Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR 2020)
Applications and Systems

Argumentation as a Framework for Interactive Explanations for Recommendations

Antonio Rago[1] , **Oana Cocarascu**[1] , **Christos Bechlivanidis**[2] and **Francesca Toni**[1]
[1]Department of Computing, Imperial College London, UK
[2]Department of Experimental Psychology, University College London, UK
{a.rago15, oc511}@imperial.ac.uk, c.bechlivanidis@ucl.ac.uk, ft@imperial.ac.uk

Abstract

As AI systems become ever more intertwined in our personal lives, the way in which they explain themselves to and interact with humans is an increasingly critical research area. The explanation of recommendations is thus a pivotal functionality in a user's experience of a recommender system (RS), providing the possibility of enhancing many of its desirable features in addition to its *effectiveness* (accuracy wrt users' preferences). For an RS that we prove empirically is effective, we show how argumentative abstractions underpinning recommendations can provide the structural scaffolding for (different types of) interactive explanations (IEs), i.e. explanations supporting interactions with users. We prove formally that these IEs empower feedback mechanisms that guarantee that recommendations will improve with time, hence rendering the RS *scrutable*. Finally, we prove experimentally that the various forms of IE (tabular, textual and conversational) induce *trust* in the recommendations and provide a high degree of *transparency* in the RS's functionality.

1 Introduction

Recommender systems (RSs) (Resnick and Varian 1997) are increasingly popular methods for helping users discover items which may be of relevance, according to some preferences, by exploiting vast data sources that humans alone could never fully utilise. Desirable features of RSs (Tintarev and Masthoff 2015) include: *effectiveness*, guaranteeing recommendation accuracy with regards to users' preferences; *transparency*, explaining how recommendations are made to users; *scrutability*, permitting feedback from users based on these explanations; *trust*, increasing users' confidence in the RS; and *satisfaction*, increasing users' enjoyment in using the RS. Arguably, the main focus of the literature over the past decade has been the relentless pursuit of effectiveness via ever more complex models (Dacrema, Cremonesi, and Jannach 2019), leaving the (as important) features relating to users' *experience* of the RS somewhat neglected.

Meanwhile, there has been an unprecedented push of late towards the *explainability* of AI systems from academia, industry and governments, e.g. see (Cath et al. 2018). Recent studies have also shown that the use of complex and uninterpretable models is at times both unnecessary and counterproductive, in RSs specifically (Dacrema, Cremonesi, and Jannach 2019) and in AI in general (Rudin 2019). Further, as argued in (Abdul et al. 2018), the core trends of explainability in AI are not on "usable, practical and effective transparency that works for and benefits people" and true progress in this direction can only be made via interdisciplinary works. Within the area of RSs, recent works have begun to focus more on explanatory desirable features, giving users the option to provide more information about their preferences via feedback mechanisms to achieve scrutability, even at the expense of effectiveness, e.g. see (Balog, Radlinski, and Arakelyan 2019). In (Rader, Cotter, and Cho 2018), different types of explanation are used to explain *Facebook*'s newsfeed algorithm and all are shown to have beneficial effects on transparency, thus justifying the variety of explanations emerging in the RS literature, e.g. graphical (Rago, Cocarascu, and Toni 2018), tabular (Vig, Sen, and Riedl 2009) or conversational (Balog, Radlinski, and Arakelyan 2019). Further justification is given by (McInerney et al. 2018), who show via a bandit-based method that different users respond best to different explanations.

We use *argumentation*, as understood in AI (e.g. see (Simari and Rahwan 2009)), to provide a general, unifying framework for RS explanations, combining aspects from Knowledge Representation and Reasoning, Human-Computer Interaction and RSs. Argumentation has been shown to be an excellent means towards explainability, in RSs (e.g. see (Chesñevar, Maguitman, and González 2009)) and beyond (e.g. see (Cyras et al. 2019a; Madumal et al. 2019)). In the context of RSs, various forms of explanations, e.g. conversational (Cocarascu, Rago, and Toni 2019) or linguistic (Cyras et al. 2019b), may be drawn from argumentation frameworks extracted from RSs (Rago, Cocarascu, and Toni 2018), unearthing reasons for or against recommendations. We define an RS with a diverse explanatory repertoire for the *aspect-item* frameworks of (Rago, Cocarascu, and Toni 2018). Our RS results from a new method for extracting *argumentation explanations* in the form of *bipolar argumentation frameworks* (Cayrol and Lagasquie-Schiex 2005), serving as the underlying scaffolding for the explanations with which users may interact in multiple ways. We undertake empirical, theoretical and experimental evaluations of the RS to show that it satisfies four of the aforementioned desirable features (effectiveness, scrutability, transparency and trust). Finally, we discuss the implications of this study, e.g. various avenues for future work.

2 Background

Bipolar argumentation frameworks (BFs) (Cayrol and Lagasquie-Schiex 2005) are triples $\langle \mathcal{X}, \mathcal{L}^-, \mathcal{L}^+ \rangle$ with $\mathcal{X}$, a set of *arguments*, and $\mathcal{L}^-, \mathcal{L}^+ \subseteq \mathcal{X} \times \mathcal{X}$, *attack* and *support* relations, resp. For any $a \in \mathcal{X}$, the *strength* of a (wrt $\langle \mathcal{X}, \mathcal{L}^-, \mathcal{L}^+ \rangle$) is given by $\sigma(a)$, for $\sigma : \mathcal{X} \to \mathbb{I}$ a *(strength) function* and $\mathbb{I}$ a set equipped with a preorder $\leq$ (Baroni, Rago, and Toni 2018).

We focus on RSs that can be understood as aspect-item frameworks (A-Is) (Rago, Cocarascu, and Toni 2018), which are 6-tuples $\langle \mathcal{I}, \mathcal{A}, \mathcal{T}, \mathcal{L}, \mathcal{U}, \mathcal{R} \rangle$ such that:

- $\mathcal{I}$ is a finite, non-empty set of *items*, e.g. in the movie context, *Breakfast at Tiffany's* (m_0) or *My Fair Lady* (m_1);
- $\mathcal{A}$ is a finite, non-empty set of *aspects*, disjoint from $\mathcal{I}$, e.g. *Audrey Hepburn* (a_1) or *drama* (g_1), and $\mathcal{T}$ is a finite, non-empty set of *types*, where each aspect in $\mathcal{A}$ has a single type in $\mathcal{T}$ (possibly shared with other aspects), e.g. a_1 is of type *actor* while g_1 and g_2 (*romance*) are of type *genre*;
- $\mathcal{L} \subseteq (\mathcal{I} \times \mathcal{A}) \cup (\mathcal{A} \times \mathcal{I})$ is a symmetric binary relation, where for any $i \in \mathcal{I}, a \in \mathcal{A}$, $(i, a) \in \mathcal{L}$ and $(a, i) \in \mathcal{L}$ indicate that item i holds aspect a, e.g. m_0 holds aspects a_1, g_1 and g_2 but m_1 holds only a_1 and g_1;
- $\mathcal{U}$ is a finite, non-empty set of *users*;
- $\mathcal{R} : \mathcal{U} \times (\mathcal{I} \cup \mathcal{A}) \to [-1, 1]$ is a partial function of *ratings*, e.g. a user may give m_1 a very positive rating of 1 but a_1 a moderately negative rating of -0.5.

In the remainder of the paper we will assume as given an arbitrary A-I $\mathcal{F} = \langle \mathcal{I}, \mathcal{A}, \mathcal{T}, \mathcal{L}, \mathcal{U}, \mathcal{R} \rangle$, unless otherwise specified, and we will use the following notations: for any $t \in \mathcal{T}$, $\mathcal{A}_t$ denotes $\{a \in \mathcal{A} |$ the type of a is $t\}$; $\mathcal{X}$ denotes $\mathcal{I} \cup \mathcal{A}$ and is referred to as the set of *item-aspects*; $\mathcal{L}(x) = \{y \in \mathcal{X} | (y, x) \in \mathcal{L}\}$ denotes the set of *linked item-aspects* of $x \in \mathcal{X}$; for $i \in \mathcal{I}$, $\mathcal{L}_t(i)$ denotes $\{a \in \mathcal{L}(i) | a \in \mathcal{A}_t\}$. We will assume that $\mathcal{L}(x) \neq \varnothing$ ($\forall x \in \mathcal{X}$) and that $\mathcal{L}_t(i) \neq \varnothing$ ($\forall i \in \mathcal{I}$ and $\forall t \in \mathcal{T}$), i.e. that all item-aspects have linked item-aspects and that all items have linked aspects of each type.

As in (Rago, Cocarascu, and Toni 2018), we will define an RS for users characterised by a *profile*. Here, the profile π_u of $u \in \mathcal{U}$ consists of: a *'collaborative filtering'* constant $\phi^u \in]0, 1]$ (by excluding ϕ^u=0, we impose that collaborative filtering is never disregarded); $\forall t \in \mathcal{T}$ a *'type importance'* constant $\mu_t^u \in]0, 1]$ (by excluding $\mu_t^u = 0$, we impose that all types of aspect are considered); $\forall v \in \mathcal{U}$ such that $u \neq v$, a *'similarity'* constant $\omega_v^u \in [0, 1]$, indicating how similar users u and v are, and thus how much $v's$ ratings should be taken into account in making recommendations to u. Intuitively, increasing a constant increases how much the corresponding feature is taken into account for that user.

