ISSN: 0067-2904

Dual-Stage Social Friend Recommendation System Based on User Interests

Sammer A. Qader*, Ayad R. Abbas
Department of Computer Science, University of Technology, Baghdad, Iraq

Received: 3/8/2019 Accepted: 22/10/2019

Abstract

The use of online social network (OSN) has become essential to humans' lives whether for entertainment, business or shopping. One system that is used extensively for this purpose is friend recommendation system (FRS) which recommends users to other users in professional or entertaining online social networks. In this paper, a Dual-Stage Friend Recommendation (FR) model is proposed. The model applies dual-stage methodology on unlabeled data of 1241 users collected from OSN users via online survey platform featuring user interests and activities based upon which users with similar social behavioral patterns are recommended to each other. The model employs techniques including user-based collaborative filtering (UBCF) approach in stage one and graph-based approach friend-of-friend recommendation (FOF) in stage two. The model offers a solution to common problems of FRS, such as data sparsity, using a dimensionality technique called non-negative matrix factorization (NMF) to create a dense representation of the collected data and reduce its sparsity, in addition to providing seamless integration with other FRSs. The evaluation of the FR model shows positive correlation of Pearson correlation coefficient (PCC) as compared the outcomes of using Cosine similarity and Euclidean distance as a baseline.

Keywords: Recommendation system, Friends Recommendation System, Collaborative Filtering, Content-based Filtering, User Interests

"نظام توصية صديق اجتماعي ثنائي المرحلة بالأعتماد على اهتمامات المستخدم"

سامر عبد الوهاب قادر محمد*، أياد روضان عباس
قسم علوم الحاسوب, الجامعة التكنولوجية, العراق, بغداد

الخلاصة

أصبح استخدام الشبكة الاجتماعية عبر الإنترنت (OSN) ضروريًا لحياة البشر سواء للترفيه أو العمل أو التسوق. نظام واحد يستخدم على نطاق واسع لهذا الغرض هو نظام توصية الأصدقاء (FRS) الذي يوصي المستخدمين للمستخدمين الآخرين في الشبكات الاجتماعية عبر الإنترنت المهنية أو مسلية. في هذه الورقة ، تم اقتراح نموذج توصية صديق ثنائي المرحلة. يطبق النموذج منهجية ثنائية المراحل على بيانات لـ (1241) مستخدمًا تم جمعها من مستخدمي شبكات التواصل الاجتماعية عبر منصة استطلاع عبر الإنترنت تعرض اهتمامات المستخدمين وأنشطتهم بناءً على المستخدمين ذوي الأنماط السلوكية الاجتماعية المماثلة التي ينصح بها بعضهم البعض. يستخدم النموذج تقنيات بما في ذلك نهج التصفية التعاونية القائمة على المستخدم (UBCF) في المرحلة الأولى وتوصية صديق لصديق القائمة على الرسم البياني (FOF) في المرحلة الثانية. يقدم النموذج حلاً للمشاكل الشائعة لـ FRS مثل تباين البيانات باستخدام تقنية الأبعاد تسمى عامل المصفوفة

*Email: 111817@student.uotechnology.edu.iq

غير السلبي (NMF) لإنشاء تمثيل كثيف للبيانات التي تم جمعها وتقليل تباينها وكذلك توفير تكامل سلس مع FRSs الأخرى. يُظهر تقييم نموذج FR الارتباط الإيجابي لمعامل ارتباط بيرسون (PCC) مقارنة بين نتائج استخدام تشابه جيب التمام والمسافة الإقليدية كخط أساس .

1. Introduction

The widespread use of social media applications is becoming increasingly important as more users enjoy sharing their experiences and daily activities via rating and reviewing products, posting opinions, expressing mood, and even making new friendships. Another very popular use of social media that has gained crucial necessity is telecommunication; millions of users use video and audio calls extensively to connect with their friends and loved ones. The demand for using social applications generates a huge amount of data that users cannot deal with, and therefore, the need for a data mining system that automatically suggests products or users will not only improve the experience of social media users, but also make it more effective and meaningful.

One popular data mining system that fits this need is the recommendation systems (RS) which employs techniques of machine learning to recommend products or people. There are several types of RS; one type is an item recommendation system that suggests products (e.g., movies, books, music, and so on) to customers by taking others users rating products in the past and features of customers interests to create a new recommendation list, such as that applied in Amazon. Another common recommendation system is the Friends recommendation system (FRS). The goal of this type of systems is to recommend friends with similar interests to each other, conditioned on the particular context. While there exists a plethora of studies in the literature on how to recommend friends to social media users, it remains an active and challenging problem. Most of the friend suggestion systems used in social media services such as MySpace, Facebook, LinkedIn, and Twitter, recommend friends based on network graphs using PYMK features (People You May Know). This method, used primarily by MySpace on November 15, 2008, suggested users to other users depending on what their common friends share on a social network [1].

There are three main techniques used in RS [2, 3]; Content-based filtering (CBF), collaborative filtering (CF), and hybrid technique. The CBF technique is based on the description of the item that is recommended to the person based on the one that he or she preferred in the past. The CF, also called "people-to-people correlation" [2] is the most popular RS technique [4] that uses a large amount of data, featuring people behavior, preferences, and activities, and predicts what people might like based on how similar their activities, preference, and behavior to their friends. The CF technique has two primary approaches: Memory-based (neighborhood-based) and model-based. The memory-based approach finds comparative clients (or items) for the dynamic client or item utilizing likeness estimation strategy, and after that, it aggregates the evaluation of these neighbors as the prediction [5]. Memory-based CF contains two strategies: item-based (IBCF) and user-based (UBCF). While the Model-based CF provides adequate strategies for making a model of the dataset, it allows the off-line processing for the foremost through likeness calculations [6].

The third approach is the hybrid technique, which is a combination of CF and CBF techniques that uses the advantages of one technique to fix the disadvantages of the other [7]. During building a new RS, there are various challenges and boundaries facing those systems [3, 8], such as those of cold start, sparsity, limited content analysis, and over-specialization [9]. Also, understanding user interests plays an important role in implementing personalized recommendations. Therefore, companies use third-party applications and services to obtain more information from users as it can help mitigate problems like the cold-start problem. One can define the user's interests as the activities that users enjoy doing and the topics that they like to spend extra time learning about [10], such as specific topics (movies, books, sports, video games, and music), or even joining social groups. All these factors can be collected to make friends recommendation decision among social users.

