Multi-Criteria Service Recommendation Based on User Criteria Preferences

Liwei Liu
Center for Service Research
The University of Manchester
+44 161 306 3319, UK
liwei.liu@postgrad.mbs.ac.uk

Nikolay Mehandjiev
Center for Service Research
The University of Manchester
+44 161 306 3319, UK
n.mehandjiev@mbs.ac.uk

Dong-Ling Xu
Manchester Business School
The University of Manchester
+44 161 306 3319, UK
ling.xu@mbs.ac.uk

ABSTRACT

Research in recommender systems is now starting to recognise the importance of multiple selection criteria to improve the recommendation output. In this paper, we present a novel approach to multi-criteria recommendation, based on the idea of clustering users in "preference lattices" (partial orders) according to their criteria preferences. We assume that some selection criteria for an item (product or a service) will dominate the overall ranking, and that these dominant criteria will be different for different users. Following this assumption, we cluster users based on their criteria preferences, creating a "preference lattice". The recommendation output for a user is then based on ratings by other users from the same or close clusters. Having introduced the general approach of clustering, we proceed to formulate three alternative recommendation methods instantiating the approach: (a) using the aggregation function of the criteria, (b) using the overall item ratings, and (c) combining clustering with collaborative filtering. We then evaluate the accuracy of the three methods using a set of experiments on a service ranking dataset, and compare them with a conventional collaborative filtering approach extended to cover multiple criteria. The results indicate that our third method, which combines clustering and extended collaborative filtering, produces the highest accuracy.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval -- *Clustering, Information filtering, Selection process*

General Terms

Measurement, Performance, Design, Experimentation.

Keywords

Recommender systems, Multiple Criteria Decision Making, multi-criteria recommender systems, clustering, service

1. INTRODUCTION

The Internet has brought about exponential growth of information about services such as hotels, movie screening, theatre performances, etc. This information is available from service review and ranking sites as well as direct e-commerce sites which aim to sell services to online users. Rather than supporting users in their choice, the abundance of this information has caused the *information overload* problem, where customers can't find what they want in a sufficiently short time and are often lost during the searching process. The sheer volume of available information also makes it hard for users to judge its reliability and trustworthiness.

Recommender systems (RS) have been developed as an effective solution to this problem. A Recommender System is defined as "*any system that produces individualized recommendations as output or has the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options*" [1]. They are routinely used by e-commerce websites to help consumers make their purchasing decision. Their use can increase e-commerce sales by: (a) converting browsers into buyers, (b) increasing cross-sell by suggesting additional products and (c) building customer loyalty through "creating a value-added relationship between the site and the customer" [2].

With the development of recommender systems, the recommended objects range widely, from books, movies, music to TV programs, web pages and so on. Most of the existing RS on the market are based on a single numerical rating that represents user's opinion about the item as a whole. Two types of entities, *users* and *items* are used for the recommendation, thus giving it its two classical dimensions $Users \times Items \rightarrow R_0$, R_0 is the set of possible predicted values for the overall impression of the item [3]. However, describing an item through multiple attributes and taking into account user feedback about these can help to make more effective recommendations [4], especially in services area. Indeed, service selection can be considered broadly identical to product selection, yet service selection tends to be more personal because of the closer involvement of the customer in the service delivery [5]. For example, in recommending a restaurant, different people will value differently the different aspects of the service, such as the quality of the food, the speed of the service, and the environment of the meal. Instead of a single rating, we can consider ratings on each of these criteria. Thus we have an overlap between the Multiple Criteria Decision Making (MCDM) methods and recommendation methods, referred to as multi-criteria recommender system. Manouselis and Costopoulou point out that MCDM methods can facilitate the recommendation process [4].

However, MCDM approaches are mostly used for high-value decisions since they require time for users to select criteria weights and rank products according to their criteria. This is not the case with the Recommender Systems, where any recommendation should be calculated transparently from the user, using only past data. This suggests that effective RS support

RecSys'11, October 23–27, 2011, Chicago, Illinois, USA.

would need innovative ways of using the rankings regarding multiple criteria. This is indeed the focus of recent work in multi-criteria recommender systems, which aims to use all available criteria to achieve more effective recommendations [6, 7].

In contrast, we start from the position that, for the majority of users, only a few selection criteria will impact their overall rating. For instance, if the selection of a hotel can potentially be impacted by five criteria: location, cleanliness, rooms, service and value, some users may consider location as most important, whilst other users may prefer cleanliness or service. To use these differences for deriving better predictions, we propose a novel approach which clusters users based on the criteria they prefer, creating a "Preference Lattice". Recommendations for a user are made using rankings from users with similar criteria preferences, which are clustered in the same or in closely related clusters. The novel approach of clustering into a "Preference Lattice" is then instantiated by three methods also created by us: (a) an aggregation function of the criteria, (b) using the total item ratings for the recommendation, rather than the rankings of each criteria, and (c) combining our clustering approach with the results from *collaborative filtering (CF)*, a widely used recommendation approach described in Section 2. At the end, we evaluate the accuracy of these three methods by comparing their results with those produced by the traditional CF approach, which is used as the baseline, and also two main multi-criteria approaches in existence: one extending CF with Euclidean distance metric, and one extending CF with average similarities of criteria [3].

