Temporal Rating Habits: A Valuable Tool for Rating Discrimination

Pedro G. Campos1,2  
fernando.diez@uam.es  
1Universidad Autónoma de Madrid  
Francisco Tomás y Valiente 11  
28049, Madrid, Spain

Alejandro Bellogín1  
alejandro.bellogin@uam.es  
2Universidad del Bio-Bio  
Av. Collao 1202  
4081112, Concepción, Chile

ABSTRACT
In this paper, we describe the experiments conducted by the Information Retrieval Group at the Universidad Autónoma de Madrid (Spain) to tackle the Identifying Ratings (track 2) task of the CAMRa 2011 Challenge. The experiments performed include time-frequency probabilistic strategies, heuristic collaborative filtering (CF) and a model-based CF approach. Results show that probabilistic classifiers based on temporal behavior of users have better performance than traditional recommendation-based strategies, thus reflecting that temporal information is a valuable source for the identification or discrimination of user ratings.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Information Filtering, Retrieval Models, Selection Process; I.5.1 [Pattern recognition]: Models

General Terms
Algorithms, Performance, Experimentation

Keywords
Context-Aware Recommender Systems, Movie Recommendation, Probability Models

1. INTRODUCTION
Information about context can help improving personalization-related tasks [1]. The Challenge on Context-aware Movie Recommendation 2011 (CAMRa2011) provides an interesting opportunity to test recommendation approaches on real data. We focus on the Identifying Ratings track, which consists of determining which members of a household made some specific “unidentified” ratings, once information about which users belong to each household is known, along with the movie ratings assigned by the households. In this case, there are two dimensions of contextual information. On the one hand, household information, which may allow us to take advantage of knowing the existence of a relationship among some users (although the actual relation remains unknown), and on the other hand, temporal data, since each rating has an associated timestamp, which allows to track users’ concept drift. Nonetheless, other interesting information which has been used previously in different recommendation strategies such as movies features (title, genre), user demographics and other social relationships, are not available in this challenge. This fact makes more difficult to define relations between each item in question and the users to be allocated.

Considering the above issues mentioned, we conducted a series of experiments with different models, in order to better predict to whom each “unidentified” rating belongs to, which we describe in this work. The remainder of this paper is structured as follows. Section 2 describes the main characteristics of the available data for the competition. Section 3 presents a brief review of models that could be used, considering information available in the dataset. Section 4 details the models used for making the predictions. Section 5 presents the results obtained, along with the evaluation methodology followed and required by the challenge. We finalize with some concluding remarks and devised additional approaches to experiment in Section 6.

2. DATASET ANALYSIS

2.1 General Description
CAMRa 2011’s MoviePilot Dataset consists of a training set of 4,536,891 timestamped ratings from 171,670 users on 23,974 items on a timespan from July 11, 2009 up to July 12, 2010, and two test sets (one for each competition track): track 1 containing 4482 ratings from 594 users on 811 items on a timespan from July 15, 2009 up to July 10, 2010 and track 2 containing 5450 timestamped ratings from 592 users on 1706 items on a timespan from July 13, 2009 up to July 11, 2010. Since we focus only on track 2, from now on we only analyze data related with that track.

Figure 1 shows the rating, community and catalog growth of training data (upper side) and testing data for the track 2 (lower side) through time. It may be seen that data growth follows a similar proportion on both data splits. It is also available, for some users, information about which household a user belongs to. Table 1 shows the size distribution of households in the dataset. 2-sized households represent
the 93.8% of all the households, whilst 3-sized and 4-sized households represent the 4.8% and 1.4% respectively. Note that, although there are 602 users which are members of some household (they appear in the training set), only 592 users have data in the test set for this track.

### 2.2 Frequency-based Analysis

Taking into account that we do not know whether the household’s relationships correspond to friends, siblings, couples, etc., and that no other information is provided, we focused our analysis on temporal trends which may help us on completing the task at hand. We performed a descriptive study of the given characteristics on the training data and we observed a phenomenon repeated in several of the users belonging to different households. In Figure 2, it is shown the rating hour probability mass functions (PMFs from now on) of two users in the first household. We can observe here that there is a clear disparity between the hours employed by each of the household’s members for rating movies. The user u40426 has a probability near 1 (0.93) to rate movies in the period from 18:00 to 19:00. On the contrary, user u311738 rates movies starting at 20:00 and later on, that is, mostly by night. Similar circumstances to the one described above are repeated along the data set, suggesting time-aware strategies might be useful.