In this paper, a Dual-Stage FR model is proposed. Initially, the dataset is converted into a binary (0, 1) multi-dimensional sparse matrix, where each row represents a feature vector for a particular user. Then, matrix dimensionality is reduced using Non-negative Matrix Factorization (NMF) to resolve the matrix sparsity issue. After that, an initial graph of relations among similar users is constructed (stage one). Finally, friend-of-friend algorithm is applied to the initial graph to enhance the overall recommendation process (stage two) by assuming that, for instance, U1and U2 are friends with similar

interest, and U2 and U3 are friends with similar interest too, then U1 and U3 are possibly friends as a result.

2. Related Work

Many studies were organized and carried out in the friend recommendation area. Different friend suggestion systems were proposed based on the similarity, using users' profiles information, geographical location of users, and graph-based similarity. This section briefly presents some studies handled in recent years by various techniques.

In 2013, Akbar *et al*. proposed a FRS method that uses Artificial Bee Colony (ABC) algorithm on a dataset of 1000 users extracted from YouTube to recommend users to each other by analyzing the topological features of the graph network generated from the dataset. ABC algorithm is used to learn the graph weights by optimizing four parameters. The method is compared to classic machine learning algorithms such as K-nearest Neighbor, Support Vector Machine and Multilayer perceptron, and it yielded an accuracy of 77% [11] .

In 2014, Eirinaki *et al*. introduced a trust-aware framework for user suggestion which examined the dynamics and semantics of the friend-enemy relationship (implicit, explicit) connections amongst users based on reputation mechanism. The framework is divided into three phases. Firstly, after data preparation that is distributed on the network, the explicit and implicit connection is established that provides the signs of trust among the users. Secondly, reputation degrees are calculated. Finally, user recommendations for positive and negative are formed. Two datasets are used: Epinion dataset and Wikipedia vote network. Epinion dataset has a product review interactions collected from 132000 users and over 136 million user-to-user positive and negative statements. The introduced method achieved an accuracy of 0.9 and 0.7 for Epinion and Wikipedia datasets, respectively. The advantage of such a system is that it attracts similar users together while it repels different users from each other [12].

In 2015, Zhao *et al*. presented a study of the social network user relationships and behaviors. The authors introduced a FRS by using hybrid algorithms (Clustering Algorithm and Factorization Machine (FM)). The reasons for using the FM algorithm are to solve the data sparseness problem. Also, it classified users and made it simple to recognize their characteristics and interests. The model was trained by using Markov Chain Monte Carlo (MCMC) algorithm. The data used in that study is Tencent Weibo that contains 2320895 users information. The algorithm achieved a root mean square of error (RMSE) value of 0.5015 [13].

In 2015, Huang *et al*. presented a two-stage FRS by using multimedia information, Flickr tags feature, and friendship network. In stage one, it produces a friend list based on the relationship of different OSN. In stage two, a co-clustering method on the user, tag, and image information is applied to create groups. Then, it builds a more precise RS to improve the system outcomes from stage one. The authors collected data of 10000 users from Flickr with their photos and tags. The total number of photos collected was 543,754. The co-clustering method achieved a consistent precision value of 0.28 in comparison to other methods [14].

In 2015, Hasan *et al*. suggested a novel FRS which is based on utilizing individual user's behavior on social network sites. First, it measured the recurrence of the activities done by the clients and the upgrading of the data according to users' activities. Then, the persons' behavior classification strategy was applied by Frequent Pattern Growth algorithm (FP-Growth) to find out the required behavior (common and exceptional). Finally, the multilayer threshold for FRS was used. The data used is a collection of users and their relationships extracted from Facebook social network. The model achieved an accuracy of 94% [15].

In 2016, Wu *et al*. introduced a FRS based on location preference, in which the temporal, spatial, and social relationships are considered. Firstly, the Markov chain algorithm was used to calculate the user's friendship similitude on the social network. Then, user's area inclination similitude within the real world was calculated based on the history check-in information. The experimental results were based on using a dataset consisting of 604138 user relationships. The check-in data showed could the possibility of suggesting friends with both similar companionship and area inclination to clients within the large-scale [16].

In 2017, Ding *et al*. proposed a FRS based on matrix factorization approach by extracting latent architectural models from the input network utilizing convolution neural networks algorithm. After that, it uses the Bayesian ranking algorithm to make the user suggestion. The work uses two datasets:

Epinion dataset and Slashdot dataset, each of which has approximately 3000 user reviews. After evaluation, the results showed 0.97 AUC value for Epinion dataset and 0.96 for Slashdot dataset [17].
In 2018, Kumar *et al*. presented a graph-based FRS using two CF strategies: number of mutual users and influence factor. Then, it assigned a score number to each possible friend to find the higher similarity between users based on the highest score number. The datasets used are Stanford SNAP which consists of 4039 and 81,306 users from Facebook and Twitter, respectively. The accuracy of the model was 97.2%[18].

3. Dual-Stage FR Model Design

The FR Model is a combination of UBCF and graph-based FR. The framework consists of dual-stages. In the first stage, an initial graph is constructed based on user interest's similarity. In the second stage, a FOF graph-based method is applied to the initial graph, and a final friend recommendation list is built. The Dual-stage friend proposal can suggest friends to users who have similar interests based on user's behavior characteristics. The following steps show the general process of the proposed model, while Figure-1 illustrates these steps as a block diagram.

Step_1. Filling in the Online Survey: Users are asked to fill in a biographic (name, age, study filed, email, etc.) and interest information (movie and music genre types, sport type, favorite book, etc.) to create user profile and database.

Step_2. Preprocessing Data: the dataset is converted into a binary (0, 1) multi-dimensional sparse matrix, where "zero" indicates that the user dislikes or is not interested, while "one" indicates that the user likes or is interested in the particular feature or activity.

