Doc2Vec model is applied to users' information and places information, further a matrix is designed with the embeddings obtained and a Neural Network with one hidden layer is used in the first model variant. A system based on distributed representations for texts is proposed in the second model variant.

The Labsemco-UAEM team presents a research called *A Recommendation System for Tourism Based on Semantic Representations and Statistical Relational Learning* [14] that proposes a text representation method different from the lexical co-occurrence methods in text. This method extracts the linguistic features in the text, specifically the lexical and semantic signals of synonymy-antonymy. Furthermore, the ComplEx model is used to generate recommendations based on the relationships between users and places.

Theses related works tackle principally accuracy problem of recommender systems because it is the main challenge in Rest-Mex competition, however the scalability problem is other main challenge for RS in digital age, due to high information volumes must be process with minimum execution times by recommendation engines [15]. For this reason, RI techniques and data structures of inverted index are utilized to reduce the recommendations' compute cost in this research.

3. Methodology

The dataset, the information retrieval techniques and recommender systems techniques used to develop a hybrid recommender model based on information retrieval for Mexican tourism text in Rest-Mex 2022 are presented in this section. The methodology design is introduced in the Figure 1, which contents five main components that are the dataset, preprocessing, information retrieval techniques, recommender systems techniques and evaluation. These components are described each one below.

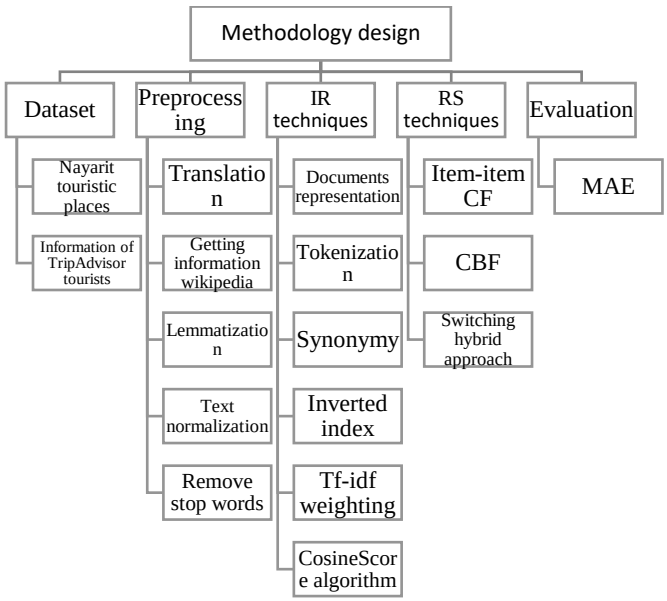

Figure 1: Methodology design.

3.1. Dataset

The dataset has 2,263 instances with 2,011 tourist and 18 touristic places from Nayarit, México. This dataset was obtained from the tourists who shared their satisfaction degree on TripAdvisor between 2010 and 2020. Each class of satisfaction is an integer between 1 and 5, where 1 is very bad, 2 is bad, 3 is neutral, 4 is good and 5 is very good. The recommendation system task goal of Rest-Mex 2022 is to predict the class satisfaction that a tourist may have when recommending a destination. The classes

distribution on training dataset is as follows, the class 1 has 45 instances, the class 2 has 53 instances, the class 3 has 167 instances, the class 4 has 457 instances, and the class 5 has 860 instances. The training dataset has in total 1,582 instances that represent around 70 % of the total instances. The classes distribution on test dataset is as follows, the class 1 has 20 instances, the class 2 has 24 instances, the class 3 has 72 instances, the class 4 has 196 instances, the class 5 has 369 instances. The test dataset has in total 681 instances that represent around 30 % of the total instances. The class imbalance represents a big challenge for this task of Rest-Mex 2022.

Each instance consists of two information parts. (1) The user information contents the tourist's gender, the tourist place that the tourist recommends a visit, the place of origin of the tourist, the recommendation date, type of trip (family, friends, alone, couple and business) and history of the places the tourist has visited with his/her opinions or reviews on each of these places. (2) The place information contents a brief text description of the place and a set of characteristics as the tourism type (adventure, beach, relaxation, among others), the atmosphere type (family, private or public), it is free or paid, among others [3].

