Artificial Intelligence 296 (2021) 103506

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Argumentative explanations for interactive recommendations [☆]

Antonio Rago [a,*], Oana Cocarascu [b], Christos Bechlivanidis [c], David Lagnado [c], Francesca Toni [a]

[a] Department of Computing, Imperial College London, UK
[b] Department of Informatics, King's College London, UK
[c] Department of Experimental Psychology, University College London, UK

ARTICLE INFO

Article history:
Received 17 April 2020
Received in revised form 11 March 2021
Accepted 15 April 2021
Available online 21 April 2021

Keywords:
Argumentation
Explanation
User interaction
Recommender systems
User evaluation

ABSTRACT

A significant challenge for recommender systems (RSs), and in fact for AI systems in general, is the systematic definition of explanations for outputs in such a way that both the explanations and the systems themselves are able to adapt to their human users' needs. In this paper we propose an RS hosting a vast repertoire of explanations, which are customisable to users in their content and format, and thus able to adapt to users' explanatory requirements, while being reasonably effective (proven empirically). Our RS is built on a graphical chassis, allowing the extraction of argumentation scaffolding, from which diverse and varied *argumentative explanations* for recommendations can be obtained. These recommendations are interactive because they can be questioned by users and they support adaptive feedback mechanisms designed to allow the RS to self-improve (proven theoretically). Finally, we undertake user studies in which we vary the characteristics of the argumentative explanations, showing users' general preferences for more information, but also that their tastes are diverse, thus highlighting the need for our adaptable RS.

1. Introduction

Recommender systems (RSs) [55] aim to help users discover items that may be of interest, most often by ranking or predicting ratings for them [2]. The most widely used types of methods for RSs are 'collaborative filtering' (considering similarities between users or items to determine recommendations for users), 'content-based filtering' (operating on information about items, e.g. their features) and those which are 'knowledge-based' (using constraints obtained as user requirements), while 'hybrid' methods use a combination thereof. These methods are able to navigate vast datasets in a way which would never be possible for human users alone, so long as we are happy to entrust them with the task of item recommendation. However, these systems suffer from scalability, data sparsity, 'cold start' problems and, most importantly in the context of this paper, a lack of explanations for their recommendations. The last problem is an issue because if the reasons behind recommendations are not explained to users then they may be unable to provide effective feedback to the RS to help it *adapt* to the users' preferences, which in turn may cause users' unwillingness to follow the recommendations in the future.

[☆] This paper is part of the Special Issue on Explainable AI.

[*] Corresponding author.

E-mail addresses: a.rago15@imperial.ac.uk (A. Rago), oana.cocarascu@kcl.ac.uk (O. Cocarascu), c.bechlivanidis@ucl.ac.uk (C. Bechlivanidis), d.lagnado@ucl.ac.uk (D. Lagnado), ft@imperial.ac.uk (F. Toni).

https://doi.org/10.1016/j.artint.2021.103506

Indeed, transparency of RSs has been shown to be strongly linked to user satisfaction [33]. The lack of transparency has been exacerbated in the recent past by a trend towards ever more complex models in RSs [27], with explanatory aspects of systems being somewhat neglected. This problem is not specific to RSs alone, and is shared with AI in general [57].

There has of late been a drive towards explainability in AI from academia, industry and government.[1] Notwithstanding these efforts, empowering explainability is no simple task. One first issue is the trade-off between accuracy and explainability [29], and some, e.g. Balog et al. [8], have posited that it is worth sacrificing the former (somewhat) in place of the latter. Another issue is the fact that there is no general solution for what makes a *good* explanation, with considerations including, but not limited to, the information it should include and how it is delivered to users (see [62] for a taxonomy). Indeed, different users may require and benefit from different forms of explanations, and thus explanations need to be adaptable to the human users to whom they are directed [69,49]. A further issue is whether explanations can empower users to interact with the system, in particular to provide feedback on outputs by the system. The social sciences clearly indicate that explanation is a social process for humans [49] and human explanations for recommendations are still seen to be of better quality than machine-generated explanations and therefore inspire more trust in the users, though making the generated explanations richer may alleviate some of this deficiency [44]. Also, recent work in RSs advocates the capability of integrating feedback as crucial for user trust [8]. We aim to address some of these concerns here, by defining an adaptable, hybrid RS which is equipped with explanations that are: faithful to the method for calculating recommendations, customisable to diverse explanatory requirements and amenable to elicit feedback from humans within interactions aimed at gaining users' trust and improving the RS, while being reasonably effective.

In this paper we give a hybrid *Aspect-Item* RS (overviewed in Fig. 1), named this way because it relies upon an underpinning *graphical chassis* linking items and their aspects (or properties/features). This chassis houses the information for recommendations and is the basis for the *argumentative scaffolding* from which *argumentative explanations* (of various kinds and formats) are automatically generated to support interactive recommendations for users, including the opportunity for giving feedback on recommended items and their aspects. The recommendations result from a hybrid method for calculating predicted ratings from ratings given by the user and by similar users. The argumentative scaffolding amounts to *tripolar argumentation frameworks* (TFs), in the spirit of argumentation in AI (see [6,10] for recent overviews), but extending classical abstract [30] and bipolar [20] argumentation frameworks by including a 'neutralising' relation (labelled 0) in addition to the standard 'attack' (labelled -) and 'support' (labelled +) relations. These relations are extracted so that they meet logical requirements based on how predicted ratings for items and aspects affect one another. This synchronisation between the RS and its explanations is a crucial advantage of our approach, since users can trust that the explanations describe *how* recommendations were generated. Argumentative explanations include, amongst others, conversational and visual explanations. These explanations form the basis for interactions with users to explain recommendations and receive feedback that can be accommodated into the RS to improve its behaviour. Thus, not only are our explanations varied and diverse, but they also account (in a limited sense) for adaptable recommendations over time.

Concretely, we map the graphical chassis underpinning our RS onto user-tailored TFs determined by predicted ratings for the user, giving a dialectical interpretation of the factors influencing a recommendation, and taking advantage of argumentation's natural amenability for representing human-like reasoning. To extract the TFs, we understand the predicted ratings computed by our RS as a *gradual semantics* for the TF, exhibiting a desirable property of *weak monotonicity*, which can thus be seen as the driving force behind the extraction of the TF, in the spirit of Baroni et al. [11], but for a novel form of argumentation framework and for a novel property; these choices of framework and property lead to an argumentative scaffolding faithful to the underlying RS. We illustrate how the user-tailored TFs can then be used to generate a range of explanations for recommendations by our RS, varying in their characteristics. We show by means of illustrations that these explanations can be constructed to elicit feedback in user interactions, leading to positive, intuitive effects on the quality of future recommendations, building on the effectiveness of the recommendations prior to interactions; we assess effectiveness empirically, on a number of publicly available datasets, in comparison with various baseline algorithms. Finally, we provide user evaluations of various forms of explanation from our RS, and show that users' preferred characteristics in an explanation vary, justifying the flexibility and customisability naturally afforded by our method.

Our contributions can be summarised as follows:

- We define a novel, hybrid RS which utilises argumentation technology to provide customisable explanations of recommendations while being reasonably effective in making recommendations, in comparison with publicly available RSs lacking explanatory capabilities; argumentation plays a crucial role in our RS, in that argumentation frameworks (in the form of TFs) provide a faithful counterpart of the RS (in the spirit of what is advocated by Ignatiev [40]) from which relevant information can be harboured to systematically obtain explanations in a wide range of styles.
- We showcase various forms of explanations that can be drawn from the underpinning argumentative scaffolding (i.e. from the TFs), emphasising how their content, format and feedback mechanisms can be varied to satisfy variations in user preferences.

