A COLLABORATIVE FILTERING APPROACH COMBINING CLUSTERING AND NAVIGATIONAL BASED CORRELATIONS

Ilham Esslimani, Armelle Brun and Anne Boyer
KIWI Team, Université Nancy2, LORIA
615 rue du Jardin Botanique, 54600 Villers-Lès-Nancy, France

Keywords: Recommender systems, Collaborative filtering, Clustering, Usage analysis, Navigational patterns.

Abstract: Recommender systems are widely used for automatic personalization of information on web sites and information retrieval systems. Collaborative Filtering (CF) is the most popular recommendation technique, but several CF systems still suffer from problems like data rating availability and space dimensionality for neighborhood selection. In this paper, we present a new CF approach (PSN-CF) that uses usage traces to model users. These traces are used to estimate ratings that will be employed to generate clusters. Then, the PSN-CF evaluates navigational correlations between users within these clusters. Predictions are performed in a following step. The performance of PSN-CF is evaluated in terms of accuracy and time processing on a real usage dataset. We show that PSN-CF highly improves the accuracy of predictions in terms of MAE. Moreover, the use of clustering and positive sequences before computing the navigational correlations contributes to an important reduction of time processing.

1 INTRODUCTION

The development of Internet engendered an important proliferation of information resources. Thus, the need of tools for automatic personalization of information becomes heightened. Recommender systems are widely used for this purpose thanks to their ability to analyze users behaviors and guide them towards relevant resources that suit their preferences.

Collaborative Filtering (CF) is a recommendation technique that identifies similarities between users, based on their ratings in order to select neighbors and compute predictions for the active users.

Despite the success of recommender systems and collaborative filtering in many application areas, some research questions still remain. Some of these questions concern the requirement of explicit rating data to compute correlations between users. As explicit rating data is not always available, one challenge for recommender systems is to take into consideration another type of data that could represent efficiently users behaviors. In this context, usage traces can be a relevant source of data.
Moreover, another challenge for CF systems is the reduction of space dimensionality. Indeed, as the number of users and resources tend to increase, it turns out that the required time for computing correlations and generating neighborhoods also increases. Therefore, employing clustering techniques is one way to reduce time required for processing these correlations.

So, the research problem we are interested in, is related to the integration of usage traces in CF systems. Thus, by using these traces, how can we estimate implicit ratings, how can we select nearest neighbors and evaluate correlations between users, finally how can we improve the accuracy of predictions.

In this paper, we attempt to explore these issues and propose some solutions. We suggest a new CF approach called PSN-CF, that exploits navigational patterns to model users. This approach integrates users clustering and a new navigational based technique to enhance the performance of CF.

This paper is organized as follows. We describe in the second part some research studies related to clustering based recommender systems and analysis of usage traces. In the third part of the paper, we present the PSN-CF approach. The fourth part describes the experimentation. Then, the results of the model experimentations are put forward in the fifth part and finally we discuss these results and present a conclusion.

Esslimani I., Brun A. and Boyer A.
A COLLABORATIVE FILTERING APPROACH COMBINING CLUSTERING AND NAVIGATIONAL BASED CORRELATIONS.
DOI: 10.5220/0001841303640369
In *Proceedings of the Fifth International Conference on Web Information Systems and Technologies (WEBIST 2009)*, page
ISBN: 978-989-8111-81-4

2 RELATED WORK

2.1 Clustering based Recommender Systems

In the context of recommender systems, clustering algorithms are generally used to identify clusters of similar users, sharing preferences. Clustering methods have been integrated in several CF based recommender systems in order to reduce dimensionality or to alleviate the sparsity and scalability problems. To overcome the sparsity problem, (Xue et al., 2005) use a CF system based on a K-means clustering in order to smooth the unrated data for individual users according to the clusters. For the same issue, (Jiang et al., 2006) suggest a cluster-based collaborative filtering based on an iterative clustering method that exploits the inter-relationship between users and items. In this model, both users and items are clustered using the K-means algorithm, then a predicted rating is generated over user classes and item classes.

As regards the problem of scalability, (Conner and Herlocker, 1999) choose to partition items by experimenting various clustering algorithms. Predictions are then computed independently within each partition. With the same perspective (George and Merugu, 2005) use a collaborative filtering approach based on a weighted co-clustering algorithm that involves simultaneous clustering of users and items.