3.2. Preprocessing

The langdetect library of Python is utilized to identify users' reviews in English language, then the googletrans library is used to translate users' reviews in English language to Spanish language in the first preprocessing step. The Wikipedia library is used to get information on the places related to the users' reviews in the second preprocessing step because the dataset only contains information on the 18 places from Nayarit, Mexico. The Spacy library is used to lemmatize the dataset in the third preprocessing step. The text normalization is applied to switch capital letters to lowercase letters, remove accents and remove punctuation marks as [,.;:#$!¡?%\n\f=*@\|] in the fourth preprocessing step. A stop words list of Spacy library is employed to remove stop words on the dataset in the fifth preprocessing step. The preprocessing result allows to tokenize the dataset documents in a more efficient way to select the terms of each document in a standardized way to develop an inverted index as a data structure for the IR techniques.

3.3. Information retrieval techniques

A vector space model is generated due to this model represents a set of documents as vectors in a common vector space and each documents' term represents a dimension of vector space [9]. Besides, this model is fundamental for several information retrieval operations as scoring documents on a query, document classification and document clustering. As well, the query representation as vector is very important for find the documents vectors most similarities to query's vector using cosine similarity. The formula of cosine similarity is the next.

$$sim(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{(\vec{V}(d_1) \,||\vec{V}(d_2))}, \tag{1}$$

Where the similarity between two documents d_1 and d_2 is computed with the cosine similarity and their vector representation $\vec{V}(d_1)$ and $\vec{V}(d_2)$, further the numerator represents the dot product of the vectors, while the denominator is the product of their Euclidean lengths. The term frequency - inverse document frequency (tf-idf) weighting scheme is utilized to assign to term t a weight in document d given by the next formula.

$$tf - idf_{t,d} = tf_{t,d} \times idf_t, \tag{2}$$

Where the weight is (1) highest when t occurs many times within a small number of documents, (2) it is lower when the term occurs fewer times in a document or occurs in many documents and (3) it is

lowest when the term occurs in virtually all documents. The term frequency tf and inverse document frequency $idf_{t,d}$ are normalized with the logarithm function by the next formulas.

$$tf_{t,d} = 1 + \log(tf_{t,d}) , \tag{3}$$

$$idf_t = \log\left(\frac{N}{df_t}\right), \tag{4}$$

A vector space model with tf-idf weighting scheme and cosine similarity is implemented with Python using the CosineScore algorithm of Manning [9]. Where the touristic places are represented as queries and users' reviews are represented as documents. Then documents terms of users' reviews are gotten for the tokenization process and these terms are indexed in an inverted index data structure, besides a synonyms dictionary is used to index the synonyms of terms in the inverted index.

3.4. Hybrid recommender model based on information retrieval

A hybrid recommender model is generated applying the recommendation techniques item-item collaborative filtering, content-based filtering and switching hybrid approach. This model is described in the Figure 2, where the switching hybrid approach is used in the following way. The item-item CF is employed to users that content enough information in theirs reviews on touristic places, on the other hand a CBF is employed to users that don't have reviews to tackle the cold star challenge. An inverted index is developed by each user in the item-item CF, while an inverted index is developed with all documents of users' reviews in CBF. Both techniques use a document of tourist place recommended to a tourist as query, then the *k* documents of tourist's reviews most similarities to query are gotten to average their ratings and generate the degree of satisfaction (between 1 and 5) that the tourist will have when visiting that place. Finally, the recommendations predicted are evaluated by MAE metric.

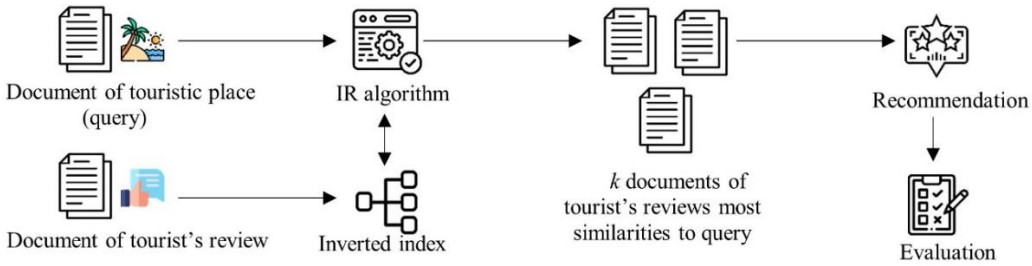

Figure 2: Hybrid recommender model based on information retrieval. This picture is designed with resources of Flaticon.com [16].