[1] For example, see, respectively, the recent survey [35], IBM's *AI Explainability 360* launched in August 2019, and the European Commission's Ethics Guidelines for Trustworthy AI (8 April 2019), available at https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

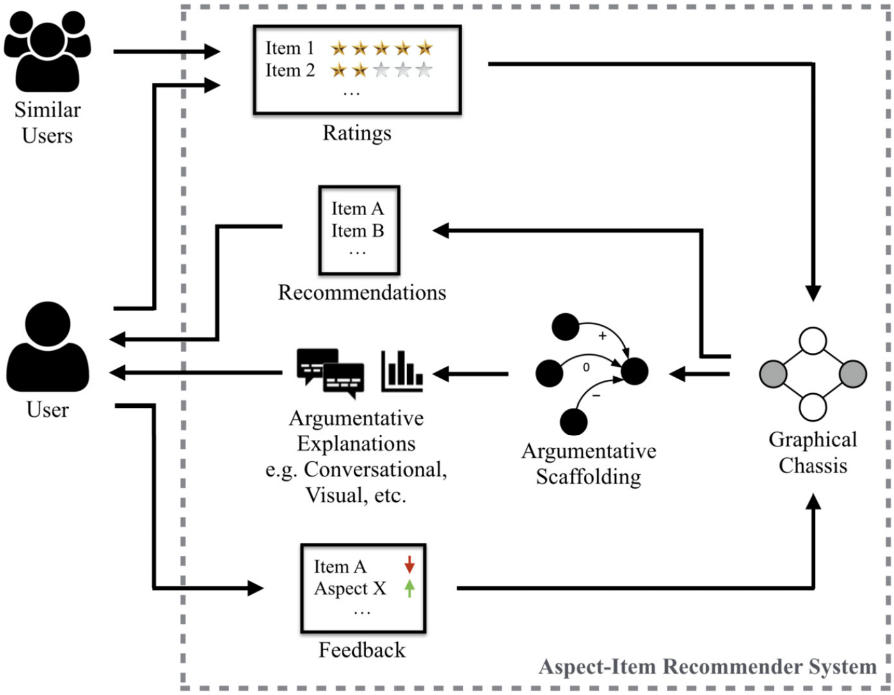

Similar Users
Item 1
Item 2
...
Ratings
Item A
Item B
...
Recommendations
User
Argumentative Explanations e.g. Conversational, Visual, etc.
+
0
−
Argumentative Scaffolding
Graphical Chassis
Item A
Aspect X
...
Feedback
Aspect-Item Recommender System

Fig. 1. Overview of our hybrid RS.

- We conduct user studies on take-up of some of these forms of explanation, showing that users prefer more information in explanations but also that their preferences on explanatory format are diverse.
- We provide a novel perspective on the use of argumentation in AI: whereas conventionally *semantics* for argumentation frameworks and *properties* thereof are chosen post-hoc, in order to analyse the frameworks and the conclusions drawn from them, we start with the choice of semantics (the RS's predicted ratings) and use properties as a driving force for obtaining argumentation frameworks suitable as a basis for explanations.

The paper is organised as follows. In Section 2 we provide background on (explainable) RSs, argumentation as understood in AI, as well as existing work on using argumentation to provide explanations for RSs. In Section 3 we define the graphical chassis on which our RS is based and provide our method for calculating predicted ratings therein, along with some theoretical analysis of the predicted ratings' behaviour. In Section 4 we assess the RS's effectiveness empirically (in comparison with a number of baselines), in a popular domain for RSs (movie recommendations), with various datasets: the Netflix challenge dataset[2] as reported in [54], the MovieLens Development dataset and the MovieLens 100K benchmark dataset [36]. In Section 5, we map instances of the RS (and thus the recommendations thereof) onto user-tailored TFs, and show theoretically that they make sense from an argumentation viewpoint, represent the predicted ratings faithfully and indicate intuitive changes that can be made to the RS via interactions with users. Thus, this section provides solid theoretical underpinnings for our explanations, and the basis for their high customisability. In Section 6 we then give illustrations of the varying characteristics of explanations which may be delivered to users, including the feedback mechanisms with which they may interact, before, in Section 7, providing some experiments on the preferences of users with regards to some of these explanatory characteristics. In Section 8 we conclude, in particular pointing to future work.

This paper builds upon and extends [54] and [53] as follows. Sections 2 and 3 extend the same sections in [54]: we have added discussions, proofs and examples throughout. In particular, in Section 3 we have added a theoretical analysis of the RS with Propositions 1 and 2. Section 4 is mostly new and includes empirical analysis of the RS's performance with three datasets. Section 5 significantly extends the same section in [54] with further theoretical analysis of the TFs' properties: Definition 8, Propositions 3, 4 and 5, and the corresponding discussions are all new. Section 6 is completely new. Finally, Section 7 is mostly new, with the second user study adapted from [53].

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

2. Background

2.1. Recommender systems

Common methods used for making recommendations in RSs are 'latent factor models' and 'neighbourhood models' between items or users. Latent factor models, based on matrix factorization, describe the items as vectors of factors inferred from data. Neighbourhood models have been used to support various collaborative filtering algorithms for RSs. These models include non-negative matrix factorization models [47], Singular Value Decomposition [16,71], Slope One techniques [45], and Co-clustering, a simultaneous clustering of users and items [34]. In addition, collaborative filtering and content-based filtering can be combined to give hybrid models [18,19]. The Netflix Prize competition[3] showed that matrix factorization models are superior to nearest-neighbour models, such as KNN [4], as many of the best performing algorithms in the competition were based on matrix factorization [42,68]. Whilst these models are scalable and effective, they are not easily explainable, as the way they represent factors makes them non-interpretable. These issues are shared with many of the recent advances in deep learning methods for RSs, e.g. [73,37,82]. These methods, while performing well with regards to effectiveness, rely on neural networks and so cannot easily provide explanations as to how recommendations are generated, particularly to lay-users (though some efforts towards providing explanations have been made, e.g. using attention mechanisms [59]). We share the views of Rudin [57], namely: if interpretable models perform acceptably in a task, the additional explainability that they possess may prove to be a worthwhile advantage over black-box methods. We will compare the performance of our RS with some of these existing techniques.

The trend towards explainability in RSs was ahead of that in general AI methods, possibly hastened by RSs' end-user-facing nature. A thorough review of explainable RSs was undertaken by Zhang & Chen [78], in which the authors assess explanations' *types*, i.e. how the explanation is displayed to the user, as well as considering how to evaluate explanations and also their applications. These concerns are a particular focus of this paper, since we introduce a method (based on argumentation) for generating explanations of various types (which we understand as resulting from combinations of different *content* and *format*), satisfying various evaluation measures within the application of movie recommendation. The possible explanations for a particular RS strongly depend on the method for calculating recommendations, particularly as regards explanatory content. For example, content-based RSs are naturally amenable to explanations of a similar nature, e.g. as in [1], and similarly for collaborative filtering RSs, e.g. as in [70]. This is not the case for the format of the explanation, however, as RSs in the same category may adopt wildly different formats. Moreover, explanations in the same format may result from very different methods. For example, textual explanations may be produced via templates utilising underlying features, e.g. as in [80] (as we also do later in the paper), or via neural methods operating on reviews, e.g. as in [25].

Some explainable RSs, e.g. that of Xian et al. [76], deploy knowledge graphs in a form of scaffolding similar to that which we propose, however there is no argumentative reasoning underpinning the predicted ratings for the items. The RS of Wang et al. [74] also uses a *collaborative knowledge graph* (i.e. with information from other users) with attention mechanisms to calculate the recommendations, also allowing explanations to be generated using a subgraph of the knowledge graph. A key difference between these RSs and our approach is that from their explanations it is difficult to say exactly *how* recommendations were made, while our explanations are transparently extracted from the RS. A graphical, method-agnostic method for explaining RSs linguistically is given by Musto et al. [50], which uses a similar framework to ours but with the rationale for recommendations generated independently from the algorithm making the recommendations. Our explanations, meanwhile, adhere fully with the predicted rating calculations and are thus faithful to the underlying RS.

Some of the most popular types of explanation (often equipped with feedback mechanisms) are *conversational*, perhaps driven by the popularity of AI assistants such as Apple's *Siri*. The theoretical framework for conversational search and recommendation of Radlinski & Craswell [52] considers the types of interaction between the user and the system, providing application contexts for each. *Vote Goat* [28] converses with users to provide recommendations via a speech-based natural language interface using *Dialogflow*. The *System Ask, User Respond* system [79] provides conversational recommendations using a multi-memory attention network, giving results once the system has a high confidence in the item. The RS of Sun & Zhang [63] uses a belief tracker to keep track of states with LSTM (Long Short-Term Memory) networks before reinforcement learning is used to determine whether to ask for facet value pairs or to give recommendations. The conversational RS of Sepliarskaia et al. [60] uses a *static preference questionnaire* to avoid cold start problems and to elicit user preferences (cold start is a problem that arises in RSs when sufficient information about users or items is lacking). A similar method [23] achieves a 25% increase in accuracy using only two questions in the restaurant recommendation context. Meanwhile, Balog et al. [8] render recommendations scrutable with rule-based, pairwise explanations where users are allowed to correct each individual clause in an explanation. Finally, Aliannejadi et al. [3] combine templates and neural methods: questions are generated offline with users in a Human Intelligence Task before a neural model is used to select between them in conversational interactions, finding that an increase in user interaction improves recommendations. Our RS admits conversational explanations as a special type, drawn, like other types of explanations, automatically from the same underpinning argumentation scaffolding.

