# Personalization using Memory-based CF approach for Marketing of Footwear

Khaing Wah Wah Linn; Khin Lay Thwin
*University of Computer Studies, Hpa-an*
*khaingwah2linn@gmail.com*

## Abstract

*Personalization systems compute recommendations based on user information. There are many kinds of personalization techniques such as collaborative filtering, content-based filtering, rule-based approach etc. Collaborative filtering (CF) has become an important data mining technique to make personalized recommendations for books, web pages or movies, etc. Collaborative filtering (CF) selects advertisement for customers based on the opinions of other customers with similar past preferences. Memory-based CF predicts users preference based on their similarity to other user in the database.*

*This paper presents memory-based CF to predict for footwear advertisement. Pearson Correlation Coefficient Similarity algorithm is used to find the similarity between users.*

## 1. Introduction

In order to draw users' attention and to increase their satisfaction towards online information search results, search engine developers and vendors try to predict user preference based on the user behavior. Personalization systems are implemented in commercial and non-profit web sites to predict the user preferences. For commercial web sites, accurate predictions may result in higher selling rates. The main functions of personalization systems include analyzing user data and extracting useful information for further predictions. Personalization systems are designed to allow users to locate the preferable items quickly and to avoid the possible information overloads.

Among the various approaches, collaborative filtering (CF) is the process of filtering for information or patterns using collaborative techniques involving multiple users. Collaborative filtering approach allows personalization for e-commerce by exploiting similarities and dissimilarities among users' preferences.

The CF predicts the preference score of a user for items he/she had not previously evaluated. To predict the preference score, CF systems use a user-item data set which is comprised of numerical preference scores with a fixed range.

CF can be divided into two types [6]. They are memory-based and model-based CFs. The memory-based CF refers to an all user-items matrix of a data set for predicting preference scores. It evaluates the similarity between each user or item, generates nearest neighborhoods, and predicts preference scores with nearest neighborhoods. The evaluation of similarity is the most essential step, and the evaluated similarity is used as a weight for predicting preference scores and as a measure for generating nearest neighborhoods [7], [10], [4], [5].

The organization of this paper is as follows: Section 2 presents the related work of the system. Section 3 discusses about the Personalization systems and different types of personalization approaches. In section 4, Memory based Collaborative Filtering approach is described. Section 5 presents the proposed system design and Section 6 has the system implementation.

## 2. Related Work

Collaborative recommender systems offer personalized recommendations of articles to users based on information about similarities among users' tastes. Most available systems rely on simple methods to represent similarities among users, e.g. the linear correlation coefficient for a given pair of users. The task in CF is to predict the preference of a particular user (active user) based on a database of users' preferences. There are two general classes of CF algorithms: memory-based methods and model-based methods [8].

Memory-based algorithm is the most popular prediction technique in CF applications. The basic idea is to compute the active user's vote on a target item as a weighted average of the votes given to that item by other like-minded users. The Pearson correlation, which was first introduced in [9], has been widely and successfully used as a similarity measure between users. Memory-based methods have the advantages of being able to rapidly incorporate the most up-to-date information and relative accurate prediction [8], but they suffer from the poor scalability for large number of users.

Model-based CF, in contrast, uses the user's preference database to learn a model, which is then used for predictions. The model can be built off-line over a matter of hours or days. The resulting model is very small, very fast, and essentially as accurate as memory-based methods. The reported model-based CF algorithms include Bayesian networks [8], clustering techniques [8] singular value decomposition with neural network classification [2], induction rule learning [1], and Personality Diagnosis [3]. Model-based methods may prove impractical for environments in which knowledge of consumer preference changes, since the training time is very long.

## 3. Personalization

Personalization is the process of customizing the content and the structure of web sites in order to provide users with the information they are interested in, without asking for it explicitly.

Most data mining approaches to personalization can be viewed as extensions of collaborative filtering. In these approaches the patterns discoveries algorithms take as input the historical rating or navigational profiles of past users and generate aggregate user models. The user models, in turn, can be used, in conjunction with the profile of an active user, to predict future user behavior or generate recommendations.

Learning from data can be classified into memory based learning and model based learning depending on whether the learning is done online while the system is performing the personalization tasks or offline using training data. Model based approaches perform the computationally expensive learning phase offline during the online deployment stage. Memory based systems simply memorize all the data and generalize from it at the time of generating recommendations. This paper presents Personalization system in memory based approach. Personalization systems fall into three basic categories:

- Rule based Personalization System
- Content based Filtering
- Collaborative based Filtering

### 3.1 Rule based Personalization System

Rule based filtering systems rely on manually or automatically generated decision rules that are used to recommend items to users. The primary drawbacks of rule based filtering techniques, in addition to the usual knowledge engineering bottleneck problem emanate from the methods used for the generation of user profiles.

### 3.2. Content based Filtering

In content-based filtering system, a user profile represents the content descriptions of items in which that user has previously expressed interest. The content descriptions of items are represented by a set of features or attributes that characterize that item. The primary drawbacks of content based filtering systems are their tendency to overspecialize the item selection since profiles are solely based on the user's previous rating of items.

