Who are the Most Influential Users in a Recommender System?

Mohammad Amin Morid
Department of Computer Engineering & Information Technology
Amirkabir University of Technology
Tehran, Iran
morid@aut.ac.ir

Mehdi Shajari
Department of Computer Engineering & Information Technology
Amirkabir University of Technology
Tehran, Iran
mshajari@aut.ac.ir

Alireza Hashemi Golpayegani
Department of Computer Engineering & Information Technology
Amirkabir University of Technology
Tehran, Iran
sa.hashemi@aut.ac.ir

ABSTRACT
Collaborative filtering (CF) is a popular method for personalizing product recommendations for e-commerce applications. In order to recommend a product to a user and predict her preference, CF utilizes product evaluation ratings of the like-minded users. This process of finding the like-minded users causes a social network to be formed among all users. In this social network, each link between a couple of users presents an implicit connection between them. Here, there are some users who have more connections with others and are called the most influential users. This paper attempts to model and analyze the behavior of these users by employing data mining techniques. First, the most important features which present a user's influence were selected with a linear regression method, and then, the modeling was performed by a decision tree. Based on our results, the most influential users are users who show more interest to rate more than average number of items with low frequency. Moreover, other most influentials are users who rate in moderation items which have been seen in moderation. In addition, these items are rated with good degree of agreement with other users' rates on the items. We achieved a high accuracy with this model.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.5 [**Information Storage and Retrieval**]: On-line Information Services; I.2.6 [**Artificial Intelligence**]: Learning.

General Terms
Algorithms, Experimentation, Performance.

Keywords
Recommender Systems, Social Networks, Collaborative Filtering, Most Influential Users, Data Mining.

1. Introduction
With the increasing popularity of e-commerce, customers are faced with a large number of products and advertisements. Navigating web pages to find a desired product consumes valuable time. An intelligent mechanism should be developed to guide a customer to the desired product without wasting time navigating irrelevant web pages. One of the tools which can be used to support this idea is a recommender system. Recommender systems explore the interests of a user's past activities and rely on decisions made by like-minded people [8].

Recommender systems may use variety of techniques to infer a customer's preferences. Collaborative filtering (CF) is one of the most successful recommendation techniques and is currently used in different applications specially recommending products such as movies, music, and books [2,7].

CF uses a database of customer ratings to identify items that a user is likely to prefer. The preference likelihood to an item is estimated based on the recorded preference of other customers to the item and the degree of similarity between these customers and the target customer [7].

By employing this technique there will be some users that their ratings are used by the recommender system more frequently than the other users. These users have more influence on the recommender system's performance and are called the most influential users. A key goal of the present research is to analyze and model the behavior of these users.

2. Collaborative Filtering
Collaborative filtering consists of two major steps. At first it computes the similarity between users, and in the second step, the technique will predict a user's preference to an item. For the computation of similarity between users, the Pearson coefficient is the most common approach as shown in the following equation [4]:

$$r_{ij} = \frac{Cov(i,j)}{\sigma_i \sigma_j} = \frac{\sum_k (S_{ik} - \overline{S_i})(S_{jk} - \overline{S_j})}{\sqrt{\sum_k (S_{ik} - \overline{S_i})^2} \sqrt{\sum_k (S_{jk} - \overline{S_j})^2}} \quad (1)$$

where S is the customer representative. So, S_{ik} is the rating value of the ith customer for kth item. The r_{ij} is the correlation coefficient between customers i and j, and $\overline{S_i}$ is the average preference score of customer i. The correlation coefficient r_{ij} will be close to 1 if two customers i and j have similar preferences and will be close to -1 if they have opposite preferences. Values close to zero convey no correlation.

Next step is the estimation of the customer preference to an item. The predicted preference score, P_{ik} , of customer i to item k can be calculated as shown in Equation 2. Here, *Rater(k)* presents the set of customers who scored item k. In this equation, a predicted preference value of customer i to item k is calculated by taking the weighted average of all the ratings for item k plus the average rating scores of customer i to other items [3].

$$P_{ik} = \bar{S}_i + \frac{\sum_{l \in Rater(k)} (S_{lk} - \bar{S}_l) r_{il}}{\sum_{l \in Rater(k)} |r_{il}|} \quad (2)$$

where l is a user who has rated item k. The correlation r_{il}, specifies the impact of the user l's preferences on the user i's predicted preference which results in a more similar user i causes a higher impact to the predicted preference of user i.