2.2 Analysis of Usage Traces

Several studies describe the impact of usage traces on the recommendation process in predictive systems. Analysis of usage traces is mainly related to the area of Web Usage Mining (WUM) which aims at observing users behaviors while interacting with a system. This observation refers to direct traces as explicit ratings and annotations, or non direct traces like bookmarking, frequencies of visits, visited links, etc. from which users preferences can be inferred.

Frequent patterns mining, Longest Common Subsequences (LCS) technique and Markov models, are some of the WUM approaches that tend to harness the navigational activities in order to analyze users behaviors. (Gery and Haddad, 2003) describe the attempt of frequent patterns mining as the discovering of time ordered sequences that have been followed by past users in order to predict future resources.

Discovering of Longest Common Subsequences (LCS) is another technique that has been applied in WUM domain in order to analyze the potential links between navigational paths and users profiles. Basically, this technique is one dynamic programming method, it aims at identifying the longest common subsequence relating to two given sequences. (Jalali et al., 2008) suggest an LCS based architecture for classifying navigational patterns and generating predictions to users. In (Banerjee and Ghosh, 2001) an algorithm based on LCS technique is proposed for clustering users by using their navigational data. This clustering approach uses the similarities between two navigational paths based on the LCS and the time spent on resources contained in this LCS.

Another approach that uses sequential links for navigational activities is Markov chain model. In accordance with (Eirinaki et al., 2005), the sequential dependencies of navigational behaviors of users are modeled by Markov Models ; the conditional probability of one resource, considering users navigational traces is computed.

3 DESCRIPTION OF PROPOSED APPROACH

In this paper, we propose a new collaborative filtering model which exploits on one hand clusters of users based on a similarity matrix of users ; this step allows a reduction of dimensionality. On the other hand it uses the navigational patterns, so as to refine the result of this clustering and produce recommendations.

Figure 1 describes the different layers of our model called “Pam clustering on Similarities and Navigational based-CF” (PSN-CF) comparing to the classical clustering based CF model. The classical clustering based CF (dashed lines) uses directly the rating matrix (User x Item) in order to generate clusters and compute predictions. At the opposite, the PSN-CF model (solid lines) uses a similarity matrix (User x User) computed by using the rating matrix, so as to create clusters of users. At the same time, PSN-CF applies a selection of “Positive Sequences” from the original users sequences, these sequences contain the preferred accessed resources. Then PSN-CF computes the navigational correlations between users, based on Positive Sequences, within individual clusters. The Positive Sequences are used to improve the time processing over the stage of computing navigational correlations. Finally, the new generated neighbors are used to compute predictions.

The following sections present in details the different mechanisms used by PSN-CF model at each layer.

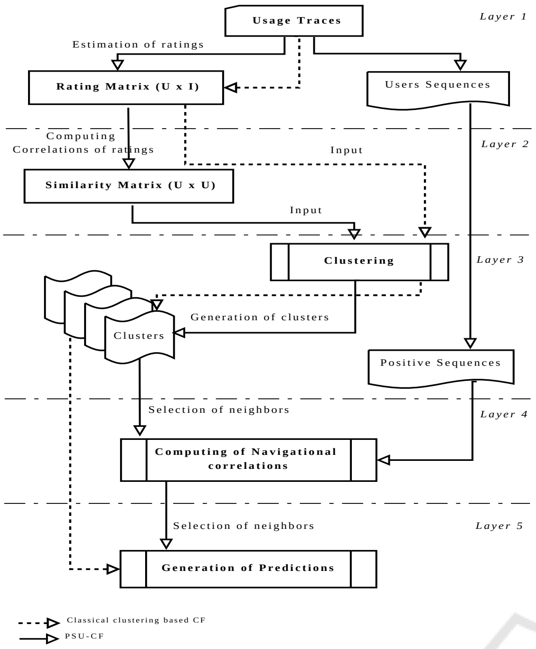

Figure 1: Global scheme comparing the PSN-CF approach to the classical clustering based CF.