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

Evaluation of (explanations in) RSs is overviewed by Tintarev & Masthoff [65], which also identifies desirable features of RSs (see also [66]), including: *effectiveness*, e.g. increasing a system's accuracy with regards to users' preferences; *transparency*, i.e. explaining how a system works and showing how it predicts ratings; *scrutability*, i.e. allowing feedback based on these explanations; and *trust*, i.e. correcting a system based on user feedback. None of the systems surveyed in [65] fulfilled all four of these aims. Of those which aimed to improve scrutability, the systems in [16,26] both use template-based responses based on factors affecting the recommendation. A study on the relationships between such desirable features is undertaken by Balog & Radlinski [7], highlighting strong dependencies amongst them. Meanwhile, investigations have been performed [38,33] into how users respond to different types of explanation for recommendations. The considered explanations therein are wide ranging: Herlocker et al. [38] found that visual explanations depicting histograms of the number of neighbours with negative, neutral and positive ratings performed best, while tag-based explanations, i.e. those which highlight aspects of recommended items, performed poorly. In the same work, a tabular method giving statistics on ratings of items performed well while simple statistical explanations such as an overall average rating (from all users) did not. Gedikli et al. [33] assess explanation types with respect to the desirable features of RSs of Tintarev & Masthoff [66], showing that different explanation types have different effects on users. For example, their statistical explanation of the overall number of positive ratings was the most efficient (users rated the explanation quickly) but it performed poorly with regards to effectiveness (specifically regarding the error in a user's rating given after the explanation). Indeed, McInerney et al. [48] utilise bandits to determine which explanations users are most receptive to. Our aim in this paper is not to determine which explanations perform best but to show that argumentative scaffolding supports numerous explanation types. Meanwhile, Kulesza et al. [43] assess the importance of the soundness and completeness of explanations, showing that users value both. The importance of soundness motivates our work, since our explanations are faithful to the RS without approximations, while that of completeness (or, at least, the amount of information included in an explanation) will be examined once more in our user study.

With respect to analysis of the feedback capability of RSs, a comparison of the different types of implicit feedback (e.g. skipping a song) is given by Schnabel et al. [58], showing how the feedback quality can effect results. Zhao et al. [81] compare implicit and explicit (e.g. correcting predicted ratings) feedback for 6 different types of RS, showing that implicit feedback engages users more but the effect is mixed as it induces both positive and negative engagements, and that incorporating both implicit and explicit feedback is often optimal. In this paper we only consider explicit feedback. We leave as future work the question as to whether our RS could be made to infer implicit feedback from user actions.

2.2. Argumentation and its use in recommender systems

Abstract argumentation frameworks (AFs) [30] are pairs consisting of a set of arguments $\mathcal{X}$ and a binary relation between arguments $\mathcal{L}^-$, representing attacks. Formally, an AF is any $\langle \mathcal{X}, \mathcal{L}^- \rangle$ where $\mathcal{L}^- \subseteq \mathcal{X} \times \mathcal{X}$. *Bipolar argumentation frameworks* (BFs) [20] extend AFs by considering two separate binary relations between arguments: attack $\mathcal{L}^-$ and support $\mathcal{L}^+$. Formally, a BF is any $\langle \mathcal{X}, \mathcal{L}^-, \mathcal{L}^+ \rangle$ where $\langle \mathcal{X}, \mathcal{L}^- \rangle$ is an AF and $\mathcal{L}^+ \subseteq \mathcal{X} \times \mathcal{X}$. Various other types of argumentation frameworks have been proposed in the literature, including *tripolar frameworks*, as in [31], and *generalised argumentation frameworks*, as in [9], both allowing for additional dialectical relations (in addition to attack and support). The argumentation frameworks we use in the paper for explanation can be seen as a special instance of these latter types of frameworks.

Argumentation frameworks are assigned semantics, which may be given in terms of so-called extensions (as in [30]) or in terms of a gradual *strength function* (as overviewed by Baroni et al. [12]). In this paper, we will treat predicted ratings by our RS as a form of the latter type of semantics. Given an argumentation framework with arguments $\mathcal{X}$, for any $x \in \mathcal{X}$, the *strength* (or gradual evaluation) of x is $\sigma(x)$, where $\sigma : \mathcal{X} \rightarrow \mathbb{I}$ is a *strength function* and $\mathbb{I}$ is set equipped with a preorder $\leq$ [12]. In this paper, we will use $\mathbb{I} = [-1, 1]$.

The requisite behaviour of a suitable strength semantics for a given application may be characterised by properties, e.g. *strict monotonicity* [12], which states that an argument's strength depends monotonically on its base score and on the strengths of its attackers and supporters. We will mention other properties studied in the literature later in Section 5.

Several argumentation-based RSs have been proposed in the literature. For example, some [22,17,64] use *defeasible logic programming* (DeLP) [32] to enhance recommendation technologies with argument-based analysis. DeLP supports defeasible reasoning dialectically, can handle incomplete and contradictory information, and uses a comparison criterion to solve conflicting situations between arguments. Chesñevar et al. [22] model user preferences as facts, strict rules and defeasible rules. Along with background information, user preferences can be used in a DeLP program to make recommendations which are modelled as arguments in favour of or against a particular decision. Teze et al. [64] enhance the argument-based RS of Chesñevar et al. [22] to allow for an argument comparison criterion on users' preferences to be encoded by means of conditional expressions, thus enabling the argument preference criterion to be selected rather than being treated as a fixed component. The movie RS of Briguez et al. [17] relies on a set of predefined postulates describing the conditions under which a movie should be recommended to a given user; these conditions can be translated into DeLP rules. Examples of postulates are *a user may like a movie if the actor of the movie is one of the user's favourites* or *a user may like a movie if the movie is liked by a group of similar users*. Explanations are extracted from the dialectical tree supporting a recommendation. Our argumentation-based explanations, meanwhile, are generated automatically from data without any need for knowledge to be manually incorporated. Bedi & Vashisth [14] define a hybrid RS in which argumentation is used to repair recommendations by correcting rule-based arguments, e.g. 'the actor is popular', in user interactions; this RS can thus be deemed to be adaptive to users' preferences. Note that in our RS arguments are based on whether a user likes, or is predicted to like, items or aspects, rather than rules. Argumentation can also be used to differentiate between techniques in the hybridisation process when making recommendations, as shown by Rodríguez et al. [56], where recommendations from rule-based RSs are combined using argumentation. Finally, Toulmin's model of argumentation [67] is used by Naveed et al. [51] in developing a formalisation for explanations for recommendations, demonstrating mockup explanations without explicitly defining an RS. User studies are then performed showing that different levels of argumentation-based explanation are preferred by different users, providing evidence for the need for argumentative scaffolding supporting customisable explanations in our proposed RS. However, the argument structure in [51] is limited to a chain of evidential reasoning based on the support relation, rather than 'debates' built from three dialectical relations (in TFs) as we do.

3. Predicted ratings in the aspect-item recommender system

We define an RS where *items* (e.g. movies) are associated with *aspects*[4] (e.g. comedy), which in turn have *types* (e.g. genre), and *users* may have provided *ratings* on some of the items and/or aspects. The associations form an underlying *graphical chassis* within the *aspect-item framework* underpinning the RS:

Definition 1. An *aspect-item framework* (A-I in short) is a tuple $\langle \mathcal{I}, \mathcal{A}, \mathcal{T}, \mathcal{L}, \mathcal{U}, \mathcal{R} \rangle$ such that:

- $\mathcal{I}$ is a finite, non-empty set of *items*;
- $\mathcal{A}$ is a finite, non-empty set of *aspects* and $\mathcal{T}$ is a finite, non-empty set of *types* such that for each aspect $a \in \mathcal{A}$ there is a (unique) type $t \in \mathcal{T}$ with t the type of a; for any $t \in \mathcal{T}$, we use $\mathcal{A}_t$ to denote $\{a \in \mathcal{A} |$ the type of a is $t\}$;
- the sets $\mathcal{I}$ and $\mathcal{A}$ are disjoint; we use $\mathcal{X}$ to denote $\mathcal{I} \cup \mathcal{A}$, and refer to it as the set of *item-aspects*;
- $\mathcal{L} \subseteq (\mathcal{I} \times \mathcal{A}) \cup (\mathcal{A} \times \mathcal{I})$ is a symmetric binary relation;
- $\mathcal{U}$ is a finite, non-empty set of *users*;
- $\mathcal{R} : \mathcal{U} \times \mathcal{X} \to [-1, 1]$ is a partial function of *ratings*.