Step_3. Reducing Dimensionality: sparsity is reduced by converting the sparse vectors into denser vectors using dimensionality reduction NMF algorithm. Reducing dimensionality will reduce the sparsity and keep those important, relevant features.

Step_4. Stage one of UBCF approach begins by computing the similarity among all users to find the most similar users under the threshold (delta = 0.5), then an initial graph is built. The threshold is an adjustable value; it can span from 0 to 1. The value of delta indicates that users with similarity of 50% or more are considered similar. Nodes will serve as users and the connection between those nodes will serve as the similarity scores.

Step_5. Stage two of the graph-based approach begins when a user does not get more than five user's recommendation from step 3 (where five is defined as a threshold called gamma). A user will get recommendations from friends of a friend that has a high similarity score in friends list by using FOF algorithm.

Step_6. A new score is calculated for the user recommended in stage two by taking the maximum of the old similarity score of stage one (S1) and the old similarity score of their friend of friends (S2), i.e. a new score = max(S1,S2). Finally, the final friends' recommendation graph is created accordingly.

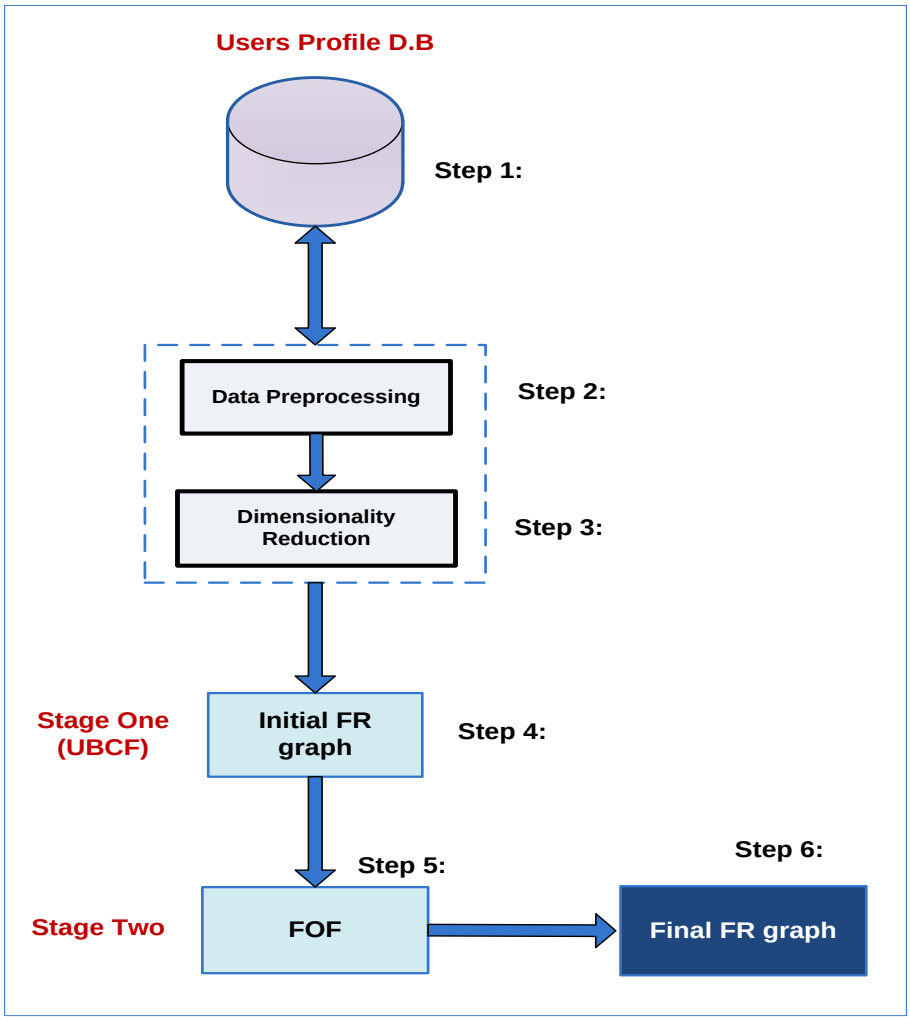

Figure 1- General process of the proposed model

4. Non-negative Matrix Factorization NMF

The Vector Space Model (VSM) is a common knowledge retrieval model, which describes a data collection matrix since vector space matrices are initially high dimensional and sparse. Treating a model with a large space dimensionality on a dataset usually requires a vast time and space complexity, and often leads to overfitting problems. Thus, reducing the dimensions can reduce the noise. Also, dimensionality reduction reduces the unnecessary parts of the data and finds those surprisingly very closely relevant in one's smaller subspace, then one can easily apply a simple learning algorithm [19].

The NMF decomposes a non-negative matrix (A) into two non-negative matrices (W and H); one of the decomposed matrices can be viewed as the basis vectors (W). The dimensionality reduction can be achieved by projecting the input vectors onto the lower-dimensional space which is formed by these basis vectors, as shown in Figure-2. Also, Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) are popular techniques for dimensionality reduction based on matrix decomposition [20]; however, the cost of their calculation will be limited when the matrices become high. The NMF method is distinguished from the other methods, e.g. PCA and SVD, by its non-negativity constraints. These constraints lead to a parts-based representation because they allow only the additive, not subtracted, combinations. Also, the NMF computation is based on the simple iterative algorithm; it is, therefore, advantageous for applications involving large matrices.

