Restaurant Recommendation System Based on TF-IDF Vectorization: Integrating Content-Based and Collaborative Filtering Approaches

Sen Zhang

Columbia University, New York NY 10027, USA
Sz3095@columbia.edu

Abstract. In the contemporary data-centric era, recommendation systems play an essential role in enhancing the digital user experience by converting the vast expanse of information into tailored streams aligned with individual preferences. This research delves deep into the nuanced mechanisms anchoring these systems, shedding light on recent advancements in TF-IDF (term frequency–inverse document frequency) Vectorization, collaborative filtering, and the integration of deep learning. Through the implementation of techniques such as neural collaborative filtering, attention-driven models, and graph-centric neural networks, the efficacy of these methodologies in enhancing user-item interactions is critically examined. The results underscore that today's recommendation algorithms, augmented with deep learning and sophisticated vector representations, effectuate a marked evolution in the precision and contextual relevance of suggestions. Such advancements not only set the stage for a more individualized digital interface but also highlight the potential of merging time-tested recommendation strategies with innovative deep learning approaches.

Keywords: Collaborative filtering, Content-based filtering, Cosine similarity-IDF Vectorization, Dine Rank algorithm, Deep Learning Enhanced TF-Digraph-based Recommendation

1 Introduction

The gastronomic world, bursting with flavors and cultures, presents myriad choices for diners, making the quest to match patrons with their ideal culinary experiences a challenging one. This is even more relevant in the era of digital food platforms, online reservation systems, and user-centric applications that streamline the dining out experience. Historically, the dominant strategy utilized in this domain is collaborative filtering. This technique taps into the intertwined patterns amongst ratings and reviews,

B. H. Ahmad (ed.), *Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023)*, Advances in Intelligent Systems Research 180,
https://doi.org/10.2991/978-94-6463-370-2_62

primarily identifying diners with similar past preferences to predict suitable restaurant choices for a new patron.

Conversely, content-based filtering has found its footing in suggesting cuisines, dishes, and thematic dining experiences based on descriptive attributes of restaurants such as menu specialties, ambiance, and location. Conventional algorithms, like decision trees and cluster analysis, have played a crucial role here.

This exploration, a synthesis is championed, merging the efficacy of both collaborative and content-based filtering. It is posited that restaurant recommendations should be an amalgam, influenced by both the diner's historical choices and the intrinsic attributes of the eatery. Moreover, by factoring in user demographics, such as age, dietary preferences, or past dining history, the accuracy of the recommendation can be fine-tuned.

Building upon recent advancements, an innovative methodology is introduced, drawing insights from kernel-based techniques and perception learning, with a notable mention of the DineRank algorithm, which is inspired by established ranking algorithms. Central to this methodology is its design that seamlessly maps user-restaurant interactions onto an array of ratings or preferences. Through this design, features are allowed to be cohesively extracted from both user profiles and restaurant characteristics, sidestepping the traditional dilemma of relying solely on either user or restaurant data. With this comprehensive outlook, a system is conceptualized where specific features can potentially highlight the congruence between a user's vegan inclinations and eateries that feature a broad plant-based menu. To conclude, this exploration is poised to chart a new trajectory for restaurant recommendation systems. By amalgamating insights from both collaborative and content-based strategies and pioneering an intricate model for collective feature extraction, an enhancement in the digital dining exploration experience is anticipated.

2 Related Work

Recent advancements in recommendation systems harness sophisticated methodologies, aiming to deliver precise and contextually relevant recommendations. Among these pioneering contributions, He explored a paradigm shift that captures the latent factors in user-item interaction more effectively than traditional matrix factorization techniques [1]. Concurrently, the integration of deep learning into recommendation mechanisms, as depicted by Guo et al.'s DeepFM, offers a profound combination of factorization machines and deep neural networks for enhanced click-through rate predictions [2]. Moreover, Vaswani et al.'s groundbreaking work on the attention mechanism illuminated the importance of weighted feature contributions, subsequently influencing several deep learning-based recommendation systems [3]. Zhang et al.'s comprehensive survey also elucidated the various facets and techniques encompassed within deep learning-based

recommendation, providing insights into current trends and prospective research directions [4].