When analyzing the rating date from each user, it is also possible to detect some interesting facts. Figure 3 shows how many ratings are made by users through time. The left frame shows that the mean user rating window size (i.e. timespan at which users make ratings) is very small (just a few days). The center and right frames also show that the vast majority of ratings are incorporated during the first days of participation of a user. Considering that users start their participation on different days, this information can be helpful in our task. We also noted that there are differences on which day of the week each user rates movies, although, for the sake of clarity, we prefered to leave those figures out of this paper.

The analysis of the raw rating value frequency alone also
gives us some clues about user behaviors. Figure 4 shows an example of two PMFs of rating values, corresponding to two pair of users in different households. The one on the left emphasizes the fact that user u322924 (thick lined) rarely gives ratings higher than 90 points. On the contrary, user u880228 (dashed lined) usually gives ratings higher than 90 points. The example on the right has a stronger discrimination. The dashed user rates with less than 10 points most of the time. On the contrary, the thick lined user tends to rate over 60 points.

The analysis presented above suggests us to take into account the following dimensions in order to identify raters:

- **The hour of the day** in which a user rates movies more frequently (H).
- **The day of the week** in which a user rates movies more frequently (W).
- **The date of rating** (D).
- **The number of ratings** given by a user (R).

These findings motivated us to use probability-based models in order to classify users in a household depending upon the described dimensions. Previously to their application, we studied possible classifiers to be used, as well as more traditional recommendation models which could serve as baselines.

### 3. RELATED WORK

#### 3.1 Classification Models

With respect to classification, there are many different paths to explore. Inductive learning can be defined as those methods trying to induce a general rule from a set of observed instances. Frequently used and well known are, among others, the information theoretic-based ones (e.g. Rocchio classifier [19]), decision trees (like the C4.5 algorithm [18]), the evolved algorithm of the famous Quinlan’s ID3 [17]), instance-based methods (including Nearest Neighbor (kNN)-based models [5]), and probabilistic classifiers (e.g. Naïve Bayes or simple Bayesian classifier [6]). As we can observe, depending on the context under study it is possible to choose between many different models to classify items. One key circumstance in the decision about which model to use is the existence or not of prior knowledge. Prior knowledge plays a major role. Usually the way to classify is highly dependent on the existence of examples (used as training data), and the features of the data under study. When prior knowledge of the patterns to be classified is available, existing methods differentiate between supervised and unsupervised learning. In the first case the training data are provided with the classes to which the examples belong to. In the other case no prior classification exists, and the user will set hypothesis about the number of classes to be generated (as for example by means of clustering techniques [9]).

When dealing with simpler Bayesian classifiers, independence is also an important aspect to be considered. When features are independent, binary models usually leads to simplified linear classifiers. The more independent we can make the classifiers, the simpler the classifier can be. Naïve classifiers are those under the assumption that the attributes are independent given the class. In [6], it is shown that these simple Bayesian classifiers can be optimal under zero-one loss (i.e. the misclassification rate). There exist innovative proposals as the one described in [13], to extend the modeling flexibility of Naïve Bayes models by introducing latent variables to relax some of the independence assumptions in these models.

In the case of the MoviePilot dataset, there is a vast quantity of training data, all of it with proper class assignments. As showed in the brief study made above, there is a lot of information about the a priori behavior of the users being part of each of the households, by means of the characteristics known about them, that is, their features. Features should be as much discriminant as possible. However, in this case, there are not many possibilities to extract features useful for the analysis from the data. Some of the ones analyzed are directly extracted from the data itself (as for example the hour of the day (H)), but others are derived ones (e.g. the date of rating (D)).

Because of their usefulness, many disciplines use classification methods based on supervised learning. For example, Pang et al. [15] employed three different machine learning methods (Naïve Bayes, maximum entropy classification and SVM) to classify documents by overall sentiment, using movie reviews as data. In another context, Herrera et al. [11] used both kNN and Naïve Bayes methods for in-
Recommender Systems apply techniques from statistics and knowledge discovery to the problem of recommending items to users of a system [20]. The most common used approach is collaborative filtering (CF) [21], which tries to predict the utility of items for a target user based on items previously liked by the target and/or other users. The information about user interest is called the user profile. There are also content-based (CB) models [16], which search for items similar to other items that the user liked in the past, using descriptions of the items and the user profile. Thus, CB recommendation models require an explicit description of items. Due to limitations from both approaches, there are also hybrid approaches [4], which combine elements from both CF and CB. Due to the fact that the Moviepilot dataset do not contain descriptions about items, the only choice herein is CF.