Note that each aspect has a unique type, but of course different aspects may have the same type. Thus, $\mathcal{T}$ implicitly partitions $\mathcal{A}$, by grouping together all aspects with the same type. Note also that we replicate identical aspects for different types, e.g. we consider *Quentin Tarantino* the director to be a separate aspect to *Quentin Tarantino* the actor (justified by the fact that they may be rated differently). Finally, we assume that ratings, when defined, are real numbers in the [-1,1] interval. Other types of ratings can be translated into this format, for example a rating $x \in \{1, 2, 3, 4, 5\}$ can be translated into a rating $y \in [-1, 1]$ using $y = ((x-1)/2) - 1$.[5]

The $\mathcal{I}$, $\mathcal{A}$, $\mathcal{T}$ and $\mathcal{L}$ components of an A-I may be visualised as a graph (thus the term 'graphical chassis' for an A-I), as illustrated in Fig. 2 for the movie domain (we ignore for now the labels of the nodes in the graph, but note that we assume a mapping from elements of $\mathcal{X}$ and nodes).

In the remainder of the paper (unless otherwise specified) we assume as given an arbitrary A-I $\mathcal{F} = \langle \mathcal{I}, \mathcal{A}, \mathcal{T}, \mathcal{L}, \mathcal{U}, \mathcal{R} \rangle$.

Definition 2. The set of *linked item-aspects* of $x \in \mathcal{X}$ is $\mathcal{L}(x) = \{y \in \mathcal{X} | (y, x) \in \mathcal{L}\}$. We also use $\mathcal{L}_t(i)$, for $i \in \mathcal{I}$, to denote $\{a \in \mathcal{L}(i) | a \in \mathcal{A}_t\}$.

For the example shown in Fig. 2, the set $\mathcal{L}_{actor}$(Catch Me If You Can) comprises Leonardo DiCaprio and Tom Hanks.

The primary goals of RSs [2], formulated for A-Is, are: (i) Prediction - for a user $u \in \mathcal{U}$, $\forall i \in \mathcal{I}$ such that $\mathcal{R}(u, i)$ is undefined, compute a *predicted rating* $\mathcal{P}^u_{\mathcal{I}}(i)$; and (ii) Ranking - for a user $u \in \mathcal{U}$, compute a *ranking* on $\{i \in \mathcal{I} | \mathcal{R}(u, i)$ is undefined$\}$. In this paper, we focus on prediction, given that rankings or *top-N* recommendations can be derived from predictions. Before giving our method for predicting ratings, we define users' *profiles*.

Definition 3. The *profile* π^u of user $u \in \mathcal{U}$ consists of:

- a *'collaborative filtering'* constant $\phi^u \in [0, 1]$;
- $\forall t \in \mathcal{T}$ a *'type importance'* constant $\mu^u_t \in]0, 1]$;
- $\forall v \in \mathcal{U}$ such that $u \neq v$, a *'similarity'* constant $\omega_{u,v} \in [0, 1]$.

Intuitively, ϕ^u defines how much u wishes collaborative filtering to be taken into account, and a larger ϕ^u will give other users' ratings more prevalence in the calculations of predicted ratings and, conversely, a smaller ϕ^u will give the content-based components more prevalence. Thus, ϕ^u allows users to control the hybrid nature of our RS. Also, μ^u_t defines how important type t is to u and how much u wants aspects of type t to be taken into account, and larger values of μ^u_t will

[4] Note that we refer to properties/features as *aspects* in line with the terminology in [54].

[5] Note that very few ratings (if any) may be specified by $\mathcal{R}$ at the start (giving rise to the cold start problem we mentioned earlier in Section 2). The cold start is a commonly occurring problem in the RS literature. This could be addressed, for instance, by recommending the most popular movies with respect to (highest) ratings.

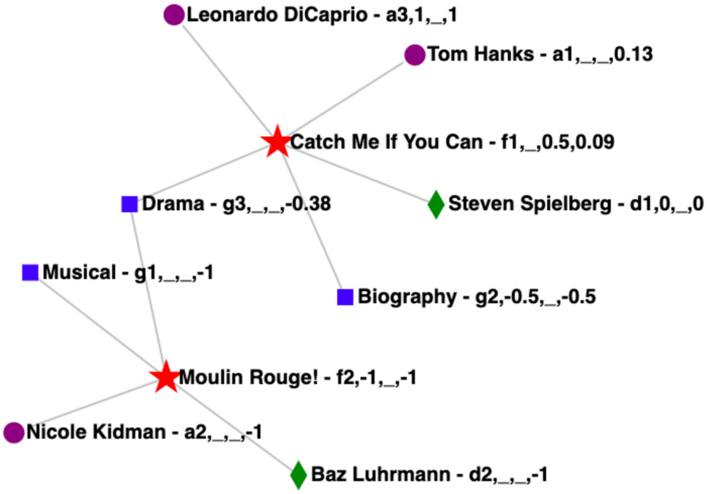

Leonardo DiCaprio - a3,1,_,1
Tom Hanks - a1,_,_,0.13
Catch Me If You Can - f1,_,0.5,0.09
Drama - g3,_,_,-0.38
Steven Spielberg - d1,0,_,0
Musical - g1,_,_,-1
Biography - g2,-0.5,_,-0.5
Moulin Rouge! - f2,-1,_,-1
Nicole Kidman - a2,_,_,-1
Baz Luhrmann - d2,_,_,-1

Fig. 2. Example components of an A-I visualised as a graph, with items given by red stars and types: genres (whose aspects are blue squares), actors (whose aspects are purple circles) and directors (whose aspects are green diamonds). Each node's label is of the form (Name - x, $\mathcal{R}(u,x)$, $\mathcal{R}(v,x)$, $\mathcal{P}^u_{\mathcal{X}}(x)$), with $\mathcal{U}=\{u,v\}$, _ standing for 'undefined' and assuming a mapping from each node to its item-aspect. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

give these aspects, and the user's own ratings on items which are linked to them, a higher impact. We assume that $\mu^u_t > 0$ for any $t \in \mathcal{T}$ and $u \in \mathcal{U}$, i.e. that all aspects are taken into account at some degree for all users.[6] Finally, $\omega_{u,v}$ defines how similar u and v are, and how much $v's$ ratings should impact the calculations. Note that all other users' ratings can be considered, by ensuring that all the similarities are greater than zero, or only the k nearest (by similarity) neighbours, by setting the similarities to all other users (aside from the k nearest) to zero. In Section 4, we have set the number of similar users to 20, as this worked reasonably well in the empirical analysis; however other values could be used and we leave finding the optimal value to future work.

Our method for calculating predicted ratings of items, based on users' profiles, makes use of the following notion of *weighted average rating*:

Definition 4. For any $u \in \mathcal{U}$ and any $i \in \mathcal{I}$, let $\Upsilon^u(i) = \{v \in \mathcal{U} \setminus \{u\} | \mathcal{R}(v,i)$ is defined$\}$ be the set of users other than u who have rated item i. Then, the *weighted average rating* $\rho^u : \mathcal{I} \to [-1,1]$ is obtained as follows, for $u \in \mathcal{U}$ and $i \in \mathcal{I}$:
if $\Upsilon^u(i) \neq \emptyset$ and $\sum_{v \in \Upsilon^u(i)} \omega_{u,v} > 0$
then $\rho^u(i) = \frac{\sum_{v \in \Upsilon^u(i)} \omega_{u,v} \mathcal{R}(v,i)}{|\Upsilon^u(i)|}$
else $\rho^u(i)$ is undefined.

Thus, the weighted average rating of an item for a user is undefined when no other user or no other similar users have given any ratings for the item. As mentioned earlier, the choice of which users may be deemed similar is made by an appropriate choice of the constants $\omega_{u,v}$.

The predicted rating for an item is given in terms of the predicted rating for aspects, defined as follows.

