Item-Based Explanations for User-Based Recommendations

Marius Kaminskas
Frederico Durão
and Derek Bridge
Department of Computer Science
Insight Centre for Data Analytics
University College Cork
Ireland
Email: marius.kaminskas|fred.durao|derek.bridge@insight-centre.org

***Abstract*—Explanations can increase user satisfaction with recommender systems. While it is relatively easy to explain the recommendations of a content-based or an item-based collaborative recommender system, user-based collaborative recommendations are harder to explain. In this work, we adopt an approach from the literature that generates *explanation rules* for user-based collaborative-filtering recommendations. These rules are item-based: for example, "If you liked *Toy Story* then you might also like *Finding Nemo*". We modify the approach by proposing two new, alternative measures of explanation rule quality. We evaluate the two new measures in a user study and show that users prefer explanation rules whose antecedents are both accurate and unique with respect to the recommended item.**

***Keywords*–Recommender systems; Explanations; Collaborative filtering.**

I. Introduction

An explanation of a recommendation is any content, additional to the recommendation itself, that is presented to the user with the goal of increasing (among other things) transparency, trust in the system, and decision-making effectiveness [1]. The problem that we examine in this work is how to produce effective explanations (ones that help the user make a good decision) for recommendations made by user-based collaborative filtering (CF) recommender systems.

User-based CF recommender systems were among the first recommenders, and they remain important, e.g., as part of larger ensembles of recommenders. They find the active user's nearest-neighbours and use the neighbours' ratings to predict the active user's rating for items that are in the neighbours' profiles but not in the active user's profile. It is relatively easy to explain the recommendations of content-based recommenders, e.g., by displaying meta-descriptions (such as features or tags) that the active user's profile and the recommended item have in common [1]. Item-based CF recommendations are also amenable to explanation, e.g., by displaying items in the user's profile that are similar to the recommended item [2]. User-based CF recommendations, on the other hand, are harder to explain. Displaying the identities of the active user's neighbours is unlikely to be effective (and may not be ethical) because, when these systems are deployed at scale, the user will not know the neighbours; displaying their profiles is unlikely to be effective too, since even the parts of their profiles they have in common with the active user will be too large to be readily comprehended.

This paper adopts the approach of Bridge & Dunleavy [3], who proposed an explanation generation algorithm for user-based CF recommendations. The algorithm produces explanations in the form of *explanation rules*: for example, "If you liked *Toy Story* then you might also like *Finding Nemo*". The antecedent of an explanation rule (in this case, *Toy Story*) characterizes a subset of the active user's tastes that are predictive of the recommended item, which appears in the consequent of the rule (in this case *Finding Nemo*). In this paper, we refer to such explanations as being in an *item-based style* [4]. They are a familiar style of explanation, since they are used by amazon.com [2].

However, the Bridge & Dunleavy algorithm has a popularity bias (see next section). For this reason, in this paper we propose two new, alternative measures of explanation rule quality that can be used in the algorithm's objective function. The remainder of the paper is structured as follows: Section II describes Bridge & Dunleavy's rule generation algorithm. Section III proposes two new, alternative measures of quality for use in the algorithm. Section IV extends the way in which candidate opinions are obtained from neighbours' profiles. Section V presents both offline experiments and a user study. Section VI reviews related work.

II. Generating Explanation Rules

The algorithm for generating explanation rules presented in [3] constructs explanations in a way that is similar to the mining of association rules (ARs) [5]. Unlike in AR mining, the literals constituting explanation rules represent *item opinions* rather than just the items. Given a set of items I, an item opinion is represented by a tuple $(i, opinion)$, such that $i \in I$ and $opinion \in \{dislike, neutral, like\}$. We will often write just i in place of $(i, opinion)$ allowing context to make clear which is intended. Since most CF datasets contain item ratings on a $1-5$ scale, Bridge & Dunleavy convert the numerical ratings into opinions using a rating threshold, with items rated lower than 3.0 considered *disliked* by a user, items rated as 3.0 assigned a *neutral* opinion, and items rated higher than 3.0 considered *liked*.

