Chapter 5. Collaborative filtering via customer demands

Although various approaches for making recommendations have been proposed, such as WRFM-based CF method (referred to Chapter 3) and WRFMCP method (referred to Chapter 4), both methods suffer from limitations. Unlike the CBF method, CF is sensitive to whether users generally in general prefer a product. CBF makes recommendations by analyzing the description of the items that have been rated by the user. Several researchers are exploring hybrid methods of combining CF and CBF to smooth out the disadvantages of each (Basu et al., 1998; Claypool et al., 1999; Good et al., 1999). This work uses customer demands derived from the frequently purchased products in each industry as valuable content information to integrate CF for making recommendations. This work also combines customer demands and past purchasing preferences to reduce the sparsity of the customer-item matrix, and names extended-preferences to improve recommendation accuracy. Customer demands are then included as a factor in making recommendations for re-ranking candidate products.

5.1 Customers demands

CBF methods make recommendations by analyzing the description of the items that user has rated and the descriptions of items to be recommended. Since descriptions of items are not easy to obtain in this work, customer demands for each industry are used for recommendations as available content information. Most customers belonging to the same industry have similar specific demands. These demands are determined by simply statistics to calculate the frequently purchased products of customers in each industry. If the number of frequently purchased product is greater than a given threshold, θ then the industry specifically demands the items. Content-based recommendations can generally deal with new items unseen by others (Balabanovic and Shoham, 1997). This work was limited to the content information available, therefore could not deal with new items. But new customer problems can still be solved. When a new customer wants to buy products, his industry will already be known, allowing the system to recommend products to her/him. The element r_{ij} of the customers-demand matrix $\boldsymbol{CD}$ represents whether the ith customer had specifically demanded the jth product. If the ith customer has specific demands to jth product, r_{ij} is 1; otherwise, it is 0. The similarity of customers demands ($Corr_{cd}$) among customers can also be measured by computing the Pearson correlation coefficient.

5.2 Customers' extended preference

In a real domain, customers may purchase very few product items and thus the customer-item matrix ***R*** is generally sparse. But the fact that a customer has not bought a product does not imply that the customer does not need or is not interested in that product. Therefore, this work proposed a denser matrix based on the spirit of hybrid works of sequential combination to improve the recommendations accuracy. Limited to available content information, that product features are not provided in this work. Accordingly, customer demands are integrated to purchasing preferences to reduce the sparsity of customer-item matrix ***R*** in this work. Herein, the denser matrix is named the extended-preferences matrix, ***EP***. The element r_{ij} of the extended-preference matrix ***EP*** represents whether the *i*th customer had purchased or had specifically demanded the *j*th product. If the *i*th customer has specific needs or had purchased the *j*th product, r_{ij} is 1; otherwise it is 0. The similarity of extended preferences ($Corr_{ep}$) among customers can also be measured by computing the Pearson correlation coefficient.

5.3 Combining WRFM and customer demands

The WRFM-method (referred to Chapter 3) is proposed to identify effective recommendation rules for customers with high lifetime value or loyalty. However, as described previous, these methods belong to collaborative filtering recommendations. Accordingly, this work explores a hybrid method combining WRFM-based CF and customer demands, termed WRFMCD. Figure 9 illustrates this method.