3 Recommender System

An RS, in its essence, provides recommendations to a user based on how items satisfy certain preferences or requirements. In our case, we predict how highly the user would rate an item based on various other forms of predicted ratings in the A-I, which are propagated via linked item-aspects. We define these various other forms of predicted ratings before doing so for items' predicted ratings, using throughout the supporting notions given in Table 1.

Definition 1. *Let $u \in \mathcal{U}$. The* weighted average rating $\rho^u : \mathcal{I} \to [-1, 1]$ *is defined as follows, for $i \in \mathcal{I}$: if $\Upsilon^u(i) \neq \varnothing$ and $\sum_{v \in \Upsilon^u(i)} \omega_v^u > 0$ then $\rho^u(i) = \frac{\sum_{v \in \Upsilon^u(i)} \omega_v^u \mathcal{R}(v,i)}{|\Upsilon^u(i)|}$; otherwise, $\rho^u(i)$ is undefined.*

Intuitively, weighted average ratings are approximations of how the user would rate an item based on similar users' ratings, and they are undefined when no other (similar) users have given any ratings for the item. Weighted average ratings are used to determine aspects' predicted ratings in the absence of user's ratings, as follows (again, using Table 1):

Definition 2. *Let $u \in \mathcal{U}$. The* predicted aspect rating $\mathcal{P}_{\mathcal{A}}^u : \mathcal{A} \to [-1, 1]$ *is defined as follows, for $a \in \mathcal{A}$:*

if $\mathcal{R}(u, a)$ is defined then $\mathcal{P}_{\mathcal{A}}^u(a) = \mathcal{R}(u, a)$; else

if $\Lambda^u(a) = \Lambda^{-u}(a) = \varnothing$ then $\mathcal{P}_{\mathcal{A}}^u(a) = 0$; else

if $\Lambda^u(a) = \varnothing$ then $\mathcal{P}_{\mathcal{A}}^u(a) = \phi^u \frac{\sum_{i \in \Lambda^{-u}(a)} \rho^u(i)}{|\Lambda^{-u}(a)|}$; else

if $\Lambda^{-u}(a) = \varnothing$ then $\mathcal{P}_{\mathcal{A}}^u(a) = \frac{\sum_{i \in \Lambda^u(a)} \mathcal{R}(u, i)}{|\Lambda^u(a)|}$; else

$$\mathcal{P}_{\mathcal{A}}^u(a) = \left[\frac{\sum_{i \in \Lambda^u(a)} \mathcal{R}(u, i)}{|\Lambda^u(a)|} + \phi^u \frac{\sum_{i \in \Lambda^{-u}(a)} \rho^u(i)}{|\Lambda^{-u}(a)|}\right] / [1 + \phi^u]$$

Note that this definition is well-defined. In particular, in the third case, $\rho^u(i)$ is necessarily defined for every $i \in \Lambda^{-u}(a)$ (by definition of Λ^{-u}, see Table 1). Similarly, in the fourth case, $\mathcal{R}(u, i)$ is necessarily defined and, in the final case, $\rho^u(i)$ and $\mathcal{R}(u, i)$ are defined. Predicted aspect ratings are then used to calculate predicted ratings for items:

Definition 3. *Let $u \in \mathcal{U}$. For $i \in \mathcal{I}$ and $t \in \mathcal{T}$, let the* contribution *of t towards i's predicted rating for u be $c_t^{u,i} = \frac{\sum_{a \in \mathcal{L}_t^u(i)} \mathcal{P}_{\mathcal{A}}^u(a)}{|\mathcal{L}_t^u(i)|}$. Then, the* predicted item rating $\mathcal{P}_{\mathcal{I}}^u : \mathcal{I} \to [-1, 1]$ *is defined as follows. For any $i \in \mathcal{I}$, if $\mathcal{R}(u, i)$ is defined then $\mathcal{P}_{\mathcal{I}}^u(i) = \mathcal{R}(u, i)$, else: $\mathcal{P}_{\mathcal{I}}^u(i) = \frac{\sum_{t \in \mathcal{T}} \mu_t^u c_t^{u,i}}{\sum_{t \in \mathcal{T}} \mu_t^u}$.*

Intuitively, a user's rating on an item is its predicted rating, else the latter is the weighted average of the types' contributions, each weighted by its importance. A contribution itself is a prediction of how the user would rate an item's linked aspects of a certain type as a whole, e.g. all its actors or genres, each being an average of our predictions of how the user might rate each aspect. Note that the notion $\mathcal{P}_{\mathcal{I}}^u$ is well-defined, in particular given our assumptions that $\mathcal{L}_t(i) \neq \varnothing$ and $\mu_t^u \neq 0$, for any $i \in \mathcal{I}$ and $t \in \mathcal{T}$ (see Section 2).

For illustration see Figure 1 (treating all edges and their symmetricals as elements of $\mathcal{L}$). Here, we have given ratings for three items (m_1, m_2 and m_3), which are used to calculate predicted ratings for the three aspects (a_1, g_1 and g_2), which are in turn used to calculate a predicted rating for m_0.

We will use $\mathcal{P}_{\mathcal{X}}^u(x)$, called the *predicted rating of an item-aspect*, to denote $\mathcal{P}_{\mathcal{I}}^u(x)$ or $\mathcal{P}_{\mathcal{A}}^u(x)$ for $x \in \mathcal{I}$ or $x \in \mathcal{A}$, resp.

Note that our predicted aspect and item ratings differ from those in (Rago, Cocarascu, and Toni 2018) mainly in that we do not use collaborative filtering on each item directly to affect its predicted rating, instead incorporating it indirectly via the similar users' ratings' effect on predicted ratings of

Notion	Description	Formula
$\Upsilon^u(i)$	the set of users other than u who have rated item i	$\{v \in \mathcal{U} \backslash \{u\} \mid \mathcal{R}(v,i)$ is defined$\}$
$\Lambda^u(a)$	the set of items linked to a with ratings from u	$\{i \in \mathcal{L}(a) \mid \mathcal{R}(u,i)$ is defined$\}$
$\Lambda^{-u}(a)$	the set of items linked to a without ratings from u but with defined weighted average ratings	$\{i \in \mathcal{L}(a) \mid \rho^u(i)$ is defined$\} \backslash \Lambda^u(a)$
$\mathcal{L}^u(x)$	the set of item-aspects *affecting* x for u	$\{y \in \mathcal{X} \mid (y,x) \in \mathcal{L}^u\}$
$\mathcal{L}^u_t(i)$	the set of aspects affecting i of type t for u	$\{a \in \mathcal{L}^u(i) \mid a \in \mathcal{A}_t\}$
$\mathcal{L}^-(x)$	the set of *attackers* of x	$\{y \in \mathcal{X} \mid (y,x) \in \mathcal{L}^-\}$
$\mathcal{L}^+(x)$	the set of *supporters* of x	$\{y \in \mathcal{X} \mid (y,x) \in \mathcal{L}^+\}$

Table 1: Supporting notions, with $u \in \mathcal{U}$, $i \in \mathcal{I}$, $a \in \mathcal{A}$, $t \in \mathcal{T}$ and $x \in \mathcal{X}$.

the item's linked aspects. This will allow concise explanations of recommendations, aiding transparency.

4 Argumentation Explanations

In the spirit of (Rago, Cocarascu, and Toni 2018), we extract argumentation explanations for recommendations drawn from predicted ratings by directing the relations between item-aspects (indicating where one's predicted rating affects another's) and then translating the directed relations into appropriate argumentation relations.

Definition 4. *The* directed A-I *for* $u \in \mathcal{U}$ *is* $\mathcal{F}^u = \langle \mathcal{I}, \mathcal{A}, \mathcal{T}, \mathcal{L}^u, \mathcal{U}, \mathcal{R} \rangle$*, where:*
$\mathcal{L}^u = \{(i,a) \in \mathcal{L} \mid \mathcal{R}(u,a)$ *is undefined* $\wedge \exists v \in \mathcal{U}$ *such that* $\mathcal{R}(v,i)$ *is defined where if* $v \neq u$ *then* $\omega^u_v \neq 0\} \cup \{(a,i) \in \mathcal{L} \mid \mathcal{R}(u,i)$ *is undefined*$\}$
For any $i \in \mathcal{I}$*, let* $r^u(i)$ *be* $\mathcal{R}(u,i)$ *if defined, else* $\phi^u \rho^u(i)$ *if defined, and otherwise undefined. The* BF corresponding to $\mathcal{F}^u$ *is* $\langle \mathcal{X}, \mathcal{L}^-, \mathcal{L}^+ \rangle$ *(with* $\mathcal{X} = \mathcal{I} \cup \mathcal{A}$*, as before) such that:*
$\mathcal{L}^- = \{(i,a) \in \mathcal{L}^u \mid r^u(i) < 0\} \cup \{(a,i) \in \mathcal{L}^u \mid \mathcal{P}^u_{\mathcal{A}}(a) < 0\}$;
$\mathcal{L}^+ = \{(i,a) \in \mathcal{L}^u \mid r^u(i) > 0\} \cup \{(a,i) \in \mathcal{L}^u \mid \mathcal{P}^u_{\mathcal{A}}(a) > 0\}$.

