User Preferences for Hybrid Explanations

Pigi Kouki  
UC Santa Cruz  
pkouki@soe.ucsc.edu

James Schaffer  
U.S. Army Research Laboratory  
james.a.schaffer.ctr20@mail.mil

Jay Pujara  
UC Santa Cruz  
jay@cs.ucsb.edu

John O’Donovan  
UC Santa Barbara  
jod@cs.ucsb.edu

Lise Getoor  
UC Santa Cruz  
getoor@soe.ucsc.edu

ABSTRACT

Hybrid recommender systems combine several different sources of information to generate recommendations. These systems demonstrate improved accuracy compared to single-source recommendation strategies. However, hybrid recommendation strategies are inherently more complex than those that use a single source of information, and thus the process of explaining recommendations to users becomes more challenging. In this paper we describe a hybrid recommender system built on a probabilistic programming language, and discuss the benefits and challenges of explaining its recommendations to users. We perform a mixed model statistical analysis of user preferences for explanations in this system. Through an online user survey, we evaluate explanations for hybrid algorithms in a variety of text and visual, graph-based formats, that are either novel designs or derived from existing hybrid recommender systems.

KEYWORDS

explanations, hybrid recommendations, hybrid explanations

1 INTRODUCTION

Successful recommender systems are integral to many applications ranging from movie recommendations to e-commerce. Effective recommendations must be both accurate and explainable [9, 10, 13, 16]. Hybrid recommender systems, which use a combination of signals such as social connections, item attributes, and user behavior, demonstrate improved recommendation accuracy [4]. However, compared to non-hybrid counterparts, hybrid models are more complex, and present many challenges when explaining recommendations to users. In this work, we study how to provide useful hybrid explanations that capture informative signals from a multitude of data sources without the burden of understanding the complex hybrid model.

Explanatory approaches for traditional, non-hybrid systems rely on a single explanation style (e.g., collaborative, content, knowledge, utility, or social explanation styles) [7]. Existing work has explored user preferences for these single-style explanations [2, 8].
2 RELATED WORK

Several previous studies have considered explanations for recommender systems, demonstrating explanations can increase persuasiveness [8] and satisfaction [25, 26]. To better understand user preferences, research has compared explanation styles [2], combined different explanation styles [24], and considered dimensions such as personalization [25, 26], tags [30], ranking [17], and natural language presentations [5]. In addition to user preferences and design criteria for explanations, several GUIs and visualizations have been proposed for recommendations, including concentric circles (Peer-Chooser [11, 18]), clustermaps (TalkExplorer [29]), Venn diagrams (SetFusion [21]), and paths among columns (TasteWeights [3, 12]). Extending these ideas, we focus on hybrid explanations and study user preferences across designs and interfaces.

3 HYBRID EXPLANATIONS

The first challenge in understanding hybrid explanations is developing a system to translate a hybrid recommender model into a set of component signals and the utility of each component in the model. Here, we use HyPER [14] which makes use of probabilistic soft logic (PSL) [1], a generic probabilistic reasoning framework. We summarize the HyPER model, then describe our approach to converting the model’s output to visualizable explanations. Our approach is compatible with any hybrid system with similar output.

The HyPER model is specified using a series of rules in first-order logic syntax defined by the modeler. The rules are used to define a probability distribution over recommendations. For example, to implement user-based collaborative filtering recommendations, the following rule is included in the PSL model:

\[ w_{su} : \text{SIMILARUsers}(u_1, u_2) \land \text{LIKES}(u_1, i) \Rightarrow \text{LIKES}(u_2, i). \]  

Each rule is associated with a weight, learned from training data, capturing the importance of the rule to the model. In addition to the user-based collaborative filtering rule above, HyPER also includes item-based collaborative filtering rules with mean-centering priors, as well as rules for social-based recommendations using friendships, and content-based rules for using item attributes.

