Curator: A Game with a Purpose for Collection Recommendation

Greg Walsh, Jennifer Golbeck
Human-Computer Interaction Lab
University of Maryland, College Park, MD
jgolbeck@umd.edu

ABSTRACT
Collection recommender systems suggest groups of items that work well as a whole. The interaction effects between items is an important consideration, but the vast space of possible collections makes it difficult to analyze. In this paper, we present a class of games with a purpose for building collections where users create collections and, using an output agreement model, they are awarded points based on the collections that match. The data from these games will help researchers develop guidelines for collection recommender systems among other applications. We conducted a pilot study of the game prototype which indicated that it was fun and challenging for users, and that the data obtained had the characteristics necessary to gain insights into the interaction effects among items. We present the game and these results followed by a discussion of the next steps necessary to bring games to bear on the problem of creating harmonious groups.

Author Keywords
human computation, games with a purpose, serious games

ACM Classification Keywords
H5.m Information interfaces and presentation: Miscellaneous

INTRODUCTION

Collection recommender systems [2] are similar to existing recommender systems but instead of recommending individual items to the user, they recommend groups of items that work well together as a whole unit. There are many factors to consider when creating a collection. Obviously, the quality of each item matters. The size of the collection, diversity of items, and potentially the order of items all have an impact. However, one of the most challenging features to consider is the interaction effects between items.

CHI 2009, April 4 - 9, 2009, Boston, Massachusetts, USA.

Regardless of the domain, some items work well together and others do not. These co-occurrence effects are one of the most important factors in the success or failure of many collections.

It can be a complex task to evaluate co-occurrence effects. Even two items that both have high individual item ratings may not work well together. Someone might have a deep love for chocolate and also for pickles, but not for the two together. This is a rather intuitive effect when considering pairs, but gets more complicated when considering the quality of larger sets of items such as a triple.

For example, chocolate bars and graham crackers are a fine combination; marshmallows and chocolate bars are also; and marshmallows and graham crackers are as well. None of these pairs are poor but neither are they exceptional. However, the combination of all three into a smore makes a much beloved snack for many people. The combination of all three items is better than would be indicated by looking at the three pairs. On the other hand, three items that are very good pairwise can make a bad triple. Consider building a research team of two professors and one graduate student. The professors may work well together, and each may work well with the student. However, all three may have trouble working together. The presence of a student may bring out some tension between the faculty members about who is in control, and the student may have trouble balancing work or contradictory instructions from the faculty.

Similar scenarios can be made moving up from groups of three to four, and so on. While it is useful to look at the compatibility of groups of two or even three items, this approach quickly becomes computationally difficult, requiring $O(n^k)$ comparisons for groups of size k.

Even with extensive data on users' preferences for items and groups of items, this space is vast enough that general rules will almost certainly be necessary for collection recommender systems to be successful. To derive these rules, too, will require a large set of data. One way to obtain that data is through collection-oriented games with a purpose (GWAPs). Just as human users have been able to successfully label millions of images with descriptive tags [4], they can also build small collections that can be used to learn patterns about what

works well together and what does not.

There have been GWAPs designed around eliciting user preferences [1], where users see two items and select the best, but eliciting preferences about what items go well together requires a new game design. In this paper, we present a new class of games, *collection games*, for creating combinations of items that work well together. We then present a prototype game and results from a pilot experiment. We then outline future steps for implementing a large scale game for understanding collection preferences.

RELATED WORK

Games with a Purpose have been used successfully for many tasks that humans can easily solve but computers cannot. Perhaps the greatest success and most widely used example is the ESP game / Google Image Labler where hundreds of thousands of players have contributed tens of millions of labels [4]. Describing the content of media has indeed been the focus of these games. The ESP game generates labels for images; TagATune [3] gathers labels for songs; Peekaboom [5] has players identify objects within an image.

Users' preferences, the focus of our work, have also been addressed in some games. Matchin [1] is the best example. This game presents users with two photos and awards points when they agree on which is "more beautiful". This game has been very successful and has yielded interesting data. Most relevant to this work, the authors also used a variation on the SVD algorithm to produce an image recommender system based on preferences users express in the game. Many of the insights from this work are applicable to our task.

However, our domain has different requirements. To apply the Matchin game technique directly to collection recommenders would involve showing users two collections and having them pick the better one. The number of combinations is so vast that it is unlikely even a popular game would produce useful results. For example, with only 1,000 base items to group together, there are 1 billion combinations of three. Direct comparison as a ranking mechanism on a set of 1 billion would require far more game play than is reasonable. We require a game that considers many more combinations at once and where we can gather data both about the combinations people make (and agree upon) and what combinations are not made. This problem requires a new class of game to gather data about what items work well in groups.