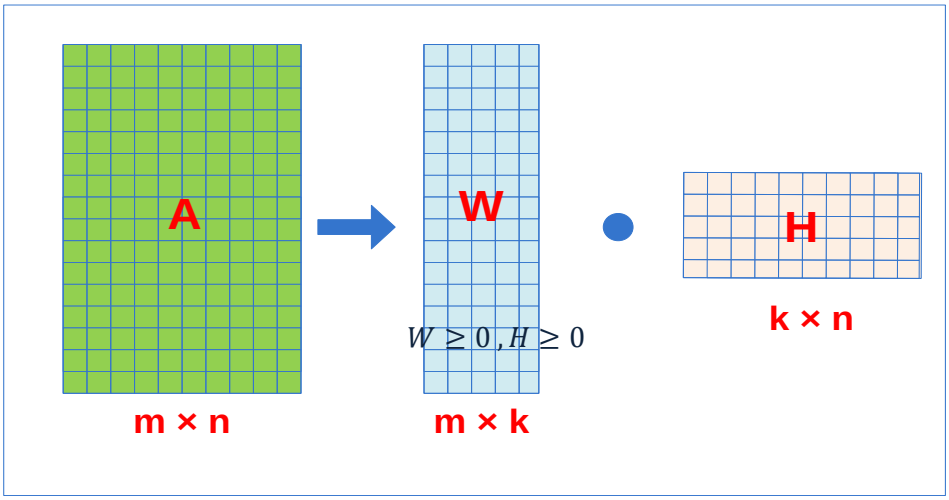

Figure 2- Definition of NMF

The mathematical calculation in NMF [20] can be determined as follow:

Let matrix *A* be the product of matrices *W* and *H*,

$$A = W.H \tag{1}$$

For factorization matrix (*A*), matrices (*W*) and (*H*) are randomly initialized with nonnegative values. Next, we define a squared error function (Cost_function) to make the product of both matrices equals to the matrix (*A*):

$$min_{W,H} = \|A - WH\|_F \tag{2}$$
$$Subject\ to\ W \geq 0, H \geq 0$$

Next, a multiplicative update rule is applied to both of them iteratively, until W and H are stable, as follows:

$$H_{[i,j]}^{n+1} \leftarrow H_{[i,j]}^{n} \frac{((W^n)^T V)}{((W^n)^T W^n H^n)_{[i,j]}} \tag{3}$$

And,

$$W_{[i,j]}^{n+1} \leftarrow W_{[i,j]}^{n} \frac{(V(H^{n+1})^T)_{[i,j]}}{(W^n H^{n+1}(H^{n+1}))^T{}_{[i,j]}} \tag{4}$$

5. Similarity Measurements

5.1 Cosine Similarity

One of the common similarity measures is cosine similarity, which is to consider the item\user interest as a feature vector of an n-dimensional space and calculate their similarity as the cosine of the angle between two user's vectors [2, 21]. If the Cosine similarity for two users' vectors (A, B) is smaller than the angle, then the similarity is high. Let us explain this in more details:-

For understanding the concept of cosine-based similarity, the definition of dot product is explained. For example, vectors *a* and *b*.

$\vec{\mathbf{a}}$= [a_1, a_2, a_3…….] **and,** $\vec{\mathbf{b}}$= [b_1, b_2, b_3…….]

where, (a_n, b_n) are the elements of the features vector values and (n) is a vector dimension.

$$\vec{\mathbf{a}}.\vec{\mathbf{b}} = \sum_{i=1}^{n} a_i b_i = a_1\, b_1 + a_2\, b_2 + \cdots + a_n\, b_n \tag{5}$$

The description of the dot product is the sum of element-wise multiplication between the two vectors. For instance, the dot product of vectors $\vec{a} = (0,5)$ and $\vec{b} = (3,0)$ is:

$$\vec{\mathbf{a}}.\vec{\mathbf{b}} = 0 * 3 + 5 * 0 = 0$$

It can be noted that the input is two vectors, and the output is a single value, but not another vector. The geometric description of the dot product is:

$$\vec{a}.\vec{b} = \|\vec{a}\|\|\vec{b}\| \cos\theta \tag{6}$$

where $\|\vec{a}\|$ and $\|\vec{b}\|$ are the norms of vector ($\vec{a}$) and vector ($\vec{b}$), respectively. In this context, each user is represented as a vector, where each vector has the user interest rates, and then the cosine similarity equation is utilized as follows:

$$\mathrm{Cos}(\vec{a},\vec{b}) = \frac{\vec{a}.\vec{b}}{\|\vec{a}\|\times\|\vec{b}\|} \tag{7}$$

and the equation of cosine distance will be as follows:

$$\textbf{Cosine Distance} = 1 - \cos(\vec{a},\vec{b}) \tag{8}$$

5.2 Pearson's Correlation Coefficient (PCC)

The second similarity metric is Correlation-Based, in which the similarity between two users' vectors, (a) and (b), are measured by Pearson's correlation coefficient (PCC). PCC can be helpful in data analysis and modeling to better guess the relationships between variables [22]. PCC ranges from -1 to +1. The statistical relationship between the two vectors is assigned to their correlation, by using a PCC. The association could be positive, indicating that both variables are in the same direction, or negative, indicating that both variables are in the opposite direction. Correlation can also be zero, indicating that the variables are uncorrelated [23]. The similarity using PCC for user-based between vectors $\vec{a}$ and $\vec{b}$ is described as follows:

$$p_{a,b} = \frac{cov(a,b)}{\sigma a\ \sigma b} \tag{9}$$

where the $\mathrm{cov}(a,b)$ is the covariance of vectors $\vec{a}$ and $\vec{b}$, while σa and σb represent the standard deviation of vectors $\vec{a}$ and $\vec{b}$, respectively. More formally, PCC formula can be written as follows.

5.3 Euclidean Distance

The third similarity metric is the Euclidean distance. The formula that is measuring the distance between vectors $\vec{a}$ and $\vec{b}$ is represented as follows:

$$dist(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{10}$$

where $dist(a,b)$ represent the similarity between the two users vectors $\vec{a}$ and $\vec{b}$, and it is the square root of the sum of squared differences between identical features of the two vectors.

6. Graph Measurement

In graph theory, a graph can be either dense or sparse depending on how close or far the number of edges from the maximal number of edges. A graph with a high number of edges is called a dense. On the contrary, a graph with few edges is called a sparse graph [24]. Graph density can be measured for an undirected graph using:

$$\boldsymbol{Density} = \frac{2\,|E|}{|V|(|V|-1)} \tag{11}$$

, and for a directed graph using:

$$\boldsymbol{Density} = \frac{|E|}{|V|(|V|-1)} \tag{12}$$

where the E represents the number of edges and V is the number of nodes or vertices in the graph.