The remainder of the paper is organized as follows: we begin by introducing the general collaborative filtering approach and its two existing extensions for multi-criteria recommender systems, which are later used for comparison with our methods. Then, in Section 3, we introduce our novel approach to clustering users according to their criteria preferences, and then provide the details of the three methods which instantiate our clustering approach. The experiments we conducted to evaluate the proposed methods are detailed in Section 4, including the results from these experiments. We compare our work with existing similar work in Section 5. Section 6 contains the conclusions of the paper and also the future work we intend to explore.

2. BACKGROUND

2.1 Collaborative Filtering Approach

Recommender systems have become an important research area since mid-1990s. They predict unknown ratings of items (products or services) and recommend the items with highest ratings. The recommendation approaches are usually classified into three categories: collaborative filtering, content-based and hybrid approaches [8], [9], [10], [11], [12]. *Collaborative Filtering (CF)* approach, the most common recommendation approaches, recommends to users the items liked by other "similar" users. The similarity is identified on the basis of similar past rankings [11]. The assumption is that users who had common interests in the past tend to have similar tastes in the future [13]. GoupLens, Ringo, Amazon.com et al are all successful CF approach for prediction [14, 15]. According to [16], CF can be grouped into two classes: memory-based and model-based. Memory-based algorithms essentially are "heuristics that make rating predictions based on the entire collection of previously rated items by the users" [10]. This is used as one of the methods in our paper later on.

Identifying those users who have similar taste to the active user in the past is crucial for successful application of CF. The two most frequently used approaches are Pearson correlation and cosine-based approach. The similarity based on Pearson correlation is calculated as follows:

$$sim(u,u') = \frac{\sum_{i\in I}(r_{u,i}-\overline{r_u})(r_{u',i}-\overline{r_{u'}})}{\sqrt{\sum_{i\in I}(r_{u,i}-\overline{r_u})^2}\sqrt{\sum_{i\in I}(r_{u',i}-\overline{r_{u'}})^2}} \quad (1)$$

Where I represents the set of common items rated by user u and user u'. The value of any of the two approaches ranges from -1 to $+1$. The greater the value, the more similar these two users are. Thus, -1 means that the two users have exact opposite taste, and $+1$ means they have exactly the same taste.

2.2 Multi-Criteria Recommender System

Traditional recommender systems use a single rating as input, usually an overall numerical ranking by user i of item j. However, in some applications, this kind of recommendation does not meet users' personalized needs and multi-criteria ratings are considered. E-commerce sites such as *epinions.com* rate a digital camera using multiple criteria such as ease of use, durability, battery life, photo quality and shutter lag. Instead of giving one total rating (also referred as overall rating) to show the preference, users need to express their opinion on an item through rating its multiple attributes in multi-criteria recommender system. In addition to the total rating, multiple criteria ratings provide more information about user preferences from different aspects than the traditional recommender system. And thus the rating function changes to $Users \times Items \rightarrow R_0 \times R_1 \times \ldots \times R_k$, where R_0 represents the total rating if there is any, and R_c represents the rating of each possible criterion c $(c = 1,2,\ldots,k)$ [3]. The total rating shows how well the user likes the item overall, and the criteria ratings provide the insight and explain which aspects of the item he or she likes. Multi-criteria recommender systems predict the overall rating for an item based on past ratings regarding both the item overall and individual criteria, and recommend to the user the item with the best overall score (the detailed prediction formula shown in Section 3.3). According to [3], multi-criteria system provides more information about user's preferences than a single-rating system. And by adopting a decision theory, multi-criteria systems can provide rich tools for system designer to build more interesting systems as well [17]. Thus, the algorithm for a multi-criteria recommender system can be extended from a single-rating recommender system.

Following this approach, Adomavicius and Kwon present two approaches to leverage multi-criteria ratings through extending single-rating CF [3]. One is computing the overall user similarity through aggregating the similarities calculated from each individual criterion (referred as *AvgSimCF*),

$$sim_{avg}(u,u') = \frac{1}{k+1}\sum_{c=0}^{k} sim_c(u,u') \quad (2)$$

Where there are $k+1$ criteria [3]. After that the approach proceeds in the usual CF manner.

The other approach is aiming for a more holistic calculation of user similarity through multidimensional distance metrics. Each rating is presented in a vector format, such as $r_{u,i} = (r_0, r_1, \ldots, r_k)$, and r_0 is the overall rating that user u has rated item i, while r_k presents the rating of criterion k. The distance between two metrics and the two users' similarity is inversely related. For example, using Euclidean Distance to calculate the distance between two users (referred as *EuclideanCF*):

$$d_i(u,u') = \sqrt{\sum_{c=0}^{k}(r_c - r_c')^2} \quad (3)$$

Where k is the total number of the criteria, and r_c is the rating of user u on criterion c. The similarity between user u and u' is denoted as the inverse of the distance:

$$sim(u,u') = 1/(1 + \frac{1}{I}\sum_{i \in I} d_i(u,u')) \quad (4)$$

Where I is the number of items rated by both user u and u' [3].