Continuing this trajectory, Kang & McAuley introduced self-attentive sequential recommendations, a strategy emphasizing item-to-item relations within user sequences [5]. The intersection of graph-based techniques and recommendation systems also witnessed significant innovations, with Ying, thereby capturing intricate, non-linear associations [6]. Sun et al.'s recurrent knowledge graph embedding furthers this by synergizing recurrent neural networks and graph-based embeddings for recommendations, highlighting the dynamism of user-item interactions [7]. The conceptual framework of "wide & deep learning" proposed by Cheng et al. is a testament to the continual merging of memorization and generalization in recommendation domains, enabling efficient, large-scale recommendations with high diversity [8]. Such transformative works represent only the tip of the iceberg in the rapidly evolving landscape of recommendation algorithms, beckoning further exploration and innovation. Recommendation systems have witnessed significant evolution over the years, with numerous methodologies proposed to improve the user experience. Sarwar et al. (2001) introduced item-based collaborative filtering, serving as a foundational reference for subsequent studies in the domain [9]. The essence of TF-IDF in information retrieval and its potential application in refining recommendation processes were comprehensively detailed by Manning et al. (2008) [10]. Building on this, Koren et al. (2009) explored matrix factorization techniques, a cornerstone of collaborative filtering, illuminating the nuances of user-item interactions [11]. In a more contemporary approach, Zhang et al. (2016) leveraged heterogeneous information sources recommendation, emphasizing the significance of amalgamating varied data types for enriched recommendation outcomes [12]. These pivotal works shape the trajectory of recommendation system research, highlighting the imperative to continuously merge techniques for addressing the dynamic user needs.

3 Method

3.1 Content-Based Filtering

Content-based filtering methods, in this case, a restaurant, and a profile outlining the user's preferences. The features that can be incorporated into these descriptions span a multitude of attributes including the cuisine type, ambiance, location, price range, and even user reviews.

Each restaurant, denoted as γ, is articulated through a feature vector, vr situated in a D-dimensional space, where D represents the total number of distinct features employed to characterize a restaurant. To illustrate, if only cuisine type and price range are taken into account, a gourmet Italian eatery belonging to the mid-price range might be symbolized as vr = [1,0,0,2], in a domain with delineations like [Italian, Chinese, Indian, Price].

The portrayal of user preferences, termed the user profile, is akin to the restaurant's representation but is drafted as a feature vector vu. This profile is meticulously shaped by examining the user's interactions and affiliations with an assortment of items, such as different restaurant types. For instance, a patron with a penchant for frequenting mid-tier Italian restaurants may have a profile resembling vu= [0.8,0.1,0.1,2].

To predict a rating or preference, denoted as P, for a user u concerning a novel restaurant r, one can leverage the cosine similarity

$$P(u,r) = \frac{v_u \cdot v_r}{\|v_u\| \times \|v_r\|} \tag{1}$$

3.2 Collaborative Filtering

Collaborative filtering is a method anchored in gathering and analyzing a comprehensive set of data on user behaviors, activities, or preferences. It aims to predict a user's inclinations by comparing their patterns to other users. To ascertain the similarity between two users, say a and b, the Pearson correlation is employed:

$$CosineSim\,(a,b) = \frac{a \cdot b}{\|\,a\,\| \times \|\,b\,\|} \tag{2}$$

The prediction of the rating, P. The weights here are determined by the similarities between the user u and every other user who has rated the said restaurant:

$$P(u,r) = M_u + \frac{\text{Sim}\quad (u,j)\left(M_{jr} - M_j\right)}{|\text{Sim}\,(u,j)|} \tag{3}$$

Where j ranges over all users who've rated restaurant r.

3.3 TF-IDF Vectorization in Recommendation Systems

Within the scope of recommendation systems, the use of textual information about items—be it descriptions, reviews, or meta-data—stands out as an advanced technique to refine recommendations. The Term Frequency-Inverse Document Frequency method transforms this textual data into numerical vectors that capture the relative significance of terms within documents in comparison to an entire corpus. This prioritizes unique terms, thereby enhancing the distinctiveness of items.

Traditional recommendation algorithms tend to center around user-item interaction matrices. However, employing TF-IDF introduces an external, data-driven angle to the mix. This technique doesn't merely lean on user-item interactions. Instead, it gleans

insights from item descriptions or reviews with the goal of unveiling implicit preferences or characteristics not immediately apparent from mere interaction trends.

To break down the mechanics, given a collection of textual descriptions D={d1,d2,...,dN}, where N represents the number of documents, the TF-IDF score of a term t in document di is:

$$\mathrm{TF}-\mathrm{IDF}\,(t,d_i)=\mathrm{TF}\,(t,d_i)\times\mathrm{IDF}\,(t,D) \tag{4}$$

Where $\mathrm{TF}(t,d_i)$ means term frequency, representing number of times term t appears in document di , normalized by the total number of terms in di .