7. Experimental Result

In this section, experiments are conducted on the present dataset (1241 users). Dual-Stage for friend's recommendation system based on user's interests is proposed, and the experimental details are as follows:

7.1 Dataset

The data used for the dual-stage FR model is obtained via surveying active social media users. An electronic survey adopted from a previous study [3] is used to collect inputs from users on their social media activities and preferences. For example, users are asked about their favorite movie or music genre, books, video games, or the social communities to which they belong. The questions provide binary numerical values of 0 or 1, where 1 represents the presence of this particular feature (like) and 0 otherwise (dislike). There are seven main questions that inform 99 features. These features will serve as binary (0, 1) vector representation of users' interests. The total number of users collected is 1241, as shown in Table-1. A summary of the current dataset for user interest is given in Figure-3.

Table 1-Dataset Statistics

Statistics	Quantity
Number of users	1241
Users (male)	584
Users (female)	657
Users age Range	17 - 84

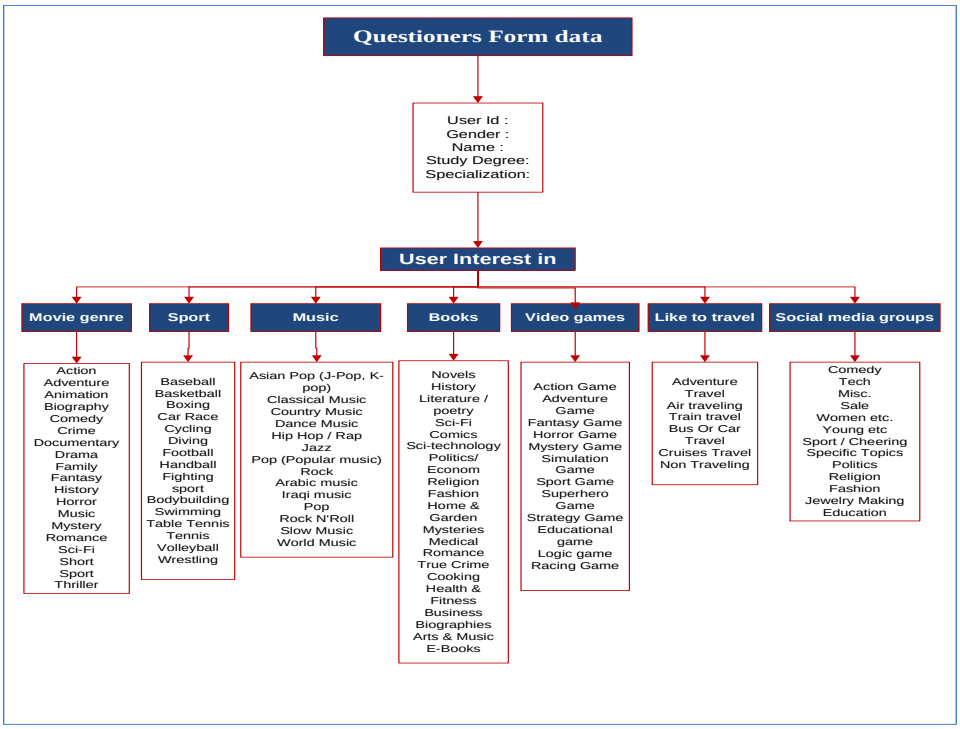

Figure 3- User interest summary

7.2 Preprocessing

The goal of this step is to create a binary (0, 1) 99-dimension vector representation for the features. Hence, we use the methods of find and replacement to replace strings of "Yes" or "Like" to 1 and "No" or "Dislike" to 0. Table-2 shows an example of users data collected from the online survey platform after preprocessing.

Table 2-Example of user's data in sparse representation

ID	Action	Adventure	Animation	...	Racing Game	Adventure Travel	Air traveling
Usr1	0	1	0	...	0	0	0
Usr2	1	1	0	...	0	0	1
Usr3	1	0	0	...	0	0	0
Usr4	1	0	0	...	1	0	0
Usr5	1	0	1	...	0	0	0

7.3 Data Sparsity Avoidance

One of the challenges that face RS and FRS is data sparsity. Data coming from the online survey platform is feature-rich, yet it is sparse. After data preprocessing, measurement of sparsity of data showed that 86% of the data are sparse. To avoid sparsity, non-negative matrix factorization (NMF) is used, as shown in algorithm (1).

Algorithm 1 : NMF Algorithm

INPUT: Sparse Matrix A, number iteration
Output: W,H (Dim reduce)

A ∈ in R^d (m, n)
H ∈ R^d (m, k)
HW ∈ R^d (k, n)
iter : Iteration
A : Input matrix
H : Required matrix H
W : Required matrix W
Initialize(W, H) randomly
Cost_function(W, H) = (A - dot_product(H, W))^2 → (Squared Error)
For iter in number_iteration:
  Update H
    H = H × ((A ×W^T) / (H×W × W^T)))
  Update W
    W = W×((H^T× A) / (H^T× H× W)))
END For
Return (W, H)

The significance of the NMF algorithm is to mitigate the sparsity of the data and replace it with more a compact, denser representation. NMF reduces the sparsity from 86% to 44.9%. Figure-4 shows the results before and after using NMF.