Both these approaches claim to provide better results than the single-ranking systems, and in Section 4 we compare their results with the output of our approach which is presented next.

3. PREDICTION BASED ON CLUSTERING USER PREFERENCES

Here we present our novel approach to multi-criteria recommendation system, which clusters users according to the criteria they consider valuable in judging the overall quality of the item. This approach is different from the approaches reviewed above which directly extend single-rating to multi-criteria algorithms, and is also different from the approaches which apply MCDM methods based on user input to facilitate recommendation of suitable items.

We are inspired by the fact that different users assign different importance to different selection criteria, *i.e.* different attributes of the item being selected. This is especially true for items which require a higher degree of user involvement, such as services. And we hypothesize that those users who prefer the same criteria have a higher degree of similarity in their selection compared to those who don't. For example, when providing feedback for a hotel service, the users are asked to rate the hotel on five criteria and provide one overall rating. The five criteria are: *location (L)*, *cleanliness (C)*, *rooms (R)*, *service (S)* and *value (V)*. User1, a young traveler, is likely to prefer *location* when he chooses the hotel, and hence his rating for "location" will be most indicative of his overall satisfaction with the hotel. But User2, who loves cleanliness and attention, would likely prefer a hotel with a very *clean* environment, and attentive *service*. If a hotel is far away from the city center or attractions, but is very clean and with good service, User1 may give it a ranking of 2 or 3 out of 5, since the location of this hotel fails his expectations. But User2 who prefers cleanliness will probably give this hotel a high rating of 4 or 5. Indeed, users who value location would generally provide a lower rating to this hotel than the users who prefer cleanliness and service due to their expectation of these criteria. There are some theories supporting this, such as Assimilation Contrast Theory [18].

3.1 Significant Criteria

In a multi-criteria recommender system, the ratings of an item given by a user contain one total rating and a number of criteria-specific ratings. The assumption driving our approach is that the overall rating is usually highly correlated to those criteria ratings which are significant for the individual[19]. Under the assumption that the overall rating has a certain relationship with criteria ratings, the criteria that can affect the overall rating significantly are those criteria that are important for the user, also referred to as *significant criteria* in this paper. Thus, the overall rating is viewed as dependent on ratings of significant criteria.

The research issue arising out of this view is how to determine which criteria are significant. We can ask the user to specify the criteria they pay more attention to, or to give the weights of criteria, or to make pair-wise comparisons of criteria, and then use analytic hierarchy process (AHP) to decide the weights of criteria, such as [20]. These techniques are aligned with multi-criteria decision making in that they ask for additional information from users. This makes the level of user involvement too high for a RS, changing the balance between costs and benefits through a high level of intrusiveness[21].

Instead of asking the users to specify their significant criteria, an aggregation function can be used to estimate overall ratings of the items a user has rated by the underlying criteria ratings of those items. The function is expressed as $r_0 = f(r_1, r_2, \dots, r_k)$ [3]. And from this aggregation function, we can obtain the significant criteria for this user, and also their importance in predicting the total rating. In our paper, the statistical technique "linear least squares regression" is applied as the aggregation function to estimate overall ratings for each user. For user u, the overall ratings r_0 of the items he has rated before are estimated by the criteria ratings: $r_0 = \sum_{c=1}^{k} w_c r_c + \varepsilon$, where w_i presents the weight of criterion rating r_c, and ε is the error term that accounts for the variability in r_0 that cannot be explained by the linear relationship between independent variables and dependent variable. If $r_1, \dots, r_k$ and r_0 are linearly related, then we must have $w_c \neq 0$. T-test is applied to test whether we can conclude that $w_c \neq 0$. Only those criteria which pass the significance level (e.g. $p - value < 0.1$) are recorded as the significant criteria for a user.

3.2 Clustering Users into a Preference Lattice

After obtaining the significant criteria for each user, we can cluster users based on their significant criteria. The users who share similar significant criteria preferences are very likely to have similar purchasing behaviors. Following the example above, User1 prefers location, and he will be positioned within *Location* cluster, while User2 will be placed in *Cleanliness* cluster. However, following this cluster, there will be another cluster with two or more than two significant criteria, such as *Value, Rooms* and *Location*, or *Cleanliness* and *Service*, et al. Thus we use a cross table to depict user criteria preferences in order to gain better understanding, an example is shown in Table 1.

Table 1. User Criteria Preference Cross-table

	V	R	L	C	S
User1			×		
User2				×	
User3	×	×	×		
User4				×	×

According to the classical clustering, data points partitioned into one cluster are similar to each other, but dissimilar with respect to the data points in other clusters. In our case, the users within *Location* cluster certainly share much similarity with the users from *Value, Rooms* and *Location* cluster. In fact, our clusters are partially ordered in a lattice [22], which we call a "preference lattice".

A partial order set means "p is related to q under R", where R is a binary relation that contains all the pairs of points related to each other under R [23]. And p and q are comparable iff $p \leq q$ or $q \leq p$.