4. Experiments and results

The first experiment is realized with the training dataset that has 1,582 instances that represent around 70 % of the total instances. The hybrid recommender model based on information retrieval is used by each instance to generate its prediction. The results of hybrid recommender model are evaluated by MAE metric, this metric was selected by Rest-Mex 2022. Besides, this metric is usually employed when there are many outlier districts and it measures the distance between the vector of predictions and the vector of target values [17]. The MAE formula is the next [18].

$$MAE = \frac{\sum_{u \in U} \sum_{i \in testset_u} |rec(u,i) - r_{u,i}|}{\sum_{u \in U} |testset_u|}, \tag{5}$$

Where MAE computes the average deviation between computed recommendation scores $rec(u,i)$ and actual rating value $r_{u,i}$ for all evaluated users $u \in U$ and all items in their testing sets $testset_u$.

Four hundred tests are performed to identify the best values to the parameters *k1* and *k2*, that correspond to the *k* documents of tourist's reviews most similarities to query for the IR algorithm. The *k1* parameter is utilized to CBF and *k2* parameter is utilized to item-item CF. The best three results for this experiment are presented in the Table 2.

Table 2
The best three results for the first experiment on the training dataset.

Test number	K1	K2	MAE
1	14	13	0.7111251580278128
2	15	18	0.7117572692793932
3	13	18	0.7117572692793932

The second experiment is realized with the test dataset that has 681 instances that represent around 30 % of the total instances. The hybrid recommender model based on information retrieval is used with the parameters identified on the best three results for the first experiment on the training dataset. Then, three output submissions are submitted to the Rest-Mex 2022, the results are presented in the Table 3. Where, MAE is the main metric for the evaluation of the competency and other metrics as accuracy, f-measure, macro recall and macro precision were harnessed too. The first result with *k1=13* and *k2=18* got a MAE of 0.6981707317 that represents the second place in the Rest-Mex 2022, the difference with the first place was 0.004814684 of MAE. The second result with *k1=14* and *k2=18* got a MAE of 0.699695122 that represents the third place in the competency, besides the macro recall obtained of 0.2548259978 represents the best result on the macro recall metric in the Rest-Mex 2022 [4].

Table 3
The results for the second experiment on the test dataset.

Team	MAE	Accuracy	F-measure	Macro Recall	Macro Precision
LKEBUAP_RUN_k1_13_k2_18	0.6981707317	46.64634146	0.2215145883	0.242524662	0.2303778647
LKEBUAP_RUN_k1_14_k2_18	0.699695122	45.88414634	0.2196485875	0.2548259978	0.2280366323

5. Conclusions

This research presented a hybrid recommender model based on information retrieval for Mexican tourism text in Rest-Mex 2022. Where, a vector space model with tf-idf weighting scheme and cosine similarity is implemented with Python. Besides, a hybrid recommender model is generated applying the recommendation techniques item-item collaborative filtering, content-based filtering and switching hybrid approach. Then, two experiments were realized. The first experiment utilized the training dataset to identify the best parameters to the information retrieval algorithm. The second experiment used the test dataset with the three best parameters identified in the first experiment to submit three output submissions to Rest-Mex 2022. Furthermore, the second place and third place were gotten in the recommendation system task of the competency. Too, the best result of macro recall metric was gotten. Finally, this work will benefit the recommender system community and Mexican tourism.

6. References

[1] UNTWO: Glossary of tourism terms | UNTWO, https://www.unwto.org/glossary-tourism-terms, last accessed 2022/05/19.

[2] UNWTO: Why Tourism?, https://www.unwto.org/why-tourism, last accessed 2022/05/19.