### 3.3. Collaborative Filtering

Based on the assumption that users with similar past behaviors have similar interests, this system recommends items that are liked by other users with similar interests. This approach relies on a historic record of all user interests such as can be inferred from their ratings of the items on a website (products or web pages). Rating can be explicit (explicit ratings, previous purchases, customer satisfaction questionnaires) or implicit (browsing activity on a website). Computing recommendations can be based on memory based or model based learning phase to model the user interests.

## 4. Memory based Collaborative Filtering

In memory based learning, all previous user activities are simply stored, until recommendation time, when a new user is compared against all previous users to identify those who are similar, and in turn generate recommended items that are part of these similar users' interests. Memory based models are fast in training/learning, but they take up huge amounts of memory to store all user activities, and can be slow at recommendation time because of all the required comparisons.

$v_{i,j}$= vote of user *i* on item *j*

$I_i$ = items for which user *i* has voted

Mean vote for *i* is

$$\overline{v}_i = \frac{1}{|I_i|} \sum_{j \in I_i} v_{i,j}$$

Eqn. 1

Predicted vote for "active user" *a* is weighted sum

$$p_{a,j} = \overline{v}_a + \kappa \sum_{i=1}^{n} w(a,i)(v_{i,j} - \overline{v}_i)$$

Eqn. 2

### 4.1 Pearson's Correlation Coefficient

A correlation is a measure of the strength of the linear relationship between two measurable variables. The Pearson correlation coefficient, represented as r, gives the strength and direction of this relationship. The closer r is to 1 or to -1 then the stronger the linear relationship between the two variables.

- To demonstrate the concept of a correlation two simple sets of data, representing the scores for variable X and Y will be used.
- X = 2, 6, 9, 10
- Y = 5, 6, 4, 12
- Both the x and y scores are increasing (with one exception in the y set). This means variable X and Y are positively correlated. The correlation coefficient r when calculated will lie between 0 and 1.
- If the x scores were increasing and the y scores decreasing then there would be a negative relationship and the r would lie between 0 and -1.
- To calculate Pearson's correlation coefficient five sums need to be found: the sums of X , Y , $X^2$, $Y^2$ and XY.
- The sums are best worked out by using a table format and totaling each column. The symbol Σ is used to represent each sum or total.

In order to compute the Pearson's Correlation Coefficient r:

$$\boxed{r = \frac{SP}{\sqrt{(SS_x \times SS_y)}}} \qquad \text{Eqn. 3}$$

Where,

$$SP = \sum XY - \frac{\sum X \times \sum Y}{N}$$

$$SS_x = \sum X^2 - \frac{(\sum X)^2}{N}$$

$$SS_y = \sum Y^2 - \frac{(\sum Y)^2}{N}$$

## 5. Proposed System

This system presents the personalization system for online footwear store. Collaborative Filtering approach is used in proposed system. It uses memory-based CF to get the footwear advertisements. Memory-based collaborative approach is easy to implement. New data can be added and it will increment automatically. It does not consider the contents of the items being recommended. It scales well for co-rated items. The main difficulty is the speed of the system since it calculates it calculates every time user is log in. Therefore when it is compared to the Model based approach in speed; Memory-based algorithm has processing overhead.

Memory-based filtering uses the entire or a sample of user item database to generate the prediction. It uses the following steps:

- calculate the similarity or weight between two user
- produce a prediction for the active user by taking the weight
- The active user and a set of weights calculated from the user database.
- The predicted vote of the active user for item ***i***, $p_{a,i}$ is a weighted sum of the votes of the other user:

$$P_{a,i} = \bar{r}_a + \frac{\sum_{u \in U} (r_{u,i} - \bar{r}_u) \cdot w_{a,u}}{\sum_{u \in U} |w_{a,u}|},$$

Eqn. 4

where U = number of users in the CF database

w(a,u)= similarity between each user i and active user

Proposed system design is shown in Figure 1. When the new user comes in, the system builds the initial profile and saved in the database. This recommendation is calculated based on the transactions of each user, compared to the other user groups. Personalization set is calculated based on the similarity between current user and other user groups.

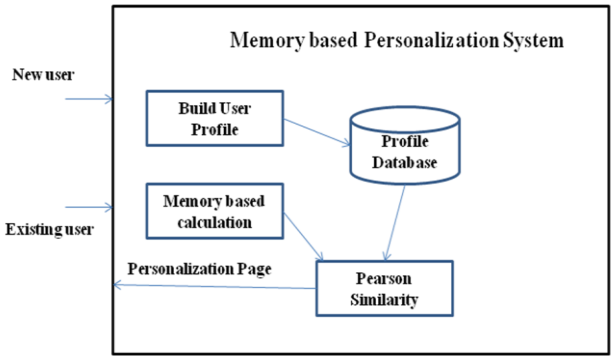


**Figure 1**: Proposed Personalization Architecture

## 6. System Implementation

This system is implemented as web-based footwear sales system. Footwear advertisements will come up as a result of personalization system. Microsoft ASP .Net with C# is used to develop the system. Microsoft Access 2003 is used to store the footwear data and customer records. Customer records include customer's demographic profile and sales profile. Personalization process is performed when the customer logged into the system. It finds the nearest customer

### 6.1. Finding Similar Persons

In this system, when the user is logged in, the system finds the similar users by Pearson's Correlation Coefficient. It is based on user's vote on each product item. In this system, there is no voting feature for the item. It is assumed that user likes when he / she bought it. If two users bought the same item or same items, then they are said to be similar. Table 1 shows the bought item strings of users.