Next, a process known as grounding is used to combine the model with data and instantiate a set of propositions. For example, given a dataset with user ratings and a social network, user similarities, item similarities, social relationships, and item attributes are generated as evidence. Together, these ground rules are used by PSL to define a probabilistic graphical model which ranks unseen user-item pairs.

Running inference in HyPER generates recommendations captured by the \text{LIKES} predicate. After inference is complete, we select the top \( k \) items for each user. Then, for each of the top \text{LIKES}(u, i), we produce associated groundings used during the inference process. For example, in a restaurant recommendation setting, suppose that \text{Mary’s} top recommended restaurant is \text{Crudo} (i.e., the predicted value of the unobserved variable \text{LIKES}(\text{Mary, Crudo}) has the highest value among all other predicted values). While inferring the value of \text{LIKES}(\text{Mary, Crudo}), HyPER generated the following ground rules:

\[
\begin{align*}
\text{FRIENDS}(\text{Mary, Cindy}) \land \text{LIKES}(\text{Cindy, Crudo}) & \Rightarrow \text{LIKES}(\text{Mary, Crudo}) \\
\text{PERUVIAN}(\text{Limon, Crudo}) \land \text{LIKES}(\text{Mary, Limon}) & \Rightarrow \text{LIKES}(\text{Mary, Crudo}) \\
\text{SIMILARUsers}(\text{Mary, John}) \land \text{LIKES}(\text{John, Crudo}) & \Rightarrow \text{LIKES}(\text{Mary, Crudo}) \\
\text{SIMILARItems}(\text{Crudo, LaMar}) \land \text{LIKES}(\text{Mary, LaMar}) & \Rightarrow \text{LIKES}(\text{Mary, Crudo})
\end{align*}
\]

4 PRESENTATION OF EXPLANATIONS

Given hybrid explanations from HyPER, the next step is designing an interface to present these explanations to users. At a high level, the goal of any presentation style is to improve the user experience. We study the effect of different explanation presentation styles on user experience. To this end, we identify several dimensions for designing interfaces:

- **Presentation (Pres.):** Natural language (Table 1), rule-based (equation 3), or graphical visualizations.
- **Weighting (Wgt):** Whether or not explanation weights are displayed.
- **Grouping (Group):** Whether or not explanations are grouped by style. Each rule can have many groundings in a dataset. For example, do users prefer to be shown the explanation “Mary’s friend Cindy likes Crudo; Mary’s friend Josh likes Crudo” or grouping explanations, “Mary’s friends Cindy and Josh like Crudo.”?
- **Information Density (Dens.):** Amount of information shown.
- **Aggregation (Aggr.):** Whether or not rules are aggregated in the grouping case. Do users prefer explanations of the type “Mary’s friends Cindy and 4 others like Crudo.” or “Mary’s friends Cindy, Josh, Rosie, George, Michael like Crudo.”?
- **Meta-explanations (Meta):** Amount of high-level metadata in explanations (i.e., user similarity, item average rating, and user average rating).
- **Visualization (Visual):** There are different ways of visualizing hybrid explanations, such as concentric circles [11, 18], Venn diagrams [21], and pathways among columns [3]. For Venn diagrams we used three intersections as suggested in [29]. Pathways among columns was tested with and without display of rules/reasoning.

In addition to the effect of different presentation styles on user experience, we also studied whether users have a specific preference over the ranking of the different explanation styles. For example, do users prefer to see social explanations before content-based ones?

5 EVALUATION

We constructed a set of 13 different treatments (Table 2) based on the explanation dimensions described in Section 4. Testing all possible subsets of dimensions was prohibitive, so we chose subsets that we judged as the most informative for explanation design. We
We also included a baseline treatment (BASE) where we presented a synthetic experiment where all users were shown the exact same mockups and details of the study can be found in [15]. We ran Interface mockups were shown to participants in random order. All participants were shown the explanations from Table 1 (hiding the explanation style) and asked them to rank these styles from the most to least important. We randomized the order different explanation styles were shown to avoid any order-related biases.

