Machine learning based recommendation system on movie reviews using KNN classifiers

Machine learning based recommendation system on movie reviews using KNN classifiers

J Ananda babu[1*], D R Vinay[1], B V Kumaraswamy[2], and Chethan Chandra S Basavaraddi[3]
[1]Department of Information Science and Engineering, Malnad College of Engineering, Hassan, Karnataka, India
[2]Koch Industries, Hassan, Karnataka, India
[3]Department of Computer Science and Engineering, Kalpataru Institute of Technology, Tiptur, Karnataka, India
Email: *abj@mcehassan.ac.in

Abstract: Recommender systems are the systems that are designed to recommend items to the consumer depending on several different criteria. These systems estimate the most possible product that the consumers are most likely to buy and are of interest to. Companies like Netflix, Amazon, etc. use recommender services to allow their customers to find the right items or movies for them.In the current system recommendations, the content of ltering and collective ltering typically fall into two groups. The method is formerly Periment in our paper in all methods. We take film features such as stars, directors, for content-based ltering. Movie definition and keywords as inputs use TF-IDF and doc2vec for measuring the film resemblance. For the first time, Input to our algorithm is the film ranking encountered by users, and we use neighbours nearest K, as Factorization of matrix to estimate film scores for consumers. We find that teamwork functions better than content. Predictive error and estimation time ltering.
Keywords: Recommendation System, K-neighbours, TF-IDF, Companies, customers and content-based filtering.

1. Introduction

We use the machine to build a custom film scoring system based on the previous movie of the consumer ratings. In movies, people have a different taste, and it's not a single item that we can see when we Google a movie. Our film scoring framework allows consumers to find movies that they like, no matter how diverse their preferences are Perhaps it is.Content-based ltering recommends the same functionality in products. The term recurrence backwards record recurrence weighting method (tf-idf) in the recovery of data [1&2] and word2vec in the regular language process [3] are normal in substance based ltering procedures. We go beyond the use of related films by tf-idf Predict movie ratings and add doc2vec, word2vec extension, to retrieve details in the film summary context. The Bayesian classifier [4][5], the decision tree, the neural networks, etc. [6] are other methods. Content-based filtering has the bonus of addressing the issue of cold start if there were not enough users or if it wasn't scored for the content. It is therefore restricted to characteristics directly connected to products and it involves a broad method of data collection. Automatic characteristics are particularly costly and costly to remove Multimedia data including videos, video and audio streams [7]. Often the standard of products is not differentiated. Another example is a well-known and a critically regarded

film, whether they share the same kind of film. Features like common film definitions of sentences.Collaborative ltering solves certain content-based ltering problems {suggests products close to consumers and avoids the need, using the underlying user preference constructs, to gather data on each object. Two are relevant Collaborative filtering approaches: the neighbourhood model and simulations of latent factor. Model of the neighbourhood mends the nearest items or the highest-ranking items of the closest person. The latent factor construct, including factoring matrix, Movie and consumer features latent space [8-10]. By adding non-linear kernels, we are changing the methodology Update and review the efficiency of schemes. The factorization of the matrix can be modified to include time dependency to record time-change shifts for users [11]. In [12] articles discussed food packet distribution system data prediction using data mining techniques. In [13] discussed about privacy of the healthcare system using cloud and blockchian trending techniques for content Deduplication. The Block Chain Based technique discussed for applying the security on Food Beverages [14]. In [15] executed a guess mechanized construction as Filtered Wall (FW) and it separated discarded substance from OSN customer substances.

2. Methods and Materials

We use TF-IDF to measure the weighted similarity between words bag, using the simple content of the ltering. Evolution the meaning of the word't,' but the frequency of term't' is positively associated with the number His capacity to discriminate between documents is inversely related to all documents. So the frequency of the term is determined t in document d, weighted in all records by reverse of frequency t:This measures the occurrence of the word t in the document d, as opposed to t in all documents: tf-idf (t) = tf (t; d) frequency (d)= tf(t;d) raises the document jDj 1+jd:t2dj, and jd:t 2 dj is the number of documents where t appears. In this manner, we calculate the frequency of the word t in the documents d.