An attack (support) from one item-aspect to another indicates that the former weakens (strengthens, resp.) the *argument* that the user would rate the latter highly. We use r^u to extract attacks or supports from items as it determines their effects on linked aspects' predicted ratings (see Definition 2), whereas $\mathcal{P}^u_{\mathcal{A}}$ does so from aspects to linked items' predicted ratings (see Definition 3), avoiding circularity issues.

An illustration is shown in Figure 1, for an A-I where $\mathcal{I} = \{m_0, m_1, m_2, m_3\}$, $\mathcal{A} = \{a_1, g_1, g_2\}$, $\mathcal{T} = \{actor, genre\}$, and $\mathcal{L} = \{(x,y), (y,x) \mid (x,y) \in \mathcal{L}^- \cup \mathcal{L}^+\}$. Given ratings from a user u: $\mathcal{R}(u,m_1) = 1$, $\mathcal{R}(u,m_2) = 0.4$, $\mathcal{R}(u,m_3) = -0.8$, and constants: $\mu^u_{genre} = 0.8$, $\mu^u_{actor} = 0.5$, the predicted ratings are shown on the graph. Here, due to their positive ratings, m_1 and m_2 support the arguments that the user likes aspects that these items hold while the relation from the negatively-rated m_3 is one of attack. Similarly, the positive predicted ratings for a_1 and g_1 lead to supports

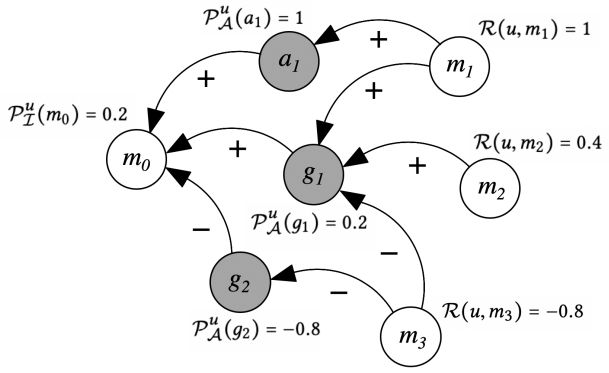

Figure 1: Example argumentation explanation, visualised as a graph, with arguments corresponding to items (aspects) shown in white (grey, resp.) and given and predicted ratings indicated. In the movie domain, item-aspects could be: m_0 - *Breakfast at Tiffany's*, a_1 - *Audrey Hepburn*, g_1 - *drama*, g_2 - *romance*, m_1 - *My Fair Lady*, m_2 - *The Birds* and m_3 - *The Umbrellas of Cherbourg*.

towards m_0 while g_2's negative predicted rating means it attacks m_0.

In the spirit of (Rago, Cocarascu, and Toni 2018), predicted ratings of item-aspects can be understood as functions for determining the strengths of item-aspects (arguments) in argumentation explanations, exhibiting desirable dialectical properties as gradual semantics of the underpinning BFs. Here, we consider a novel property of *fluid monotonicity*, which is instrumental for driving the feedback that a user may provide in interactions shaped by argumentation explanations (see Section 5). This property is formulated for two BFs, with the same arguments, equipped with respective strength functions; it sanctions that changes to the strength of some attacker/supporter of an argument result in a change, in the same "direction", of the argument's strength:

Definition 5. *Let* $B = \langle \mathcal{X}, \mathcal{L}^-, \mathcal{L}^+ \rangle$ *and* $B' = \langle \mathcal{X}, \mathcal{L}^{-\prime}, \mathcal{L}^{+\prime} \rangle$ *be BFs. Let* σ *and* σ' *be strength functions wrt* B *and* B' *resp. Then,* B, B'*,* σ, σ' *satisfy* fluid monotonicity *if, for any* $x, y \in \mathcal{X}$ *where* $x \in \mathcal{L}^-(y) \cup \mathcal{L}^+(y)$*, if* $\sigma'(x) > \sigma(x)$ *then* $\sigma'(y) > \sigma(y)$ *and if* $\sigma'(x) < \sigma(x)$ *then* $\sigma'(y) < \sigma(y)$*.*

This property draws inspiration from *reinforcement* (Amgoud et al. 2017) and *strict monotonicity* (Baroni, Rago, and Toni 2019), stating that the weakening (strengthening) of any argument weakens (strengthens, resp.) an argument it attacks or supports. An instance of this property, for specific choices of argumentation explanations B, B' and definitions of σ, σ' in terms of predicted ratings, is satisfied, as follows:

Proposition 1. *Let* $u \in \mathcal{U}$ *and* $x \in \mathcal{X}$*. Let* $\mathcal{F}' = \langle \mathcal{I}, \mathcal{A}, \mathcal{T}, \mathcal{L}, \mathcal{U}, \mathcal{R}' \rangle$ *be such that* $\mathcal{R}'(v, z_1) = \mathcal{R}(v, z_1)$ *for all* $v \in \mathcal{U} \backslash \{u\}$ *and all* $z_1 \in \mathcal{X}$*, and* $\mathcal{R}'(u, z_2) = \mathcal{R}(u, z_2)$ *for all* $z_2 \in \mathcal{X} \backslash \{x\}$*. Let* $\mathcal{P}^{u\prime}_{\mathcal{X}}$ *and* $r^{u\prime}$ *be the predicted ratings in* $\mathcal{F}'$*. Finally, for any* $y \in \mathcal{X}$ *where* $x \in \mathcal{L}^-(y) \cup \mathcal{L}^+(y)$*, let* σ^u *and* $\sigma^{u\prime}$ *be such that:*
• $\sigma^u(x) = r^u(x)$ *and* $\sigma^{u\prime}(x) = r^{u\prime}(x)$ *if* $x \in \mathcal{I}$*, otherwise* $\sigma^u(x) = \mathcal{P}^u_{\mathcal{A}}(x)$ *and* $\sigma^{u\prime}(x) = \mathcal{P}^{u\prime}_{\mathcal{A}}(x)$*;*
• $\sigma^u(y) = \mathcal{P}^u_{\mathcal{X}}(y)$ *and* $\sigma^{u\prime}(y) = \mathcal{P}^{u\prime}_{\mathcal{X}}(y)$*.*
Then, $\mathcal{F}, \mathcal{F}'$*,* $\sigma^u, \sigma^{u\prime}$ *satisfy fluid monotonicity.*

Proof. (Sketch) Since $x \in \mathcal{L}^-(y) \cup \mathcal{L}^+(y)$, Def. 4 gives that either $x \in \mathcal{I}$, $y \in \mathcal{A}$ (case 1) or $x \in \mathcal{A}$, $y \in \mathcal{I}$ (case 2). In case 1, Defs. 2 and 4 show that $\mathcal{R}(u,y)$ is not defined and $r^u(x)$ is defined, the latter giving that $\exists v \in \mathcal{U}$ such that $\mathcal{R}(v,x)$ is defined, thus the first two conditions of Def. 2 do not hold. From Def. 2, a change in $\mathcal{R}'$ such that $r^{u\prime}(x) > r^u(x)$ in the three other conditions gives $\mathcal{P}^{u\prime}_{\mathcal{A}}(y) > \mathcal{P}^u_{\mathcal{A}}(y)$. The inverse effect holds for a change such that $r^{u\prime}(x) < r^u(x)$. In case 2, Def. 4 shows that $\mathcal{R}(u,y)$ is not defined hence $\mathcal{P}^u_{\mathcal{I}}(y) = \frac{\sum_{t\in\mathcal{T}} \mu^u_t c^{u,y}_t}{\sum_{t\in\mathcal{T}} \mu^u_t}$, by Def. 3. Let the type of x be t_x. The change $\mathcal{P}^{u\prime}_{\mathcal{A}}(x) > \mathcal{P}^u_{\mathcal{A}}(x)$ has no effect on μ^u_t, $\forall t \in \mathcal{T}$, nor on $c^{u,y}_t$, $\forall t \in \mathcal{T} \backslash \{t_x\}$, thus the only parameter affected is $c^{u,y}_{t_x} = [\sum_{a\in\mathcal{L}^u_{t_x}(y)} \mathcal{P}^u_{\mathcal{A}}(a)]/|\mathcal{L}^u_{t_x}(y)|$. So the change $\mathcal{P}^{u\prime}_{\mathcal{A}}(x) > \mathcal{P}^u_{\mathcal{A}}(x)$ gives $c^{u,y\prime}_{t_x} > c^{u,y}_{t_x}$ and thus $\mathcal{P}^{u\prime}_{\mathcal{I}}(y) > \mathcal{P}^u_{\mathcal{I}}(y)$. The inverse effect holds for a change such that $\mathcal{P}^{u\prime}_{\mathcal{A}}(x) < \mathcal{P}^u_{\mathcal{A}}(x)$. Thus, the proposition holds in both cases. □

This property is useful for our RS as it characterises how the item-aspects' predicted ratings in the RS affect one another, but we also posit that it is intuitive from an argumentation viewpoint in general when an argument's semantic meaning is *fluid*. For instance, in our RS setting, an argument x may signify that the user (strongly) likes its corresponding item-aspect if its strength is (extremely) positive or that the user (strongly) dislikes the item-aspect if its strength is (extremely) negative. Thus, if we reduce the predicted rating of some supporter y of x, weakening it (lowering its potential to increase its linked item-aspects' predicted ratings), past the midpoint where y's semantic meaning is modified, its support towards x becomes attack, and continuing to reduce y's predicted rating actually strengthens the attack against x (and its potential to reduce its linked arguments' predicted ratings). The inverse effect occurs for an attacker being weakened. For example, in Figure 1, weakening an argument, e.g. reducing the rating of m_1, m_2 or m_3, reduces the predicted rating of arguments it attacks or supports, e.g. g_1, thus strengthening the attack or weakening the support. Note that strengthening (weakening) an aspect inherently increases the likelihood of an item which holds (does not hold, resp.) the aspect being recommended.

