A new approach to identify similar users based on customer reviews

Dimple Chehal¹, Parul Gupta² and Payal Gulati³

¹Research Scholar, Department of Computer Engineering, J.C. Bose University of Science and Technology, YMCA, Faridabad
²,³Assistant Professor, Department of Computer Engineering, J.C. Bose University of Science and Technology, YMCA, Faridabad

E-mail: dimplechehal@gmail.com

Abstract. Recommender System suggests items of interest to users based on their preferences. These preferences are gauged through various sources such as purchase history, ratings, reviews and browsing behaviour. Collaborative filtering and content based filtering are the two widely used techniques that help in generating recommendations to the target user(s) by identifying similar users to target user or similar items to items of interest. Through this paper a new method to identify similar users based on the similarity of reviews has been proposed.

Keywords- User similarity, collaborative filtering, user reviews, recommender system, subjectivity analysis

1. Introduction
Recommender Systems (RS) build user preferences and suggest products/services to the consumer based on his/her preferences [1]. They are widely used by online giants such as Amazon, Netflix, YouTube and eBay so as to increase their sales and provide a better user experience [2]. They help in facilitating the decision process and converting potential customers to actual customers. In order to construct user preferences various factors are taken into account, for instance, user ratings, feedbacks or reviews, purchase and browsing patterns. A fine analysis of these inputs helps the system to predict user preference for unseen items and output a list of preferred items and thus allure the potential customers. Hence, this area has a commercial aspect associated to it thereby making it presence in various fields of computer such as data analytics [3] and machine learning [4]. Collaborative filtering (CF), content based filtering (CBF) and Hybrid filtering as shown in figure 1, are the three main techniques used in RS [2, 5, 6]. In CF technique, users similar to the target user are identified and the preferences of these similar users are then suggested to the target user. But, in CBF technique, a user’s past purchase history is used and items similar to the ones purchased by him in the past are suggested to him. Also, description of item is also taken into account in this technique. Hybrid as the name suggests is a combination of the above two techniques in was introduced to overcome the limitations of these techniques. The proposed method finds similarity between users based on the reviews provided by them for a product. The idea is to populate a list of similar users i.e. if two users provide similar reviews for similar products then these users are similar. For instance, it doesn’t always hold true that an apparel mentioned by a user as ideal for a petite lady would be suitable to others with the same body type. Due to this difference in preference, it becomes important to identify customers who have the similar preferences.
The rest of the paper is organized as follows: In section 2, background work is discussed. In section 3, proposed method has been described. In last section i.e. section 4, paper is concluded along with future scope.

2. Related Work
This section examines the related work in the field of RS, including collaborative filtering, content based filtering and hybrid filtering. Liangqiang Li et al. [8] studied the purchase-review behaviour of online customers and defined a new method introducing aspects to explore the customer’s opinions. Wen et al. [2] focused on the sparsity problem in collaborative filtering recommender system. They incorporated cosine similarity with Matrix factorization technique to overcome this sparsity issue [18, 19]. To calculate user similarity based on provided ratings, cluster based approaches were used by [9, 10]. To identify similar users based on user-item rating matrix, cosine similarity, adjusted cosine similarity, Pearson correlation coefficient (PCC), constrained Pearson correlation coefficient, sigmoid function based PCC, Jaccard coefficient, Mean Square difference, PIP (Proximity-Impact-Popularity) were used as stated in [11-17]. Yun et al [20] proposed a recommendation system based on subjectivity analysis on after purchase reviews. The authors in [11] designed a new user similarity measure to overcome the problem of co-existence of rated items for two different users by introducing the measure of item similarity in it. Matsunami et al [5] calculated similarity of users purchasing cosmetic products, based on online reviews provided by them. The authors chose their own similarity measure as compared to Pearson coefficient as the user review was drilled down to aspect level and sentiment score was calculated for each aspect resulting in large size of the rating matrix as compared to the traditional m (product) x n (users) size.

3. Proposed Method
In this paper, a new method has been proposed, as shown in figure 2, to predict a user’s preference for an item based on the reviews given by him. For each item, the proposed method predicts a sentiment score for each user.