3. Dataset Exploration

We use the Movie Lens Dataset on Kaggle 1, which includes more than 45 thousand films and 26 million users. The details were split into two sets: the first consists of an analysis of films and features including Cash, money, cast, etc. We have 45,433 films after eliminating duplicates in the results. The top ten are Table 1 the famous films are measured using the weighted score of the IMDB 2. This dataset is divided randomly into 80% Training set and 20% content-based ltering evaluation set are shown in Table 1

Table 1: Representation of first 10 data in dataset directory

Movie Title	Avg Votes	Num Votes	Weighted Score
The Shawshank Redemption	8.5	8358.0	8.445871
The Godfather	8.5	6024.0	8.425442
Dilwale Dulhania Le Jayenge	9.1	661.0	8.421477
The Dark Knight	8.3	12269.0	8.265479
Fight Club	8.3	9678.0	8.256387
Pulp Fiction	8.3	8670.0	8.251408
Schindler's List	8.3	4436.0	8.206643
Whiplash	8.3	4376.0	8.205408
Spirited Away	8.3	3968.0	8.196059
Life Is Beautiful	8.3	3643.0	8.187177

The other used dataset is the user-film scores, which include a user ID, a film ID and the 1-5 user ranking. This data is represented by users in one direction and films in the other; the matrix is very thin since most of the films contain just a limited fraction of all the users. This dataset is evaluated using methods like the nearest neighbour and matrix factorization.

Instead of a classic fiction issue of thumb up/thumb down, we treat the job as one of continuous classification, which is more excessive and codes more detail. We will map to expected ranks and recommend the top-rated products for users at recommendation time. We use a random subset of 100,000 rating data for our project due to computational constraints. We randomly divided the ranking pairs into a training set of 80% and a test set of 20%. The sum of the square error between the approximate ratings is minimised.

4. Results and Discussion

We use movie similarity matrices developed by TF-IDF and word2vec for content-based ltering to predict film ratings of the consumer (Formula (1)). We determine their weighted number to combine the two TF-IDF matrices. We can see that the RMSE is smaller than the weight of the w1 = 0:7; w2=0:3 as seen at gure 1 in the assessment of the results in a training set consisting of 80% random chosen consumer movie rating pairs. With these weights on the test range, we run the prediction algorithm. Figure 3 portrays our algorithm's success in predicting movie consumer scores. Each bin is a pair of user movies with a speci c ranking. The blue bars are the part of our forecast The blue bars
The true rating is inside − bis0:75, and the green bars reflect a section outside of − bis0:75 the true rating for the expected component of our algorithm. We see that for user film pairs with ratings greater than 3 our algorithm is successful.For doc2vec we use Paragraph Vector (PV-DM) distributed memory version and add some noise terms To boost robustness overall. Different starting rates have been checked (Table 3) and è = 0:025 Reduces to 0.927 RMSE. The analysis rate begins at 0.025 and decreases linearly to 0.001. RMSE values and histogram predictions are shown in Figure 01 and Figure 02.