In the next section, we capitalise on this behaviour by providing users with the means to interact with the RS via feedback mechanisms which affect predicted ratings intuitively.

Note that our argumentative explanations differ from those in (Rago, Cocarascu, and Toni 2018) since: (i) we use BFs rather than *tripolar* argumentation frameworks, which also include a *neutralising* relation for diluting effects on positive or negative predicted ratings (this third relation is not required in our methodology for interacting with users); (ii) intuitive behaviour which drives feedback, such as fluid monotonicity, cannot be guaranteed for the predicted ratings in the RS in (Rago, Cocarascu, and Toni 2018).[1]

[1]However, it satisfies a property of *weak balance* (Rago, Cocarascu, and Toni 2018), which characterises attacks (supports) as links between an item-aspects such that if we isolate the *affecter* as the only item-aspect affecting the *affectee*, the former reduces (increases, resp.) the latter's predicted rating wrt its neutral midpoint. Our σ^u also satisfies weak balance (omitted for lack of space).

5 Interactive Explanations

We now define methods for extracting interactive explanations (IEs) from the argumentation explanations of the previous section.[2] The IEs are generated following positive or negative recommendations to the user (for or against items, resp.) based upon (positive or negative, resp.) predicted items ratings. Thus, for example, the RS may recommend the user movie m_0 (indicated simply as m_0) or not (indicated simply as $\neg m_0$). We define three IEs, of varying formats, namely *tabular*, *textual* and *conversational*, offering a glimpse into the flexibility afforded by using argumentation explanations as scaffolding for IEs. The IEs are designed so as to vary the information provided in order to cater for different users with diverse explanatory requirements, i.e. targeting *width* or *depth* in the argumentation explanation, and also how the information is provided, i.e. *statically* where all information is given at once or *dynamically* where information is provided progressively in a conversation.

All our IEs are equipped with feedback mechanisms for users to provide more information about their preferences, triggered by one of two cases where there is a discrepancy between the recommendation and the user preferences: case ↓, where an item is recommended but the user would have liked it not to be, and case ↑, where an item is not recommended but the user would have liked it to be. These situations could be rectified by simply giving a *corrected rating* on the item, in the spirit of data augmentation, but we define the IEs in such a way that more information is elicited from the user regarding the reasons for the discrepancy.

5.1 IE1: Tabular

Our first IE is tabular in nature and is of a similar format to the explanations of (Vig, Sen, and Riedl 2009). For case ↓ (↑), the user is presented with a set of the strongest supporters (attackers, resp.) in a table with their types and predicted ratings. IE1 may be seen as making use of the width in the argumentation explanations as many of the aspects affecting an item may be considered, but nothing deeper in the argumentation explanation. We choose (up to) three aspects to avoid overloading users, as suggested by (Pu and Chen 2007), and provide the information statically in one interaction. Formally, IE1 for an item i_0 consists of the following supporters for case ↓:[3]

$a_{max1} = argmax_{a\in\mathcal{L}^+(i_0)} \mathcal{P}_{\mathcal{A}}(a)$

$a_{max2} = argmax_{a\in\mathcal{L}^+(i_0)\backslash\{a_{max1}\}} \mathcal{P}_{\mathcal{A}}(a)$ [if $|\mathcal{L}^+(i_0)| > 1$]

$a_{max3} = argmax_{a\in\mathcal{L}^+(i_0)\backslash\{a_{max1},a_{max2}\}} \mathcal{P}_{\mathcal{A}}(a)$ [if $|\mathcal{L}^+(i_0)| > 2$]

and the following attackers for case ↑:

$a_{min1} = argmin_{a\in\mathcal{L}^-(i_0)} \mathcal{P}_{\mathcal{A}}(a)$

$a_{min2} = argmin_{a\in\mathcal{L}^-(i_0)\backslash\{a_{min1}\}} \mathcal{P}_{\mathcal{A}}(a)$ [if $|\mathcal{L}^-(i_0)| > 1$]

$a_{min3} = argmin_{a\in\mathcal{L}^-(i_0)\backslash\{a_{min1},a_{min2}\}} \mathcal{P}_{\mathcal{A}}(a)$ [if $|\mathcal{L}^-(i_0)| > 2$]

For illustration, in the example in Figure 1, an IE1 for m_0 for case ↓ may consist of supporters a_1 and g_1, their types and their (translated) predicted ratings:

[2]From now on, we focus on recommendations to $u \in \mathcal{U}$ and drop the superscript u throughout. Also, we let any parameter marked prime ($'$) be the new instance after some indicated change.

[3]Here $argmax_{s\in S} f(s) = \{s \in S | \forall t \in S \backslash \{s\} : f(t) \leq f(s)\}$ and $argmin_{s\in S} f(s) = \{s \in S | \forall t \in S \backslash \{s\} : f(t) \geq f(s)\}$.

Audrey Hepburn	actor	5/5 stars
Drama	genre	3/5 stars

while an IE1 for $\neg m_0$ for case ↑ utilises its attacker g_2:

Romance	genre	0.5/5 stars

The user can then interact with the explanations by rectifying ratings for supporter/attackers. In the illustration, for case ↓ the user may decide to lower the rating of (one of) a_1 and g_1, and in case ↑ the user may increase the rating of g_2.

5.2 IE2: Textual

Our next IE is textual, amounting to a linguistic description of the reasons behind a recommendation, delivered statically. It differs from IE1 not just in style, but also in content, by targeting depth, rather than width, in the argumentation explanation. For case ↓ (↑), we first state the type t_{max} (t_{min}, resp.) that had the biggest positive (negative, resp.) effect, followed by the most prominent supporting (attacking, resp.) aspect a_{max} (a_{min}, resp.) of that type, followed by its most prominent supporting (attacking, resp.) item i_{max} (i_{min}, resp.). Formally, IE2 is as follows for cases ↓/↑ for a recommendation i_0/$\neg i_0$:

The recommender system inferred that you would/would not like item i_0 due to t_{max}/t_{min}. It reached this conclusion as:
[if $\mathcal{R}(u, a_{max})$/$\mathcal{R}(u, a_{min})$ is defined:] + *you liked/disliked a_{max}/a_{min}.*
[else:] + *it inferred that you like/dislike a_{max}/a_{min} because:*
[if $\mathcal{R}(u, i_{max})$/$\mathcal{R}(u, i_{min})$ is defined:] + *you liked/disliked i_{max}/i_{min}.*
[else:] + *similar users liked/disliked i_{max}/i_{min}.*

where: $t_{max} = argmax_{t \in \mathcal{T}} c_t^{i_0}$, $t_{min} = argmin_{t \in \mathcal{T}} c_t^{i_0}$, $a_{max} = argmax_{a \in \mathcal{L}^+_{t_{max}}(i_0)} \mathcal{P}_{\mathcal{A}}(a)$, $a_{min} = argmin_{a \in \mathcal{L}^-_{t_{min}}(i_0)} \mathcal{P}_{\mathcal{A}}(a)$, $i_{max} = argmax_{i \in \mathcal{L}^+(a_{max})} r(i)$ and $i_{min} = argmin_{i \in \mathcal{L}^-(a_{min})} r(i)$.

For illustration, in the example in Figure 1, an IE2 for m_0 for case ↓ may be *The recommender system inferred that you would like the movie Breakfast at Tiffany's due to its actors. It reached this conclusion as it inferred that you like Audrey Hepburn because you liked My Fair Lady.*

The user may then interact with the RS in the following ways for cases ↓/↑, resulting in the indicated changes:
• *I don't care about a movie's t_{max}/t_{min}* - Reduce constant such that $\mu'_{t_{max}} < \mu_{t_{max}}$ / $\mu'_{t_{min}} < \mu_{t_{min}}$.
• *I dislike/like a_{max}/a_{min}* - Assign rating $\mathcal{R}'(u, a_{max})$ / $\mathcal{R}'(u, a_{min})$ such that $\mathcal{P}'_{\mathcal{A}}(a_{max}) < 0 < \mathcal{P}_{\mathcal{A}}(a_{max})$ / $\mathcal{P}'_{\mathcal{A}}(a_{min}) > 0 > \mathcal{P}_{\mathcal{A}}(a_{min})$.
• *But I dislike/like i_{max}/i_{min}* [available if $\mathcal{R}(u, a_{max})$/$\mathcal{R}(u, a_{min})$ is not defined] - Assign rating such that $\mathcal{R}'(u, i_{max}) < 0 < r(i_{max})$/ $\mathcal{R}'(u, i_{min}) > 0 > r(i_{min})$.
• *I don't care about what other users think* [available if $\mathcal{R}(u, a_{max})$/ $\mathcal{R}(u, a_{min})$ and $\mathcal{R}(u, i_{max})$/$\mathcal{R}(u, i_{min})$ are not defined] - Reduce constant such that $\phi' < \phi$.