5.1 Setup

Interface mockups were shown to participants in random order. All the mockups and details of the study can be found in [15]. We ran a synthetic experiment where all users were shown the exact same mockups that were manually generated. Each mockup presented a hybrid explanation for a random user called “Mary” for the restaurant “Crudo”. For each mockup, we elicited answers for a set of user experience questions, corresponding to understandability, system satisfaction, and perceived persuasiveness (Table 3).

Subjective metrics relating to recommender systems have shown to be strongly correlated (e.g., [22]), thus, confirmatory factor analysis was used to group the question items into a latent user experience variable to allow for simpler presentation of results and eliminate measurement error. A Cronbach’s alpha [6] of 0.89 indicates good internal reliability of the constructed user experience (UXP) factor. Average variance extracted (AVE) was 0.64, indicating good convergent validity (AVE > 0.5).

Our experiment considered how participants’ individual characteristics could affect user experience scores for each treatment. We asked pre-study questions related to visualization familiarity (VF, also shown in Table 3). Analysis showed co-variance between visualization familiarity and user experience was less than 0.5, which indicates good discriminant validity between the constructs.

Users were also given a task to rank different explanation styles according to their preference using a drag and drop interface. Specifically, we showed users the explanations from Table 1 (hiding the style) and asked them to rank these styles from the most to least important. We randomized the order different explanation styles were shown to avoid any order-related biases.

5.2 Results

Figure 2 shows a plot of the mean user experience (factor loadings fixed to 1). To test for differences between the within-subjects treatments, we used structural equation modeling (SEM) [28], which can accommodate latent variables during significance testing, thus eliminating measurement error. We specified two factor models: the first with all within-subjects variables loaded onto a single factor (null hypothesis: no differences between treatments); the second with a factor specified for each of the 13 treatments (hypothesis: treatments cause a change in UXP). The model with a factor specified for each treatment achieved better fit. We used the Akaike Information Criterion (AIC) to estimate the quality of each model (a lower AIC indicates better comparative fit) and achieved

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BASE</td>
<td>no exp.</td>
<td>no</td>
<td>no</td>
<td>NA</td>
<td>no</td>
<td>NA</td>
</tr>
<tr>
<td>AGGR</td>
<td>english</td>
<td>no</td>
<td>yes</td>
<td>low</td>
<td>yes</td>
<td>low</td>
</tr>
<tr>
<td>GROUP</td>
<td>english</td>
<td>no</td>
<td>yes</td>
<td>high</td>
<td>no</td>
<td>low</td>
</tr>
<tr>
<td>RULE</td>
<td>rule</td>
<td>no</td>
<td>no</td>
<td>low</td>
<td>no</td>
<td>high</td>
</tr>
<tr>
<td>WGT</td>
<td>english</td>
<td>yes</td>
<td>no</td>
<td>low</td>
<td>no</td>
<td>high</td>
</tr>
<tr>
<td>PLAIN</td>
<td>english</td>
<td>no</td>
<td>no</td>
<td>low</td>
<td>no</td>
<td>high</td>
</tr>
<tr>
<td>NO-GR</td>
<td>english</td>
<td>no</td>
<td>no</td>
<td>med</td>
<td>no</td>
<td>low</td>
</tr>
<tr>
<td>MED-IN</td>
<td>english</td>
<td>no</td>
<td>yes</td>
<td>med</td>
<td>no</td>
<td>low</td>
</tr>
<tr>
<td>LOW-IN</td>
<td>english</td>
<td>no</td>
<td>yes</td>
<td>low</td>
<td>no</td>
<td>low</td>
</tr>
</tbody>
</table>

Table 2: Dimension values for each treatment for the different types of explanations tested.