3. Social Network on CF Recommender System

In CF recommender systems, relationships between users are formed by what they rate in common and the way they rate. In other words, a relationship between users is established when a group of users rate from a common pool of items which represents the similarity of their taste. By expanding this idea, a social network between the users of a CF recommender system can be inferred. In this social network, nodes represent users and an edge between two nodes expresses an implicit connection between the corresponding users. This connection is formed according to their non-zero similarity based on Pearson correlation coefficient. For more clarification, Figure 1 shows a scenario where a recommendation is being computed for the target (user, item) pair (u_t,m) by using CF algorithm. All the users who have rated at are inspected so that at most k users (see Equation 2) most similar to u_t can be selected as neighbors. Since opinions of each of these neighbors are consulted to calculate the prediction a directed link can be imagined from u_t to each of the selected neighbors. Here, the originating node of a link represents u_t, and the destination node of the same edge represents one of her selected neighbors [6]. By expanding this scenario for all (user, item) pairs the promising social network will be constructed.

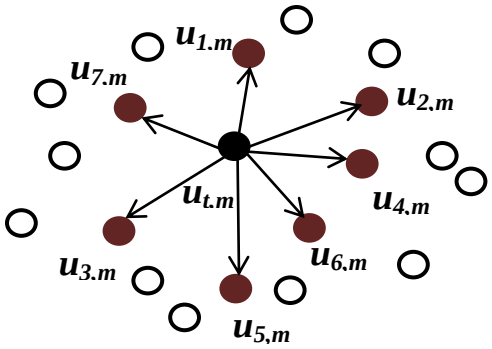

Figure 1. Social network of user u_t with her neighbors

3.1 Most Influential users in a social network

In the introduced social network, there might be a set of users who help a large number of other users to receive recommendations. These users are known as most influential users [6]. Most influential users in a social network graph such as Figure 1 are the nodes which are the destination for many directed links.

3.2 Computing users influence

In order to detect the most influential users, we need a method to identify influence value for all users and then select users with highest value. To do so, a novel method called Loo-Based Influence was introduced in [6]. The method models influence of a user in a recommender system by observing what happens when the user is absent from the system. The more people are affected by the absence of the target user, the more influential she is. To do so, the Leave One Out (Loo) strategy is employed. The target user's rating profile will be removed from the recommender system's data set, and then, its effects on rest of the users will be analyzed. Therefore, influence of a user as described in [5] is computed by the following equation:

$$Influ_{u_i} = \sum_j w_{a_j} \Pr(C_{a_j} = 1 | \tilde{u_\iota}) \quad (3)$$

where, w_{a_j} indicates the probability of a_j being rated, and is computed as the fraction of users who has rated the item. C_{a_j} is a binary random variable indicating shift on a_j after u_i is removed. Therefore, if recommendations on a_j are computed for total n_{a_j} of users and $\acute{n}_{a_j|\tilde{u}_i}$ of them see changes in their recommendations, $Pr(C_{a_j} = 1|\tilde{u_\iota})$ is computed as:

$$Pr(C_{a_j} = 1 | \tilde{u_\iota}) = \frac{\acute{n}_{a_j|\tilde{u}_i}}{n_{a_j}} \quad (4)$$

More details about this equation can be found in [5].

4. Analysis of the most influential user's behavior

In this section, we model the behavior of the most influential users in order to infer the reason of their great influence. To do so, we used the public data of the movie recommendation website MovieLens.org. MovieLens data comprises movies rated from 1 to 5 for about 1690 movies by 950 users. The total data is 100 thousand user ratings and some additional information about users (occupation, age, marriage, etc.) and movies (genre, production date, director, etc.). Moreover, in all of our experiments 80 percent of the data set has been used for training and the rest for testing phase.