Definition 5. For any user $u \in \mathcal{U}$ and aspect $a \in \mathcal{A}$, let $\Lambda^u(a) = \{i \in \mathcal{L}(a) | \mathcal{R}(u,i)$ is defined$\}$ be the set of linked items with ratings from u and let $\Lambda^{-u}(a) = \{i \in \mathcal{L}(a) | \rho^u(i)$ is defined$\} \setminus \Lambda^u(a)$ be the set of linked items with defined weighted average ratings but without ratings from u. Then, the *predicted aspect rating* $\mathcal{P}^u_{\mathcal{A}} : \mathcal{A} \to [-1,1]$ for a is obtained as follows, for $u \in \mathcal{U}$ and $a \in \mathcal{A}$:

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

$$\text{if} \quad \Lambda^u(a) = \Lambda^{-u}(a) = \emptyset \text{ then } \mathcal{P}^u_{\mathcal{A}}(a) = 0; \text{ else}$$

$$\text{if} \quad \Lambda^u(a) = \emptyset \text{ then } \mathcal{P}^u_{\mathcal{A}}(a) = \phi^u \frac{\sum_{i \in \Lambda^{-u}(a)} \rho^u(i)}{|\Lambda^{-u}(a)|} / [1 + \phi^u]; \text{ else}$$

[6] This is a departure from [54], where $\mu^u_t \in [0,1]$. Note that, although we restrict μ^u_t to be non-zero, it can be infinitesimally small. Note also that although these constants cannot be negative, negative influences between item-aspects, as indicated in the argumentative scaffolding as described in Section 5, may still arise.

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

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

Intuitively, the predicted aspect rating weights the average ratings on linked items from the user and from similar users based on ϕ^u, but is overridden by a rating on the aspect itself from the user. Aspects without ratings (from the user or similar users) have the neutral predicted aspect rating of zero.

We finally use the predicted aspect ratings to calculate the predicted item ratings, as follows.

Definition 6. For any $u \in \mathcal{U}$, the *predicted item rating* $\mathcal{P}^u_{\mathcal{I}} : \mathcal{I} \to [-1, 1]$ is obtained as follows, for any $i \in \mathcal{I}$:

$$\text{if} \quad \mathcal{R}(u,i) \text{ is defined then } \mathcal{P}^u_{\mathcal{I}}(i)=\mathcal{R}(u,i); \text{ else}$$

$$\text{if} \quad \rho^u(i) \text{ is undefined then } \mathcal{P}^u_{\mathcal{I}}(i)=\frac{\sum_{t\in\mathcal{T}}\mu^u_t[\sum_{a\in\mathcal{L}_t(i)}\mathcal{P}^u_{\mathcal{A}}(a)]/|\mathcal{L}_t(i)|}{\sum_{t\in\mathcal{T}}\mu^u_t}; \text{ else}$$

$$\mathcal{P}^u_{\mathcal{I}}(i)=\frac{\phi^u\rho^u(i)+\sum_{t\in\mathcal{T}}\mu^u_t[\sum_{a\in\mathcal{L}_t(i)}\mathcal{P}^u_{\mathcal{A}}(a)]/|\mathcal{L}_t(i)|}{\phi^u+\sum_{t\in\mathcal{T}}\mu^u_t}.$$

The predicted item rating is again overridden by a rating from the user. This calculation weights the average ratings on the item from similar users with ϕ^u against the predicted aspects ratings from each of the linked aspects using their corresponding μ^u_t. Thus, aspects with a positive, negative or neutral predicted ratings have positive, negative or neutralising, respectively, effects on items to which they are linked. Note that our method can be seen as a form of hybrid RS as it combines collaborative filtering with content-based factors.

In the remainder of the paper, for simplicity we use $\mathcal{P}^u_{\mathcal{X}}(x)$ to refer to $\mathcal{P}^u_{\mathcal{I}}(x)$ or $\mathcal{P}^u_{\mathcal{A}}(x)$ depending on whether $x \in \mathcal{I}$ or $x \in \mathcal{A}$, respectively. We also refer to $\mathcal{P}^u_{\mathcal{X}}$ as the *predicted rating* of an item-aspect.

As an illustration, consider the A-I with $\mathcal{I}$, $\mathcal{A}$, $\mathcal{T}$ and $\mathcal{L}$ as in Fig. 2, $\mathcal{U} = \{u, v\}$ and $\mathcal{R}$ such that: $\mathcal{R}(u, a_3) = 1$, $\mathcal{R}(u, d_1) = 0$, $\mathcal{R}(u, f_2) = -1$, $\mathcal{R}(u, g_2) = -0.5$ and $\mathcal{R}(v, f_1) = 0.5$. Assume that $\phi^u = \mu^u_{actors} = \mu^u_{genres} = \mu^u_{directors} = 1$ and $\omega_{u,v} = 0.5$. Then, by Definitions 5 and 6, the predicted rating for the item-aspects that u has rated is equal to these ratings, e.g. $\mathcal{P}^u_{\mathcal{A}}(u, a_3) = \mathcal{R}(u, a_3) = 1$ (similarly for d_1, f_2 and g_2). For a_1, $\Lambda^u(a_1) = \emptyset$ and $\Lambda^{-u}(a_1) = \{f_1\}$ thus $\mathcal{P}^u_{\mathcal{A}}(a_1) = \phi^u \times \omega_{u,v} \times \mathcal{R}(v, f_1)[1 + \phi^u] = 1 \times 0.5 \times 0.5/[1 + 1] = 0.125$. For x any of a_2, d_2 and g_1, $\mathcal{P}^u_{\mathcal{A}}(u, x) = \mathcal{R}(u, f_2) = -1$ since $\Lambda^u(x) = \{f_2\}$ and $\Lambda^{-u}(x) = \emptyset$. For g_3, $\Lambda^u(g_3) \neq \emptyset$, $\Lambda^{-u}(g_3) \neq \emptyset$ and $\mathcal{P}^u_{\mathcal{A}}(g_3) = [\frac{-1}{1} + 1\frac{0.5*0.5}{1}]/[1 + 1] = -0.375$. Finally, for f_1:

$$\phi^u\rho^u(f_1) = \phi^u \times \omega_{u,v} \times \mathcal{R}(v, f_1) = 1 \times 0.5 \times 0.5 = 0.25;$$

$$\mu^u_{actors}[\sum_{a\in\mathcal{L}_{actors}(f_1)} \mathcal{P}^u_{\mathcal{A}}(a)]/|\mathcal{L}_{actors}(f_1)| = 1 \times [0.125 + 1]/2 = 0.563;$$

$$\mu^u_{genres}[\sum_{a\in\mathcal{L}_{genres}(f_1)} \mathcal{P}^u_{\mathcal{A}}(a)]/|\mathcal{L}_{genres}(f_1)| = 1 \times [-0.5 - 0.375]/2 = -0.438;$$

$$\mu^u_{directors}[\sum_{a\in\mathcal{L}_{directors}(f_1)} \mathcal{P}^u_{\mathcal{A}}(a)]/|\mathcal{L}_{directors}(f_1)| = 1 \times [0]/1 = 0;$$

$$\mathcal{P}^u_{\mathcal{I}}(f_1) = \frac{0.25 + 0.563 - 0.438 + 0}{4} = 0.09.$$

These predicted ratings, alongside the given ratings, are visualised in Fig. 2.

We now prove theoretically that the predicted ratings satisfy properties which render the RS's behaviour suitable for making recommendations and, as we will show later, capable of supporting explanatory feedback processes. These properties are implicitly formulated for two A-Is: the first, $\mathcal{F} = \langle \mathcal{I}, \mathcal{A}, \mathcal{T}, \mathcal{L}, \mathcal{U}, \mathcal{R} \rangle$, is the starting point; the second, $\mathcal{F}' = \langle \mathcal{I}', \mathcal{A}', \mathcal{T}', \mathcal{L}', \mathcal{U}', \mathcal{R}' \rangle$, is the result of some modification to the components of $\mathcal{F}$. These modifications may take different forms, e.g. new ratings for item-aspects or additional item-aspects and links. No matter what the modifications are, they are of interest because they lead to modified predicted ratings (as indicated in the propositions below).

Our first proposition shows that increasing an aspect's predicted rating can only have a positive effect on the predicted ratings of the items to which it is linked. We posit that this is intuitive; for example, if we do not know if a user likes a movie but we receive information that they like one of its actors, we would expect our prediction of the user's rating on the movie to increase.