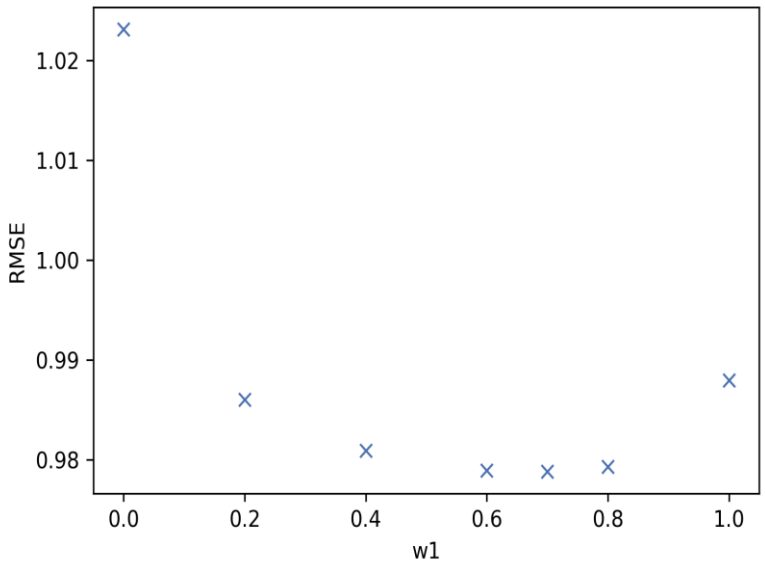

Fig 1: RMSE values difference

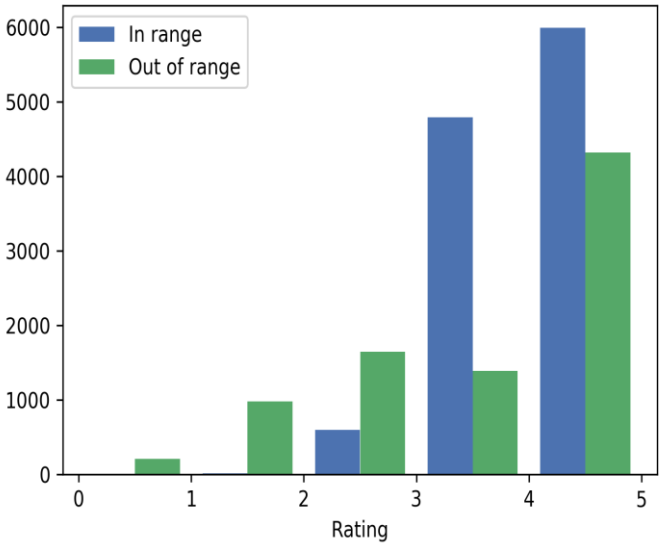

Fig 2: Histogram predictions using TF-IDF

With this model, we can also ask keywords or film titles for the similarity scores from the movie for related films Overview. The 5 most related films to Sky fall are seen in the Table 2.

Table 2: learning rate testing and similar movies score

Similar Movies	Sim Score
Octopussy	0.8358
Transporter	0.8298
Safe house	0.8287
Unlocked	0.8106
Undercover Man	0.8062
Push	0.8033
Sniper 2	0.8007
Patriot Games	0.8005
2047 Sights Death	0.7952
Interceptor	0.7951

Learning Rate	RMSE
0.020	0.927403
0.025	0.927003
0.030	0.927009
0.035	0.927358
0.040	0.927994
0.05	0.929483
0.1	0.938572
0.2	0.950744
0.4	0.970215

We perform item by item and user to user CF using KNN for collaborative ltering. We found that the total RMSE test decreases as k increases. We found the best K value. The RMSE test stabilises around k = 20 and does not decrease monotonously. For item-to-item CF k=43 results the lowest RMSE of 0.9079 and k=28 yields the highest user-to-user RMSE of 0.9203, while other k>20 values produce the same set of results in 3 decimal points, making results reasonably stable. The item-to-item CF-workout RMSE is 0.2900 with the evaluation RMSE 0.9079 with k = 43. With k = 28, the CF training RMSE for user-to-user is 0.2830 and the RMSE measure 0.9203. Any degrees of over-taking suggest the distinctions between preparations and test RMSE. The difference between the Test and the RMSE train persists, however. The ranking histogram from the KNN model reveals in Figure 3