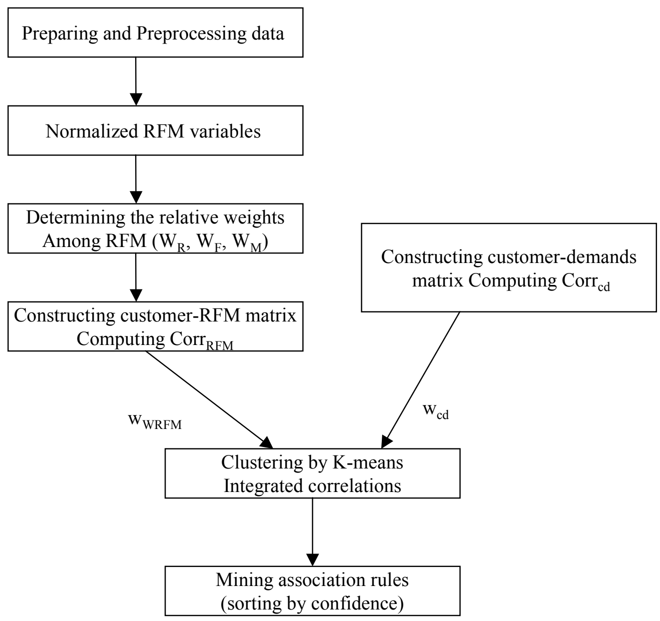

Figure 9. WRFMCD method for product recommendation

The WRFMCD method first establishes a customer-WRFM and a customer-demands matrix. Then, the WRFM-based and the customer-demands correlation coefficients are computed using Pearson correlation coefficient, respectively. Subsequently, K-means clustering is used to group customers with similar CLV and demands based on weighted correlation coefficients. Finally, the association rule mining approach is applied to extract recommendation rules from each group derived from K-means clustering. Each candidate products is sorted by an associated confidence value, and the top-*N* highest ranked candidate products are thus assembled in a recommendation set. The following subsections detail the WRFMCD method.

5.3.1 Grouping customers with similar CLV and customer's demands by K-means

This method clusters customers by integrating their CLV and specific demands (as described in Chapter 5.1). Customers' specific demands are determined by simply statistics to frequent purchased pattern for an industry. Customers belong to an industry have same specific demands. This work begins by establishing the customer-WRFM matrix and specific demands. For the first, the RFM value are normalized and then multiplied by the relative importance of the RFM variables. Eq. 2 (referred to Chapter 3.1.1) is used to compute the WRFM-based Pearson correlation coefficient, $corr_{WRFM}$, while Eq. 1 (referred to Chapter 2.3.2.2) is used to compute the specific demand correlation coefficient, $corr_{cd}$. The integrated correlation coefficient is then derived according to Eq. 4.

$$Corr_{integrated}(c_i, c_j) = w_{WRFM} \times corr_{WRFM}(c_i, c_j) + w_{cd} \times corr_{cd}(c_i, c_j) \quad (4)$$

w_{WRFM} and w_{cd} represent the relative importance (weights) of the elements of CLV and customer-demands, respectively. If the $w_{WRFM} = 0$ then the customer demands is used for recommendations; if the $w_{WRFM} = 1$, then the method becomes WRFM-based CF method.

K-means technique is employed to cluster customers according to the integrated correlation coefficients. In general, such a coefficient between the centroid c_j of a cluster and a customer c_i is measured using Eq. 4. The centroid is represented by both the average WRFM values and the average specific demands of the customers in the cluster. Customers are assigned to a cluster with maximum integrated correlation coefficient. The weights of parameters w_{WRFM} and w_{cd} are used to yield an integrated correlation coefficient. The proper weighting values of w_{WRFM} and w_{cd} can be determined by performing analytical experiments to evaluate the quality of recommendations under different weight combination (for example, w_{WRFM} equals 0.8 and a weight of customer demands equals 0.2).

5.3.2 Recommendation phase

Association rule mining is used to extract a set of recommendation rules from the transactions associated with each cluster. A cluster is generated by grouping customers according to the weighted correlation coefficients of CLV and demands. Then, the set of recommendation rules extracted from cluster C_j is then used to select the top-N candidate products that is to be recommended to customer u. Here, the demands of customers are also used as a factor in recommendation, termed the *adjusted* WRFMCD (A-WRFMCD) method. Each weight is multiplied by the confidence value to fit customer demands and re-rank the candidate recommended products. This work assumes mutual independence between products a and b, explaining why the recommendation value is equal to the probability of buying b, given a, $P_r(b|a)$ multiplied by the degree of demands for product b. The resulting formula is:

$$Recommendation\ value\ (b) = P_r(b|a) * d_i$$

Preliminary experimental results show that the recommendation quality is better by setting d_i = 1, if the customer specifically demands the product b; d_i = 0.5, otherwise. The recommended product that matches the customer demands thus gains high priority in the ranking among candidate products.

5.4 Combining WRFM and extended preferences

The WRFMCD method combines customer lifetime value and customer demands to recommend products to customers, but does not consider customers' purchase preferences. However, the fact that a customer did not buy a product does not imply that he has no need for or interest in it. Therefore, this work proposes another hybrid method by which customers are clustered by integrating the dimensions of customer lifetime value and extended-preferences, named WRFMEP method. This method adopts relative weighting to adjust the importance of customer lifetime value and extended preferences in the clustering. The association rule mining approach is applied to extract a set of recommendation rules from each group derived from K-means clustering.