Proposition 1. *Let $u \in \mathcal{U} \cap \mathcal{U}'$, $i \in \mathcal{I} \cap \mathcal{I}'$ be such that $\mathcal{R}(u, i)$ is not defined, and $a \in \mathcal{L}(i) \cap \mathcal{L}'(i)$. Further, let $\mathcal{P}^u_{\mathcal{A}}(a)$ and $\mathcal{P}^u_{\mathcal{A}}{}'(a)$ be, resp., the predicted aspect ratings of a for u in $\mathcal{F}$ and $\mathcal{F}'$ and $\mathcal{P}^u_{\mathcal{I}}(i)$ and $\mathcal{P}^u_{\mathcal{I}}{}'(i)$ be, resp., the predicted item ratings for u of i in $\mathcal{F}$ and $\mathcal{F}'$. Then,*

Table 1
Datasets statistics: number of items ($|\mathcal{I}|$), users ($|\mathcal{U}|$), aspects ($|\mathcal{A}|$), and individual aspects (|directors|, |actors| and |genres|), as well as maximum (max) and average (avg) number of aspects/actors/genres per film.

	Netflix	MovieLens Dev.	MovieLens 100k
$\|\mathcal{I}\|$	240	7225	670
$\|\mathcal{U}\|$	4113	610	943
$\|\mathcal{A}\|$	538	7881	1267
|directors|	101	1795	275
|actors|	419	6066	974
|genres|	18	20	18
avg $\|\mathcal{A}\|$	7.85	7.43	6.62
avg |actors|	4.23	4.03	3.67
avg |genres|	2.60	2.34	1.92
max $\|\mathcal{A}\|$	14	18	10
max |actors|	11	5	5
max |genres|	3	10	5

- *if* $\mathcal{P}^u_\mathcal{A}{}'(a) > \mathcal{P}^u_\mathcal{A}(a)$ *then* $\mathcal{P}^u_\mathcal{I}{}'(i) > \mathcal{P}^u_\mathcal{I}(i)$*;*
- *if* $\mathcal{P}^u_\mathcal{A}{}'(a) < \mathcal{P}^u_\mathcal{A}(a)$ *then* $\mathcal{P}^u_\mathcal{I}{}'(i) < \mathcal{P}^u_\mathcal{I}(i)$*.*

Proof. By inspection of the two latter cases of Definition 6, as only $\mathcal{P}^u_\mathcal{A}(a)$ has changed. □

Our second proposition concerns the effect of changing the user's or similar users' ratings on a movie on the predicted ratings of the aspects it holds: increasing the former can only have a positive effect on the latter. We believe that this is also intuitive behaviour; for example, if we do not know if a user likes an aspect representing a genre but we receive information that they or similar users like a movie of that genre, we would expect that our prediction of the user's rating on the aspect representing the genre to increase.

Proposition 2. *Let* $u \in \mathcal{U} \cap \mathcal{U}'$*,* $a \in \mathcal{A} \cap \mathcal{A}'$ *be such that* $\mathcal{R}(u, a)$ *is not defined, and* $i \in \mathcal{L}(a) \cap \mathcal{L}'(a)$*. Further, let* $\rho^u(i)$ *and* $\rho^u{}'(i)$ *be, resp., the weighted average rating of i for u in* $\mathcal{F}$ *and* $\mathcal{F}'$*, and* $\mathcal{P}^u_\mathcal{A}(a)$ *and* $\mathcal{P}^u_\mathcal{A}{}'(a)$ *be, resp., the predicted aspect ratings of a for u in* $\mathcal{F}$ *and* $\mathcal{F}'$*. Then,*

- *if* $\mathcal{R}'(u, i) > \mathcal{R}(u, i)$ *or* $\rho^u{}'(i) > \rho^u(i)$ *then* $\mathcal{P}^u_\mathcal{A}{}'(a) > \mathcal{P}^u_\mathcal{A}(a)$*;*
- *if* $\mathcal{R}'(u, i) < \mathcal{R}(u, i)$ *or* $\rho^u{}'(i) < \rho^u(i)$ *then* $\mathcal{P}^u_\mathcal{A}{}'(a) < \mathcal{P}^u_\mathcal{A}(a)$*.*

Proof. We focus here on the first bullet (the second can be argued similarly). Since $\mathcal{R}(u, a)$ is not defined, the first case of Definition 5 does not apply. The same is true of the second case as either $\mathcal{R}(u, i)$ or $\rho^u(i)$ is defined, and thus $\Lambda^u(a) \neq \emptyset$ or $\Lambda^{-u}(a) \neq \emptyset$, resp. If $\mathcal{R}(u, i)$ is defined then the fourth or the fifth case applies, in which case we can see that $\mathcal{R}'(u, i) > \mathcal{R}(u, i)$ gives $\mathcal{P}^u_\mathcal{A}{}'(a) > \mathcal{P}^u_\mathcal{A}(a)$. Likewise, if $\rho^u(i)$ is defined then the third or the fourth case applies, in which case we can see that $\rho^u{}'(i) > \rho^u(i)$ gives $\mathcal{P}^u_\mathcal{A}{}'(a) > \mathcal{P}^u_\mathcal{A}(a)$. □

4. Empirical evaluation

We evaluate the A-I RS empirically on three datasets: a subset of the Netflix challenge dataset[7] as reported in [54], the MovieLens Development (Dev.) dataset and the MovieLens 100K benchmark dataset [36]. For all datasets, to obtain the aspects associated with $\mathcal{T} = \{$genre, actor, director$\}$, we use The Movie Database (TMDb) API.[8] Like others, e.g. [8,41,77,72], we focus on the movie domain as a popular, exemplary domain for RSs, but choose three different datasets for variety in the evaluation.

The datasets contain ratings on a five star scale from 1 to 5, with both integral (Netflix, MovieLens 100K) and half-star increments (MovieLens Dev.). The users in the MovieLens datasets have rated at least 20 movies[9] whereas the users in the Netflix dataset have rated at least 10 movies [54]. In our experiments, we keep only those actors and directors that have appeared in at least two movies. If, after this filtering, there exist movies with only actors and directors that have been discarded, then these movies are also discarded from the datasets. Statistics of the resulting datasets are shown in Table 1.

We calculate the users' ratings for each aspect of each type, namely *genre*, *actor*, and *director*. For example, to obtain a user's rating for each aspect of type genre, we multiply the ratings' matrix of the user with the movie genre matrix. Similarly, we calculate the users' ratings for each actor and for each director as found in our database. To determine the

[7] https://www.netflixprize.com/.
[8] http://www.themoviedb.org.
[9] https://grouplens.org/datasets/movielens/.

similarity between any two (different) users, we use the cosine distance between the users' ratings for all aspects of type genre. Formally, $\omega_{u,v} = \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{||\boldsymbol{u}|| \cdot ||\boldsymbol{v}||}$ where $\boldsymbol{u}$ and $\boldsymbol{v}$ are vectors representing user u's and user v's ratings, respectively, for each aspect a. In the experiments, for each user u, we use u's 20 most similar users, leaving studies investigating the variation of the similarity constant (ω) to future work. For all datasets, we use the following constants for the profiles of all users: $\phi = 0.7$, $\mu_{actor} = 0.1$, $\mu_{director} = 0.1$, $\mu_{genre} = 0.1$. We consider these constants to represent the default configuration, giving more weight to similar users to allow for a fair comparison with the algorithms we use as baselines. We leave the optimisation of these constants (e.g. per user) to future work.

The experimental setting is defined as follows. We perform five-fold cross-validation using, in turn, 80% of each user's ratings as training data and the remaining ratings as test data. This is in line with most related works, using a large majority of ratings as training and 20-25% of the ratings as testing [8,41,77]. Furthermore, we run additional experiments to test the robustness of our method when dealing with limited training data. This is in line with what happens in reality, where new users may not have rated many items and thus there may not be many ratings available to the RS when making recommendations. To evaluate the performance of our method with limited training data, we also partition the datasets into five parts and conduct five experiments but by using, in turn, only 20% of each user's ratings as training data (while the remaining 80% of the ratings constitute the test data). We compare against the following recommendation algorithms as implemented in the Surprise library [39]:

- **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;
- **NMF**: Non-negative Matrix Factorization, a collaborative filtering algorithm [47];
- **Slope1**: Slope One [45], based on 'popularity differential' between items for users by finding the average rating differential;
- **CoClust**: Co-clustering [34], an algorithm built on simultaneous clustering of users and items.