4.1 Feature selection for influence modeling

In order to model the most influential users' behavior, we need to specify the model's input which should be a set of users' attributes. To achieve this, some potential attributes which seem to affect a user's influence is selected. Then a linear regression model is applied on the dataset to model the influence value of all users. Afterward, according to the regression model's output, attributes with negligible weight is leaved out and the rest of attributes are chosen for modeling the behavior of the most influential users. We selected 24 potential attributes which are defined as follow:

Number of ratings: This is the total number of movies which have been seen and rated by a user. Any user who has rated many movies has greater chance to have common movies with other users and so might be useful for recommendation of the non-common movies to them (See Equation 2).

Degree of agreement with others: This measure shows on average how much a user agrees to the average opinion of others. It can be computed by following equation:

$$DegAgr_{u_i} = \frac{1}{|I_{u_i}|} \sum_{a_j \in I_{u_i}} \left| S_{u_i,a_j} - \overline{S_{a_j}} \right| \quad (5)$$

where I_{u_i} is the set of movies that user u_i has rated. Low value of *DegAgr* for any user shows her similar interest with others which

increases her chance to be within the nearest neighbors of other users in recommendation process (See Equation 2).

Rarity of the rated items: This measure shows the rarity of movies which are seen by a user and is computed as follow:

$$Rarity_{u_i} = \frac{1}{|I_{u_i}|}\sum_{a_j \in I_{u_i}} \frac{1}{Freq(a_j)} \quad (6)$$

where $Freq(a_j)$ is the number of users who have rated movie a_j. Incentive of using this factor is that two users who rated obscure movies share a common interest that is unique [6].

Average of user's rating: This shows the rating average of a user.

Standard deviation in user's rating: This amounts to the degree a user's ratings deviate from her rating average.

Frequency of the rated items: This measures the average rating frequency of the rated movies and is computed by following equation:

$$AvgFreq_{u_i} = \frac{\sum_{a_j \in I_{u_i}} Freq(a_j)}{|I_{u_i}|} \quad (7)$$

This factor shows weather a user is interested in popular and high frequently rated movies or non-popular and low frequently rated movies.

Interest to different genres: This identifies the level of a user interest in different movie genres. It can be computed as the fraction of the rated movies in each 18 genres. From the above mentioned attributes some of them are also mentioned in [6].
By considering the above factors as the model input and user's influence value as the output, a linear regression model is applied on the dataset. The prediction performance is shown Table 1.

Table 1. Regression model performance on computing user's influence

Root Mean Squared Error	Mean Absolute Error	R^2
6.13	3.06	0.78

High prediction accuracy of the model shows significant correlation between the inputs and the output. By studying the weights of each of the 24 factors, useless factors with negligible weights are omitted. Finally, 4 attributes remain as the most effective factors on the influence value of a user. These attributes are: Number of ratings (*Total*), Popularity of the rated movies (*AvgFreq*), Rarity of the rated movies (*Rarity*) and Degree of agreement with others (*DegOfAgree*).

4.2 Most influential user's behavior modeling

This section attempts to model the behavior of the most influential users based on the selected attributes from the previous section. To do so, different data mining models can be employed such as decision trees, neural networks, association rules and so on. Due to our purpose, mining models such as neural networks which have poor interpretability cannot be used. In addition, since the significance level of the model's components is important, association mining is not pleasant. Therefore, a decision tree with information gain measure is employed for modeling the behavior of the most influential users. More details about this method can be found in [1].
Calculated influence value for all users fall into the rang 0 to 80. Users with an influence value more than 30 are considered as the most influential and lower than 5 as a regular user. Since we are using decision tree model, it is needed to discretize the continuous value of input attributes. To do so, the attributes are discretized according to entropy of their distribution. For example, Figure 2 depicts the distribution scheme of all regular and most influential users in the MovieLens according to their *Total* and *AvgFreq* values.