[3] M. Á. Álvarez-Carmona, R. Aranda, S. Arce-Cardenas, D. Fajardo-Delgado, R. Guerrero-Rodríguez, A. P. López-Monroy, J. Martínez-Miranda, H. Pérez-Espinosa, A. Y. Rodríguez-González: Overview of Rest-Mex at IberLEF 2021 : Recommendation System for Text Mexican Tourism. Sociedad Española para el Procesamiento del Lenguaje Natural (2021). https://doi.org/10.26342/2021-67-14.
[4] M. Á. Álvarez-Carmona, R. Aranda, S. Arce-Cardenas, D. Fajardo-Delgado, R. Guerrero-Rodríguez, A. P. López-Monroy, J. Martínez-Miranda, H. Pérez-Espinosa, A. Y. Rodríguez-González: Overview of Rest-Mex at IberLEF 2022: Recommendation System for Text Mexican Tourism. Sociedad Española para el Procesamiento del Lenguaje Natural (2022).
[5] E. Çano, M. Morisio: Hybrid Recommender Systems: A Systematic Literature Review. Intell. Data Anal. 21, 1487 (2017). https://doi.org/10.3233/IDA-163209.
[6] D. Jannach, P. Resnick, A. Tuzhilin, A., Zanker, M.: Recommender Systems — Beyond Matrix Completion. Commun. ACM. 59, 94–102 (2016). http://dx.doi.org/10.1145/2891406.
[7] P. M. Alamdari, N. J. Navimipour, M. Hosseinzadeh, A. A. Safaei, A. Darwesh: A Systematic Study on the Recommender Systems in the E-Commerce. IEEE Access. 8, 115694–115716 (2020). https://doi.org/10.1109/ACCESS.2020.3002803
[8] V.G. Morales Murillo, D.E. Pinto Avendaño,F. Rojas López, F., J.M. Gonzales: A Systematic Literature Review on the Hybrid Approaches for Recommender Systems. Comput. y Sist. 26, 357–372 (2022). https://cys.cic.ipn.mx/ojs/index.php/CyS/article/view/4180
[9] C. D. Manning, P. Raghavan, H. Schütze: An Introduction to Information Retrieval. Cambridge University Press, Cambridge (2009).
[10] O. Kaššák, M. Kompan, M. Bieliková: Personalized Hybrid Recommendation for Group of Users : Top-N Multimedia Recommender. Inf. Process. Manag. 52, 459–477 (2016). https://doi.org/10.1016/j.ipm.2015.10.001
[11] N. Idrissi, A. Zellou,O. Hourrane,Z. Bakkoury,E.H. Benlahmar: A New Hybrid-Enhanced Recommender System for Mitigating Cold Start Issues. In: ICIME 2019: Proceedings of the 2019 11th International Conference on Information Management and Engineering. pp. 10–14. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3373744.3373746.
[12] F. Tahmasebi, M. Meghdadi, S. Ahmadian, K. Valiallahi: A hybrid recommendation system based on profile expansion technique to alleviate cold start problem. Multimed. Tools Appl. 80, 2339–2354 (2021). https://doi.org/10.1007/s11042-020-09768-8
[13] J. Arreola, L. Garcia, J. Ramos, A. Rodríguez: An Embeddings Based Recommendation System for Mexican Tourism. Submission to the REST-MEX Shared Task at IberLEF 2021. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). pp. 110–117. , Málaga, España (2021). http://ceur-ws.org/Vol-2943/restmex_paper1.pdf.
[14] E. Morales, D. Torres, A. Ehrlich, M. Toledo, B. Martínez, J. Hermosillo: A Recommendation System for Tourism Based on Semantic Representations and Statistical Relational Learning. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). pp. 134–148. , Málaga, España (2021). http://ceur-ws.org/Vol-2943/restmex_paper4.pdf.
[15] M. Nilashi, O. Ibrahim, K. Bagherifard: A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques. Expert Syst. Appl. 92, 507–520 (2018). https://doi.org/10.1016/j.eswa.2017.09.058.
[16] Flaticon: Iconos vectoriales y stickers - PNG, SVG, EPS, PSD y CSS, https://www.flaticon.es/.
[17] A. Géron: Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems. O'Reilly, CA 95472 (2019).
[18] D. Jannach, M. Zanker, A. Felfernig, G. Friedrich: Recommender Systems An Introduction. Cambridge University Press, New York, NY, USA (2011).