COLLECTION GAMES

A game for understanding preferences about collections requires a new design because of the large space. Direct comparison of two collections is not sufficient because pairwise consideration would take millions of rounds to consider each combination even once. However, combining items without constraint - the closest parallel to labeling GWAPs - would, in most domains, lead to frustrating game play. Making a match would be difficult, and even the simple task of selecting a combination to make would be daunting if all items are considered. Thus, we have created a class of games that allows for multiple matches on a constrained set.

In *collection games*, players see sets of items that they must combine into groups. The items are drawn from a large pool of possible items. Players group the items together into collections and submit their choices. They are awarded points for collections that match (this may include points for partial matches as well as perfect matches).

Game Structure and Play

CURATOR is a two-player game output agreement game played online. It has been designed as a prototype *collection games*. Users connect to a lobby where they are randomly paired with another player who is also waiting in the lobby. The players are then taken to the first round of the game. In the round, they are shown two sets of items and asked to group them into collections. When both players have finished making collections, they move to the scoring phase. Both players are awarded points for each matching collection.

At the end of each round, players see which collections matched and which did not. For both, they see their choices and their partner's choices. To score well, players must not only submit collections that match their own preferences, but also consider the preferences of their partner. This review phase helps players gather some insights about their partner's tastes without direct communication.

At the end of each round, players can earn a bonus. For each matching collection, the users rate the quality of that collection. Although they may have put the collection together, that does not mean they thought it was a particularly good combination, but perhaps it was the best given the options in the round. For each set of matching ratings, the players receive extra points. The other player's ratings are never shown in the bonus round to prohibit development of strategies for cheating in this phase. Checks are also used to make sure players do not always assign the same rating to every combination.

There are five rounds in a game after which high scores are recorded and displayed on a leaderboard.

Few Cheating Strategies

As in most other GWAPs [1, 3, 4], players are matched randomly, so the chance that two people who have worked out a strategy will be paired together is very low.

Even if they were to be paired, there are few strategies available for cheating. There is no communication

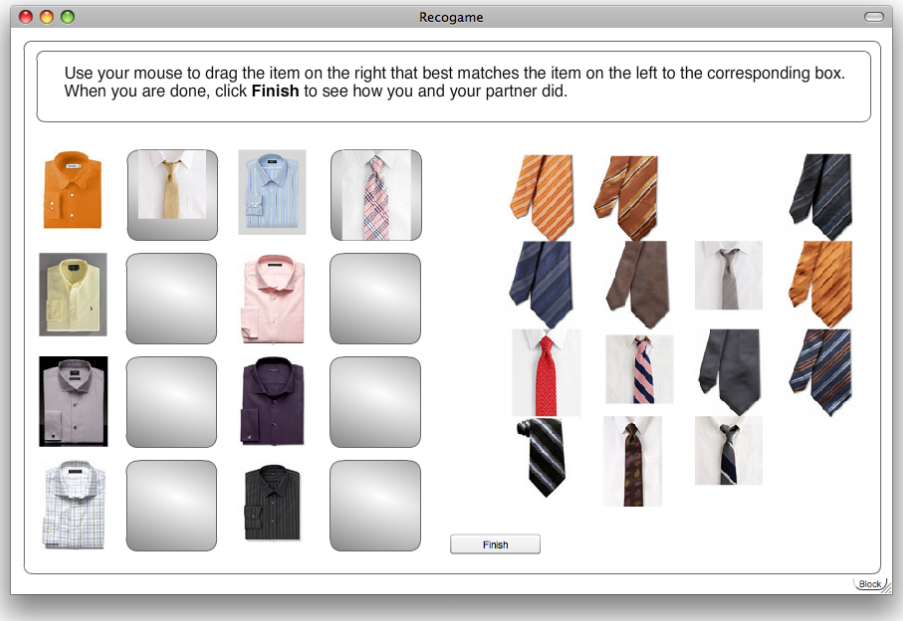

Figure 1. A round of Curator underway with two matches already in place.

between players, so they cannot discuss the moves to make. Since there are a fixed set of items to choose from, players may not cheat by using a predetermined answer, such as they may use the same tag on every image in the ESP game. Items in both groups are displayed in random order for both players. Thus, players cannot use strategies like always matching the first items in each column, second items, and so on.

Available Analyses

In a game like our prototype, with 8 items to be paired with any of 16 other items, there are $\frac{16!}{8!} = 518,918,400$ combinations possible in any single round of play. By requiring players to create combinations they think will work well, we gain data in two ways. First, and most obviously, we can see which combinations they make and which combinations they agree on. These can be analyzed with methods traditionally used for output of GWAPs. Once a combination has been made enough times, it can be considered a valid and interesting combination for study. The number of times it is made is also interesting. Ranking mechanisms and collaborative filtering techniques, such as those from [1] discussed above, could also be used on the sets of combined items.