5.3 IE3: Conversational

Our final IE uses the same information as IE2, i.e. exploiting the depth of the argumentation explanation, but structured in a dynamic, conversational protocol between the user and the RS, potentially resulting in changes in predicted ratings. The protocols for both cases ↓/↑, as shown in Figure 2, consist of:

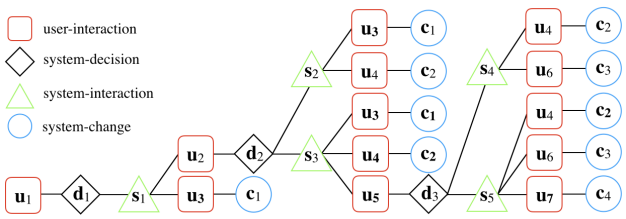

Figure 2: IE3 protocol.

• *user-interactions* - statements from the user to request reasons or provide information about preferences (labelled u_i);
• *system-decisions* - decisions by the system in response to user-interactions (labelled d_i);
• *system-interactions* - statements from the system to elicit more information about the user's preferences (labelled s_i);
• *system-changes* - adjustments to the system after the user-interactions to effect some change in the RS (labelled c_i).

We now define the components of the protocols formally.

The **user-interactions** for cases ↓/↑ are as follows:
u_1: *I disliked/liked i_0, why did the recommender system infer that I would/would not like it?*
u_2: *Why did the system infer that I like/dislike its t_{max}/t_{min}?*
u_3: *I don't care about a movie's t_{max}/t_{min}.*
u_4: *Not true, I dislike/like a_{max}/a_{min}.*
u_5: *Why did the system infer that I like/dislike a_{max}/a_{min}?*
u_6: *But I dislike/like i_{max}/i_{min}.*
u_7: *I don't care about what other users think.*

The **system-decisions** for cases ↓/↑ are as follows:
d_1: Output s_1 with t_{max}=$argmax_{t \in \mathcal{T}} c_t^{i_0}$/$t_{min}$=$argmin_{t \in \mathcal{T}} c_t^{i_0}$.
d_2: Let a_{max}=$argmax_{a \in \mathcal{L}^+_{t_{max}}(i_0)} \mathcal{P}_{\mathcal{A}}(a)$/$a_{min}$=$argmin_{a \in \mathcal{L}^-_{t_{min}}(i_0)} \mathcal{P}_{\mathcal{A}}(a)$. If $\exists \mathcal{R}(u, a_{max})$ / $\mathcal{R}(u, a_{min})$, output s_2, else output s_3.
d_3: Let i_{max}=$argmax_{i \in \mathcal{L}^+(a_{max})} r(i)$ / i_{min}=$argmin_{i \in \mathcal{L}^-(a_{min})} r(i)$. If $\exists \mathcal{R}(u, i_{max})$/$\mathcal{R}(u, i_{min})$ output s_4, else output s_5.

The **system-interactions** for cases ↓/↑ are as follows:
s_1: *The recommender system inferred you would/would not like the item due to t_{max}/t_{min}.*
s_2: *The recommender system came to the conclusion that you would/would not like this item's t_{max}/t_{min} because you liked/disliked a_{max}/a_{min}.*
s_3: *The recommender system came to the conclusion that you would/would not like this item's t_{max}/t_{min} because it inferred that you would like/dislike a_{max}/a_{min}.*
s_4: *The recommender system inferred that you would/would not like a_{max}/a_{min} because you liked/disliked the i_{max}/i_{min}.*
s_5: *The recommender system inferred that you would/would not like a_{max}/a_{min} as similar users liked/disliked the i_{max}/i_{min}.*

The **system-changes** for cases ↓/↑ are as follows:[4]
c_1: Reduce constant such that $\mu'_{t_{max}} < \mu_{t_{max}}$ / $\mu'_{t_{min}} < \mu_{t_{min}}$.
c_2: Assign rating $\mathcal{R}'(u, a_{max})$/$\mathcal{R}'(u, a_{min})$ such that $\mathcal{P}'_{\mathcal{A}}(a_{max}) < 0 < \mathcal{P}_{\mathcal{A}}(a_{max})$/$\mathcal{P}'_{\mathcal{A}}(a_{min}) > 0 > \mathcal{P}_{\mathcal{A}}(a_{min})$.
c_3: Assign rating such that $\mathcal{R}'(u, i_{max}) < 0 < r(i_{max})$ / $\mathcal{R}'(u, i_{min}) > 0 > r(i_{min})$.
c_4: Reduce constant such that $\phi' < \phi$.

[4]Note that these system-changes and the conditions under which they become available are equivalent to those of IE2.

In IE3, the first decision by the system (d_1) considers the attacking and supporting aspect arguments of the item argument itself. Firstly, the arguments are grouped by their types, e.g. *actors* vs. *genres* in Figure 1, and the user is given the opportunity to adjust how much the most prominent type is taken into account (u_3/c_1). If the user requests more information about why this inference was made (u_2) the system then determines the most prominent individual aspect in this type (d_2), e.g. (assuming that genre was the most prominent type) the arguments *drama* (g_1) vs. *romance* (g_2) in Figure 1. Depending on whether this prominent aspect's rating was given or predicted, the user is given the opportunity to change it (u_4/c_2) or in the latter case request more information about this inference (u_5). In this case, the system determines the most prominent item which led to this conclusion (d_3), i.e. item arguments which attack or support this aspect argument, e.g. *My Fair Lady* (m_1) vs. *The Birds* (m_2) in Figure 1. Depending on whether this prominent item was rated by the user or similar users, the user is given the chance to change the rating (u_6/c_3) and in the latter case reduce how much similar users are taken into account (u_7/c_4).

6 Evaluation and Discussion

In this section we assess how well the RS, equipped with the IEs, performs along some of the desirable features of (Tintarev and Masthoff 2015). The research questions (RQs) for our 3-stage evaluation are: **RQ1** (Effectiveness) - *Does the RS produce accurate (wrt users' preferences) recommendations?* **RQ2** (Scrutability) - *Does the RS adapt to user preferences?* **RQ3** ((Perceived) Transparency) - *Do users (feel they) understand how the RS works?* **RQ4** (Trust) - *Do users trust the RS's recommendations?* **RQ5** (Satisfaction) - *Are users happy with the recommendations?*

6.1 Stage 1 - Empirical Analysis

We first address RQ1, assessing how well the predicted ratings of our RS match users' given ratings in various datasets, without considering the IEs at this stage[5]. We experiment with the following datasets: a subset of the Netflix dataset as used in (Rago, Cocarascu, and Toni 2018), the MovieLens 100K benchmark dataset and the MovieLens development dataset (Harper and Konstan 2015). For all datasets, $\mathcal{T}$ = {genre, actor, director}.[6]

The Netflix dataset consists of 528 movies, each with 500 ratings on a five star (integral) scale from 1 to 5. From the total of 36968 users, we keep only those who have rated at least 10 movies, giving a total of 4772 users. The MovieLens 100K benchmark dataset consists of 1239 movies with ratings on a five star (integral) scale from 1 to 5 and 943 users in total, each having rated at least 20 movies. The MovieLens development dataset consists of 9533 movies with ratings on a 5-star scale, with half-star increments (0.5 stars - 5.0 stars), and 610 users in total, each having rated at least 20 movies. Dataset statistics can be found in Table 2.[7]

Dataset	$\lvert\mathcal{I}\rvert$	$\lvert\mathcal{U}\rvert$	$\lvert\mathcal{A}\rvert$	>100	>50	>30	10-30
Netflix	528	4772	1154	365	1077	1923	2849
ML 100K	1239	943	4328	312	510	689	254
ML Dev.	9533	610	22394	244	376	498	112

Table 2: Dataset summary with the number of items, users, aspects, and the number of users who rated more than 100, more than 50, more than 30, and between 10 and 30 movies.

We compare against the following baseline recommendation algorithms, as implemented in the Surprise library (Hug 2017) using the default configuration settings:
• **KNN**: K Nearest Neighbours, a classical collaborative filtering algorithm;
• **KNNZ**: KNN with the z-score normalization of each user;
• **SVD**: Singular Value Decomposition, an algorithm that led to the best results in the Netflix challenge[8];
• **NMF**: Non-negative Matrix Factorization, a collaborative filtering algorithm (Luo et al. 2014);
• **Slope1**: Slope One (Lemire and Maclachlan 2007), based on "popularity differential" between items for users by finding the average rating differential;
• **CoClust**: Co-clustering (George and Merugu 2005), an algorithm built on simultaneous clustering of users and items.

On the Netflix dataset, we also compare with the **AI* RS** method in (Rago, Cocarascu, and Toni 2018), but use a different experimental set-up than the latter. Indeed, in (Rago, Cocarascu, and Toni 2018), all users who have rated at least 10 or alternatively 20 movies are used as training and the users who rated less than this as testing, with a variable number of ratings (5, 7, 10) used to address the cold-start problem, i.e. where sufficient information about users/items is lacking, for the users in the testing set. Instead, given that our focus is to determine the performance of our methodology on several datasets, we split the Netflix and the other datasets into training and testing sets and address the cold-start problem as follows: for users who rated over 30 movies, we select 30 movies for testing and use those which remain for training; for users who rated 10 to 30 movies, we select 10 movies for testing and use the rest for training. We settle for this split of training and testing sets which yields a large number of user-rating pairs for testing. Also, whereas in (Rago, Cocarascu, and Toni 2018) similarities are determined based on users' genre preferences, we use cosine similarity between users based on movie ratings.