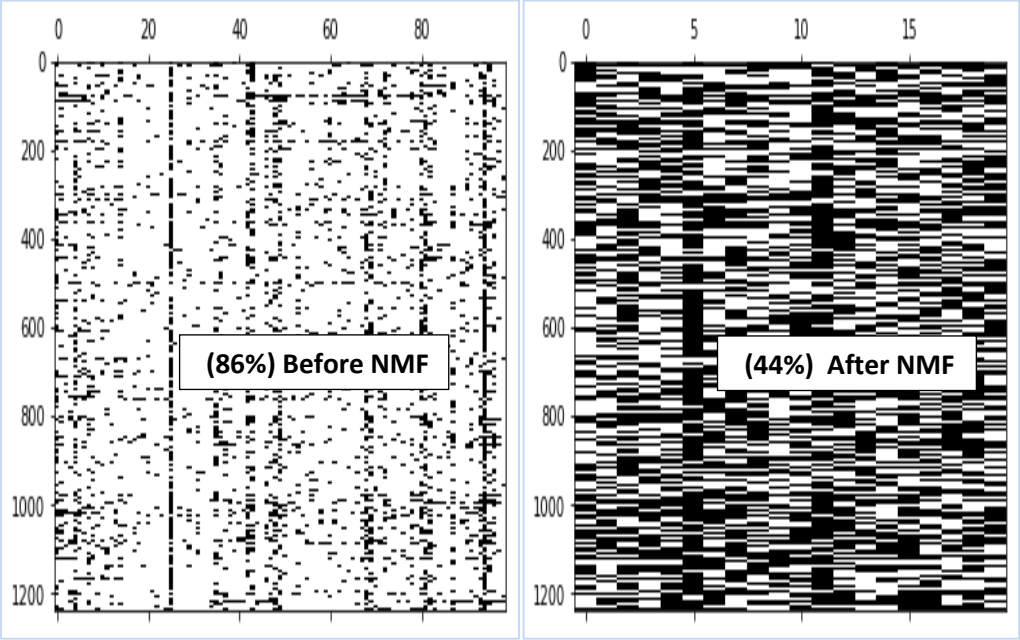

Figure 4- Sparsity before and after using NMF

7.4 Stage One FR model: Initial Graph Construction

The NMF algorithm results in creating denser vectors of 20-dimension instead of the sparse 99-dimensions. Thus, in order to start recommending users based on interests, we need to calculate the similarity between their corresponding vectors to measure which users are to be recommended to each other. The cosine distance in equation (8) measures the similarity among users' vectors to construct the initial recommendation graph, as shown in the example in Figure-5.

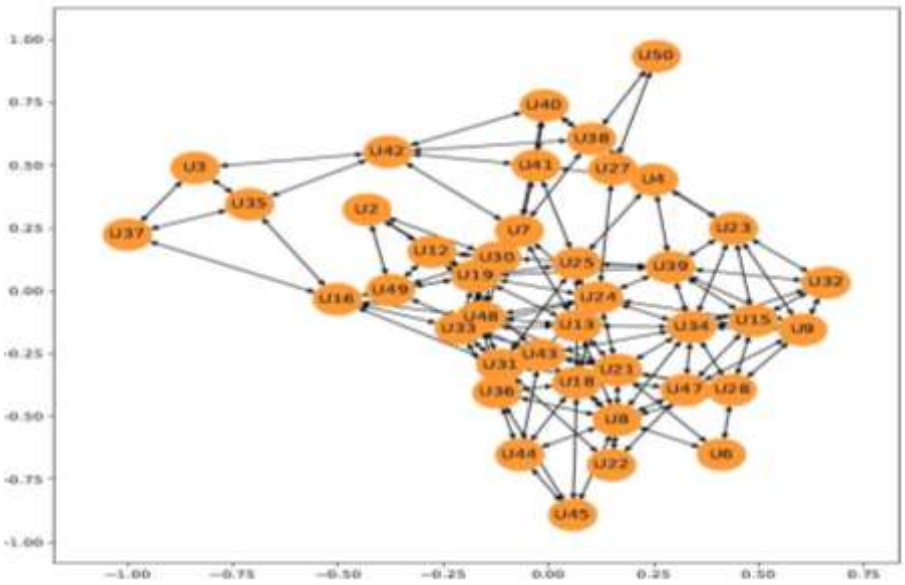

Figure 5- Small sample of stage one initial graph.

It can be noticed from Figure-6 that node U35, for example, represents user 35 who has 4 friends (user 37, user 3, user 42 and user 16). Similarly, user 50 has two users: user 38 and user 27). Another example that can be noticed is that user 32 has five friend suggestions: user 23, user 39, user 34, user 15 and user 9. Edges among all users represent the similarity scores under a threshold of 0.5. The similarity weights represent whether or not the connected users are to be recommended by the algorithm, as illustrated in algorithm (2), stage one.

Algorithm 2 : Stage One FR

Input: Spars Vectors
Output: Users Recommendation graph (Initial_graph)

```
Dim Reduce = Call NMF From Algorithm (1)
For i to N users:
    For j to N users:
        Cosin_Distance(i, j)
        IF Cosin_ Distance (i, j) < delta  and  (i ≠ j) :  (where Delta =  0.5)
            Similarity score represent the edge between two users
            Recommend (i, j)
          Initial_graph (i, j)
    End For
End For
Return  G((i, j)
```

Users with certain similarity score will be treated as nodes in the recommendation graph, while the similarity scores will serve as weights on the edges. Table-3 shows a sample user-to-user adjacency matrix of the first 10 users. It can be noticed that the diagonal of the matrix is always zero because it shows the cosine similarity of the user with itself. For the purposes of recommendation, a threshold is imposed on the cosine similarity values to 0.5. This will force users with 50% similarity or greater to be recommended to each other.

Table 3-Sample user-to-user matrix after similarity calculation

	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10
U1	**0**	0.50	0.32	0.98	0.96	0.97	0.98	0.94	0.58	0.87
U2	0.50	**0**	0.36	0.27	0.95	0.47	0.50	0.72	0.77	0.52
U3	0.32	0.36	**0**	0.79	0.80	0.75	0.80	0.75	0.60	0.85
U4	0.98	0.27	0.79	**0**	0.98	0.35	0.65	0.68	0.99	0.46
U5	0.96	0.95	0.80	0.98	**0**	0.98	0.89	0.09	0.15	0.96
U6	0.97	0.47	0.75	0.35	0.98	**0**	0.40	0.66	0.93	0.85
U7	0.98	0.50	0.80	0.65	0.89	0.40	**0**	0.76	0.90	0.83
U8	0.94	0.72	0.75	0.68	0.09	0.66	0.76	**0**	0.20	0.86
U9	0.58	0.77	0.60	0.99	0.15	0.93	0.90	0.20	**0**	0.88
U10	0.87	0.52	0.85	0.46	0.96	0.85	0.83	0.86	0.88	**0**

7.5 Stage Tow FR model: Graph Augmentation Using FOF Method

One issue that may arise from stage one is that some users may not get a sufficient number of friend recommendations due to low similarity scores. To resolve this issue, a FOF method is applied to recommend to him or her most similar friends of friends (the one with maximum cosine similarity with a friend), as shown in algorithm (3), stage two.