Having discretized item ratings into item opinions, an explanation rule R for a user u and a recommended item y is built to contain a *set* of item opinions in its antecedent and a *single* (positive) opinion of the recommended item y in its

```
Data: user profiles $U$, active user $u$, recommended item
      $y$, explanation partner $v$
Result: an explanation rule for $y$
$R \leftarrow$ if _ then $(y, like)$;
$Cs \leftarrow candidates(u, v)$;
while $Cs \neq \{\}$ do
    $Rs \leftarrow$ the set of all new rules formed by adding
      singly each candidate opinion in $Cs$ to the
      antecedent of $R$;
    $R^* \leftarrow \arg\max_{R^* \in Rs} f_{obj}(R^*)$;
    if $f_{obj}(R^*) \leq f_{obj}(R)$ then
        return $R$;
    $R \leftarrow R^*$;
    Remove from $Cs$ the candidate opinion that was
      used to create $R$;
end
return R;
```

Figure 1. Creating an explanation rule

consequent: $R : X \Rightarrow (y, like)$, where $X = \{(i, opinion) : i \in I \setminus y\}$.

Rule generation is based on identifying an *explanation partner* – the most similar neighbour of the active user u who rated the recommended item positively. Subsequently, the explanation rule is built from the item opinions *shared* by the active user and the explanation partner. Items for which the active user u and the explanation partner v share the same opinion are called *candidate opinions*:

$$candidates(u, v) = \\ \{(i, opin) : (i, opin) \in profile_u \wedge (i, opin) \in profile_v\} \quad (1)$$

where, e.g., $profile_u$ is the set of all of user u's item opinions.

Having identified the set of candidate opinions, the rule's antecedent is constructed in a greedy fashion — at each iteration, the candidate opinion which maximizes an objective function is added to the antecedent (see Figure 1).

Bridge & Dunleavy used *accuracy* as the objective function f_{obj}:

$$acc(X \Rightarrow y) = \frac{|\{u \in U : X \subset profile_u \wedge y \in profile_u\}|}{|\{u \in U : X \subset profile_u\}|} \quad (2)$$

where U is the set of all users. A rule's accuracy is equivalent to the *confidence* metric used in AR mining [6].

Bridge & Dunleavy resolved ties (equally accurate rules) using *coverage*, defined as the probability of observing the antecedent of the rule in a user's profile (equivalent to the *support* metric in AR mining):

$$cov(X \Rightarrow y) = \frac{|\{u \in U : X \subseteq profile_u\}|}{|U|} \quad (3)$$

In this work, we extend Bridge & Dunleavy's approach with two contributions. First, we observe that the objective function f_{obj} can be implemented using measures other than accuracy and coverage. We propose and evaluate two new, alternative measures. Second, we extend the candidate opinions from those of a *single* explanation partner to those of a *set* of the active user's neighbours. In the next two sections we describe the two contributions in greater detail.

III. Proposed Rule Utility Metrics

While accuracy and coverage offer an intuitive way of measuring the strength of the explanation rules, they are biased toward popular items. For instance, the movie *Star Wars* is frequently rated and therefore co-occurs in user profiles with many other (not necessarily related) movies. Relying on accuracy and coverage may lead to explanations that are trivial or irrelevant with respect to the recommended item, e.g., "If you liked *Star Wars* then you might like *Fargo*".

Intuitively, an item opinion in an explanation rule is good the more it is *unique* with respect to the recommended item. In other words, we are looking for measures that promote antecedent items that are accurate (i.e., result in high accuracy) with respect to the consequent item, but also penalize antecedent items that achieve high accuracy with (many) other consequent items.

Within AR mining, there has similarly been a quest for measures of AR interestingness, beyond confidence and support, including measures of *lift* and *conviction* [6]. However, these measures in general try to counter-act the tendency of the accuracy measure to favour rules with popular *consequents*. Hence, these measures do not achieve what we want to achieve. In our case, the consequent is a given: it is the item recommended by the user-based CF system. Our goal is to build explanation rules using measures that counter-act the tendency of the accuracy measure to favour popular items in *antecedents*.

To the best of our knowledge, the uniqueness property that we seek does not correspond to any of the existing measures of AR interestingness. We experimented with a number of these existing measures and others, such as selecting a rule whose antecedents were similar to its consequent. But none of these resulted in the selection of distinctive ('unique') antecedents. We therefore propose two new, alternative measures — one that discounts a rule's accuracy by the antecedent's popularity and the other that discounts its accuracy by the antecedent's explanatory power.