3.1 Estimating Ratings

As mentioned in section 2.2, implicit ratings can be inferred from users navigational activities. In our approach, to estimate these ratings (*Layer* 1) we choose two implicit parameters: frequencies of visiting a resource and duration of visiting a resource. Considering an active user u_a, the frequency of visiting an item i_k is the ratio of the number of visits of i_k ($N_{(u_a,i_k)}$) and the average number of visits on all items I ($\overline{N_{(u_a,I)}}$) as described in Equation (1).

$$Frequency_{(u_a,i_k)} = \frac{N_{(u_a,i_k)}}{\overline{N_{(u_a,I)}}} \qquad (1)$$

As regards duration, it is computed as the ratio of the duration of visiting an item i_k ($Drt_{(u_a,i_k)}$) and the total duration of visiting all items I ($Drt_{(u_a,I)}$) as presented in Equation (2).

$$Duration_{(u_a,i_k)} = \frac{Drt_{(u_a,i_k)}}{Drt_{(u_a,I)}} \qquad (2)$$

Once frequencies and durations are calculated, we use a formula suggested by (Castagnos, 2008), in order to compute and normalize our ratings according to the rating scale $[1-5]$ from bad to excellent.

3.2 Computing Clusters

In the context of recommender systems, clustering methods usually use “User x Item” matrices to create clusters. In our approach, we choose to employ rather a “User x User” matrix (containing similarity values between users) in order to generate clusters. Hence, the clusters are constructed based on similarities with other users instead of similarities on ratings. Moreover, in classical clustering approaches, the clusters are constructed based on co-rated items between users. At the opposite, our approach ensures the exploitation of additional items by taking into account users similarities.

3.2.1 Generation of the Similarity Matrix

In order to generate the “User x User” similarity matrix (*Layer* 2) required for the clustering step, we use the Pearson Correlation Coefficient (Herlocker et al., 1999) so as to compute the similarities between users based on the estimated ratings.

3.2.2 Users Clustering

In our approach, the goal of clustering data is to reduce the search space and the time required for computing navigational correlations and to improve the selection of neighborhoods. We choose a partitioning algorithm called PAM (Partitioning Around Medoids). The relevance of the PAM method comparing to other partitioning algorithms like K-means is its robustness (Kaufman and Rousseeuw, 1990). We choose to use a “User x User” matrix as input of the PAM algorithm. Thus, clusters are generated according to the similarities between users as presented in *Layer* 3. In the following section, we present the technique used for evaluating navigational similarities between users.

3.3 Navigational Similarities

We propose at this step the integration of navigational patterns in the process of recommendations based on a collaborative approach, so as to refine the clusters provided by the previous step (*Layer* 4).

In our model, we consider that two users u_a and u_b, who share common sequential patterns are highly correlated. Therefore, our goal is to identify for every pair of users $< u_a, u_b >$, the maximum length $L_{Kmax}(u_a,u_b)$ of a pattern among their common patterns. As described in (*Layer* 3), we select from original sequences, only “Positive Sequences”(P_s) in order to reduce the time processing required to identify these common patterns among users sessions. Thus, we retain in sequences only items with high ratings. Then, the similarity of navigation between two users is computed by using Equation 3.

This formula computes, for each pair of users u_a and u_b the correlation of navigation $SimNav_{(u_a,u_b)}$ as the ratio of the maximum length of a common frequent pattern $L_{Kmax}(u_a,u_b)$ and the minimum of maximum sizes of u_a and u_b sessions denoted $SessMax_{(u_a)}$ and $SessMax_{(u_b)}$. We note that the common frequent pattern is intra-session.

$$imNav_{(u_a,u_b)} = \frac{L_{Kmax}(u_a,u_b)}{\min(SessMax_{(u_a)}, SessMax_{(u_b)})} \quad (3)$$

This metric emphasizes the importance of the longest frequent patterns to evaluate similarities of users. The higher the length of a sequential pattern is, the more the users are correlated.

3.4 Prediction Generation

Once the navigational similarities between the active user and other users within a cluster are calculated, we employ the weighted average prediction formula used by classical CF (Herlocker et al., 1999) to compute predictions. This step corresponds to the last layer of the PSN-CF approach (*Layer* 5).

4 EXPERIMENTATION

4.1 Datasets

In order to evaluate the performance of PSN-CF, we use real usage datasets extracted from the intranet of Credit Agricole Banking Group, in particular the usage data relating to the Department of Strategies and Technology Watch.

Thus, to train our model, we use the usage data that reflects the navigational activities of users. This data has been collected during 24 months and stored in server log files. The selected dataset is related to 748 users and 3856 resources. It has been split into 80% and 20% corresponding respectively to training and test datasets. The tests have been performed on a Windows Server 2003 PC, with 2 Go of RAM and $3,4$ GHz processor (Pentium IV).

As regards clustering, in order to create clusters from matrices, we used “R”[1], an environment for statistical computing and graphics. In our experiments, 10 clusters are created.