Algorithm 3: Stage Two FR

Input: Initial_graph (i, j) , From algorithm (2)
Output: Final_Graph

```
Call stage one Algorithm (2)
Final_Graph = Associative Array ( )
For user in Initial_Graph [user]:
        IF length(Intial_Graph [user]) < Gamma:
            Closest_Friend_list = get Closest_friend[user
            Most_Similar_Friend  = min(closest_friend_list)
          Friends_of_most_similar_user_list = get from
             Initial_Graph(Most_Similar_Friend)
            FOF = Min(Friends_of_most_similar_user_list)
            While user not in Friends_of_most_similar_user_list:
                Add user and FOF to Final_Graph
                Calculate New_Score(user, FOF) =
              max(most_similar_friend, FOF)
            End While
        End IF
End For
```

For example, as shown in (Figure-6), User1 has 4 friends (U2, U3, U4, and U5), where 0.4, 0.2, 0.4, and 0.5 are their similarity scores, respectively. First, the algorithm fetches U3's friends (U22, U9, U15, U8, U12, and U11) and their similarity scores are 0.3, 0.29, 0.25, 0.27, 0.19, and 0.1, respectively. Next, it recommends U11 to U1 because it is the most similar friend to U3. Finally, the algorithm assigns a new similarity score between U1 and U11 to be the maximum of two scores. A new similarity score is required to differentiate between the users that were recommended during stage 1 and stage 2.

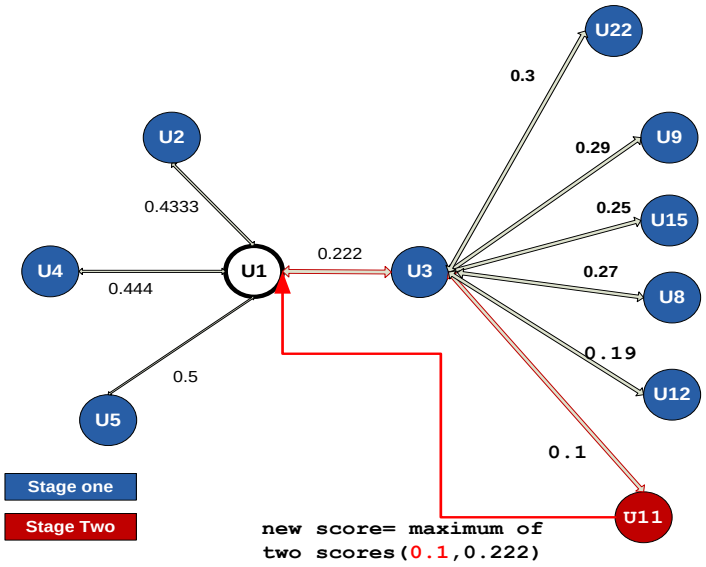

Figure 6- Stage two Recommendations (case study)

8. Evaluation

For evaluation, the present recommendation outcomes are compared by using the Pearson correlation coefficient (PCC) equation (9) between cosine similarity algorithm and Euclidean distance algorithm, as shown in equation (8) and equation (10), respectively, as baseline method to test the user's relationships. PCC helps in data modeling and analysis to predict the relationships between variables. Figure-7 shows that the PCC result of testing the current system is positive (0.47), which indicates that the present user's relationships are positively correlated.

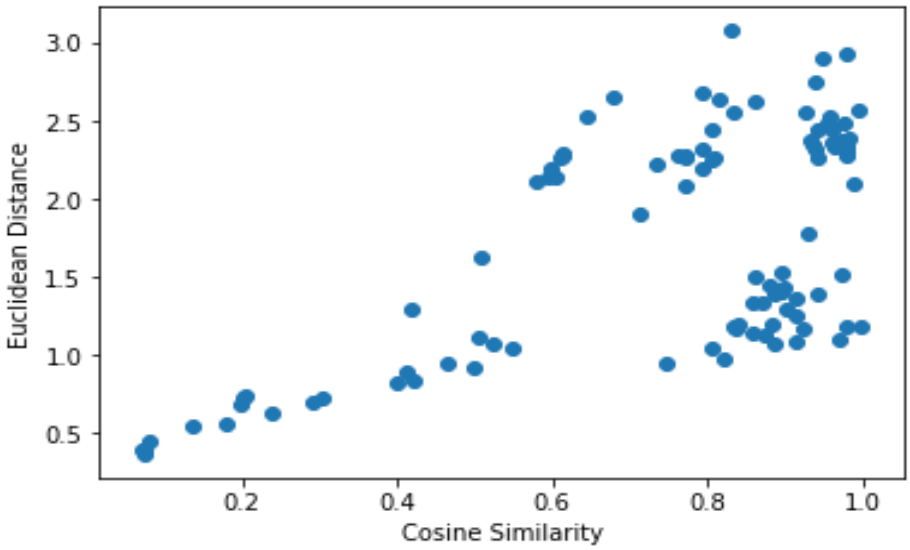

Figure 7- PCC between Cosine Similarity and Euclidean

Since the Dual-Stage FR model relies mainly on graph-based methods, summarizing the information of the initial graph is useful. As shown in Table-4, the initial graph generated from stage one algorithm (2) has a total number of vertices or nodes of 1241. The graph also has 379776 edges, which represent the connections between the users and their friend suggestions. The table also shows that the initial graph has an average degree of 306.0242 for both Indegree and Outdegree. A degree of a vertex represents the number of edges connected to the vertex. Finally, the density of the graph can be calculated using equation (11) because the initial graph is undirected. The density is computed by dividing the number of existent edges (379776) by the total number of vertices (n * (n-1)), which is 1538840, with a result that equals to 0.247.