<table>
<thead>
<tr>
<th>Visual Style</th>
<th>COL</th>
<th>visual</th>
<th>Columns + pathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPR</td>
<td>visual</td>
<td>Columns + Pathways with Reasoning</td>
<td></td>
</tr>
<tr>
<td>VENN</td>
<td>visual</td>
<td>VENN diagram</td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>visual</td>
<td>Concentric Circles</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: The latent VF and UXP factors built on participant responses to subjective questions.
Results from the ranking task are shown in Figure 3. For our analysis we converted the ranking into rating data, i.e., the item listed first was given a rating of 6 and the item ranked last received a rating of 1. Users showed the strongest preference for item average rating explanations, followed by user-based and social explanations. A repeated measures ANOVA revealed differences in rating between the explanation styles ($p < 0.001$). A Tukey post-hoc test showed no statistical difference between social, user-based, and item average rating explanations ($V < 0.10$). However, social, user-based, and item average rating were significantly better than item-based and user average rating ($V < 0.05$). User-based and item average rating were significantly better than content explanations (both $p < 0.05$).

Results from the ranking task are shown in Figure 3. For our analysis we converted the ranking into rating data, i.e., the item listed first was given a rating of 6 and the item ranked last received a rating of 1. Users showed the strongest preference for item average rating explanations, followed by user-based and social explanations. A repeated measures ANOVA revealed differences in rating between the explanation styles ($p < 0.001$). A Tukey post-hoc test showed no statistical difference between social, user-based, and item average rating explanations ($V < 0.10$). However, social, user-based, and item average rating were significantly better than item-based and user average rating ($V < 0.05$). User-based and item average rating were significantly better than content explanations (both $p < 0.05$).

Finally, we conducted analysis on the relationship between visualization familiarity and user experience. A SEM was built with visualization familiarity regressed onto user experience measurements for each treatment (Table 4). Results indicate that visualization familiarity predicts increased user experience in all treatments, except AGGR. The highest increase in user experience is seen in COL and RULE. Model fit: $N = 200$ with 174 free parameters, $RMSEA = 0.064$ ($CI : [0.058, 0.069]$), $TLI = 0.89$, $CFI = 0.90$ over null baseline model, $\chi^2 (772) = 1397$ (indicate acceptable fit, however, note that overall model fit is not an indicator of whether effects exist between variables).

### 6 DISCUSSION AND CONCLUSION

We have presented an evaluation of different visualization approaches using hybrid explanations. The results support prior findings [7] that explanations improve the user experience of recommender systems.

More specifically, Venn diagrams outperform all other visual interfaces and five of seven English interfaces, but are difficult to adapt to more than three sources. Natural language approaches were preferable to rule groundings from HyPER. Our experiments did not show a statistically significant difference across dimensions such as weights, information density, aggregation, or grouping in these explanations. This suggests that most plain English explanations may perform more or less the same in recommendation settings. Our results indicated that social, user-based, and item average rating explanations were the preferred explanation styles by users. Furthermore, we have established a reliable scale, as evidenced by Cronbach’s $\alpha$, for visualization familiarity which might be used to tailor explanation styles to individual users.

Additionally, we discovered that color-blind users ($N = 14$) rated the RULE interface higher in terms of UXP than the rest of the sampled population, with marginal significance ($\beta = 0.15$, $p = 0.051$). The mockup for this interface used a fairly intense violet color which may have been difficult to read for all but the colorblind users. While the colorblind sample was not large enough to change the results of the study, it highlights the need to accommodate these types of users when evaluating UXP for recommender systems.

The mockups that we showed to the users were manually produced and not generated by the HyPER system. As a result, the study was synthetic and did not support personalization. In future work we plan to analyze factors such as the quality of the recommendation and whether the user agrees and connects with the evidence. To this end, we plan to use HyPER’s output, implement the best performing interfaces from our study (Venn, English PLAIN) and run a more comprehensive user case study in a lab setting to better understand effectiveness, efficiency, and user satisfaction of the explanatory system. One open question is determining effective methods for ranking explanations. We plan to compare several ranking strategies by designing interactive interfaces that support personalization.