For all methods we use the default configuration settings. We report standard RS performance measures [61]: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Given that ratings are highly subjective, users who might like the same movie could give different ratings, e.g. in the case of two users who both liked a movie, one could give 5 stars (again, on a scale of 1-5 stars) to a movie whereas the other could give 4 or 4.5 stars. To accommodate variations in subjective ratings, we also convert the ratings to a binary scale such that a rating greater or equal to 3 is considered to be positive, whereas a rating less than 3 is considered to be negative. We report global Precision (as well as Recall and F_1) with respect to the binary scale. The results are shown in Table 2.

When using five fold-cross validation (top of Table 2), our RS is competitive with the baselines. There is no clear winning algorithm on the Netflix dataset while for the MovieLens datasets, SVD performs best. In the experiments with limited training data (i.e. 20% of each user's ratings as training and remaining 80% as test data, see bottom part of Table 2), in terms of MAE and RMSE, SVD performs best throughout. Our RS achieves the highest precision when considering the binary scale. We are interested in obtaining high precision rather than recall as we want to make sure the predictions we make are correct; i.e. we are interested in the fraction of relevant items retrieved out of all items retrieved rather than in the fraction of relevant items retrieved out of all relevant items. Although our method obtains higher precision compared to the baselines, it does not outperform all baselines on all metrics used. Whilst obtaining the minimum MAE/RMSE is not the main focus of this paper, it is nonetheless encouraging to see that our method does not sacrifice effectiveness significantly.

We will now show that, in addition to being reasonably effective in comparison with the other methods considered, our method is explainable, and a whole range of explanation types for its recommendations can be automatically generated from an argumentation scaffolding underpinning our A-I. Since our focus is on explainability, we leave the optimisation of (the parameters in) our RS and a comparison with other RSs (see Section 2) as future work.

5. Argumentative scaffolding

In abstract [30] and bipolar [20] argumentation, any information which may be in dialectical relationships of disagreement (attack) or, in the bipolar case, agreement (support) with other information may be considered to be an argument, and arguments (according to this loose interpretation of the term) typically have a negative or positive impact on the (gradual) acceptability of arguments they attack or support, respectively. In this spirit, item-aspects in A-Is may be seen as arguments: if a user (or another similar user) rates an item highly/lowly then this item can be seen as an argument for/against, respectively, the aspects connected with the item and, similarly, if a user rates an aspect highly/lowly then this aspect can be seen as an argument for/against, respectively, the items connected with the aspect. Moreover, if an A-I is viewed from an argumentative perspective, a user's (or similar users') opinion (rating) on an aspect/item may impact the estimation of the user's opinion (rating) of items/aspects connected with that aspect/item in the absence of actual ratings. This dialectical reading of A-Is provides the *argumentative scaffolding* from which the explanations for recommendations that are delivered to users may be extracted to facilitate fruitful interactions between the user and the RS, e.g. via conversational explanations as in [24], leveraging on argumentation's dialectical nature.

In order to fully capture the behaviour of A-Is as argumentation frameworks, a novel dialectical *neutralising* relationship is needed, in addition to the standard relationships of attack and support in bipolar argumentation frameworks, to represent item-aspects which have neither a positive nor a negative effect on other arguments but rather *neutralise* them, by moving their strength towards the neutral mid-point. For example, let us consider the case of a user liking an actor a who is in two

Table 2
Evaluation results averaged over five runs with different percentages of data used for training and testing (80% and 20%, resp., in the top part, and 20% and 80%, resp., in the bottom part). We indicate best performances in bold.

				KNN	KNNZ	SVD	NMF	Slope1	CoClust	A-I
80% training; 20% testing	Netflix	19025 pairs	MAE	0.77	0.73	**0.71**	0.76	0.74	0.75	0.90
			RMSE	1.05	**1.02**	0.99	1.05	**1.02**	1.04	1.19
			Precision	0.82	0.84	0.83	**0.85**	0.84	**0.85**	0.84
			Recall	**0.99**	0.96	0.98	0.93	0.95	0.94	0.87
			F1	**0.90**	**0.90**	**0.90**	0.89	**0.90**	0.89	0.85
	MovieLens Dev.	18706 pairs	MAE	0.74	0.69	**0.68**	0.73	0.71	0.74	0.90
			RMSE	0.98	0.93	**0.91**	0.96	0.94	0.98	1.19
			Precision	0.84	0.86	0.86	0.86	**0.87**	**0.87**	0.84
			Recall	**0.94**	0.92	**0.94**	0.90	0.90	0.87	0.85
			F1	0.89	0.89	**0.90**	0.88	0.88	0.87	0.85
	MovieLens 100K	13001 pairs	MAE	0.81	0.76	**0.74**	0.80	0.79	0.80	0.85
			RMSE	1.12	1.06	**1.02**	1.11	1.09	1.10	1.13
			Precision	0.85	**0.86**	0.85	**0.86**	**0.86**	**0.86**	**0.86**
			Recall	0.97	0.97	**0.99**	0.95	0.96	0.95	0.93
			F1	0.91	0.91	**0.92**	0.90	0.91	0.90	0.89
20% training; 80% testing	Netflix	74279 pairs	MAE	0.87	0.85	**0.77**	0.89	0.87	0.88	1.00
			RMSE	1.15	1.14	**1.05**	1.18	1.17	1.18	1.31
			Precision	0.82	0.83	0.82	**0.84**	0.83	**0.84**	**0.84**
			Recall	0.98	0.94	**0.99**	0.89	0.93	0.92	0.80
			F1	0.89	0.88	**0.90**	0.86	0.88	0.88	0.82
	MovieLens Dev.	74577 pairs	MAE	0.82	0.79	**0.74**	0.82	0.81	0.82	0.93
			RMSE	1.05	1.02	**0.97**	1.06	1.04	1.07	1.20
			Precision	0.82	0.84	0.84	0.84	0.84	0.84	**0.86**
			Recall	**0.96**	0.94	**0.96**	0.90	0.93	0.91	0.77
			F1	0.89	0.89	**0.90**	0.87	0.88	0.87	0.81
	MovieLens 100K	51637 pairs	MAE	0.91	0.88	**0.81**	0.93	0.90	0.91	0.93
			RMSE	1.22	1.18	**1.08**	1.24	1.21	1.23	1.21
			Precision	0.84	0.85	0.85	0.85	0.85	0.85	**0.86**
			Recall	**0.98**	0.98	1.00	0.94	0.97	0.96	0.91
			F1	0.90	**0.91**	**0.91**	0.89	0.90	0.90	0.88

movies, but as the sole actor in one and as one of ten in the other. Suppose that the A-I predicts a neutral (0) rating for all actors in the second movie other than *a*. These nine actors dilute (neutralise) the positive effect of *a*, impacting in turn the system's prediction on whether the user likes the aspects of type *actor* in general. Without a neutralising dialectical relation, some influences from the item-aspects on their predicted ratings would not be represented in the argumentative scaffolding. We therefore define the following argumentation frameworks.

Definition 7. A *Tripolar Argumentation Framework* (TF) is a tuple $\langle \mathcal{X}, \mathcal{L}^{-}, \mathcal{L}^{+}, \mathcal{L}^{0} \rangle$ where $\mathcal{X}$ is a set of *arguments*, and $\mathcal{L}^{-}$, $\mathcal{L}^{+}$, $\mathcal{L}^{0}$ are binary relations over $\mathcal{X}$. For $x, y \in \mathcal{X}$, we say that x *attacks* y if $(x, y) \in \mathcal{L}^{-}$, x *supports* y if $(x, y) \in \mathcal{L}^{+}$, and x *neutralises* y if $(x, y) \in \mathcal{L}^{0}$. With $\times$ as any of $-$, $+$ or 0, for any $x \in \mathcal{X}$, we will use $\mathcal{L}^{\times}(x)$ to denote $\{y \in \mathcal{X} | (y, x) \in \mathcal{L}^{\times}\}$ the *attackers*, *supporters* or *neutralisers*, resp., of x.

Note that our TFs may be seen as instances of 'tripolar frameworks' as defined in [31] and of 'generalised argumentation frameworks' as defined in [9]. Whereas these works envisage the use of relations other than attack and support, we commit (in our concrete instance) to the additional relation 'neutralise'.