A. Popularity-discounted accuracy

Our *popularity-discounted accuracy* (pda) measure is designed to balance the accuracy of a rule and the popularity of its antecedent. Specifically, we discount the rule's accuracy by the number of items that could potentially be explained by the antecedent, i.e., the number of items in the dataset (other than the recommended item) that co-occur with the antecedent in at least one user's profile:

$$pda(X \Rightarrow y) = \\ \frac{acc(X \Rightarrow y)}{|\{j \in I \setminus X \cup \{y\} : \exists u \in U, X \subset profile_u \wedge j \in profile_u\}| + 1} \quad (4)$$

Initial analysis of explanations generated using pda as the objective function in Figure 1 revealed that the explanation rules tended to contain more items in their antecedents compared to the original approach (which, for two datasets, was reported to contain no more than 3 items in the antecedent [3]). Therefore, to restrict the lengths of the rules, we included an additional constraint in the algorithm: the rule R is returned if either $f_{obj}(R^*) \leq f_{obj}(R)$ *or* if $acc(R^*) \leq acc(R)$; see Figure 2. This additional constraint ensures the quality of the rules and restricts their lengths so that they are closer to those of the original approach.

B. Uniqueness-discounted accuracy

Our *uniqueness-discounted accuracy* (uda) metric is similar to the popularity-discounted accuracy, but instead of counting the number of *all* potential explanations that could be generated from the antecedent, it counts the items that the antecedent can explain *better* (i.e., with a higher accuracy) than the target item y:

$$uda(X \Rightarrow y) = \frac{acc(X \Rightarrow y)}{|\{j \in I \setminus X \cup \{y\} : acc(X \Rightarrow j) > acc(X \Rightarrow y)\}| + 1} \tag{5}$$

Again we included the additional constraint on the rule's accuracy in the algorithm to avoid generating longer rules.

IV. Extended Candidate Opinions

In Figure 1, the candidate opinions (the set Cs) are taken from the profile of a *single explanation partner* — the most similar neighbour of the active user who liked the recommended item. However, user-based CF recommender systems generate item predictions using a larger number of nearest-neighbours.

To reflect this in the explanation generation process, we evaluate a variant of the algorithm where the candidate opinions are obtained from the profiles of *all* the active user's nearest-neighbours (where the size of this set is given by the underlying user-based CF recommender system).

In recommendation, the contribution of a neighbour to item predictions is weighted by the neighbour's similarity to the active user. We mirror this in the revised explanation generation algorithm by weighting each candidate opinion by the neighbour's similarity:

$$R^* \leftarrow \arg\max_{R^* \in Rs} f_{obj}(R^*) \cdot sim(u, v) \tag{6}$$

where u is the active user and v is the neighbour whose profile contains the candidate opinion used to obtain R^*. If the candidate opinion is contained in more than one neighbours' profiles, the highest $sim(u, v)$ is used.

The changes that we have proposed in this section and the previous one are summarized in Figure 2.

```
Data: user profiles $U$, active user $u$, recommended item
      $y$, nearest neighbours $NN$
Result: an explanation rule for $y$
$R \leftarrow$ if _ then $(y, like)$;
$Cs \leftarrow \bigcup_{v \in NN} candidates(u, v)$;
while $Cs \neq \{\}$ do
    $Rs \leftarrow$ the set of all new rules formed by adding
      singly each candidate opinion in $Cs$ to the
      antecedent of $R$;
    $R^* \leftarrow \arg\max_{R^* \in Rs} f_{obj}(R^*) \cdot sim(u, v)$;
    if $f_{obj}(R^*) \leq f_{obj}(R) \lor acc(R^*) \leq acc(R)$ then
        return $R$;
    $R \leftarrow R^*$;
    Remove from $Cs$ the candidate opinion that was
      used to create $R$;
end
return R;
```

Figure 2. Creating an explanation rule: revised

V. Experiments

Our main goal is to compare the effectiveness of the two new measures (pda and uda) against the original accuracy-based approach (acc). Each measure can be used by taking candidate opinions either from a single explanation partner (designated ep) or from the set of neighbours (designated nn), as in Section IV, resulting in a total of six alternatives: acc+ep, pda+ep, uda+ep, acc+nn, pda+nn and uda+nn.