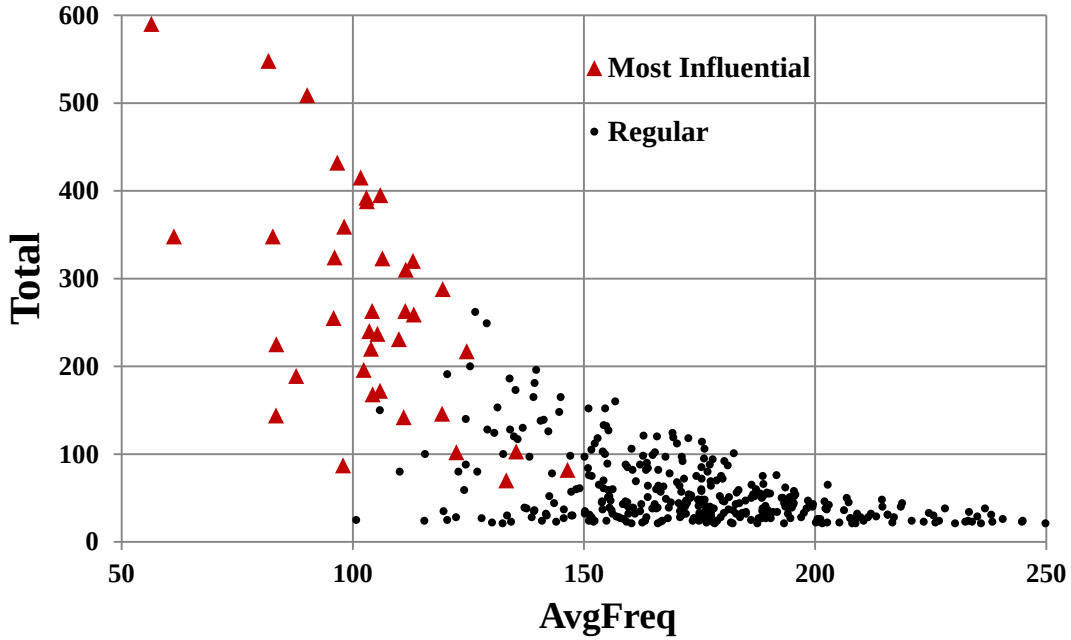

Figure 2. Distribution scheme of all users according to their Total and *AvgFreq* values

As it is seen from Figure 2, users with *AvgFreq* values lower than 100 are all most influential and bigger than 150 are all regular. Also, for *AvgFreq* values between 100 and 150, both user types can be found. Therefore, *AvgFreq* is discretized and labeled according to 100 and 150 (i.e *AvgFreq*<100 labeled as low, 100<*AvgFreq*<150 as medium, *AvgFreq*>150 as High). Similar approach is applied for the other three attributes to discretize them.

Finally, the decision tree model is applied to the discretized data set and the constructed tree is shown in Figure 3:

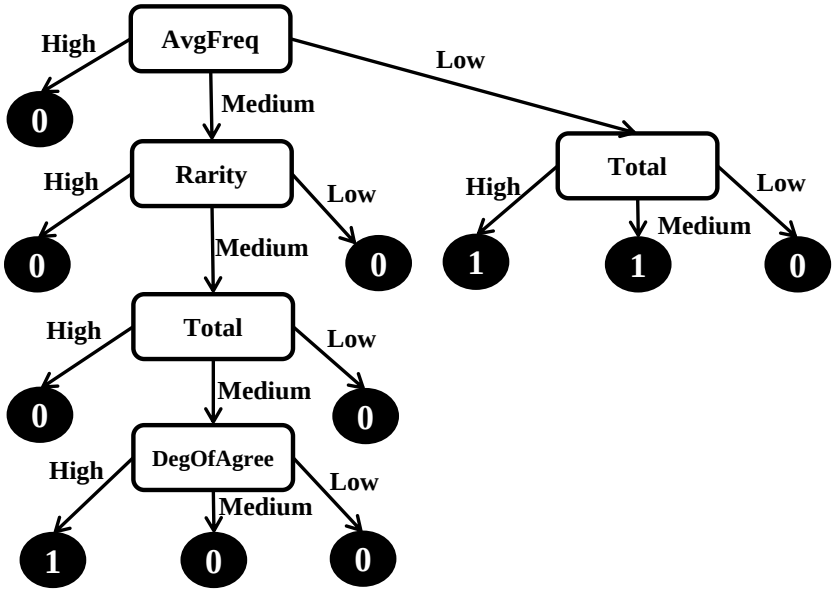

Figure 3. Final decision tree. Leafs with value 1 indicate the most influential users and 0 indicate the regular users.