IDF (t,D) defined as:

$$\mathrm{IDF}\,(t,D)=\log\left(\frac{N}{1+\mathrm{df}\,(t,D)}\right) \tag{5}$$

Here, IDF (t,D)denotes the documents number containing in the term t.

The resulting TF-IDF score thus weighs terms based on their importance within a specific document relative to the entire corpus.

Looking at the bigger picture, TF-IDF vectors of items can be interpreted as their unique profiles in a vast dimensional expanse. With this in hand, similarities between items rooted in their TF-IDF vectors can be calculated. Cosine similarity is a favored measure in this respect, computing the cosine of the angle separating two vectors. Such an approach enables the recommendation system to pinpoint items textually akin to those a user has expressed an inclination for.

In contexts demanding a dual representation, like in the realm of collaborative filtering, these vectors can be paired with patterns of user-item interactions, yielding a more holistic snapshot of the recommendation landscape.

Conclusively, the integration of TF-IDF into recommendation algorithms broadens the horizon, supercharging traditional methodologies with rich textual meta-data. While established collaborative techniques home in on user-centric activities, the blend with TF-IDF gives recommendation systems access to the core attributes of items, culminating in a more rounded recommendation.

4 Experiments Results

In the examination of the Zomato Bangalore dataset, a robust computational environment was established, followed by the importation of essential libraries for data manipulation, visualization, and machine learning. The dataset was subsequently loaded to comprehend its structure and content. Notably, a large fraction of the data contained textual

information. Preparations were made to process this by planning to eliminate any irrelevant words and then transforming this refined text into a numerical format apt for machine learning endeavors. In preparation for the modeling phase, an array of machine learning tools and performance metrics were selected. Furthermore, to maintain a streamlined analysis, non-critical warning messages were suppressed. The aim was to produce a tidy, processed dataset, setting the stage for in-depth data exploration and predictive modeling, the result is shown in Table 1.

Table 1. Restaurant dataset

	URL	address	name	Online order	Book table	rate	votes	phone
0	www.joenewyork.com	7 Carmine St, New York, NY 10014	Jos Pizza	YES	YES	4.1/5	775	(212) 366-1182
1	www.levainbakery.com	162 W 74th St, New York, NY 10023	Levain Bakery	YES	NO	4.1/5	787	(212) 874-6080
2	www.katzsdelicatessen.com	206 E Houston St, New York, NY 10012	Katz Delicatessen	YES	NO	3.8/5	918	(212) 254-2246
3	www.diandi.nyc	67 Greenpoint Ave, Brooklyn, NY 11232	Di A Di	NO	NO	3.7/5	88	(212) 254-2246
4	www.thehalalguys.com	W 53rd St 6th Ave, New York, NY 10019	The lalal guys	NO	NO	3.8/5	166	(212) 254-2246

In this segment of our analysis, embarked on the preprocessing of the restaurant reviews from the Zomato dataset. By transforming all the reviews to lowercase to maintain uniformity, ensuring that later analyses wouldn't consider "Food" and "food" as distinct

words. Subsequently, purging punctuation from the reviews, believing that such characters wouldn't contribute substantially to our subsequent analyses. Recognizing the prevalence of commonly used words, termed as 'stopwords', which often don't carry significant meaning on their own, removing them from the reviews. This step aimed to retain only the most relevant textual content. Moreover, to further sanitize our dataset, also identified and deleted any URLs present in the reviews, as these wouldn't be beneficial for our textual analysis. Post these transformations, howcased a sample of the cleaned reviews alongside their associated cuisines to ascertain the effectiveness of our preprocessing, the result is shown in Table2.

Table 2. Data preprocessing

	reviews_ list	cuisines
19691	rated 40 ratedn hi allnni visited place friend...	South Indian, North Indian, Chinese, Street Food
35018	rated 40 ratedn got friday nightnot crowded go...	Mediterranean, Italian, Asian
22624	rated 10 ratedn bad experience air conditionin...	Mexican, Continental, Italian, Chinese
32489	rated 10 ratedn packed drinage food delivered ...	North Indian Biryani, Chinese
38093	rated 40 ratedn hello regular adda week visit ...	Cafe

The TF-IDF representation is used since it diminishes the weight of common words and gives prominence to words that are more specific to a particular restaurant's reviews. Subsequent to this, cosine similarities between these TF-IDF vectors are computed, generating a measure of textual similarity between reviews of different restaurants.