CF recommendation algorithms can be further classified into heuristics and model-based ones [3]. The former make use of the entire collection of user profiles to compute predictions (e.g., in kNN-based recommendation models [10]), whilst the latter learn a model, which is then used to compute predictions (e.g., in Matrix Factorization-based recommendation models [12]).

4. PREDICTIVE MODELS

This section describes the models used for the challenge. We begin with the probabilistic models which turned out to give the best performance results. Then, we describe other more traditional (ad-hoc) recommendation models which were used to compare our results.

4.1 Probability-based Models

The findings observed from the dataset analysis motivated us to use probability-based models to classify which users were the ones who evaluated each movie as required by the challenge. We used a discriminant function based on the PMFs obtained, giving more probability to users depending on the probabilities of the previously mentioned dimensions of information, namely time and number of ratings. We describe below our two approaches.

4.1.1 A priori Model

Let us consider a set of objects $O = \{o_1, o_2, \ldots, o_m\}$ and a set of classes $\Omega = \{\omega_1, \omega_2, \ldots, \omega_r\}$, such that each object $o_i$ is member of one, and only one, class $\omega_j$. In addition, consider that these objects are described by means of the value of some numerical quantity feature, called $X$. Now, the question we want to answer herein is whether it is possible to determine which class an object $o_i$ belongs to or not, once the value $x_i$ of its feature $X$ is already known. If we assume that we know the a priori probabilities of the respective classes, a simple classification rule can be:

$$\text{Assign } o_i \text{ to } \omega_j^o = \arg \max_{\omega_j \in \Omega} P(X = x_i | \omega_j)$$

Bringing this model to our case, let $U_h$ be the set of users from household $h$, and let $R_h = \{r_1, r_2, \ldots, r_m\}$ be the set of unidentified ratings from $h$, that is, ratings that are known to be given by a user $u_j$ from $U_h$, but not knowing which particular user $u_j$ gave it. We define, based on the a priori PMFs of feature $X$, $P(X | u_j)$ (where $X$ can be any of the information dimensions described in Section 2.2):

$$\text{score}(r_i, u_j) = P(X = x_i | u_j)$$

Once the scores given to each pair $(r_i, u_j)$ are determined, the a priori-based discriminant function assigns the rating $r_i$ to the user that reached the highest probability. That is:

$$\text{Assign } r_i \text{ to } u_j^* = \arg \max_{u_j \in U_h} P(X = x_i | u_j)$$

4.1.2 Naïve Bayes Model

Now, considering we know the PMFs of feature $X$ and each class $\omega_j$, i.e., $P(X)$ and $P(\omega_j)$, by means of applying the Bayes’ theorem, we compute the corresponding probabilities of each class provided the feature $X$:

$$P(\omega_j | X = x_i) = \frac{P(X = x_i | \omega_j) P(\omega_j)}{P(X = x_i)}$$

Then, the previous classification rule is rewritten as:

$$\text{Assign } o_i \text{ to } \omega_j^* = \arg \max_{\omega_j \in \Omega} P(\omega_j | X = x_i)$$
4.2 Rating-based Models

A different discriminant can be build if, instead of inferring which user is more likely to rate a particular item, we know the actual rating value of \( \hat{r}_{i,j} \) and computing the combined probability of features \( X_k \). Therefore, in our case we compute again the previously defined scores as:

\[
\text{score}(\hat{r}_{i,j}, x_i) = P(x_i | X = x_i)
\]

Then, we can apply the same decision rule as defined in the previous model (1).

These models can be easily extended to consider a set of features \( X = \{X_1, X_2, ..., X_n\} \) describing each object \( o_i \) by computing the combined probability \( P(X_1 = x_{1i}, X_2 = x_{2i}, ..., X_n = x_{ni} | \omega_j) \). Using the conditional independence (a.k.a. naïve) assumption that each feature \( X_k \) is conditionally independent of every other feature \( X_l \) for \( k \neq l \), we can compute the combined probability as \( \prod_{k=1}^{n} P(X_k = x_{ki} | \omega_j) \) [23].