![Recommendation System Types](image)
To implement the proposed method a subset of dataset of women’s e-commerce clothing reviews has been taken from Kaggle [7]. The subset for conducting proof of concept is shown in Table 1 below.

Table 1 Example dataset.

| User ID | Item ID | Review Text                                                                 |
|---------|---------|-----------------------------------------------------------------------------|
| U1      | 1       | I purchased these in taupe, mint, and coral. They are extremely comfortable and soft. They can be rolled up to 3 different lengths. I stayed in my regular size and the fit is great. I can see why they sell out so quickly. A must for summer season. |
| U2      | 1       | These are some of the softest most comfortable shorts i own and wish i had them in more colors. i like that i an adjust the length of the cuff since it's not "hemmed" to a certain length. i ordered the pink/salmon color and they go with so much! |
| U3      | 2       | I saw this romper online and knew i needed it as i love flannels and i love rompers, it's super comfy. i bought with the intention of wearing it out, not just around the house. i think in the fall it will be cute with high socks and boots but as i just got it, i’ve been opting for tights. lots of compliments so far. i got a small because i didn't want it to be too short, it fits well. |
| U4      | 2       | I wanted to love this romper, but it just wasted right for me. i am 5'5", 135lbs, 34c, curvy/muscular frame and ordered size small. i may have liked it better with more room in the medium, decided to return. i still recommend trying this product, but it wasn't for me and my hips! |
| U5      | 3       | Love everything about it but had to get a size bigger to be long an off/ to short for a mom. if i was 20 would be fine. |
I went out on a limb ordering this romper. It's not really in my "wheelhouse" the whole romper thing, but it was a home run! I was afraid it might look too young and that I couldn't "pull it off," but the long sleeves and overall style was perfect. And my twenty something nieces were obsessed with it as well. As far as fit, I'm 5'7" 134lbs and ordered a size 4. I have a long torso and was concerned because some people said it ran short there, but I didn't have any trouble with that. I just took

U1  3
So cute and so adorable but too short for my body in size small. I'm going to try a size medium.

U2  4
Fabric is a nice weight cotton. Lining is good, sleeves are a not too tight.

U3  4
This romper is cute, well-made and true to size, but I haven't figured out how to put this on without having someone tie the back ties. Which pretty much means that going to the bathroom is not an option while you have this romper on. Not sure what the designers were thinking here. I'm returning this one.

U4  5
Mine came smelling like gasoline. Not sure why, but I would have kept it otherwise. It's a smell that will be really hard to get out. Looks like the picture.

U5  5
I love this sweatshirt! I truly did not pay much attention to it online but while in my local store one was returned and it caught my eye immediately as the flowers are embroidered in a nice substantial rope type yarn to give it a more demential effect. The torn holes here and there give the appearance of being the most loved garment in your closet...and it has become mine along with the jacket with the same embroidery...funny how 2 pieces I did not give a second thought about have become my li

U1  6
I tried this on in the ivory color because it was on sale and I thought it "might" be sort of cute. A comfy, flowy, warmer to cooler weather transition top. Little did I know how much I would fall in love with it! I tried it on over like 3 things in the fitting room, including a black strapless maxi dress I wore into the store, and it still looked great! It is comfortable, loose, and goes with pretty much anything (I've tried and still haven't found anything it looks bad over). It's not super dr

U2  6
Ordered navy in a medium and it is wide. Sometimes that's a good thing, but not this time. I'm short waisted and hoped this would be more fitted at the waist on me but it isn't. It just looked frumpy with lots of see through parts. Obviously it needs a cami but wasn't attractive on me at all. Seemed well made and probably better on someone taller.

The subsequent calculations for 5 users and 6 items are shown in the following figures and tables. The implemented system took 'Review Text' attribute for estimating the user sentiment and is implemented through Python language. The sentiment is estimated on a scale of 0-5. The architecture of the proposed method is shown in figure 2.