**Table 1**: Bought Items of users

| User | I1 | I2 | I3 | I4 | I5 | I6 |
|---|---|---|---|---|---|---|
| A | 1 | 1 | 0 | 0 | 0 | 1 |
| B | 0 | 0 | 1 | 1 | 0 | 0 |
| C | 1 | 1 | 0 | 0 | 0 | 0 |
| D | 0 | 0 | 1 | 1 | 1 | 1 |
| E | 0 | 0 | 1 | 1 | 1 | 0 |

**Pearson's Coefficient of user B and other users:**
Score (B, A) = 0
Score (B, C) = 0
Score (B, D) = 4 / 8 = 0.5
Score (B, E) = 6 / 9.485 = 0.707
If user B is log on the system, user D and user E are nearest users. Items user D and E bought, but User B has not bought yet are I5 and I6, Therefore I5 and I6 are recommended to the user.

### 7 Experimental Result

This system is tested with 250 items in the database and 100 customers registered in the system. There are 87 sales transactions in the system with 220 sales detail items. The experiment is proceeded in a desktop PC environment with Processor Intel(R) Pentium(R) Dual CPU 2.20GHz, Memory 2 GB of RAM. We have tested accuracy with different values. Memory based approach is compared with Model based approach. Accuracy is calculated as follows.

When the system is set to similarity threshold 0.35, Model based approach missed some of the recommended items (higher accuracy) of Memory based approach. When threshold is set to 0.55 some of the items missed in recommended by both algorithms. Table 2 presents the experimental results of the system.

**Table 2**: Accuracies of Memory based and Model based approaches

| Threshold | Model based | Memory based |
|---|---|---|
| 0.3 ~ 0.4 | 0.9156 | 0.96115 |
| 0.4 ~ 0.5 | 0.9156 | 0.9356 |
| 0.5 ~ 0.6 | 0.9023 | 0.9023 |

## 8. Conclusion

The primary goal of personalization is to evaluate the accuracy and effectiveness of web personalization. CF has become an important data mining technique to make personalized recommendations for books, web pages etc. CF method works by assessing the similarities among user on the basis of their overlap in requests. Then, recommending to an active user that likeminded user consumed previously. The experimental results show that Memory based algorithms work well in the personalization process.

## 9. References


[1] C. Basu, H. Hirsh, and W. Cohen, “Recommendation as Classification: Using Social and Content-based Information in Recommendation”, In Proceedings of the 1998 Workshop on Recommender Systems, pages 11-15, AAAI Press, August 1998.

[2] D. Billsus and M. J. Pazzani, “Learning Collaborative Information Filters”, In Proceedings of the International Conference on Machine Learning, 1998.

[3] [Pennock et al, 2000] D. M. Pennock, E. Horvitz, S. Lawrence and C. L. Giles, “Collaborative Filtering by Personality Diagnosis: A Hybrid Memory- and Model-Based Approach”, in Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence, pp. 473-480, Morgan Kaufmann, San Francisco, 2000.

[4] E. Garcia, “Cosine Similarity and Term Weight Tutorial”, http://www.miislita.com/information-retrievaltutorial/cosine-similarity-tutorial.html#Cosim. 2006.

[5] F. L. Wang, “Improvements to Collaborative Filtering Algorithm”, Lecture Note in Computer Science, Vol. 3314, 2004, pp. 975-981.

[6] G. Adomavicius and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," IEEE Trans. on Knowledge and Data Engineering, vol. 17, no. 6, June 2005 , pp. 734-749.

[7] J. L. Herlocker, J. A. Konstan, A. Borchers and J. Riedl, “An Algorithmic Framework for Performing Collaborative Filtering” ACM SIGIR 22nd Int. Conf. on Research and Development in Information Retrieval, 1999, pp.230-237.

[8] J. S. Breese, D. Heckerman, and C. Kadie, “Empirical Analysis of Predictive Algorithms for Collaborative

Filtering", In Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, 1998.

[9] [Resnick et al., 1994] P. Resnick, N. Iacovou, M. Sushak, P. Bergstrom, and J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews", In Proceedings of the 1994 Computer Supported Collaborative Work Conference.

[10] [6] S. Toby, "Programming Collective Intelligence: Building Smart Web 2.0 Applications", O'reilly, 2007.