In this paper we say that p and q are *close* to each other only when they are comparable [23]. And p is *closer* to q than g when $p < q < g$ or $g < q < p$.

In our example, the starting point is the cluster where all five criteria are significant. Each user belongs to one and only one cluster. And the maximum number of clusters can be determined from the number of the criteria used. Figure 1 shows the structure of our lattice drawn as Hasse diagram [24]. For simplicity, we only present part of all the connections in the diagram. There are users who have the same criteria preference within the lattice. The users under the same cluster share more similarity with each other than with those from the other clusters. In this paper, the users from *close* clusters (User4 and User2, for example) are more similar than users from incomparable clusters (User1 and User2). From the diagram, we can see that a cluster is *closer* to the clusters in higher or lower level than those adjacent clusters to the left or to the right. In our example, User3 is in a *close* cluster to User1, and so is User2 to User4. However, User1 and User2 are not close since the clusters they belong to are not comparable in the partial order.

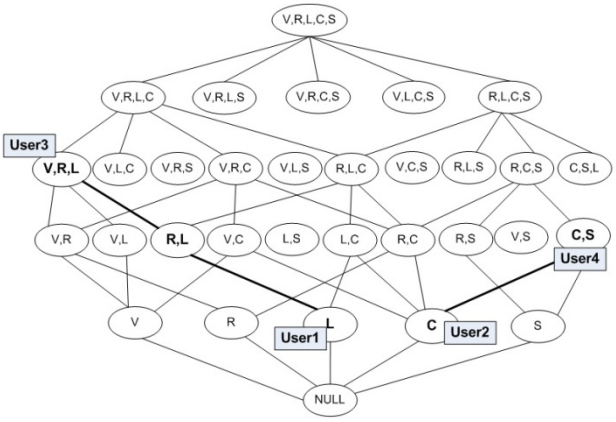

Figure 1. Criteria Preference Lattice drawn as Hasse diagram [24].

In prediction, the users from the same cluster will be given the higher priority. When there is a shortage of data points within one cluster, then we consider the users from the closest clusters in the higher and lower levels, until we gather sufficient data points. Following this innovative approach to selecting similar users, we can alleviate the sparsity problem whilst controlling the degree of recommendation accuracy.

3.3 Prediction based on Clustering

To instantiate the general approach of clustering proposed above, we have developed three different methods for prediction based on the clusters we have made. In other words, we predict the ranking of a service for a user based on past feedback about this service from users within the same cluster, or close clusters. And we use the other users' past feedback in three different ways, described in the three sub-sections below.

3.3.1 Aggregation Function of Criteria

This method assumes that the total rating has certain relationship with multiple criteria ratings, and this total rating can be generated by some aggregation function of criteria ratings. Multi-attribute Utility Theory (MAUT) is one broadly used method for a decision problem with multiple variables. Among the methods, linear function is the simplest and most popular from additive value function [7]. The total rating is predicted as below:

$$r_{u,i} = \sum_{c=1}^{k} w_c (\sum_{n=1}^{N} r_{u',c}/N) + \varepsilon \quad (5)$$

Where $r_{u,i}$ is the total rating predicted for user u on item i, w_c is the coefficient of criterion c for user u, while $r_{u',c}$ is the rating of criterion c that user u' gave to item i. N is the number of users involved in prediction, and ε is user u's error term. We have to emphasize that users u' and u are from the same or a close cluster

The aggregation function above can have two different scopes: (a) using all the criteria for the prediction with their coefficients (referred as *CluAllCriteria*), and (b) using only the significant criteria and recalculated coefficients (referred as *CluSigCriteria*). The coefficients are recalculated by applying additive value function with only the significant criteria. The former implies that although a user may have significant criteria affecting his general impression on the item, the other criteria still take some account in the final prediction, while the latter implies that we only consider the criteria with higher priority, regardless of other criteria ratings. We explore and evaluate both methods in our experiments.

3.3.2 Aggregating from Total Rating

This method is different from the method above since it focuses on the overall ratings only once the users are clustered. It assumes that the general impression of an item for one user is very similar to the general impression of the other users from the same or a close cluster. For example, the overall rating of a hotel rated by User1 (he prefers location criteria as explained in above example) is more similar to the users who prefer location, or who prefer location and some other criteria, compared to users who prefer some other criteria. The satisfaction (or dissatisfaction) with the overall item (service) which is gained by an increase (or decrease) of perception of location criterion is more than the change due to other criteria.

The overall rating can then be calculated as a simple average of the overall ratings of this item given by other users from the same or close cluster, as shown in eq. (6) below. This method will be futher referred to as *CluOverall*.

$$r_{u,i} = \frac{\sum_{n=1}^{N} r_{u',i}}{N} \quad (6)$$

Where $r_{u',i}$ is the total rating rated by user u' on item i. User u' is from the same or close clusters as user u. The close cluster is defined in section 3.2.