For the Netflix dataset, we use the following constants for all users' profiles: $\phi = 0.7$, $\mu_{actor} = 0.1$, $\mu_{director} = 0.1$, $\mu_{genre} = 0.8$. For both MovieLens datasets, we use: $\phi = 0.1$, $\mu_{actor} = 0.1$, $\mu_{director} = 0.1$, $\mu_{genre} = 0.6$. For $u \in \mathcal{U}$, $\forall v \in \mathcal{U}$ with $u \neq v$, $\omega_v^u = 0$ if v is not one of 20 most similar users to u.

Similarly to (Rago, Cocarascu, and Toni 2018), we consider predicted ratings differing from an actual rating by 1 star to be correctly predicted in order to accommodate the

[5]The code and the datasets to replicate these experiments are available at: https://github.com/CLArg-group/KR2020-Aspect-Item-Recommender-System

[6]Note that the RS is applied in the movie context in this paper but it may be deployed in, and indeed is well-suited to, other contexts, e.g. on e-commerce or music streaming platforms.

[7]Note that our method takes only minutes to compute predicted ratings for these datasets.

[8]https://www.netflixprize.com/

	Model	A	MAE	RMSE	F_1 (P/R)
Netflix (86180 pairs)	KNN	85%	0.78	1.06	**89%** (81%/**98%**)
	KNNZ	85%	0.78	1.07	88% (83%/93%)
	SVD	**88%**	**0.73**	**1.01**	**89%** (82%/**98%**)
	NMF	84%	0.80	1.09	87% (83%/91%)
	Slope1	85%	0.78	1.07	88% (83%/94%)
	CoClust	84%	0.80	1.09	87% (**84%**/91%)
	A-I* RS	77%	0.97	1.32	83% (82%/84%)
	Ours	86%	0.78	1.05	**89%** (82%/97%)
MovieLens 100K (23210 pairs)	KNN	83%	0.83	1.11	**92%** (86%/**99%**)
	KNNZ	83%	0.82	1.12	**92%** (**87%**/97%)
	SVD	**87%**	**0.76**	**1.04**	**92%** (**87%**/99%)
	NMF	83%	0.84	1.13	**92%** (**87%**/97%)
	Slope1	84%	0.82	1.10	**92%** (**87%**/98%)
	CoClust	82%	0.84	1.13	91% (**87%**/96%)
	Ours	84%	0.84	1.10	**92%** (**87%**/98%)
MovieLens Dev. (16060 pairs)	KNN	79%	0.77	1.00	93% (89%/**98%**)
	KNNZ	80%	0.74	0.99	93% (**90%**/97%)
	SVD	**82%**	**0.73**	**0.96**	**94%** (**90%**/**98%**)
	NMF	79%	0.76	0.99	93% (**90%**/96%)
	Slope1	80%	0.74	0.99	**94%** (**90%**/97%)
	CoClust	79%	0.77	1.00	93% (**90%**/96%)
	Ours	77%	0.82	1.03	93% (**90%**/96%)

Table 3: Experimental results for the three datasets with the number of user-item pairs in the testing datasets indicated.

variations seen in this type of subjective rating. In Table 3 we report **A**ccuracy (i.e. number of correct predictions compared to the total number of predictions) as well as standard RS performance measures (Silveira et al. 2019): **M**ean **A**bsolute **E**rror, **R**oot **M**ean **S**quared **E**rror given the 1-5 star scale, and **P**recision, **R**ecall, and $\mathbf{F}_1$ given the binary scale: a rating greater or equal to 3 is considered to be positive, whereas a rating less than 3 is considered to be negative.

The experiments show that our RS is competitive with standard baselines, as well as the **A-I*** **RS**. Our results differ from those of (Rago, Cocarascu, and Toni 2018) as we used more user-rating pairs for testing. Despite effectiveness not being our main focus, it is encouraging to see that it is not sacrificed in place of explainability by our method. Amongst the baselines, KNN may also be deemed to be explainable, but our focus is on argumentative explanations, which, to the best of our knowledge, have not been defined for KNN.

6.2 Stage 2 - Theoretical Analysis

We now answer RQ2 by performing a theoretical analysis showing that the IEs affect the predicted ratings intuitively and desirably. Note that an experimental evaluation of scrutability would have required a complex and extensive user study (e.g. to obtain statistically significant changes in the predicted ratings) and so we leave this for future work.

The following corollary of Proposition 1 shows that IE1's system-changes are guaranteed to affect not only i_0's predicted rating intuitively, i.e. reducing/increasing it in case ↓/↑, resp., (this could indeed be achieved with a corrected rating on i_0, inherent in the identification of the discrepancy), but also the predicted ratings of other unrated items which hold the adjusted aspect.

Corallary 1. *Increasing the predicted rating of some $a \in \mathcal{A}$ such that $\mathcal{P}'_{\mathcal{A}}(a) > \mathcal{P}_{\mathcal{A}}(a)$, increases the predicted rating of any $i \in \mathcal{I}$ where $a \in \mathcal{L}^-(i) \cup \mathcal{L}^+(i)$, i.e. $\mathcal{P}'_{\mathcal{I}}(i) > \mathcal{P}_{\mathcal{I}}(i)$.*

We now show that the changes in the predicted ratings for IE2 and IE3 are intuitive. Once again, a corrected rating on i_0 would rectify the discrepancy in the ratings and so we again show that the system-changes also affect the predicted ratings of other items intuitively. The next proposition guarantees that system-changes $c_1^{\downarrow}$ and $c_1^{\uparrow}$ (where superscripts indicate cases), which relate to the user specific constant for t_{max} and t_{min}, resp., reduce its weighted contribution towards any unrated item's predicted rating.

Proposition 2. *System-change $c_1^{\downarrow}$ ($c_1^{\uparrow}$) guarantees that for any $i \in \mathcal{I}$ where $\mathcal{R}(u,i)$ is not defined, the weighted contribution of t_{max} (t_{min}, resp.) towards i's predicted rating is reduced, i.e. $|\mu_{t_{max}} c^i_{t_{max}}{}'| < |\mu_{t_{max}} c^i_{t_{max}}|$ ($|\mu'_{t_{min}} c^i_{t_{min}}{}'| < |\mu_{t_{min}} c^i_{t_{min}}|$, resp.).*

Proof. (Sketch) For $c_1^{\downarrow}$, the system-change is $\mu'_{t_{max}} < \mu_{t_{max}}$. By Defs. 3 and 2, $\mu_{t_{max}}$ only affects $\mathcal{P}_{\mathcal{I}}$, thus $c^i_{t_{max}}{}' = c^i_{t_{max}}$, giving $|\mu'_{t_{max}} c^i_{t_{max}}{}'| < |\mu_{t_{max}} c^i_{t_{max}}|$. The proof for $c_1^{\uparrow}$ is similar. □

The following corollary of Proposition 1 shows that system-changes $c_2^{\downarrow}$ and $c_2^{\uparrow}$, which relate to a_{max} and a_{min}, resp., affect all unrated items that hold the aspect intuitively, i.e. reducing or increasing, resp., their predicted ratings.

Corallary 2. *System-change $c_2^{\downarrow}$ ($c_2^{\uparrow}$) guarantees that for any $i \in \mathcal{I}$ where $a_{max} \in \mathcal{L}^+(i)$ ($a_{min} \in \mathcal{L}^-(i)$, resp.), i's predicted rating is reduced (increased, resp.), i.e. $\mathcal{P}'_{\mathcal{I}}(i) < \mathcal{P}_{\mathcal{I}}(i)$ ($\mathcal{P}'_{\mathcal{I}}(i) > \mathcal{P}_{\mathcal{I}}(i)$, resp.).*

The following corollary of Proposition 1 shows that system-changes $c_3^{\downarrow}$ and $c_3^{\uparrow}$, which relate to i_{max} and i_{min} resp., affect all unrated aspects held by the item intuitively, i.e. reducing or increasing, resp., their predicted ratings.

Corallary 3. *System-change $c_3^{\downarrow}$ ($c_3^{\uparrow}$) guarantees that for any $a \in \mathcal{A}$ where $i_{max} \in \mathcal{L}^+(a)$ ($i_{min} \in \mathcal{L}^-(a)$, resp.), a's predicted rating will be reduced (increased, resp.), i.e. $\mathcal{P}'_{\mathcal{A}}(a) < \mathcal{P}_{\mathcal{A}}(a)$ ($\mathcal{P}'_{\mathcal{A}}(a) > \mathcal{P}_{\mathcal{A}}(a)$, resp.).*

The final proposition guarantees that system-changes $c_4^{\downarrow}$ and $c_4^{\uparrow}$, which relate to ϕ, cause a change such that, the effects of attackers (supporters) for which similar users' ratings are being used, on the predicted ratings of unrated aspects that they attack (support, resp.) is diminished.