Table 4- Graph Summary Information

Number of nodes	1241
Number of edges	379776
Average in degree	306.0242
Average out-degree	306.0242
Graph Density	0.247

9. Conclusions and Future Work

a. Conclusions

In this study, Dual-Stage FR is designed and implemented. The model brings several advantages to friend recommendation. First, it is less prone to sparse data due to incorporation of the NMF algorithm. Second, it is used with unlabeled data. Third, it combines the advantages of UBCF approach and graph-based (FoF) approach to get more accurate friend recommendation. Furthermore, the simplicity of the Dual-Stage model makes it seamlessly integrated with another FR system. One possible drawback to this model is that it can be difficult to scale on OSN with enormous number of users (millions) as it can be computationally inefficient and time-consuming. The model is evaluated by PCC (eq. 9) against Euclidean distance as a baseline and yielded a positive correlation (0.47). Also, lack of availability of data featuring social behavior information led to the design of a data collection system, specifically designated to friend recommendation purposes, which allows OSN users to include information about their interests and activities.

b. Future Work :

1. Conduction of quantitative and qualitative studies on analyzing the activities of OSN users. This can help to generate datasets featuring information on social behavior.
2. Examination of the usefulness and the applicability of the clustering methods (e.g., K-means, hierarchical, etc.), instead of calculating users similarity in stage one of FR model, to build the initial graph and compare the results to user-based CF.

References

1. Moricz, M. Dosbayev, Y. and Berlyant, M. **2010**. "PYMK: friend recommendation at myspace," in Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, pp. 999-1002: ACM.
2. Ricci, F., Rokach, L., Shapira, B. and Kantor, P.B. **2011**. "*Recommender Systems Handbook*," ed. Boston, MA: Springer Science+Business Media, LLC.
3. Chen, S., Owusu, S. and Zhou, L. "Social network based recommendation systems: A short survey," in 2013 International Conference on Social Computing, 2013, pp. 882-885: IEEE.
4. Schafer, J.B., Frankowski, D., Herlocker, J. and Sen, S. **2007**. "Collaborative Filtering Recommender Systems," (in English), *Lecture notes in computer science.*, no. 4321, pp. 291-324.
5. VAIDYA, N. **2017**. "Survey on Types and Techniques used in Recommender Systems " *International Journal of Advanced Computational Engineering and Networking*, **5**(6).
6. Mobasher, B., Burke, R. and Sandvig, J.J. **2006**. "Model-based collaborative filtering as a defense against profile injection attacks," presented at the Proceedings of the 21st National Conference on Artificial Intelligence - Vol. 2, Boston, Massachusetts.
7. Burke, R. **2007**. "Hybrid Web Recommender Systems," in *The Adaptive Web: Methods and Strategies of Web Personalization*, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 377-408.
8. Xiao, J., Wang, M., Jiang, B. and Li, J. **2018**. "A personalized recommendation system with combinational algorithm for online learning," *Journal of Ambient Intelligence and Humanized Computing*, **9**(3): 667-677, 2018/06/01.
9. Adomavicius, G. and Tuzhilin, A. **2005**. "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," (in English), *IEEE Transcation on Knowledge an ddata engeneering*, **17**(6): 734-749.

10. Zeng, Y., Wang, Y., Huang, Z., Damljanovic, D., Zhong, N. and Wang, C. "User interests: Definition, vocabulary, and utilization in unifying search and reasoning," in International Conference on Active Media Technology, 2010, pp. 98-107: Springer.

11. Akbari, F., Tajfar, A.H. and Nejad, A.F. **2013**. "Graph-Based Friend Recommendation in Social Networks Using Artificial Bee Colony," (in No Linguistic Content), *Ieee 1*1th International Conference on Dependable, Autonomic Secure, Computing, pp. 464-468.

12. Eirinaki, M., Louta, M.D. and Varlamis, I. **2014**. "A trust-aware system for personalized user recommendations in social networks," **44**(4): 409-421, 2014.

13. Zhao, Y., Yang, Y., Mi, Z. and Xiong, Z. **2015**. "Combining clustering algorithm with factorization machine for friend recommendation in social network," in 2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), 2015, pp. 887-893: IEEE.

14. Huang, S., Zhang, J., Lu, S. and Hua, X.S. **2015**. "Social Friend Recommendation Based on Network Correlation and Feature Co-Clustering," presented at the Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China.

15. Hasan, M.M., Shaon, N.H., Marouf, A.A., Hasan, M.K., Mahmud, H. and Khan, **2015**. M.M"Friend recommendation framework for social networking sites using user's online behavior," in 2015 18th International Conference on Computer and Information Technology (ICCIT), 2015, pp. 539-543.

16. Wu, M., Wang, Z., Sun, H. and Hu, H. **2016**. "Friend recommendation algorithm for online social networks based on location preference," in *2016* 3rd International Conference on Information Science and Control Engineering (ICISCE), pp. 379-385: IEEE.

17. Ding, D., Zhang, M., Li, S.Y., Tang, J., Chen, X. and Zhou, Z.H. **2017**. "Baydnn: Friend recommendation with bayesian personalized ranking deep neural network," in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 1479-1488: ACM.

18. Kumar, P. and Reddy, G.R.M. **2018**. "Friendship recommendation system using topological structure of social networks," in Progress in Intelligent Computing Techniques: Theory, Practice, and Applications: Springer, pp. 237-246.

19. Tsuge, S., Shishibori, M., Kuroiwa, S. and Kita, K. **2001**. "Dimensionality reduction using non-negative matrix factorization for information retrieval," in 2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat. No. 01CH37236), vol. 2, pp. 960-965: IEEE.

20. Lee, D.D. and H. S. Seung, H.S. **2001**. "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, pp. 556-562.

21. Fócil-Arias, C., Zúñiga, J., Sidorov, G., Batyrshin, I. and Gelbukh, A. **2017**."A tweets classifier based on cosine similarity," *CLEF*.

22. Manish, H. and Wojtek, K. **2011**. "Recommender systems for e-shops," *Vrije University, Amsterdam*.

23. Raghuwanshi, S.K. and Pateriya, R. **2019**. "Collaborative Filtering Techniques in Recommendation Systems," in *Data, Engineering and Applications*: Springer, pp. 11-21.

24. Coleman, T.F. and Moré, J.J. **1983**. "Estimation of sparse Jacobian matrices and graph coloring blems," *SIAM journal on Numerical Analysis*, **20**(1): 187-209.