4.2.2 Model-based CF

In the case of model-based CF, we selected a Matrix Factorization (MF) model baseline. It is an adaptation of the Singular Value Decomposition approach that is gaining increasing interest in the field of Recommender Systems due to its good performance [12]. In this technique, the known rating values, represented as a rating matrix \( R \), are iteratively approximated by user and item factor matrices \( P \) and \( Q \) of lower dimension (\( f \) in our notation) such that:

\[
\hat{r}_{u,i} = \sum_{j=0}^{f} P_{u,j} \cdot Q_{j,i} = p_u^T q_i
\]

One advantage of this approach is that \( P \) and \( Q \) values may be computed for all users and items using only the known values \( R \), minimizing an estimation of the difference, e.g. the Frobenius Norm: \( \min ||R - PQ||^2 \). Overfitting can be alleviated using regularization, i.e. penalizing the magnitude of the approximated vectors [12]. The common regularized formulation for collaborative filtering is inspired in minimizing the squared error on the set of ratings:

\[
\min_{p_u, q_i, \gamma} \sum_{u,i \in R} \left( r_{u,i} - p_u^T q_i \right)^2 + \lambda \left( ||p_u||^2 + ||q_i||^2 \right)
\]

Different algorithms exist to compute this kind of factorization. A widely used implementation of stochastic gradient descent was published by Simon Funk\footnote{http://sifter.org/~simon/journal/20061211.html} in the context of the Netflix Prize. In this implementation, for each known rating, the parameters are optimized by updating them in the opposite direction of the gradient of the optimization criterion, using a learning rate parameter \( \gamma \) which controls the amount of update [12, 22].

\[
p_u' \leftarrow p_u - \gamma \cdot \frac{\partial J}{\partial p_u}
q_i' \leftarrow q_i - \gamma \cdot \frac{\partial J}{\partial q_i}
\]

5. EXPERIMENTS

5.1 Implementation details

Table 2 shows the parameter values used in the implementation of the rating-based models described in Section 4.2. Note that we used an item-based kNN algorithm. In the case of the probabilistic-based models, we ran out several trials combining the different features previously defined. In the next section, we show the best results obtained with all the described algorithms.

5.2 Evaluation Methodology

We have tested the accuracy of the models in terms of the classification error rate by household, i.e. the number of correct predictions divided by the number of predictions, averaged by household, in agreement with the rules of the challenge. Let \( HH \) denote the full set of households in the challenge data, and \( f(\cdot) \) the model under evaluation. The classification error rate can be expressed by:
Table 2: Parameter values

<table>
<thead>
<tr>
<th>Model</th>
<th>Param.</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>$f$</td>
<td>10 factors</td>
</tr>
<tr>
<td></td>
<td>$\lambda$</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>$\gamma$</td>
<td>0.02</td>
</tr>
</tbody>
</table>

$$E_{HH} = \frac{1}{|HH|} \sum_{h=1}^{|HH|} \frac{1}{|HH_h|} \sum_{(o_i, \omega) \in HH_h} L(\omega, f(o_i))$$

where

$$L(\omega, \hat{\omega}) = \begin{cases} 1 & \text{if } \omega = \hat{\omega} \\ 0 & \text{otherwise} \end{cases}$$

Now, let $HH^k$ denote the set of households of size $k$ ($HH = \bigcup_k HH^k$). Then, the classification error rate per household size is:

$$E_{HH^k} = \frac{1}{|HH^k|} \sum_{h=1}^{|HH^k|} \frac{1}{|HH_h^k|} \sum_{(o_i, \omega) \in HH_h^k} L(\omega, f(o_i))$$

We computed an additional set of metrics based on precision, such as Precision at level 5 (P@5) and 10 (P@10), Mean Average Precision (MAP) and Area Under the Curve (AUC), computed on each user’s recommendation list and averaged on all test users (not on a per-household basis). Note that a recommendation list in this context can be easily constructed for a target user by ordering the full set of items, putting in the top places those items believed to have been rated by the target user. That is, we sort the items classified as rated by the target user with respect to the score obtained by the probability-based models, or the predicted rating in the case of rating-based models, putting after them the remaining items in the dataset (except for those in the user profile) with a score/rating equal to 0.