For extended candidate opinions, all experiments were conducted using a neighbourhood of 150 users. Furthermore, in all experiments, we used only the positive item opinions as candidates for rule generation (i.e., opinions of the form $(i, like)$). The positive opinions were identified by selecting items having a rating higher than 3.0. We leave the exploration of alternative rating thresholds and the possible use of negative and neutral item opinions for future work.

Explanation rules can only be evaluated using feedback from real users, since, to the best of our knowledge, there are no offline metrics that can quantify the "goodness" of an explanation. However, comparing six alternatives in a user study would result in a high cognitive load for the participants. Therefore, as a first step in the evaluation procedure, we performed offline experiments in an effort to reduce the number of approaches to be evaluated in a user study.

A. Offline experiments

In the offline experiments, we used the MovieLens 1M dataset [7]. For each user, we split her rating data into train and test items. Then, we randomly selected one highly rated item (i.e., an item with a rating of 5.0) for explanation generation. (In other words, we are explaining an item that we know the user likes.) The evaluation was performed using a 5-fold cross-validation, where each fold contains 20% of user ratings as a test set. The same set of test items was used to evaluate the six different approaches.

The quality of the explanation rules was measured using a number of metrics all of which provide a single value per-rule. Those metrics that are defined at the level of individual items (i.e., novelty and similarity) were aggregated into a rule-level score using three different strategies — taking the minimum, maximum, and average value as the rule score. The full set of metrics is as follows:

- The *overlap* with the original accuracy-based algorithm. The overlap value is computed as the number of antecedent items in the generated rule that are also present in the original (accuracy-based) version of the same rule, normalized by the length of the evaluated rule;
- The *accuracy* and *coverage* metrics (see 2,3);

- The *rule length*, defined as the number of item opinions in the rule's antecedent;
- The minimum, maximum, and average *novelty* of the items in the rule's antecedent, where the novelty of item i is $-log_2 P(i)$ where $P(i) = |\{u : i \in \mathit{profile}_u\}|/|U|$, and U is the set of all users in the dataset;
- The minimum, maximum, and average *similarity* of the items in the rule's antecedent to the item in the consequent, where $similarity(i, y) = \frac{|L_i \cap L_y|}{|L_i \cup L_y|}$ and L_i and L_y are sets of text labels describing items i and y respectively. In addition to the movie descriptors included in the MovieLens dataset (a vocabulary of 18 genres, 1.65 genres per movie on average), we scraped IMDb plot keywords for each movie and kept those labels that appeared in the profiles of at least 10 movies. This resulted in an average of 60 labels per movie.

The metrics were computed for each explanation rule and then averaged over all test cases.

We recognise that these evaluation metrics are mere proxies for what we regard as good explanations, but we believe that they can nevertheless help us to reduce the six alternatives down to a few for use in a user study.

B. Results of offline experiments

The results are shown in Figure 3, which shows the metrics computed over approximately 27,600 data points (across the 5 cross-validation folds).

The lengths of the rules for all approaches is below 4 on average. But there are rules that are longer than those reported by Bridge & Dunleavy: they reported a maximum length of 3 [3], but the difference may be because they used a different version of the dataset (MovieLens 100k), as well as the other changes described in earlier parts of this paper.

Our results indicate that, pda+ep and uda+ep, which use a single explanation partner, produce rules similar to the original acc+ep (an average overlap of 75%). The average overlap between pda+ep and uda+ep themselves (not shown in the figure), is 59%. Methods that use extended candidate opinions (acc+nn, pda+nn and uda+nn) have a smaller overlap with the original acc+ep and also with each other (an average of 50% between acc+nn and each of pda+nn and uda+nn).

Rules computed from extended candidate opinions (nn approaches) achieve higher average accuracy, but lower coverage compared to the approaches that use a single explanation partner (ep). The larger set of candidate opinions from which to choose allows the algorithms to identify item opinion patterns that are more accurate but less frequent and therefore potentially more interesting to the user.

The pda approaches produce rules with the highest novelty. This is not surprising, since pda favours rules with less popular items. Also, as expected, the extended candidate opinions approaches (nn) tend to generate rules with more novel items. The two combined, pda+nn, gives highest novelty.