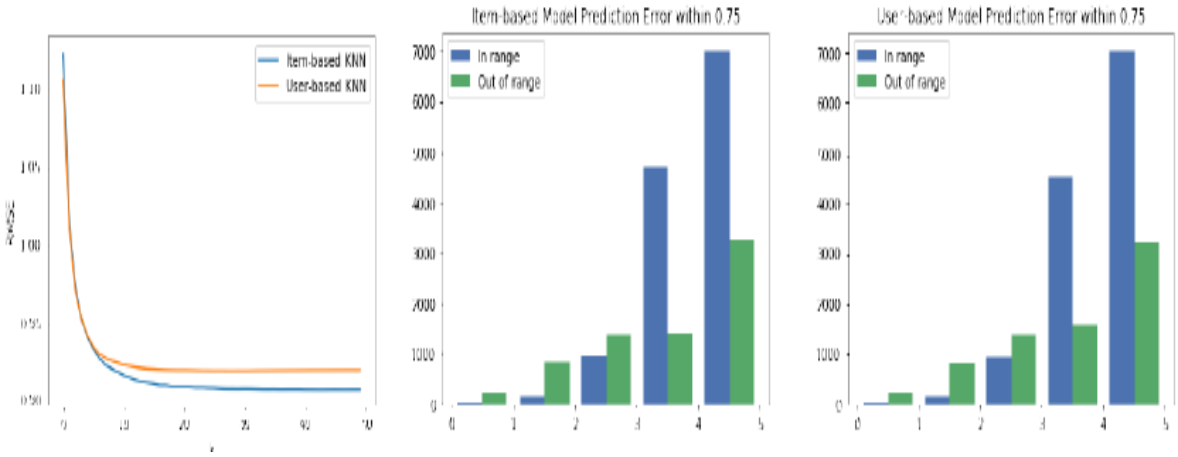

Fig 3: KNN models final predictions

5. Conclusion

In order to improve the recommendation system, we analysed both the material and interactive ltering. Overall, collective ltering succeeds in terms of RMSE tests higher than content-based ltering. Furthermore, ltering based on content is computer-based and costly rather than collaborative, since text characteristics are intensive processing. We want to resolve the biased forecast induced by the desequilibrium of the number of low ratings relative to potential job to moderate ratings. To high ratings,we also look for ways to fix KNN's over ratting issues, such as regularisation. In addition, the integration of content-based ltering and collaboratory ltering strengthens our recommendation framework. Possible approaches include integration of collective ltering of material functionality, as well as decisions and the neural network vice versa.

Reference

[1]. Reddy, M. M., Kanmani, R. S., & Surendiran, B. (2020, February). Analysis of Movie Recommendation Systems; with and without considering the low rated movies. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)* (pp. 1-4). IEEE.

[2]. Chien, C. Y., Qiu, G. H., & Lu, W. H. (2019, November). A Movie Trailer Recommendation System Based on Pre-trained Vector of Relationship and Scenario Content Discovered from Plot Summaries and Social Media. In *2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)* (pp. 1-6). IEEE.

[3]. Chen, J., Peng, J., Wang, Y., & Chen, G. (2018, October). An Implicit Information Based Movie Recommendation Strategy. In *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)* (pp. 405-410). IEEE.

[4]. Cheng, Y., Liu, N., Lu, Y., & Tang, X. (2020, May). Recurrent Knowledge Attention Network For Movie Recommendation. In *2020 3rd International Conference on Electron Device and Mechanical Engineering (ICEDME)* (pp. 648-651). IEEE.

[5]. Wang, W., Ye, C., Yang, P., & Miao, Z. (2020, June). Research on Movie Recommendation Model Based on LSTM and CNN. In *2020 5th International Conference on Computational Intelligence and Applications (ICCIA)* (pp. 28-32). IEEE.

[6]. Han, W., & Wang, Q. (2019, November). Movie recommendation algorithm based on knowledge graph. In *2019 2nd International Conference on Safety Produce Informatization (IICSPI)* (pp. 409-412). IEEE.

[7]. Vijay S.V., Narendra N. (2017). Tag Based Image Search By Social Re-Ranking In The Web Based Applications. *International Journal of MC Square Scientific Research*, 9(1), 252-259
[8]. Papadakis, H., Fragopoulou, P., Michalakis, N., & Panagiotakis, C. (2018, September). A Mobile Application for Personalized Movie Recommendations with Dynamic Updates. In *2018 International Conference on Intelligent Systems (IS)* (pp. 507-514). IEEE.
[9]. Faisal, M., Hameed, A., & Khattak, A. S. (2019, December). Recommending Movies on User's Current Preferences via Deep Neural Network. In *2019 15th International Conference on Emerging Technologies (ICET)* (pp. 1-6). IEEE.
[10]. Yuyan, Z., Xiayao, S., & Yong, L. (2019, October). A Novel Movie Recommendation System Based on Deep Reinforcement Learning with Prioritized Experience Replay. In *2019 IEEE 19th International Conference on Communication Technology (ICCT)* (pp. 1496-1500). IEEE.
[11]. Dimitrov, V. D., & Kolev, K. N. (2019, October). Development of a Mobile Movie Recommendation Application. In *2019 27th National Conference with International Participation (TELECOM)* (pp. 74-77). IEEE.
[12]. Prakash, G., & Sivasankar, P. T. (2012, February). Food Distribution and Management System Using Biometric Technique (Fdms). In International Conference on Advances in Communication, Network, and Computing (pp. 444-447). Springer, Berlin, Heidelberg.
[13]. Pandey, A., & Prakash, G. (2019). Deduplication with Attribute Based Encryption in E-Health Care Systems. International Journal of MC Square Scientific Research, 11(4), 16-24.
[14]. Prakash, G. and Nagesh Y., (2019).Secure and Efficient Block Chain Based Protocol For Food Beverages. International Journal of MC Square Scientific Research,10(3):19-30
[15]. Prakash, G., Saurav, N., & Kethu, V. R. (2016). An Effective Undesired Content Filtration and Predictions Framework in Online Social Network. International Journal of Advances in Signal and Image Sciences, 2(2), 1-8.