3.3.3 Combining Clustering and Collaborative Filtering

The third method recognizes that not all users within a single cluster will be equally generous, even if they value the same criteria. To take this into account, the third method applies extended CF within each cluster (or a set of close clusters). Thus, when we predict a rating of an item for a user, the users who have similar taste are given a higher priority than the other users, even though they are all within the same cluster. For our case, the memory-based algorithm (*cf* Section 2.1) is applied, where prediction is computed by aggregating the ratings of other users for the same item [10]:

$$\mathrm{r_{u,i}} = \bar{\mathrm{r_u}} + \mathrm{k} \sum_{\mathrm{u'} \in \mathrm{U}} \mathrm{sim'(u, u')} \times (\mathrm{r'_{u,i}} - \bar{\mathrm{r'_u}}) \quad (7)$$

Where $\mathrm{sim'(u, u')}$ is the similarity of user u and u' based on their multiple criteria ratings; U denotes the set of user u and the similar users u' from the same or close cluster who have rated the same item i before; $r_{u,i}$ presents the overall rating of user u on item i; $\bar{r_u}$ is defined as the average overall ratings of user u; and $k = 1/\sum_{u' \in U} sim'(u, u')$. And all the "neighborhoods" u' are

from similar clusters. This method uses criteria to cluster users, then overall ratings of items are used for prediction.

4. EXPERIMENTS & ANALYSIS

To evaluate the accuracy of the three methods proposed above, we conduct a set of experiments where we compare the proposed methods with each other, and with the two extensions of CF approach which we described in Section 2.2. Our experiments were implemented using MatLab 2009b. All the experiments were based on a PC with Windows XP Professional, with Intel Pentium(R) Core(TM)2 CPU, 2.13GHz and 2GB RAM.

4.1 Dataset

A feedback and use ratings data for an example service domain (hotel stays) is collected from a well-known rankings site (www.tripadvisor.co.uk). The data is suitable since a large number of feedback rankings are available from a variety of sources, and our particular set had ratings of five individual criteria, plus one overall rating. The five criteria are from *Value*, *Rooms*, *Location*, *Cleanliness* and *Service*. All the individual criteria ratings and the overall ratings ranged from 1 to 5, with 5 as the excellent.

Originally, there are 578,346 records in our dataset, with 69,269 users and 134,899 hotels. Since the raw data is too sparse to compute, we aggregate ratings of all the hotels from the same brand with the same number of stars, using the assumed equality between the levels of service and the uniformity of furnishing standards of such hotels. This is justified by the hotel franchising or branding agreements, which impose equivalent standards to levels of services, cleanliness, and furniture in all hotels belonging to the same brand. As we run regression for each user, we leave only the users who have rated more than 20 hotels. And then we filter out the hotels which only have one user's ratings. After cleaning, there are 24,264 records left, with 11,789 hotels and 1,054 users. And the sparsity level of data set is around 0.9980 ($sparsitylevel = 1 - totalrecords/(userNum * hotelNum)$). We select 90% of the data as the training set. After training the methods, we then evaluate them with the rest 10% data points (test set). Figure 2 plots the number of ratings of each criterion in our dataset. It shows the distribution of hotel ratings in our dataset.

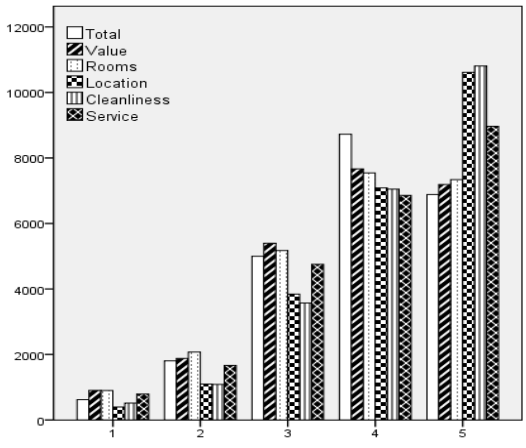

Figure 2. Distribution of Hotel Rating in Dataset

4.2 Evaluation Metric

Two metrics of evaluating the accuracy of the approaches are used in this paper. Mean Absolute Error (MAE) is one of most commonly used statistical accuracy metric [17]. It is a measure of the average absolute deviation between a predicted rating and the user's true rating.

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N} \quad (8)$$

N is the number of pairs of real ratings and predictions $< p_i, q_i >$. The lower the MAE, the more accurate predictions are [25].

The other evaluation metric we are using is ROC curve (Receiver Operating Characteristic), which emphasize the proportion of items that are not preferred that end up being recommended [26]. ROC curve measures the proportion of preferred items that are recommended (true positive), and also the proportion of the items that are recommended but actually are negatively preferred (false positive) [26]. It plots True Positive rate versus False Positive in the same graph and depicts the relative trade-offs between them [27]. The area under ROC curve (AUC) presents the accuracy evaluation measure. It ranges from 0 to 1, where 1 indicates a perfect classifier, and 0.5 presents the performance of a random classifier. Thus, if AUC is less than 0.5, then the tested approach or model has "no discriminating power" [27].

4.3 Single-rated versus Multi-Criteria Recommendation

It has been proved that multi-criteria recommendation outperforms the corresponding single-rating CF approach in the general case[3, 21]. However, even with multi-criteria ratings, different methods provide various levels of accuracy, and one method may fit a dataset better than the other one. To profile the methods known in the literature against our dataset, we apply the single rating CF approach with Pearson Correlation for user similarity, and two extended multi-criteria ratings CF approaches to our dataset. One is AvgSimCF, and the other one is EuclideanCF, as described in Section 2.2. The results are shown in Table 2.