With regard to rule antecedent similarity to the recommended item, extended candidate opinion approaches (nn) achieve a slightly higher similarity compared to the single explanation partner approaches (ep).

Overall, the higher accuracy and novelty achieved by the nn approaches lead us to believe that the use of extended candidate opinions is beneficial for the rule generation and we focus our user study on acc+nn, pda+nn and uda+nn.

C. User study

The three explanation generation approaches identified as the most promising during the offline evaluation stage were subsequently compared in a user study. For this user study, we employed the 10M version of the MovieLens dataset, rather than the 1M version used in the offline experiments, since it contains movies that are more recent, which are more likely to be recognized by the study participants [7]. To further increase the chances of user familiarity with the recommended item, we filtered the test sets (below) to include only movies produced in the year 2000 or later and having at least 100 ratings in the training set. It is important to note that we only applied the filtering to test sets, not the items appearing in antecedents of explanation rules.

Each user's item ratings were split into a train set (80%), from which antecedents can be picked, and a test set (20%), which was filtered (above) and from which one highly-rated test item (i.e., an item which we know the user likes) was picked and treated as the item to be recommended to the user. We did this for each of 100 randomly-chosen users, giving us 100 recommendations. For each recommendation, we generated three explanation rules (acc+nn, pda+nn and uda+nn). If the antecedents of the three explanation rules did not differ pairwise by at least one item, then we picked a different highly-rated item from the test set and generated its explanations. This ensures that we have no redundant survey questions, where participants are asked to judge identical explanations.

The 100 recommendations (each with three explanation rules) were partitioned across 5 questionnaires, containing 20 recommendations each. For each of the 20 recommendations, the questionnaires showed the recommended movie and the three explanation rules. The order in which the explanation rules were displayed was determined at random, e.g., sometimes acc+nn was the first of the three, sometimes the second and sometimes the last. The questionnaire asked participants to mark all explanations that they found helpful in choosing the movie recommendation. If they did not know the recommended movie or if unknown movies in the explanations prevented them from making a fair comparison, they were asked to mark an explicit option ("None of the explanations are helpful"). Hence, for each recommendation, participants can mark zero, one, two or three of the explanations as helpful.

From July to September of 2016, 50 volunteers (mostly students and researchers from Ireland and Brazil) took part in the study. Each participant responded to exactly one questionnaire through a dedicated web site, 10 volunteers per questionnaire. In order to help participants, all questionnaires had introductory guidelines for the experiment and links to synopses of the movies. The participants were also free to gather more information about the movies from any source of their choice, such as YouTube or IMDb.

D. Results of user study

Table I summarizes the responses. The maximum possible in each cell is 200: for each of the 20 recommendations up to

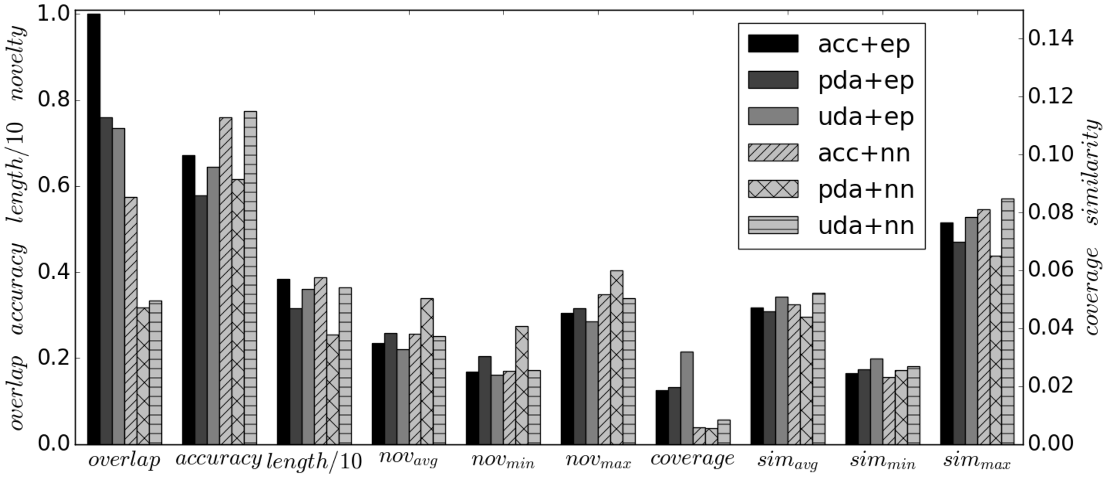

Figure 3. Offline experiments: results

TABLE I. USER STUDY: RESULTS

	Q1	Q2	Q3	Q4	Q5	Total
acc+nn	67	58	57	66	82	330
pda+nn	69	57	43	55	71	295
uda+nn	72	85	95	80	101	433
None	45	63	50	49	24	231

10 people could have found them helpful. Hence the maximum possible across the questionnaires (Q1 to Q5) is 1000.