4.2 Evaluation

Different evaluation metrics can be used in the experimentation of CF systems. The most important criterion in recommender systems is precision. The precision measures the accuracy of recommendations comparing to real votes. As a measure of precision evaluation, we used the Mean Absolute Error (MAE). This metric computes the mean of absolute errors between predicted ratings and the real ratings that are actually assigned by users.

Since items that have high prediction values are the ones that are recommended to users, we use also the HMAE (High MAE) metric (Baltrunas and Ricci, 2007) to evaluate the performance of the model. The HMAE is similar to MAE but it considers only items that are predicted with a value of 4 or 5. In our experimentation, we choose the HMAE metric to measure how our system is able to recommend relevant items to active users.

5 RESULTS

In order to analyze the performance of our approach, we evaluate the precision of predictions generated by the PSN-CF model in terms of MAE and HMAE. PSN-CF accuracy is compared to several variants of CF models so as to study the impact of clustering users, the influence of the nature of the matrix used for clustering users and the importance of using navigational patterns.
Additionally, we evaluate the impact of the use of Positive Sequences on time processing of navigational based correlations.

We note that, before the computation of predictions (*Layer* 5), we used at the same time (for all the models) two criteria to select the nearest neighbors of the active user: a threshold relating to the correlation value between the active user and other users and a minimum number of co-rated items between the active user and other users.

5.1 MAE

Table 1 presents the MAE values related to CF models when either no clustering is performed or clustering is applied on a rating matrix. We can first notice that when no clustering is performed the accuracy only slightly decreases when only navigational data is used, compared to classical CF. This confirms the idea that navigational patterns are almost as informative as rating data and may contain complementary information to ratings.

[1] http://www.r-project.org

Table 1: MAE values with and without clustering.

	Classical CF	Navigational CF
No clustering	0.7631	0.7895
K-means	0.7826	0.7971
PAM	0.7998	0.8253

Table 2: MAE values when using a similarity matrix for clustering.

Navigational CF	
K-means	0.7809
PAM	0.6747

In the case of classical CF, performing clustering on ratings (referred to as Classical Clustering based CF in Figure 1, *Layers* 1, 3 and 5) does not improve the accuracy. PAM clustering leads to the lowest accuracy (decrease of about 5%). Besides, when considering navigational data (in addition to ratings), we can notice that, as in the case of classical CF, the use of PAM clustering on a rating matrix leads to the lowest accuracy.

The PSN-CF model we propose is based on clustering applied on a similarity matrix (*Layers* 1, 2, 3, 4 and 5), we thus present in Table 2 the corresponding MAE when either PAM or K-means algorithm is performed.

From Table 2, we observe that when performing PAM clustering on a similarity matrix of users, MAE is improved by about 15% compared to clustering on a rating matrix. This configuration corresponds to the PSN-CF model. This improvement can be explained by the fact that using a similarity matrix to perform clustering does not group users only according to the way they have similarly rated items commonly seen. Users are grouped in a cluster when they are commonly similar to all the other users. Moreover, this approach has the advantage to not only consider commonly rated items, but all the items users have rated.

From Tables 1 and 2, we can notice that the K-means clustering slightly depends on the nature of the matrix: similar accuracy values are obtained for both matrices.

5.2 HMAE

Let us recall that only items with high prediction values are suggested by recommender systems to the active user. Here, we are also interested in the HMAE values when performing clustering, based on a rating matrix or a similarity matrix. These values are presented in Tables 3 and 4.

Table 3: Comparison of HMAE values with and without clustering.

	Classical CF	Navigational CF
No clustering	0.5415	0.5014
K-means	1.2851	1.2723
PAM	1.1689	1.1590

Table 4: HMAE values when using a similarity matrix for clustering.

Navigational CF	
K-means	0.5874
PAM	0.6036

We can first notice that when no clustering is performed, navigational based CF outperforms classical CF by 7% in terms of HMAE, contrary to MAE. When clustering is based on a rating matrix (Table 3), HMAE values are highly increased for both classical and navigational CF. However, at the opposite of MAE (Table 1), the use of navigational patterns in addition to ratings leads to an improvement of HMAE in all the studied models.

Table 4 presents the HMAE values related to clustering based on a similarity matrix in the case of a navigational CF. When clustering is applied on a similarity matrix, HMAE is highly decreased for both K-means and PAM clustering comparing to the results of Table 3. The lowest HMAE is obtained when K-means is used, however a similar HMAE is observed when using PAM clustering (the PSN-CF model).