Table 2. Single rating and multi-criteria ratings based recommendation

	Single rating	Multi-Criteria ratings	
Prediction Method	CF	AvgSimCF	EuclideanCF
MAE	0.8940	0.8857	0.8312

The results in Table 2 demonstrate that the extended multi-criteria ratings approaches generally provide a more accurate prediction than a single rating approach. In our dataset, the user similarity calculated through Euclidean Distance improves around 5% accuracy over the user similarity calculated through aggregating individual criteria similarities. The limited scale of improvement is due to the fact that our dataset is quite sparse, so the number of co-rated items by two users is limited. In addition to this, the aggregation of similarities (eq. (2)) allows individual criteria similarities to cancel each other out. As an extreme example, the final similarity can be reduced to 0, for example, where the similarity based on overall rating is 1; the similarity based on value criterion is minus 1; the similarity based on the "rooms" criterion is 1, and so on. Then the final similarity calculated by averaging all the criteria similarities equals to 0 and so the user will not be considered in the final prediction. However, this case

happens infrequently, in cases where there are only two co-rated items for two users, but one of the items is selected as a test set.

Given the advantages of the prediction with Euclidean distance based user similarity to the method using the average of individual criteria similarities, we apply Euclidean Distance extended multi-criteria ratings CF approach in our further experiments.

4.4 Performance of the Three Methods Based on Clustering

We run the experiments based on the processes we described above. First, we use statistical technique to find out the criteria that are preferred for each user. Multiple linear regression using the "least squares" method is applied to each user, and only the criteria that pass the significance tests are recorded as the significant criteria for that user. The significance level is set as 0.1 in our experiments. And then we use preference lattice to cluster users based on their criteria preferences. After grouping users, then we recommend services to each user from our test set through different methods by using past feedback by the users from similar groups. The results using the MAE metric are shown in Table 3.

Table 3. Prediction by different methods within clusters

	Clustering			
Prediction Basis	CluAll Criteria	CluSig Criteria	CluOverall	CluCFEuc
MAE	2.5452	1.8710	0.7176	0.6745

According to the experiments results, the combination of clustering and extended CF with Euclidean Distance provides the best prediction. The accuracy of this method is much better than the other methods within clusters we described, and also there is around 15% accuracy improvement from Euclidean Distance extended CF approach (shown in Table 2). The result is not surprising based on our assumption. Users tend to give a similar trend when they share the same of similar criteria preference. However, there are users with different requirement levels reflecting criteria ratings within clusters. And the users who share similar taste are given a higher priority in term of prediction.

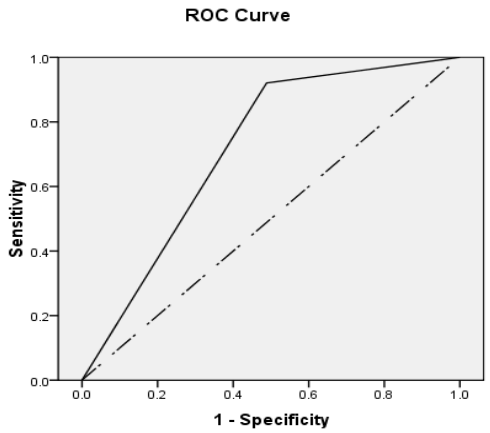

Figure 3. ROC Curve of *CluCFEuc* method.

The results of calculating the ROC Curve of the *CluCFEuc* method are shown in Figure 3. We determine that the item is relevant (successful recommendation) when its rating is more than or equal to 3. Otherwise, it is non-relevant (unsuccessful recommendation). Then we plot ROC curve of the approach that combines our clustering and extended CF. The dash line is the reference line, under which the area is less than 0.5, which presents the case of randomly guessing a class [27]. Our AUC indicates a reasonable value, which is 0.716 for 50 cut off points.

Finally, Figure 4 compares the MAE results from five methods: the SingleRating method, the AvgSimCF method, the Euclidean CF method, and two of our clustering methods: CluTotal and CluCFEue. The best prediction is provided by the combination of Clustering and Collaborative Filtering, the third method which we have developed to instantiate our clustering-based approach to multiple-criteria prediction.

The second most accurate prediction comes from another method based on our clustering approach, which uses the simple average of overall ratings of the other users on the same item as shown in Fig. 4. Even with its simplicity, its accuracy improves around 12% to the multi-criteria CF approach. These results show how effective is our clustering approach proposed here.