5.4.1 Grouping customers with similar CLV and extended preference by K-means

We begin with establishing the customer-WRFM matrix and the extended-preferences matrix. The RFM values of customers are normalized and then multiplied by the relative importance of the RFM variables at first. The integrated correlation coefficient is then derived according to Eq. 5.

$$Corr_{integrated}(c_i, c_j) = w_{WRFM} \times corr_{WRFM}(c_i, c_j) + w_{ep} \times corr_{ep}(c_i, c_j) \quad (5)$$

The K-means technique uses integrated correlation coefficients to cluster customers. In general, such a coefficient between the centroid c_j of a cluster and the customer c_i is measured by Eq. 5. The centroid here is represented by both the average WRFM values and the average extended preference of the customers in the cluster. Customers are assigned to a cluster with maximum integrated correlation coefficient. The weights of parameters w_{WRFM} and w_{ep} are used to yield an integrated correlation coefficient. The proper weighting values of w_{WRFM} and w_{ep} can be determined by performing analytical experiments to evaluate the quality of recommendations under different weight combination (for example, w_{WRFM} equals 0.8 and w_{ep} equals 0.2).

If w_{WRFM} = 0, then the extended-preferences is used to recommend product to customers, that we termed CFEP method. If w_{WRFM} = 1, then the method becomes WRFM-based CF method.

5.4.2 Recommendation phase

Association rule mining is used to extract a set of recommendation rules from the transactions associated with each cluster. A cluster is generated by grouping customers according to the weighted correlation coefficient of CLV and extended-preferences. We begin by identifying cluster C_j to which a customer, *u* belongs. Next, the set of recommendation rules extracted from C_j is used to select the top-*N* candidate products to be recommended to customer *u*. Similar to the A-WRFMCD method, the *adjusted* WRFMEP (A-WRFMEP) method uses the customer demands to adjust and re-rank the candidate recommended products.

5.5 Experimental setup

The proposed methods, WRFMCD, WRFMEP are compared with several other methods, including WRFM-based, preference-based CF methods. Based on the hybrid methods of sequential combination, this work implements a CFEP (collaborative filtering extended-preferences) method which combines the preference-based CF method and extended preferences for making recommendations. The first step of CFEP is to establish an extended-preferences matrix (referred to Chapter 5.2). Second, customers are grouped based on similarity in extended preferences. The association rule mining technique is then employed to derive recommendation rules from each group. Additionally, this work implements an EP-based *k*-NN method to make recommendations. The EP-based *k*-NN method employs the nearest-neighbor to recommend products to a target customer based on those customers having similar extended-preferences. The CFEP and EP-based *k*-NN are also compared with the proposed methods.

WRFMCD, WRFMEP, CFEP are compared with A-WRFMCD, A-WRFMEP and A-CFEP to confirm the usefulness of re-ranking candidate products based on customer demands. Like the A-WRFMCD and A-WRFMEP methods, the adjusted CFEP (A-CFEP) method also uses the customer demands to adjust and re-rank the candidate recommendation products. Since most hybrid works first use CBF to reduce the sparsity problem to support CF method, the proposed method makes recommendations based on extended-preferences to improve recommendation accuracy. To verify that methods based on extended-preferences not only improve the overall quality of recommendation, but are also useful in sparse information, this work conducts experiments to evaluate whether methods with extended-preferences is better than those methods without extended-preferences for customers who made few purchases.

5.6 Experimental results

5.6.1 Hybrid methods

5.6.1.1 Evaluation of WRFMCD method

The WRFMCD method considered different weightings on the dimensions of CLV and customer demands. The analytical experiment used the training set as the analytical data set (65%) to derive recommendation rules and 10% to determine the proper weightings, w_{WRFM} and w_{cd} (w_{cd} = 1 - w_{WRFM}). Each candidate products was sorted by associated confidence value, where the top-*N* highest ranked candidate products were selected as the recommendation set. Accordingly, the WRFMCD method achieved the best recommendation quality when w_{WRFM} = 0.3 and w_{cd} = 0.7. Based on the analytical results, Table 17 summarizes the experimental results of WRFMCD and A-WRFMCD methods on the testing set (25% data set) by setting w_{WRFM} = 0.3 and w_{cd} = 0.7 to derive top-*N* recommendations. Overall, the analytical results suggest that the A-WRFMCD method is better than the WRFMCD method.