Also, the performance results according to two main classification performance measures, i.e., Precision and Recall (Equation 8 and Equation 9), are presented in Table 2.

$$precision = \frac{true\ positives}{true\ positives + fales\ positives} \quad (8)$$

$$recall = \frac{true\ positives}{true\ positives + fales\ negatives} \quad (9)$$

Where *true positives* is the number of most influential users that were classified correctly, *false positives* is the number of regular users that were classified incorrectly and *false negatives* is the number of most influential users who were classified incorrectly.

Table 2. Decision tree model performance

User Type	Precision	Recall
Most Influential	100	85
Regular	98	100

Table 2 shows that the model's performance is remarkable, which conveys the suitability of the model for the data. By analyzing the decision tree, different features of the most influential users can be determined. First, we study those leafs with 1 value at the tree's first level (i.e. *AvgFreq =Low* and *Total=Medium* or *High*) which contains more than 90 percent of the most influential users. These are the users who have seen more than average number of movies with low frequency. Note that, a low frequent movie maybe a new movie in online movie store or a non-popular movie. Another group of the most influential users are concentrated at lowest level of the decision tree (*DegOfAgree=High*) which contains about 10 percent of all most influential users. These are the users who have seen a moderated number of movies with *AvgFreq* equal to *Medium* (i.e., they are not seen frequently and at the same time they are not rarely seen movies). Here, each movie is rated with good degree of agreement with other users' rates on the movie.

5. Conclusions and Future Work

This paper was an attempt to analyze and model the behavior of the most influential users in the recommender systems social networks. It was found that most of these users are who have seen more than average number of items with low frequency. Since, there are not many alternatives for a recommender system on low frequently rated items, system uses these users' rates. Therefore, it causes their high rate of influence on the recommender system. In addition, a small portion of the most influential users in a recommender system are the users who rate in moderation items which have been seen in moderation (i.e. items which are not seen frequently and at the same time they are not rarely seen items). Moreover, these items have been rated with good degree of agreement (i.e. similarity) with other users' rates on the items.

Attacking recommender systems is a new issue in these systems, which has attracted the researchers' attention. Here, an attacker tries to manipulate a recommender system in order to change its recommendation's output according to her wish. To do so, attackers employ different methods in order to affect the recommendation process of other users (i.e., by being within their nearest neighbors). If an attacker succeeds, her profile will be used many times by the recommender system, which causes her to be an influential user. Therefore, we believe that there should be a relationship between the attacking power and the influence of a user in recommender systems. Moreover, as our current study, we have conducted several experiments that confirm the existence of this relationship. In addition, we purpose to distinguish the users who are truly influential from the attackers who try to higher their

influence. Our experiments on this area have shown good results in detecting different attacks on recommender systems (e.g. average attack, random attack, bandwagon attack and etc.). These researches and their results will be published in our future papers.

6. References

[1] Han, J. and Kamber, M. 2006. *Data mining: concepts and techniques (2nd ed.)*, Morgan Kaufmann.

[2] Kim, J. W., Lee, B. H., Shaw, M. J., Chang, H., and Nelson, M. 2001. Application of decision tree induction techniques to personalized advertisements on Internet storefront. *International Journal of Electronic Commerce*, 5, 3, 45–62.

[3] Lee, H.J., Kim, J.W. and Park, S.J. 2007. Understanding collaborative filtering parameters for personalized recommendations in e-commerce. *Electronic Commerce Research*, 7, 3, 293-314.

[4] Mild, A., and Natter, M. 2002. Collaborative filtering or regression models for Internet recommendation systems. *Journal of Targeting, Measurement and Analysis of Marketing*, 10, 4, 304–313.

[5] Rashid, A.M. 2007. Mining Influence in Recommender Systems. *Ph.D. Dissertation*, University of MINNESOTA, USA.

[6] Rashid, A.M., Karypis, G. and Riedl, J. 2005. Influence in Ratings-Based Recommender Systems: An Algorithm-Independent Approach. *Proceedings of SIAM International Conference on Data Mining.*

[7] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. 2001. Item-based collaborative filtering recommendation algorithms. *Proceedings of WWW10*, Hong Kong, 285–295.

[8] Suwimo, S. and Esichaiku, V. 2001. Web Personalization Techniques for E-commerce. *Lecture Notes in Computer Science*, 2252, 1, 36–44.