Proposition 3. *System-changes $c_4^{\downarrow}$ and $c_4^{\uparrow}$ guarantee that for any $i \in \mathcal{I}$, where $\rho(i)$ is defined, and any $a \in \mathcal{A}$, where $i \in \mathcal{L}^-(a) \cup \mathcal{L}^+(a)$ and $\mathcal{R}(u,a)$ is not defined, the contribution of i to the predicted rating of a is reduced, i.e. $|\phi' \rho'(i)| < |\phi \rho(i)|$.*

Proof. (Sketch) For $c_4^{\downarrow}$ and $c_4^{\uparrow}$, the system-change is $\phi' < \phi$. By Defs. 3 and 2, ϕ only affects the predicted aspect rating, thus $\rho'(i) = \rho(i)$, giving $|\phi' \rho'(i)| < |\phi \rho(i)|$. □

Thus, all system-changes enact some intuitive alteration in the predicted ratings. We then posit that the more a user interacts with our RS, the more the predicted ratings will

align with their preferences and the RS is hence scrutable. The changes also accommodate dynamic preferences (e.g. if a user's rating on an aspect or item shifts over time then c_2 or c_3, resp., may capture these shifts accordingly), an important phenomenon in RSs (Chen, Zhang, and Qin 2019).

6.3 Stage 3 - Experimental Analysis

The final stage of our evaluation comprises a preliminary user study (again in the movie domain) using crowdsourcing in which RQ3, RQ4 and RQ5 are considered. We experiment with methods for assessing the transparency, trust and user satisfaction of the RS. In general, there are many challenges with RS user studies, for example, it is difficult to measure user experience without allowing users to consume items (Loepp et al. 2018). However, depending on the domain and the information provided, participants sometimes seem to approximate the actual value of recommendations, e.g. movies, reasonably well (Loepp et al. 2018). Even so, when interacting with users via crowdsourcing, replicating the real world situation for recommendation is difficult to achieve as the consequences of decisions are minimal and there is an incentive to complete tasks quickly, whereas in the real-world, users are genuinely interested in the RS output and, possibly, the mechanism that produces it. This incentive to complete tasks quickly may give an advantage to IE1 and IE2 over IE3 in our setup, since they provide a similar amount of information in a concise format, leading to faster, though perhaps less natural, interactions.

We used Amazon's *Mechanical Turk* to conduct an experiment where 75 users were asked to rate movies on a scale of 1-5 stars, along with the option that they have not seen the movie. Once we had presented 70 movies to each user, we calculated recommendations for the users who had rated at least five movies. Each user was then presented with three positive recommendations, i.e. those with the highest predicted ratings, and three negative recommendations, i.e. movies with the lowest predicted ratings, calculated using our RS from Section 3. (We required at least three positive/negative ratings to show positive/negative recommendations, resp.) For these positive and negative recommendations, each user was informed that the RS was inferring that she/he likes or dislikes, resp., those movies. Users were then asked to show their level of agreement with each of the inferences. If a user showed disagreement, i.e. a discrepancy between the predicted ratings and their own, we offered two types of IE among IE1-IE3 for that recommendation; we had 51 such occurrences, as 23 users did not rate enough movies or indicated that there was no discrepancy and one user gave nonsensical responses. We performed two pairwise comparisons among IE1 vs. IE2 and IE2 vs. IE3, thus varying: the information in the IE (i.e. utilising depth vs. width in the argumentation explanations) and the way it was provided (i.e. statically vs. dynamically), counterbalancing the order of presentation to avoid order effects. We therefore showed: 17 users IE1 then IE2, 15 users IE2 then IE1, 9 users IE2 then IE3 and 10 users IE3 then IE2.

Concerning RQ3, our aim was to evaluate participants' *perceived* and *actual* understanding of the RS, whether the IEs improved upon these aspects and if the type of IE had any effect. To assess transparency in this way, we asked users before and after their first IE (so only one IE's effect was considered) how they thought the RS worked, with nine selectable options amounting to (responses condensed for lack of space): one correct response: *from similar users' and my own ratings, the RS infers how much I like aspects to find films I may like*; three partially correct responses: *the RS suggests films similar to those I liked*; *using my ratings, the RS builds my profile to find films I may like*; and *the RS suggests films which were rated highly by other users who gave similar ratings to me*; and five incorrect responses: *the RS suggests films that were the most popular among users*; *the RS randomly selects a few films from a large database containing films' titles, posters, directors and actors*; *the RS suggests films that have similar plot to those I liked*; *the RS suggests films with posters similar to those I liked*; and *I don't know*. Figure 3 shows for all types of IE there were significant increases in the number of users who selected the correct response (p<.001), supporting the claim that our IEs aid transparency, but that there was little to discern between the types of IE in this measure (p>.720). A direct comparison between static and dynamic explanations shows no difference in perceived understanding ($t(19)$=1.14, p=.267). Although the textual explanation appears significantly higher than the tabular one ($t(31) = 2.09$, $p < .045$), this may be a random result due to multiple comparisons. For perceived transparency, we asked users if they felt like they understood how the RS worked before and after the IEs but there was no statistically significant change here (p>.547).

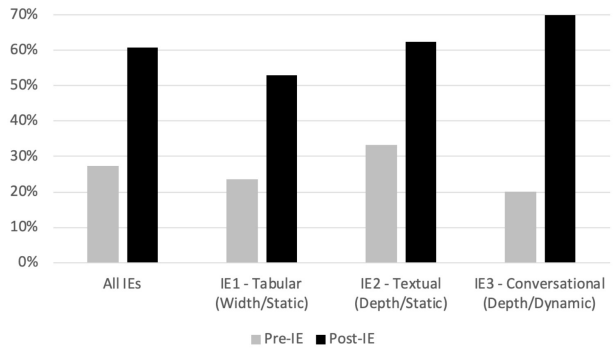

Figure 3: RQ3 results showing the proportion of users giving correct responses when asked how they think the RS works. Note that the sample size varied, with IE3 being the smallest.

For RQ4 and RQ5, we wanted to evaluate how the (types of) IEs affected users' trust in and satisfaction with the RS. Across explanation types there was a significant increase in trust ratings from before the explanation was presented (but after the recommendations were given) to after it was presented ($t(51) = 2.46$, $p = .017$). As before, we observed no significant advantage for any explanation type (p>.122). We also asked the users at the end of the test which they preferred of the two IEs (along with an option of no preference) wrt trust and user satisfaction, which gave direct pairwise comparisons between depth/width of information in the explanation and its static/dynamic delivery. For the two RQs,

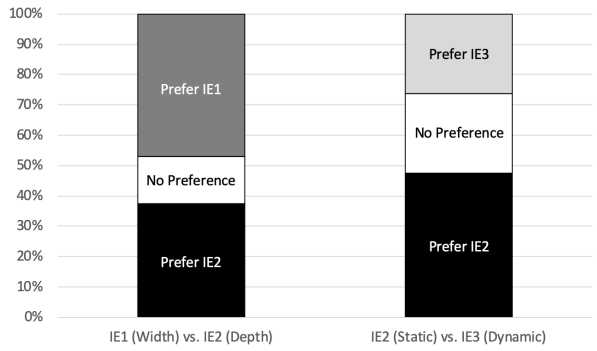

Figure 4: RQ5 results from the pairwise comparisons.

users for the most part gave the same response for both questions, with only 10% differing; due to space limitations we therefore only show the results for RQ5. (The order in which IEs were delivered to the users had little effect on these results and so this was ignored.) Figure 4 shows there was a slight preference for IE1 over IE2, i.e. width over depth in the argumentation explanation, and (somewhat surprisingly) a preference for IE2 over IE3, i.e. static, textual explanations over those which are dynamic and conversational. The sample size was fairly small but feedback from the users regarding the latter comparison referred to a preference for having all the information up front. However, the most clear finding was that the users' preferences varied significantly regarding width/depth in the IEs and their static/dynamic nature, highlighted by the fact that few users opted for the no preference option. An interesting point was that in the 19 instances of IE3, 18 included two or more user interactions, so the preference for non-conversational IEs was not due to a lack of interaction. This also shows that despite the incentive to complete the tasks quickly, conversational explanations actually induced more interaction with users.

7 Related Work

Several argumentation-based RSs have been devised given the natural amenability of argumentation for representing human-like reasoning. Some (e.g. (Briguez et al. 2014)) use Defeasible Logic Programming (García and Simari 2004). In the hybrid RS of (Bedi and Vashisth 2015), recommendations are *repaired* using argumentation to align with user preferences, giving adaptive recommendations. In (Rodríguez et al. 2017), a hybrid, rule-based RS uses argumentation to differentiate between different recommendation techniques, which is shown to outperform other hybridisation methods. A formalisation for explanations based on Toulmin's model of argumentation (Toulmin 1958) is given in (Naveed, Donkers, and Ziegler 2018), before being examined thoroughly with user studies showing that different levels of argumentation explanation are most acceptable for different users, an important motivation for our work.

RS explanations which most closely align with ours from a structural point of view are those utilising knowledge graphs, e.g. those used by (Xian et al. 2019; Wang et al. 2019), which are similar to A-I frameworks but also include user nodes and heterogeneous relations. More flexibility is therefore afforded with the scope of the explanations, but there is no argumentative reasoning underpinning the RSs.