Table 17. Analytical results of WRFMCD method under different N (top-N)

	WRFMCD method			A-WRFMCD method		
Top-N	Precision	Recall	F1-metric	Precision	Recall	F1-metric
Top-4	0.323	0.327	0.312	0.324	0.331	0.314
Top-10	0.447	0.639	0.516	0.453	0.658	0.528
Top-20	0.453	0.647	0.522	0.467	0.672	0.539
Top-30	0.451	0.648	0.521	0.459	0.664	0.528
Top-40	0.415	0.626	0.488	0.418	0.638	0.494
Top-50	0.412	0.624	0.483	0.417	0.628	0.489

5.6.1.2 Evaluation of WRFMEP method

The WRFMEP method considered different weightings on the dimensions of CLV and extended-preference. The analytical experiment used the training set as the analytical data set (65%) to derive recommendation rules and 10% to determine the proper weightings, w_{WRFM} and w_{ep} (w_{ep} = 1 - w_{WRFM}). If the w_{WRFM} = 0, then the method is CFEP; otherwise, the method is the WRFM-based CF method. Each candidate products are sorted by associated confidence value, where the top-*20* highest ranked candidate products are selected as the recommendation set. The analytical result is shown in left-side of Table 18. The WRFMEP method achieved the best recommendation quality when w_{WRFM} = 0.3 and w_{ep} = 0.7. The right-hand side of Table 18 shows the analytical result of A-WRFMEP. Overall, when $w_{ep} > w_{WRFM}$, the F1 metric of WRFMEP method exceeds that obtained using the WRFM-based CF and the CFEP methods. The A-WRFMEP method also outperformed the WRFMEP method. Based on the analytical results, further experiments were conducted to evaluate the WRFMEP method by setting w_{WRFM} = 0.3 and w_{ep} = 0.7.

Table 18. Analytical results of WRFMEP method (top-*20*)

	WRFMEP method			A-WRFMEP method		
W_{WRFM}	Precision	Recall	F1-metric	Precision	Recall	F1-metric
0	0.413	0.621	0.507	0.432	0.652	0.514
0.1	0.448	0.662	0.528	0.474	0.675	0.542
0.2	0.451	0.665	0.532	0.473	0.675	0.542
0.3	0.457	0.669	0.533	0.474	0.677	0.543
0.4	0.457	0.667	0.533	0.474	0.677	0.543
0.5	0.457	0.664	0.531	0.474	0.675	0.541
0.6	0.453	0.664	0.528	0.469	0.674	0.540
0.7	0.452	0.664	0.527	0.465	0.675	0.539
0.8	0.448	0.658	0.522	0.460	0.673	0.535
0.9	0.442	0.654	0.518	0.459	0.669	0.533
1	0.436	0.647	0.511	0.452	0.663	0.531

5.6.2 Verifying the importance of extended-preferences

Experiments are conducted to compare the EP-based k-NN, WRFMEP, CFEP with KNN-based, WRFM-based CF and preference-based CF methods, respectively, to verify the importance of extended-preferences. Moreover, WRFMEP method was also compared with the WRFM-based CF, WRFMCP, WRFMCD methods. The training set (75%) included product items purchased by customers during a specified period and was used to extract recommendation rules by association rule mining. The analytical data set (25% data set) was used to verify the quality of the recommendations. Methods were compared by varying the N, the number of recommendation items.

Table 19 summarizes the recommendation quality obtained using these various methods. From the analytical results, the F1-metrics of CFEP exceeds those of the preference-based CF method. Moreover, the F1-metrics of WRFMEP exceeds those of the WRFM-based CF method, as well as WRFMCP and WRFMCD method. EP-based k-NN also provides better recommendations than the KNN-based method. Generally, the performance ranking of these methods with extended-preferences is WRFMEP $\succ$ CFEP $\succ$ EP-based k-NN method; while the ranking of these methods without extended preferences is WRFMCP method $\succ$ WRFM-based CF method $\succ$ preference-based CF method $\succ$ KNN-based method. This ranking implies that extended-preferences, combining customer demands and purchased preferences are useful for improving the quality of recommendation.

Table 19. Analytical results of various methods to verify the importance of extended-preferences

Top-N	Preference-based CF method	CFEP	WRFM-based CF method	WRFMCD	WRFMCP	WRFMEP	KNN-based (k=100)	EP-based k-NN (k=100)
Top-4	0.335	0.294	0.333	0.312	0.342	0.323	0.286	0.298
top-10	0.476	0.497	0.499	0.516	0.486	0.497	0.487	0.490
top-20	0.518	0.518	0.524	0.522	0.533	0.535	0.515	0.515
top-30	0.502	0.525	0.504	0.521	0.525	0.533	0.498	0.518
top-40	0.496	0.500	0.484	0.488	0.496	0.513	0.467	0.482
top-50	0.473	0.495	0.477	0.483	0.489	0.505	0.422	0.467

5.6.3 Verifying the importance of re-ranking candidate products

Experiments were conducted to compare various methods by using the 75% training set, 25% testing set to verify the proposed approaches via varying the *N*, the number of recommendation items. Table 20 shows the F1-metrics of various methods under different top-*N* recommendations. Generally, the F1-metrics of both adjust methods exceed those methods without re-ranking candidate products. Re-ranking candidate products according to customer demands that offers a promising method for improving recommendation accuracy.