3.1. Calculate sentiment score of a user for an item
The first step in the proposed method is to calculate the sentiment score for each review given by a user for an item. Sentiment score reveals the user's real experience with an item as it is based on the actual textual review provided by him. For this purpose, Python library vaderSentiment has been used to find the compound score for each feedback. The compound score calculated by vaderSentiment ranged from 0-1 which has been mapped to a scale of 0-5. The calculated sentiment scores with the help of this library are shown in figure 3.
3.2. Finding User similarity

Based on the sentiment scores calculated above, user similarity is calculated. If the review provided by two users is similar i.e. if the sentiment scores calculated for two users are close to each other, then the two users are said to be similar, otherwise the two users are not counted as similar. The relative formula to find how much users are related to each other is as follows:

\[
\text{SimilarityScore}(U_i, U_j) = \sum_{k=1}^{n} |\text{Sent}i_{U_{i,k}} - \text{Sent}i_{U_{j,k}}| \quad (1)
\]

Where, \(\text{Similarity}(U_i, U_j)\) is the similarity value between two users \(U_i\) and \(U_j\). \(k\) is the number of items rated by both \(U_i\) and \(U_j\). \(\text{Sent}i_{\text{max}}\) refers to the maximum sentiment score for any review sentence i.e. 5 and \(\text{Sent}i_{U_{i,k}}\) is the sentiment score for item \(k\) by user \(U_i\), \(\text{Sent}i_{U_{j,k}}\) is the sentiment score for item \(k\) by user \(U_j\). The calculation for user similarity using Python is shown in figure 4.

Figure 3 Sentiment score of each review.

Figure 4 User Similarity score calculation through Python.

For two users if there is no commonly reviewed item, then the user similarity score is zero for all such users. The user similarity score for our example dataset is shown in Table 2.
Table 2 User similarity scores for all users.

| Similarity Score provided | U1  | U2  | U3  | U4  | U5  |
|---------------------------|-----|-----|-----|-----|-----|
| U1                        | 0   | 9.144 | 0   | 0   | 3.3305 |
| U2                        | 9.144 | 0   | 3.9875 | 0   | 0   |
| U3                        | 0   | 3.9875 | 0   | 4.645 | 0   |
| U4                        | 0   | 0   | 4.645 | 0   | 2.1115 |
| U5                        | 3.3305 | 0   | 0   | 2.1115 | 0   |

Table 3 lists down the similar users for a user. For instance, for user U1, similarity score with user U2 and U5 is calculated as they commonly reviewed item 1, 3 and 6. Hence, User 1 is similar to user User 2 and User 5. Similarity list for all users is shown in Table 3.

Table 3 List of similar users.

| Users | Similar Users |
|-------|---------------|
| U1    | U2, U5        |
| U2    | U1, U3        |
| U3    | U2, U4        |
| U4    | U3, U5        |
| U5    | U1, U4        |

3.3. Predicting sentiment score of users for items

In this part, the sentiment score of each user for each item is predicted by using the sentiment score and similarity score found in above steps. The formula to predict user sentiment score is as follows:

\[ E(y, k) = \sum_{U \in R} \text{Senti}_{U \cdot k} + \frac{\sum_{U \in R} \text{SimilarityScore}(y, U)}{\text{Senti}_{\text{max}}} \]

(2)

where, \( E(y, k) \) is the predicted sentiment score of user y for item k, R is the list of related users found in step 3.2, \( \text{Senti}_{U \cdot k} \) is the sentiment score of user U (in this case related user of user y), \( \text{SimilarityScore}(y, U) \) is the user y’s similarity score to user U and n is the total number of items. Table 4 shows the predicted sentiment score using equation (2) for all the users.

Table 4 Final Predicted Sentiment Score for all users.

| Sentiment score predicted | I1    | I2    | I3    | I4    | I5    | I6    |
|---------------------------|-------|-------|-------|-------|-------|-------|
| U1                        | 4.516 | 0.416 | 3.3455 | 4.91 | 5.01 | 4.561 |
| U2                        | 4.8405 | 5.06 | 5.03 | 3.9195 | 0.438 | 5.1 |
| U3                        | 4.718 | 4.8945 | 0.618 | 5.1 | 2.4025 | 4.763 |
| U4                        | 0.234 | 4.865 | 3.1635 | 3.7155 | 4.907 | 0.234 |
| U5                        | 4.5839 | 4.4579 | 4.7804 | 0.1814 | 1.9659 | 4.8799 |

4. Conclusion
In this paper, a new approach to identify similar users by finding their sentiment for an item hidden in textual reviews is proposed. The proposed system first calculates the user sentiment score for each item and then finds the user similarity with other users who have reviewed the same set of items. At the end, using both the above scores, the sentiment score for each item by each user is then predicted. This approach can be used to utilise the hidden sentiment stored in the form of text in user reviews as an input to collaborative filtering technique. As an improvement, in the future, this work can be tested on huge datasets to verify if the method is scalable.