Even if no improvement is obtained in terms of HMAE compared to no clustering, a large improvement is obtained in terms of MAE and the computation time of neighbors is decreased. The following section is dedicated to the study of this computation time.

5.3 Time Processing

We are now interested in the time processing required during the phase of computing the navigational correlations within clusters, by either applying the selection of Positive Sequences or not.

Results show that the models that do not perform clustering require on average a computation time 4 times higher for computing the navigational correlations. We observe also that the selection of Positive Sequences contributes to an important reduction of time processing. Indeed, the processing time decreases by about 8% when no clustering is used and from 16% to 30% for clustering-based models. Let us note that the use of Positive Sequences by navigational based models decreases accuracy at the worst case by about 1.85% in terms of MAE and by 2.34% in terms of HMAE. When PAM clustering is used either an improvement or a stability is observed in terms of accuracy.

6 CONCLUSIONS

In this paper, we presented a new Collaborative Filtering approach, named PSN-CF, that exploits navigational patterns. PSN-CF is structured in different layers as described in Figure 1.

Unlike classical predictive systems based on usage patterns, PSN-CF is user-based and attempts to identify behavioral correlations between users. The originality of PSN-CF consists in the exploitation of navigational patterns in the context of CF, thus no explicit preferences need to be provided by users. Additionally, PSN-CF exploits the concept of Positive Sequences so as to assess navigational correlations based on users preferred resources with the objective of reducing time processing required for computing these correlations.

PSN-CF has been evaluated both in terms of MAE and HMAE and has been compared to other CF models. The experimentation shows the high interest of using the PAM clustering based on a similarity matrix, on the accuracy of the CF system in terms of MAE. Moreover, the use of clustering based on similarities is also benefit in terms of HMAE. Experiments also showed the relevance of using both navigational patterns and estimated rating data for generating accurate high predictions. Last, the experiments showed the advantage of selecting Positive Sequences that is a trade-off between the optimization of time processing of navigational correlations and reduction of accuracy.

As a future work, we intend to first exploit additional methods that allow the reduction of dimensionality, second evaluate the impact of its combination with the navigational based CF on the accuracy of recommendations. Additionally, we plan to extend PSN-CF in the direction of social networks and examine the possibilities of modeling potential links between users in the context of behavioral networks.

ACKNOWLEDGEMENTS

We would like to thank Mr. Jean Philippe Blanchard and acknowledge the financial support to this project provided by the Credit Agricole Banking Group.

REFERENCES

Baltrunas, L. and Ricci, F. (2007). Dynamic item weighting and selection for collaborative filtering. In *Web mining 2.0 Workshop, ECML-PKDD 2007*. Springer-Verlag.

Banerjee, A. and Ghosh, J. (2001). Clickstream clustering using weighted longest common subsequences. In *Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining*.

Castagnos, S. (2008). *Modélisation de comportements et apprentissage stochastique non supervisé de stratégies d'interactions sociales au sein de systémes temps réel de recherche et d'accés à l'information*. PhD thesis, Nancy 2 University, France.

Conner, M. and Herlocker, J. (1999). Clustering items for collaborative filtering. In *Proceedings of the ACM SIGIR Workshop on Recommender Systems*.

Eirinaki, M., Vazirgiannis, M., and Kapogiannis, D. (2005). Web path recommendations based on page ranking and markov models. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*. ACM Press.

George, T. and Merugu, S. (2005). A scalable collaborative filtering framework based on co-clustering. In *Proceedings of the Fifth IEEE International Conference on Data Mining*. IEEE Computer Society.

Gery, M. and Haddad, H. (2003). Evaluation of web usage mining approaches for user's next request prediction. In *Proceedings of the 5th ACM international workshop on Web information and data management*. ACM Press.

Herlocker, J., Konstan, J., Borchers, A., and Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*.

Jalali, M., Mustapha, N., Sulaiman, N., and Mamat, A. (2008). A web usage mining approach based on lcs algorithm in online predicting recommendation systems. In *Proceedings of 12th conference of information visualisation*.

Jiang, X., Song, W., and Feng, W. (2006). Optimizing collaborative filtering by interpolating the individual and group behaviors. In *APWeb*.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, New York.

Xue, G., Lin, C., and Yang, Q. (2005). Scalable collaborative filtering using cluster-based smoothing. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*.