Many neural methods exist for generating conversational interactions in RSs, e.g. (Sun and Zhang 2018; Zhang et al. 2018). One such method, *Vote Goat* (Dalton, Ajayi, and Main 2018), uses *Dialogflow* to converse with users, which would be interesting to adapt to our method. Other conversational RSs include that of (Sepliarskaia et al. 2018), which uses a *static preference questionnaire* to avoid cold-start problems and to elicit user preferences, where the selection of questions is treated as an optimisation problem. The method of (Balog, Radlinski, and Arakelyan 2019) is a tag-based approach to recommendations equipped with explanations which satisfy predetermined rules and allow users to correct the explanations. Compared with the state-of-the-art, effectiveness is comparable but somewhat sacrificed in place of greater transparency and scrutability. Finally, templated and neural methods are combined in (Aliannejadi et al. 2019) by generating questions offline with users before using a neural model for question selection, showing that increasing user interaction gives better recommendations.

8 Conclusions

We have introduced an RS, adapted from, but considerably modifying, that of (Rago, Cocarascu, and Toni 2018), which is not only competitive with regards to recommendation accuracy (effectiveness, proven empirically), but is also amenable to the extraction of argumentation abstractions that act as scaffolding supporting various forms of explanations with which users can interact (IEs). This benefits both the user and the RS as it empowers feedback about the user's preferences to be accommodated to improve recommendations, making the RS scrutable (proven theoretically). We then undertook a user study, from which we drew the following tentative conclusions, requiring further studies:

- Argumentation-based IEs of various types often improve the transparency and trust in an RS (though users didn't always perceive this to be the case for transparency).
- Despite conversational IEs consistently inducing interactions with users, users expressed no explicit preference for this type over receiving all of the information statically, raising questions about the benefits of conversational explanations (at least when they are delivered as text as opposed to being spoken, e.g. by an AI assistant).
- Users showed diverse preferences for the content of explanations and the manner in which it is delivered, highlighting the importance of supporting various styles of explanation.

This paper lays the groundwork for future studies on assessments of types of IEs and their effect on users' RS experiences. We plan to study more variants of IEs, e.g. expanding user control over the RS and thus its adaptability to their preferences or developing more natural conversational protocols spoken by an AI assistant, and deeper analysis, e.g. focusing on the levels of interaction induced. It would also be interesting to deploy the RS in real world tasks where the recommendations matter to users, not to mention in other suitable contexts, e.g. music streaming or e-commerce.

Acknowledgments

This research was partially funded by the UK Human-Like Computing EPSRC Network of Excellence. Rago, Cocarascu and Toni were also partially funded by the UK EPSRC project EP/P029558/1 ROAD2H, and Rago by an EPSRC Doctoral Prize Fellowship at Imperial College London, UK.

References

Abdul, A. M.; Vermeulen, J.; Wang, D.; Lim, B. Y.; and Kankanhalli, M. S. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, 582.

Aliannejadi, M.; Zamani, H.; Crestani, F.; and Croft, W. B. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019.*, 475–484.

Amgoud, L.; Ben-Naim, J.; Doder, D.; and Vesic, S. 2017. Acceptability semantics for weighted argumentation frameworks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, 56–62.

Balog, K.; Radlinski, F.; and Arakelyan, S. 2019. Transparent, scrutable and explainable user models for personalized recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019.*, 265–274.

Baroni, P.; Rago, A.; and Toni, F. 2018. How many properties do we need for gradual argumentation? In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*.

Baroni, P.; Rago, A.; and Toni, F. 2019. From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. *International Journal of Approximate Reasoning* 105:252–286.

Bedi, P., and Vashisth, P. B. 2015. Argumentation-enabled interest-based personalised recommender system. *Journal of Experimental and Theoretical Artificial Intelligence* 27(2):199–226.

Briguez, C. E.; Budán, M. C.; Deagustini, C. A. D.; Maguitman, A. G.; Capobianco, M.; and Simari, G. R. 2014. Argument-based mixed recommenders and their application to movie suggestion. *Expert Systems with Applications* 41(14):6467–6482.

Cath, C.; Wachter, S.; Mittelstadt, B. D.; Taddeo, M.; and Floridi, L. 2018. Artificial intelligence and the 'good society': the US, EU, and UK approach. *Science and Engineering Ethics* 24(2):505–528.

Cayrol, C., and Lagasquie-Schiex, M.-C. 2005. On the acceptability of arguments in bipolar argumentation frameworks. In *Proceedings of the 8th European conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, 378–389. Springer Berlin Heidelberg.

Chen, X.; Zhang, Y.; and Qin, Z. 2019. Dynamic explainable recommendation based on neural attentive models. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.*, 53–60.

Chesñevar, C. I.; Maguitman, A. G.; and González, M. P. 2009. Empowering recommendation technologies through argumentation. In *Argumentation in Artificial Intelligence*. 403–422.

Cocarascu, O.; Rago, A.; and Toni, F. 2019. Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, 1261–1269.

Cyras, K.; Birch, D.; Guo, Y.; Toni, F.; Dulay, R.; Turvey, S.; Greenberg, D.; and Hapuarachchi, T. 2019a. Explanations by arbitrated argumentative dispute. *Expert Syst. Appl.* 127:141–156.

Cyras, K.; Letsios, D.; Misener, R.; and Toni, F. 2019b. Argumentation for explainable scheduling. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.*, 2752–2759.

Dacrema, M. F.; Cremonesi, P.; and Jannach, D. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019*, 101–109.

Dalton, J.; Ajayi, V.; and Main, R. 2018. Vote goat: Conversational movie recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, 1285–1288.

García, A. J., and Simari, G. R. 2004. Defeasible logic programming: An argumentative approach. *TPLP* 4(1-2):95–138.

George, T., and Merugu, S. 2005. A scalable collaborative filtering framework based on co-clustering. In *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005), 27-30 November 2005, Houston, Texas, USA*, 625–628.

Harper, F. M., and Konstan, J. A. 2015. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems* 5(4):19:1–19:19.

Hug, N. 2017. Surprise, a Python library for recommender systems. http://surpriselib.com.

Lemire, D., and Maclachlan, A. 2007. Slope one predictors for online rating-based collaborative filtering. *CoRR* abs/cs/0702144.

Loepp, B.; Donkers, T.; Kleemann, T.; and Ziegler, J. 2018. Impact of item consumption on assessment of recommendations in user studies. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, 49–53.

Luo, X.; Zhou, M.; Xia, Y.; and Zhu, Q. 2014. An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems. *IEEE Transactions on Industrial Informatics* 10(2):1273–1284.

Madumal, P.; Miller, T.; Sonenberg, L.; and Vetere, F. 2019. A grounded interaction protocol for explainable artificial intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, 1033–1041.

McInerney, J.; Lacker, B.; Hansen, S.; Higley, K.; Bouchard, H.; Gruson, A.; and Mehrotra, R. 2018. Explore, exploit, and explain: personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, 31–39.

Naveed, S.; Donkers, T.; and Ziegler, J. 2018. Argumentation-based explanations in recommender systems: Conceptual framework and empirical results. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, UMAP 2018, Singapore, July 08-11, 2018*, 293–298.

Pu, P., and Chen, L. 2007. Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems* 20(6):542–556.

Rader, E. J.; Cotter, K.; and Cho, J. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, 103.

Rago, A.; Cocarascu, O.; and Toni, F. 2018. Argumentation-based recommendations: Fantastic explanations and how to find them. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, 1949–1955.

Resnick, P., and Varian, H. R. 1997. Recommender systems. *Communications of the ACM* 40(3):56–58.

Rodríguez, P.; Heras, S.; Palanca, J.; Poveda, J. M.; Duque, N. D.; and Julián, V. 2017. An educational recommender system based on argumentation theory. *AI Communications* 30(1):19–36.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5):206.

Sepliarskaia, A.; Kiseleva, J.; Radlinski, F.; and de Rijke, M. 2018. Preference elicitation as an optimization problem. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, 172–180.

Silveira, T.; Zhang, M.; Lin, X.; Liu, Y.; and Ma, S. 2019. How good your recommender system is? A survey on evaluations in recommendation. *International Journal of Machine Learning & Cybernetics* 10(5):813–831.

Simari, G. R., and Rahwan, I., eds. 2009. *Argumentation in Artificial Intelligence*. Springer.

Sun, Y., and Zhang, Y. 2018. Conversational recommender system. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, 235–244.

Tintarev, N., and Masthoff, J. 2015. Explaining recommendations: Design and evaluation. In *Recommender Systems Handbook*. 353–382.

Toulmin, S. E. 1958. *The uses of argument*. Cambridge University Press (Cambridge).

Vig, J.; Sen, S.; and Riedl, J. 2009. Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th International Conference on Intelligent User Interfaces, IUI 2009, Sanibel Island, Florida, USA, February 8-11, 2009*, 47–56.

Wang, X.; He, X.; Cao, Y.; Liu, M.; and Chua, T. 2019. KGAT: knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019.*, 950–958.

Xian, Y.; Fu, Z.; Muthukrishnan, S.; de Melo, G.; and Zhang, Y. 2019. Reinforcement knowledge graph reasoning for explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019.*, 285–294.

Zhang, Y.; Chen, X.; Ai, Q.; Yang, L.; and Croft, W. B. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, 177–186.