References

[1] Chen J, Zeng W, Shao J and Fan G 2018 Preference modeling by exploiting latent components of ratings Knowledge and Information Systems 1–27
[2] Wen H, Ding G, Liu C and Wang J 2014 Matrix factorization meets cosine similarity: Addressing sparsity problem in collaborative filtering recommender system Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8709 LNCS 306–17
[3] Fazzolari M and Petrocchi M 2018 A study on online travel reviews through intelligent data analysis Information Technology and Tourism 20 37–58
[4] Sivapalan S, Sadeghian A, Rahnama H and Madni A 2011 Recommender Systems in E-Commerce Data Mining Applications with R 2015 81–90
[5] Matsunami Y, Ueda M and Nakajima S 2018 How to Find Similar Users in Order to Develop a Cosmetics Recommender System Transactions on Engineering Technologies 337–50
[6] Yang N, Ma Y, Chen L and Yu P S 2019 A meta-feature based unified framework for both cold-start and warm-start explainable recommendations World Wide Web.
[7] Kaggle Women’s E-Commerce Clothing Reviews Dataset https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews
[8] Li L, Yuan H, Qian Y and Shao P 2018 Towards exploring when and what people reviewed for their online shopping experiences Journal of Systems Science and Systems Engineering 27 367–93
[9] Liu L, Mehandjiev N and Xu D L 2011 Multi-criteria service recommendation based on user criteria preferences RecSys’11 - Proceedings of the 5th ACM Conference on Recommender Systems 77–84
[10] Alqadah F, Reddy C K, Hu J and Alqadah H F 2015 Biclustering neighborhood-based collaborative filtering method for top-n recommender systems Knowledge and Information Systems 44 475–91
[11] Wang Y, Deng J, Gao J and Zhang P 2017 A hybrid user similarity model for collaborative filtering Information Sciences 418–419 102–18
[12] Liu H, Hu Z, Mian A, Tian H and Zhu X 2014 A new user similarity model to improve the accuracy of collaborative filtering Knowledge-Based Systems 56 156–66
[13] Ahn H J 2008 A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem Information Sciences 178 37–51
[14] Resnick P, Neophytos I, Mitesh S, Peter B and John R 1994 GroupLens: An Open Architecture for Collaborative Filtering of Netnews Proceeding of the ACM Conference on Computer Supported Cooperative Work pp 175–86
[15] Adomavicius G and Tuzhilin A 2005 Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions IEEE Transactions on Knowledge and Data Engineering 17 734–49
[16] Koutrika G, Bercovitz B and Garcia-Molina H 2009 FlexRecs: Expressing and combining flexible recommendations SIGMOD-PODS’09 - Proceedings of the International
Conference on Management of Data and 28th Symposium on Principles of Database Systems
745–57

[17] Cacheda F, Carneiro V, Fernández D and Formoso V 2011 Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems ACM Transactions on the Web 5

[18] Ardimansyah M I, Huda A F and Baizal Z K A 2018 Preprocessing matrix factorization for solving data sparsity on memory-based collaborative filtering Proceeding - 2017 3rd International Conference on Science in Information Technology: Theory and Application of IT for Education, Industry and Society in Big Data Era, ICSITech 2017 2018-Janua 521–5

[19] Da Silva J F G, De Moura N N and Caloba L P 2018 Effects of Data Sparsity on Recommender Systems based on Collaborative Filtering Proceedings of the International Joint Conference on Neural Networks 2018-July

[20] Yun Y, Hooshyar D, Jo J and Lim H 2017 Developing a hybrid collaborative filtering recommendation system with opinion mining on purchase review Journal of Information Science