As can be seen, *uda+nn* produced by far the most helpful explanations. Our other new measure, *pda*, was not successful: *pda+nn* produced the least helpful explanations. In particular, *uda+nn* explanations were selected 1.3 more times than *acc+nn* and nearly 1.5 more times than *pda+nn*. Using 99% level two-tailed Student's t-tests, we observed that, in this study, i) *acc+nn* and *pda+nn* are not statistically different (p-value = 0.333); ii) *acc+nn* and *uda+nn* are statistically different (p-value = 0.017); and iii) *pda+nn* and *uda+nn* are statistically different (p-value = 0.005). From this, we conclude it is not statistically correct to claim that *acc+nn* is superior to *pda+nn*, but *uda+nn* is superior to both.

VI. Related Work

Several papers consider the role of explanations in recommender systems. They agree that providing explanations can lead to greater user satisfaction and to acceptance of a recommended item. Justifying why an item is recommended is often welcomed by users [1] [8] [9]. Herlocker et al. report that the benefits include education, acceptance, user involvement and justification [8]. In a similar fashion, Tintarev & Masthoff outline six motivations for explanations in recommender systems: transparency, trust, scrutability, effectiveness and efficiency, persuasiveness and satisfaction [9].

Vig et al. [4] divide explanations into three main kinds: user-based (such as showing the user a histogram of their neighbours' ratings, e.g., [8]) item-based (as used in this paper and in amazon.com [2]), and feature-based (such as using attribute-value pairs [10], item content (e.g., from news items) [11], user-generated tags [4] [12], or features and opinions mined from user reviews [13] [14]). Some systems combine the different types of explanations; for example, Symeonidis et al. combine feature-based with item-based [15].

Herlocker et al. conducted a user survey to test the persuasiveness of twenty-one different styles of user-based and feature-based explanation [8]. Similarly, Gedikli et al.'s study tested, among other things, the efficiency and effectiveness of ten different styles of explanation [12]: seven of them drawn from [8], plus a user-based pie-chart and two new forms of feature-based explanation using user-generated tags. For Herlocker et al., histograms of user ratings were the most persuasive; Gedikli et al. found their tag explanations to most increase satisfaction. Neither study included explanations in the item-based style.

Bilgic & Mooney ran a user study to compare item-based explanations (which they refer to as *influence-style explanations*) with user-based and feature-based explanations [16]. In their study, a user is shown a recommendation with an explanation, and she is asked to rate the item before and after consumption. Bilgic & Mooney found that user-based explanations cause users to over-estimate the quality of items; the other two forms of explanation were found to result in significantly more accurate estimations of final ratings.

One issue that is often ignored is the transparency [1] or fidelity [3] of the explanation, i.e., the extent to which the explanation reveals the logic of the recommender. (Gedikli et al. refer to this as *objective transparency* to contrast it with *perceived transparency*, i.e., whether the user thinks that the logic has been revealed [12].) A lot of the work in this area is characterized by explanations that are divorced from the recommender. By contrast, we believe that one advantage of the Bridge & Dunleavy scheme that we have adopted in this paper is that it does have some fidelity to the operation of the underlying user-based CF recommender: both the recommendations and the explanations are based on opinions shared by the active user and her nearest-neighbours.

VII. Conclusion

We have built on the work of Bridge & Dunleavy, which generates explanation rules in the item-based style for items recommended by user-based CF recommender systems [3]. In particular, we have proposed two new, alternative measures of explanation rule quality for use in the algorithm's objective function, pda and uda. These new measures attempt to overcome the tendency of the original accuracy and coverage measure to favour popular items. We also proposed extending the set of candidate opinions from which explanation rule antecedents are constructed: instead of using opinions from a single explanation partner, we modify the algorithm to allow it to use opinions from the active user's nearest neighbours.