Precision and MAP come from the Information Retrieval field [2], and are useful to measure how good a recommender is based on a given ranking. AUC is an additional metric commonly used in the Machine Learning community for measuring the classification error and how far a particular method is from a random classifier (which obtains an AUC of 0.5) [7]. For the four metrics included, the higher, the better, considering that their maximum value is 1.0.

### 5.3 Results

Table 3 shows results obtained with the tested models (bold indicates the best column value). Recall that R represents the number of ratings used as a feature for the probability-based models, H represents the hour of the day, W the day of the week, and D the date of rating; finally, a combination of those letters represents, obviously, a combination of those features.

It can be seen that the best performing algorithm is the a priori model when using the combination of hour of the day and date of rate features (HD). It is also interesting to note that, in general, a priori models have superior performance than Bayes models, independently of the features considered. A possible explanation for this is that the independence assumption is violated. Deeper analysis is required in order to verify if the independence assumption between features is acceptable or not.

All the results involving the H feature, considered alone or combined with other features, present a value up to 0.9 except for the case of Bayes (RH) within all the households (second column in Table 3). No other algorithm outperforms this value. This fact gives us a strong evidence of the importance of this feature. Among the three time-aware features studied (H, D and W), H is the one with higher discriminant capabilities for the task analyzed in this work.

It is also remarkable the poor performance of the number of ratings feature (R). It gives the lowest values for the metric considered, even lower than the baselines, in this case, the rating-based models.

Regarding the classical recommendation models, which are based on the extrapolation of rating values, both of them present poorer results than most of the probabilistic ones. The only probabilistic models that are comparable to them are the ones based on the rating value feature. This seems to remark that discriminating user ratings based only on rating values is hard, and in fact, other features (such as the temporal ones) are better suited for this task.

Table 4 shows the best results using an additional set of metrics, based on precision such as P@5, P@10, and MAP, and AUC (area under the curve). As it may be seen, results are consistent with classification accuracy rate outcome, regarding the best performing models, and besides, the obtained results are very high for the proposed probabilistic models.

6. CONCLUSIONS AND FUTURE WORK

This paper has described methods able to identify users that made particular ratings. We focused the analysis on the study of probability mass functions of the available features describing ratings, thus developing well-performing probabi-
Table 4: Additional metrics for the task

<table>
<thead>
<tr>
<th>Model</th>
<th>P@10</th>
<th>P@100</th>
<th>MAP</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A priori (HD)</td>
<td>0.9392</td>
<td>0.9375</td>
<td>0.9604</td>
<td>0.9803</td>
</tr>
<tr>
<td>Naive Bayes (HW)</td>
<td>0.9211</td>
<td>0.9214</td>
<td>0.9345</td>
<td>0.8675</td>
</tr>
<tr>
<td>kNN</td>
<td>0.6267</td>
<td>0.4541</td>
<td>0.5509</td>
<td>0.8217</td>
</tr>
<tr>
<td>MF</td>
<td>0.6098</td>
<td>0.4468</td>
<td>0.5482</td>
<td>0.8279</td>
</tr>
</tbody>
</table>

lity-based models. The results obtained, when compared with the performance of rating-based models adapted for the task, show that an adequate combination of features allows probability models to obtain an interesting classification accuracy rate (> 90%). It is also notable the good performance of the feature hour of the day combined with date of rating or day of the week, showing that users have “temporal habits” when rating movies. It is, thus, expectable that the addition of time data awareness into rating-based models improve their results. Furthermore, this finding could help on other interesting recommendation-related tasks, e.g., detecting the best hour of the day to send recommendations to users (via mobile devices, for instance).

Regarding future work, we will test additional discriminant functions, based on clustering, SVMs, etc. Moreover, we think that the usage of classifiers specific for binary classes may improve performance on 2-sized households, whereas multi-class classifiers should be used on 3 and 4-sized households. Finally, a mixture of classifiers can be considered for further improvements on classification accuracy. We also contemplate to study the independence assumption of the considered features using, for example, Fisher’s independence analysis based on contingency tables.

7. ACKNOWLEDGMENTS

This work is supported by the Spanish Government (TIN 2008-06566-C04-02) and by the Comunidad de Madrid and Universidad Autónoma de Madrid (CCG10-UAM/TIC-5877). The authors acknowledge support from CCC at UAM.

8. REFERENCES