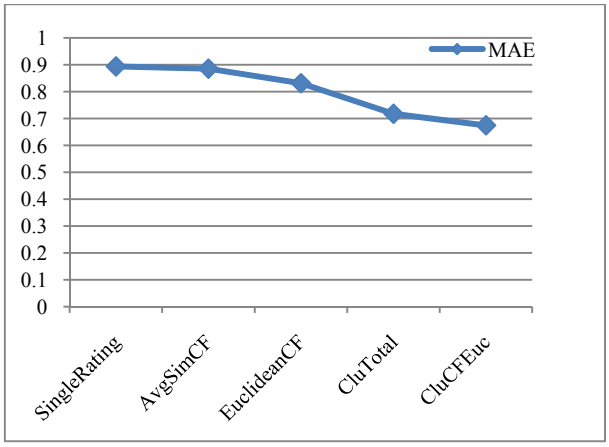

Figure 4. MAEs of different approaches.

However, the other two results from Table 3, the predictions based on clustering and additive value function are not ideal. We are currently investigating the cause for these high MAE values, our current thinking is that the problem is with the automatic coefficients we obtained by linear regression. Using these causes many predictions outside of the [1..5] range, even some with negative values. As stated in [28], regression can indeed be used for predicting a dependent variable. However, it is more often used to discover the way in which independent variables are related to a dependent variable, and to see the relative contribution of each of these variables [28]. In our experiment, when we control different numbers of user's historical records in regression, the coefficients of criteria change every time, but the significant criteria keep the same.

Even within these high numbers, we still see confirmation of the importance of the significant criteria for effective prediction, with the MAE of the significant criteria method being significantly lower than the MAE of the method which uses all criteria.

5. RELATED WORK

Multi-criteria recommender systems are gaining widespread attention from both research and industry. One of the earliest papers in this area is DIVA, which is initialized by a collaborative filtering database. For recommendation, it computes the distance between active user and the cases in the database and provides a ranking for users [29]. Manouselis and Costopoulou describe and

classify multi-criteria recommender systems on the basis of analyzing MCDA methods and recommendation process [4]. Lee and Teng [6] use collaborative filtering approach to compute the similarity of each of them separately. Subsequently, they utilized skyline queries method to identify a few good items among numerous candidates. Adomavicius and Kwon [3] believe new techniques for recommendation are needed to take full advantages of multi-criteria ratings, and they describe two approaches: a similarity-based approach and an aggregation-function-based approach. Some researchers use UTASTAR algorithm to incorporate MCDA techniques to recommendation process to provide a ranking for recommendation [27],[30].

Clustering technique has been applied in recommender system for identifying trends and reducing noise. Shepitsen *et al* [31] apply hierarchical agglomerative clustering in social tagging system. They associate users, resource, tags together. And users are clustered based on the similarity of tags. Cantador and Castells [32] present user interests with the ontology of domain concepts. Their system then takes the advantages of the relationship between concepts and cluster semantic space base do on the correlation of concepts, and users can be partitioned by projecting the concept clusters into the set of preferences of each user.

There are some works using additive value function for estimating overall rating. For example, Manouselis and Costopoulou calculate total utility by aggregating the partial utilities of all the criteria, or by weighting the predicted ratings that a user would give on each criterion [7]. Adomavicius and Kwon apply linear regression and estimate coefficients for predicting overall rating in experiments as well [3]. However, the purpose we are using regression is to find the significant criteria in order to cluster users. According to our experiments, in a very sparse dataset, especially there are much more items than users and the users have rated many items before, we would suggest not to use linear regression for the prediction, but instead, to cluster users and then apply an extended CF approach.

6. CONCLUSION & FUTURE WORK

In this paper, we are motivated by the assumption that only some of the given selection criteria dominate the user's decision, and propose a clustering method which positions users in a preference lattice dependent on which criteria they prefer. Based on the clustering, we then present three different methods for predictions which to instantiate the overall clustering approach. The three methods are:

(a) Predicting for overall rating based on the aggregation function of criteria ratings. Within this method, two directions are presented, by using all the criteria as the independent variables, or by using only significant criteria for the users to aggregate overall rating;

(b) Predicting overall rating based on the overall ratings of the same item from other users within the same or close clusters;

(c) Predicting by using extended CF approach within clusters.

We evaluate the methods with a real world service data set, and the results show that the prediction based on multiple criteria is more accurate than the prediction based on one single rating. What is more, in our dataset the extended CF, EuclideanCF method is more accurate than the other one called AvgSimCF, by 1 to 7%. By using the method we proposed, combining our clustering method with extended CF, we achieve some 12% to 17% improvement in prediction compared to using AvgSimCF, and around 12% to 15% outperforming the prediction under EuclideanCF method. These results prove our assumption that users who have the same or very similar criteria preferences share similarity with each other. Also the results indicate the effectiveness of our clustering approach.

We use linear regression function to discover the significant criteria for each user. It assumes that the overall rating is related to criteria ratings in a linear fashion, and this should be tested in our further work. This also limits the dataset to those users who have rated many items, so for our e future work, we intend to explore the performance of alternative techniques, such as non-linear regression for discovering significant criteria, which also work for users with fewer items.

7. REFERENCES

[1] Burke, R., Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction, 2002. Volume 12(Number 4): p. 331-370.

[2] Schafer, J.B., J.A. Konstan, and J. Riedl. Recommender Systems in E-Commerce. in Proceedings of the 1st ACM conference on Electronic commerce 1999.