We evaluated our proposed modifications in both an offline experiment and a user study. The offline experiment indicated the benefits of using the extended set of candidate opinions (from the nearest neighbours), resulting in rules that are both more accurate and contain more items that are novel. The online study showed that users found that explanation rules which were generated using the uda measure were far more helpful than those produced using pda and Bridge & Dunleavy's accuracy and coverage measure.

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289. The work has also been supported by the Department of Computer Science, Federal University of Bahia, Brazil.

References

[1] N. Tintarev and J. Masthoff, "Explaining recommendations: Design and evaluation," in Recommender Systems Handbook, F. Ricci, L. Rokach, and B. Shapira, Eds. Springer, 2015, pp. 353–382.

[2] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," IEEE Internet Computing, vol. 7, no. 1, 2003, pp. 76–80.

[3] D. Bridge and K. Dunleavy, "If you liked Herlocker et al.'s explanations paper, then you might like this paper too," in Procs. of the Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (Worskhop Programme of the Eighth ACM Conference on Recommender Systems). CEUR-WS.org, vol.1253, 2014, pp. 22–27.

[4] J. Vig, S. Sen, and J. Riedl, "Tagsplanations: Explaining recommendations using tags," in Procs. of the 14th International Conference on Intelligent User Interfaces. ACM, 2009, pp. 47–56.

[5] J. J. Sandvig, B. Mobasher, and R. Burke, "Robustness of collaborative recommendation based on association rule mining," in Procs. of the ACM Conference on Recommender Systems. ACM, 2007, pp. 105–112.

[6] L. Geng and H. J. Hamilton, "Interestingness measures for data mining: A survey," ACM Comput. Surv., vol. 38, no. 3, 2006, pp. 9:1–9:32.

[7] F. M. Harper and J. A. Konstan, "The MovieLens Datasets: History and Context," ACM Trans. Interact. Intell. Syst., vol. 5, no. 4, Dec. 2015, pp. 19:1–19:19. [Online]. Available: http://doi.acm.org/10.1145/2827872

[8] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," in Procs. of the ACM Conference on Computer Supported Cooperative Work. ACM, 2000, pp. 241–250.

[9] N. Tintarev and J. Masthoff, "Effective explanations of recommendations: User-centered design," in Procs. of the ACM Conference on Recommender Systems. ACM, 2007, pp. 153–156.

[10] C. Scheel, A. Castellanos, T. Lee, and E. W. De Luca, "The reason why: A survey of explanations for recommender systems," in Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation, A. Nürnberger, S. Stober, B. Larsen, and M. Detyniecki, Eds. Springer, 2014, pp. 67–84.

[11] R. Blanco, D. Ceccarelli, C. Lucchese, R. Perego, and F. Silvestri, "You Should Read This! Let Me Explain You Why: Explaining News Recommendations to Users," in Procs. of the Twenty-first ACM International Conference on Information and Knowledge Management. ACM, 2012, pp. 1995–1999.

[12] F. Gedikli, D. Jannach, and M. Ge, "How should I epxlain? A comparison of different explanation types for recommender systems," Int. J. Human-Computer Studies, vol. 72, 2014, pp. 367–382.

[13] K. Muhammad, A. Lawlor, R. Rafter, and B. Smyth, "Great explanations: Opinionated explanations for recommendations," in Procs. of the 23rd International Conference on Case-Based Reasoning, E. Hüllermeier and M. Minor, Eds. Springer, 2015, pp. 244–258.

[14] S. Chang, F. M. Harper, and L. G. Terveen, "Crowd-based personalized natural language explanations for recommendations," in Procs. of the 10th ACM Conference on Recommender Systems. ACM, 2016, pp. 175–182.

[15] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos, "Moviexplain: A recommender system with explanations," in Procs. of the Third ACM Conference on Recommender Systems. ACM, 2009, pp. 317–320.

[16] M. Bilgic and R. Mooney, "Explaining recommendations: Satisfaction vs. promotion," in Procs. of Beyond Personalization: A Workshop on the Next Stage of Recommender Systems Research at the International Conference on Intelligent User Interfaces, 2005, pp. 13–18.