Table 20. F1-methrics of various methods under different *N* (top-*N*)

Methods	WRFMCD	A-WRFMCD	WRFMEP	A-WRFMEP	CFEP	A-CFEP
Top-4	0.312	0.314	0.323	0.294	0.294	0.319
Top-10	0.516	0.528	0.497	0.508	0.497	0.506
Top-20	0.522	0.539	0.535	0.543	0.518	0.524
Top-30	0.521	0.528	0.533	0.542	0.525	0.527
Top-40	0.488	0.494	0.513	0.522	0.500	0.507
Top-50	0.483	0.489	0.505	0.512	0.495	0.496

5.6.3 Experiments on customers who purchase few product items

In previous experiments, this work focused on confirming overall recommendation accuracy, but not on sparse information. Accordingly, experiments were conducted to compare various methods for those users with purchased items not exceeding 5, 10 and 15 from the 75% training set. The numbers of customer were 73, 161 and 260, respectively.

Table 21 lists the experimental results displaying a trend similar to those of the experiments involving all of the customers. The F1 metrics of methods with extended preferences exceed those of methods without extended preferences. CFEP outperformed the typical CF method. EP-based *k*-NN performed better than the KNN-based method. Furthermore, WRFMEP outperformed the WRFMCP, WRFMCD and WRFM-based methods. The result implies that the proposed hybrid method improves the overall quality of recommendation. Additionally, making recommendations for customers who purchased few product items based on extended-preferences is better than those methods without extended-preferences. Generally, the quality of recommendation improves with increasing number of purchased items.

Table 21. Analytical results of various methods for new customers under different top-N

	Purchased items ≤ 5 (73)			Purchased items ≤ 10 (161)			Purchased items ≤ 15 (260)		
	Top10	Top20	Top30	Top10	Top20	Top30	Top10	Top20	Top30
Preference-based CF	0.3524	0.3858	0.3852	0.3655	0.3632	0.3667	0.3643	0.3637	0.3612
CFEP	0.3718	0.3861	0.3714	0.3824	0.3856	0.3807	0.4151	0.3935	0.3912
KNN-based	0.3438	0.3506	0.3201	0.3502	0.3721	0.3286	0.3523	0.3608	0.3200
EP-based KNN	0.3615	0.3688	0.3239	0.3859	0.3817	0.3306	0.4293	0.3688	0.3239
WRFM-based	0.3712	0.3618	0.3540	0.3856	0.3746	0.3729	0.4214	0.3884	0.3805
WRFMCD	0.3945	0.3914	0.3857	0.4112	0.3969	0.3835	0.4243	0.4064	0.3893
WRFMCP	0.3715	0.3662	0.3547	0.3913	0.3841	0.3803	0.4221	0.4022	0.3814
WRFMEP	0.4042	0.3935	0.3882	0.4197	0.4025	0.3881	0.4273	0.4086	0.3930

5.7 Discussions

The collaborative filtering method has been successfully used in a number of applications, but suffers several limitations. This work uses customer demands derived from frequently purchased products in each industry to integrate CF to make recommendations. This work also combines customer demands and past purchasing preferences to reduce the sparsity of customer-item matrix, named extended-preferences to improve recommendation accuracy. Customer demands is then included as a factor in making recommendations for re-ranking candidate products. This work ran several experiments to confirm the differences between methods.

According to the analytical results, generally, the performance ranking of these methods with extended-preferences is WRFMEP ≻ CFEP ≻ EP-based *k*-NN method; while the ranking of these methods without extended preferences is WRFMCP method ≻ WRFM-based CF method ≻ preference-based CF method ≻ KNN-based method. This ranking implies that extended-preferences, combining customer demands and purchased preferences are useful for improving the quality of recommendation. Furthermore, re-ranking candidate products according to customer demands that offers a promising method of improving recommendation accuracy. Finally, the results of proposed hybrid method not only improves the overall quality of recommendation, but also can be extended to recommend product items to customers who purchased few product items based on extended-preferences. And generally, the quality of recommendation improves with increasing number of purchased items.