[3] Adomavicius, G. and Y. Kwon, New Recommendation Techniques for Multicriteria Rating Systems, in IEEE Intelligent Systems. 2007. p. 48-55.

[4] Manouselis, N. and C. Costopoulou, Analysis and Classification of Multi-Criteria Recommender Systems World Wide Web, 2007. 10: p. 415-441.

[5] Sampson, S.E. and C.M. Froehle, Foundations and implications of a proposed unified services theory. Production and Operations Management, 2006. 15(2): p. 329-343.

[6] Lee, H.-H. and W.-G. Teng, Incorporating Multi-Criteria Ratings in Recommendation Systems, in IEEE International Conference on Information Reuse and Integration, 2007. 2007. p. 273-278.

[7] Manouselis, N. and C. Costopoulou, Experimental Analysis of Design Choices in Multimattribute Utility Collaborative Filtering. International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI), 2007. 21(2): p. 311-331.

[8] Karta, K., An Investigation on Personalized Collaborative Filtering for Web Service Selection. 2005.

[9] Adomavicius, G., et al., Incorporating contextual information in recommender systems using a multidimensional approach. ACM Transactions on Information Systems (TOIS), 2005. 23 (1): p. 103 - 145

[10] Adomavicius, G. and A. Tuzhilin, Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE transactions on knowledge and data engineering, 2005. 17(6).

[11] Balabanović, M. and Y. Shoham, Fab: content-based, collaborative recommendation, in Communications of the ACM. 1997. p. 66 - 72

[12] Leimstoll, U. and H. Stormer. collaborative recommender systems for online shops. in Proceedings or the 13th Americas Conference on Information Systems. 2007.

[13] Anand, S.S. and B. Mobasher, Intelligent Techniques for Web Personalization in Intelligent Techniques for Web Personalization 2005, Springer Berlin / Heidelberg. p. 1-36.

[14] Resnick, P., et al. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. in Proceedings of ACM

1994 Conference on Computer Supported Cooperative Work. 1994.

[15] Shardanand, U. and P. Maes. Social information filtering: algorithms for automating “word of mouth”. in Proceedings of the SIGCHI conference on Human factors in computing systems. 1995.

[16] Breese, J.S., 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.

[17] Lakiotaki, K., N.F. Matsatsinis, and A. Tsoukiàs, Multi-Criteria User Modeling in Recommender Systems. IEEE Intelligent Systems, 2011. 26(2): p. 64-76.

[18] Anderson, R.E., Consumer Dissatisfaction: The Effect of Disconfirmed Expectancy on Perceived Product Performance. Journal of Marketing Research,, 1973. 10(1): p. 38-44.

[19] Zhang, Y., et al., Applying probabilistic latent semantic analysis to multi-criteria recommender system. AI Communications, 2009. 22(2): p. 97-107.

[20] Liu, D.-R. and Y.-Y. Shih, Integrating AHP and data mining for product recommendation based on customer lifetime value. Information & Management, 2005. 42: p. 387-400.

[21] Adomavicius, G., N. Manouselis, and Y. Kwon, Multi-Criteria Recommender Systems, in RECOMMENDER SYSTEMS HANDBOOK. 2011, Springer. p. 769-803.

[22] Deshpandé, J.V. On Continuity of a Partial Order. in Proceedings of the American Mathematical Society. 1968.

[23] Schröder, B.S.W., Ordered Sets: An Introduction. 2003: Boston: Birkhäuser.

[24] Davey, B.A. and H.A. Priestley, An Introduction to Lattices and Order. 2nd ed. 2002: Cambridge University Press.

[25] Sarwar, B., et al. Item-based collaborative filtering recommendation algorithms. 2001. Proceedings of the 10th international conference on World Wide Web.

[26] Shani, G. and A. Gunawardana, Evaluating Recommendation Systems, in RECOMMENDER SYSTEMS HANDBOOK. 2011. p. 257-297.

[27] Lakiotaki, K., S. Tsafarakis, and N. Matsatsinis, UTA-Rec: A Recommender System based on Multiple Criteria Analysis, in The 2nd ACM conference of Recommender Systems. 2008. p. 219-225.

[28] Dancey, C. and J. Reidy, Statistics without Maths for Psychology: Using SPSS for Windows. 4th ed. 2008: Prentice Hall.

[29] Nguyen, H. and P. Haddawy. DIVA: Applying Decision Theory to Collaborative Filtering. in Proceedings of the Conference on Artificial Intelligence for Electric Commerce. 1999.

[30] Matsatsinis, N.F., K. Lakiotaki, and P. Delia. A system based on multiple criteria analysis for scientific paper recommendation. in Proceedings of the 11th Panhellenic Conference on Informatics. 2007.

[31] Shepitsen, A., et al. Personalized recommendation in social tagging systems using hierarchical clustering. in Proceedings of the 2008 ACM conference on Recommender systems. 2008.

[32] Cantador, I. and P. Castells, Multilayered Semantic Social Network Modeling by Ontology-Based User Profiles Clustering: Application to Collaborative Filtering, in Lecture Notes in Computer Science, S. Staab and V. Svatek, Editors. 2006. p. 334-349.