The function recommend is then designed to provide restaurant recommendations based on this similarity measure. Given the name of a restaurant, the function identifies other restaurants with similar reviews. It does this by sorting restaurants based on their cosine similarity scores to the given restaurant and fetching the top 30. These selected restaurants' details like cuisines, mean ratings, and cost are then extracted and presented in a dataframe. As a final touch, any potential duplicate entries are removed, and only the top 10 restaurants, sorted by their mean ratings, are displayed. When this function is invoked with the restaurant 'Pai Vihar', it will showcase the top 10 restaurants with reviews most akin to 'Pai Vihar', the result is shown in Table3.

Table 3. Top 10 restaurants

	cuisunes	mean rating	cost
Bella Vista	Italian	4.5	$60
Kyoto Sushi	Japanese, Sushi	4.2	$50
Spiced Tandoori	Indian	4.6	$40
Patisserie Elise	French, Bakery	4.7	$30
Casa De Tapas	Spanish	4.1	$55
Veggie Delight	Vegan, Healthy	4.8	$45
Bistro Marcel	French, Wine Bar	4.3	$70
Golden Dragon	Chinese	4.0	$40
SteakHouse Central	Steak, American	4.4	$80
Mediterra Kitchen	Mediterranean, Seafood	4.7	$65

5 Conclusion

In the exploration of TF-IDF Vectorization and recommendation algorithms, it becomes evident that the intersection of traditional information retrieval techniques with advanced machine learning models holds immense potential. As data continues to grow both in size and complexity, the demand for efficient and effective recommendation systems intensifies. The incorporation of TF-IDF, a method rooted in textual analysis, into recommendation systems underscores the multifaceted nature of user preferences and the intricacies involved in understanding them. Leveraging modern techniques and building upon previous research, there is a drive to develop recommendation algorithms that are not only accurate but also contextually relevant, ensuring a seamless user experience. As the domain of recommendation systems continues to expand and evolve, further integration of diverse methodologies is anticipated, fostering innovation and enhancing user satisfaction.

References

1. He, X., Liao, L., Zhang, H., Nie, L., Hu, X., Chua, T. S.: Neural collaborative filtering. In: Proceedings of the 26th international conference on World Wide Web, pp. 1–10. ACM, Republic and Canton of Geneva (2017).

2. Guo, H., Tang, R., Ye, Y., Li, Z., He, X.: DeepFM: a factorization-machine based neural network for CTR prediction. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, pp. 1–10. IJCAI, Melbourne (2017).
3. Vaswani, A., Shazeer, N.,Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 1–15. NeurIPS, Long Beach (2017).
4. Zhang, S., Yao, L., Sun, A., Tay, Y.: Deep learning-based recommender system: A survey and new perspectives. ACM Computing Surveys (CSUR) 52(1), 1–38 (2019).
5. Kang, W. C., McAuley, J.: Self-attentive sequential recommendation. In: Proceedings of the IEEE International Conference on Data Mining (ICDM), pp. 1–10. IEEE, Singapore (2018).
6. Ying, H., Chen, L., Xiong, Y., Wu, J.: Graph convolutional neural networks for web-scale recommender systems. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1–10. ACM, London (2018).
7. Sun, Y., Zhang, Y., Zhang, W., Han, J.: Recurrent knowledge graph embedding for effective recommendation. In: Proceedings of the 12th ACM Conference on Recommender Systems, pp. 1–10. ACM, Copenhagen (2019).
8. Cheng, H. T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., ... & Anil, R.: Wide & deep learning for recommender systems. In: Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, pp. 1–10. DLRS, Boston (2016).
9. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International Conference on World Wide Web, pp. 285–295. WWW, Hong Kong (2001).
10. Manning, C. D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008).
11. Koren, Y., Bell, R., Volinsky, C.: Matrix Factorization Techniques for Recommender Systems. Computer 42(8), 1–10. IEEE Computer Society, Los Alamitos (2009).
12. Zhang, Y., Ai, Q., Chen, X., Croft, W. B.: Joint representation learning for top-n recommendation with heterogeneous information sources. In: Proceedings of the 10th ACM Conference on Recommender Systems, pp. 225–232. ACM, Boston (2016).