However, the second data points we have are equally useful. While each player will only make eight pairs, and each round will result in sixteen pairs at most, we know that the remaining pairs will not have been made. On one round of data, that is not meaningful, but over time it can also lead to insights about pairs that do not work well.If we have not seen a pairing after two items have been in the same round many times, it indicates the items may not work well together. The more rounds that are played, the more information we obtain about items that have never been paired. These insights are just as valuable as the combinations which are frequently made.

In addition, the game can be reversed to ask users to create the worst possible combinations. This would provide more direct data on bad pairs and allow us to infer of items never grouped together in this "bad match" version of the game may, in fact, be good combinations.

PILOT STUDY

For a GWAP to be successful, it must be fun to play and yield useful data. To test this in our prototype game, we ran a pilot study with users playing our pilot game, Curator . For this first prototype, players only create pairs of items that work well together. In future implementations, the fixed set of items on the left will be small collections rather than single items, and players will add an additional item to each set to build larger collections.

Players chose from two versions of the game. In one, players match shirts and ties as shown in figure 1. In the other, players match shoes and handbags. Photos of these items are collected from e-commerce retailers that share their images through affiliate marketing programs. Our current prototype has approximately 100 items in each category, though a fully implemented game would have more.

All together, users with 70 unique usernames played. Many of these game rounds were played in a one hour game session held in our lab. While experimental con-

Figure 2. The two most common matches from the game

ditions were not ideal (e.g. it was possible for players to talk to one another since they were in the same room), the participants followed the rules and did not create strategies or discuss their matches as they played.

All together, players had 1,031 opportunities to match, i.e. pairs where both players created a combination. There were 157 matching selections from these opportunities, a match rate of 15.2%.

Most matches were made only once, but seven matches were made twice and two pairs were made three times. Figure 2 shows these most common matches from the shoe-and-handbag version of the game.

Of course, all pairs made by players are stored and even if the players do not agree, if a pair is frequently created it may be a good combination even if the players tend not to match it. The most common match made was between the pink shoe and bag combination shown in figure 2. These two items were combined 7 times - 6 of those instances occurred in the three rounds where both players made the combination together. There were four combinations that were made 6 times, and that includes the other combination from figure 2. Every time a player made that combination, his or her partner also made it.

The fact that these two matches are the most frequent made in the system and practically every time they are put together, both players make the same choice, indicates that they are an exceptionally strong match. While our pilot results are insufficient for statistical analysis, they provide a preliminary indication that the game can indeed produce useful, interesting, and meaningful combinations.

Results and Discussion

Qualitatively, subjects in the pilot provided interesting insights about the matches made and expertise. Interestingly, male subjects with self-confessed ignorance of what made good shoe-and-handbag pairings scored rather well. Their strategy was to match based on color alone (i.e. black shoes with black bags, blue shoes with blue bags, etc.). Much of the nuance behind creating a good match was lost even though the scores were better.

This indicates that players should accurately represent the target users of the recommender system and share their understanding of the nuances behind good matches if the data from their games is to be helpful as guides of the system.

Subjects also reported that the game was fun and challenging. While making many matches was difficult, they were enthusiastic during play when they made matches. The leaderboard was particularly motivating as subjects reported being very proud of appearing there after a successful game.

CONCLUSION AND FUTURE WORK

In this paper, we have presented a new class of games, *collection games*, where users create collections of items. We have shown that the structure of the game provides data in several ways that will be useful for developing rules for collection recommender systems. Through a pilot study with a prototype game Curator , we found that the game was fun, challenging, yet not frustrating for users. A preliminary analysis of the data showed that some pairs were very commonly used while others were never put together. More game play is necessary to obtain significant results that can be used in recommender algorithms, but these preliminary steps indicate that a GWAP can be a successful method for gathering data to help create collection recommender systems.

Future steps will require fine tuning the game play, scoring mechanism, and deploying in a domain and an environment that will receive attention and participation. User testing will also be required before final launch. Once data starts flowing in, we will begin analysis by hand as well as work with machine learning algorithms and existing recommender system techniques to discover the best ways of utilizing this data in the eventual development of collection recommender systems.

REFERENCES

1. Hacker, S., and von Ahn, L. Matchin: eliciting user preferences with an online game. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems* (New York, NY, USA, 2009), ACM, pp. 1207–1216.

2. Hansen, D. L., and Golbeck, J. Mixing it up: recommending collections of items. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems* (New York, NY, USA, 2009), ACM, pp. 1217–1226.

3. Law, E., and von Ahn, L. Input-agreement: a new mechanism for collecting data using human computation games. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems* (New York, NY, USA, 2009), ACM, pp. 1197–1206.

4. von Ahn, L., and Dabbish, L. Designing games with a purpose. *Commun. ACM 51*, 8 (2008), 58–67.

5. von Ahn, L., Liu, R., and Blum, M. Peekaboom: a game for locating objects in images.

In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems* (New York, NY, USA, 2006), ACM, pp. 55–64.