Straightforwardly, any TF $\langle \mathcal{X}, \mathcal{L}^{-}, \mathcal{L}^{+}, \mathcal{L}^{0} \rangle$ with $\mathcal{L}^{0} = \emptyset$ is a bipolar argumentation framework and if $\mathcal{L}^{+} = \mathcal{L}^{0} = \emptyset$ then the TF is an abstract argumentation framework. As in the case of abstract and bipolar argumentation frameworks, a TF may also be equipped with some *gradual strength* function σ which calculates the strength of any argument over a given interval based on the strength of the arguments in dialectical relationships with the argument, as in [31]. As in the case of abstract and bipolar argumentation frameworks, this strength function may be defined so as to satisfy desirable properties [11], including the following simple but intuitive property, which is a generalisation to the setting of TFs of one of the implications of strict monotonicity in [12]:

Definition 8. Let $\langle \mathcal{X}, \mathcal{L}^-, \mathcal{L}^+, \mathcal{L}^0 \rangle$ and $\langle \mathcal{X}', \mathcal{L}^{-\prime}, \mathcal{L}^{+\prime}, \mathcal{L}^{0\prime} \rangle$ be TFs, and let $(x, y) \in (\mathcal{L}^- \cup \mathcal{L}^+ \cup \mathcal{L}^0) \cap (\mathcal{L}^{-\prime} \cup \mathcal{L}^{+\prime} \cup \mathcal{L}^{0\prime})$. A strength function σ satisfies the property of *weak monotonicity* at (x, y) if, whenever $\sigma(x) = 0$ in $\langle \mathcal{X}', \mathcal{L}^{-\prime}, \mathcal{L}^{+\prime}, \mathcal{L}^{0\prime} \rangle$ and, $\forall z \in [(\mathcal{L}^-(y) \cup \mathcal{L}^+(y) \cup \mathcal{L}^0(y)) \cap (\mathcal{L}^{-\prime}(y) \cup \mathcal{L}^{+\prime}(y) \cup \mathcal{L}^{0\prime}(y))] \backslash \{x\}$, if $\sigma(z) = s$ in $\langle \mathcal{X}, \mathcal{L}^-, \mathcal{L}^+, \mathcal{L}^0 \rangle$ and $\sigma(z) = s'$ in $\langle \mathcal{X}', \mathcal{L}^{-\prime}, \mathcal{L}^{+\prime}, \mathcal{L}^{0\prime} \rangle$, then $s = s'$, then the following is guaranteed to hold, for $\sigma(y) = v$ in $\langle \mathcal{X}, \mathcal{L}^-, \mathcal{L}^+, \mathcal{L}^0 \rangle$ and $\sigma(y) = v'$ in $\langle \mathcal{X}', \mathcal{L}^{-\prime}, \mathcal{L}^{+\prime}, \mathcal{L}^{0\prime} \rangle$:

- if $x \in \mathcal{L}^-(y) \cap \mathcal{L}^{-\prime}(y)$, then $v' > v$;
- if $x \in \mathcal{L}^+(y) \cap \mathcal{L}^{+\prime}(y)$ then $v' < v$;
- if $x \in \mathcal{L}^0(y) \cap \mathcal{L}^{0\prime}(y)$ then $v' = v$.

Weak monotonicity characterises attacks/supports/neutralisers as links between arguments such that if we mute the affecting argument's strength, its affected argument's strength increases/reduces/does not change, resp., highlighting the negative/positive/neutral, resp., effect between the two.

We will use the interval [-1,1] as the co-domain of σ, matching the interval used for ratings, and map A-Is onto TFs so that predicted ratings of item-aspects amount to the strength of arguments. In doing so, we use TFs as the argumentative scaffolding for explanations in our RS.

Specifically, to obtain a TF from a given A-I, first we *direct* the A-I's links in $\mathcal{L}$, based on the existence of the user's and other (similar) users' ratings for item-aspects, showing which item-aspects have effects on the predicted ratings of others. These directed relations represent the direction of the inferences made in the RS, e.g. if an item-aspect has a given rating from the user, no inferences were made on its predicted rating and it thus has no inward relations. However, if the item-aspect has no rating from the user, the calculation of its predicted rating uses information from its linked item-aspects and the item-aspect may therefore have inward relations. This first step results in *directed A-Is*, defined formally as follows[10]:

Definition 9. 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 $\mathcal{L}^u = \{(i, a) \in \mathcal{L} | \mathcal{R}(u, a)$ is undefined and $\exists v \in \mathcal{U}$ such that $\mathcal{R}(v, i)$ is defined and if $v \neq u$ then $\omega_{u,v} \neq 0\} \cup \{(a, i) \in \mathcal{L} | \mathcal{R}(u, i)$ is undefined$\}$. For $x \in \mathcal{X}$, we refer to $\mathcal{L}^u(x) = \{y \in \mathcal{X} | (y, x) \in \mathcal{L}^u\}$ as the set of item-aspects *affecting* x. Also, for $i \in \mathcal{I}$ we use $\mathcal{L}^u_t(i)$ to denote the set $\{a \in \mathcal{L}^u(i) | a \in \mathcal{A}_t\}$.

For the remainder of the paper we will assume as given an arbitrary directed A-I $\mathcal{F}^u = \langle \mathcal{I}, \mathcal{A}, \mathcal{T}, \mathcal{L}^u, \mathcal{U}, \mathcal{R} \rangle$ for $u \in \mathcal{U}$, unless otherwise specified. A TF can then be obtained from $\mathcal{F}^u$ by determining the polarity of pairs in $\mathcal{L}^u$, as follows:

Definition 10. For any $i \in \mathcal{I}$, let $r^u(i)$ be $\mathcal{R}(u, i)$ if defined, else $\rho^u(i)$ if defined, and otherwise be undefined.[11] The *TF corresponding to* $\mathcal{F}^u$ is $\langle \mathcal{X}, \mathcal{L}^-, \mathcal{L}^+, \mathcal{L}^0 \rangle$ such that:

$$\mathcal{L}^- = \{(i, a) \in \mathcal{L}^u | r^u(i) < 0\} \cup \{(a, i) \in \mathcal{L}^u | \mathcal{P}^u_{\mathcal{A}}(a) < 0\};$$
$$\mathcal{L}^+ = \{(i, a) \in \mathcal{L}^u | r^u(i) > 0\} \cup \{(a, i) \in \mathcal{L}^u | \mathcal{P}^u_{\mathcal{A}}(a) > 0\};$$
$$\mathcal{L}^0 = \{(i, a) \in \mathcal{L}^u | r^u(i) = 0\} \cup \{(a, i) \in \mathcal{L}^u | \mathcal{P}^u_{\mathcal{A}}(a) = 0\}.$$

We define r^u here to represent the effect which linked items have on an aspect's predicted rating, in line with Definition 5. Then, r^u and $\mathcal{P}^u_{\mathcal{A}}$ are used to determine the polarity of an affecting argument's effects on affected arguments: if they are negative/positive/zero, we categorise the directed link as attack/support/neutraliser, resp., in the TF. It is this direct mapping from predicted ratings to the TF which ensures the fidelity of the argumentative explanations to the RS, differently from other methods, such as the graph-based approach of [50], where explanations are generated independently from the recommendations.

For illustration, the TF corresponding to the directed A-I $\mathcal{F}^u$ for user u obtained from the A-I shown in Fig. 2 is visualised in Fig. 3. Here, there are no arguments affecting f_2 since it is rated by u. Given that this rating is negative and all aspects linked to f_2 are not rated by u, f_2 attacks all such aspects. Conversely, f_1 is not rated by u but has a positive rating from v and thus f_1 supports all (linked) aspects without a rating, i.e. a_1 and g_3. The fact that f_1 is not rated by u means that all aspects linked to f_1 affect it.

Note that since an argument with zero strength (e.g. d_1 in Fig. 3) has a diluting effect on arguments it affects, we require the neutralising relation in order to ensure fidelity, to fully represent how the predicted ratings are calculated, as we briefly discussed earlier. For example, consider a movie f_1 with $n > 1$ linked aspects $\mathcal{L}(f_1) = \{a_1, \ldots, a_n\}$ such that $\mathcal{P}^u_{\mathcal{A}}(a_1) = 1$ and $\forall i > 1$ $\mathcal{P}^u_{\mathcal{A}}(a_i) = 0$, and movies f_2, f_3 with $\mathcal{L}(f_2) = \{a_1\}$, $\mathcal{L}(f_3) = \{a_2, \ldots, a_n\}$. The impact of the aspects on $\mathcal{P}^u_{\mathcal{I}}(f_2)$ should be greater than that on $\mathcal{P}^u_{\mathcal{I}}(f_1)$ (given that all of f_2's linked aspects have maximum predicted rating) - thus we need dialectical relations from $a_2, \ldots, a_n$ which reduce the strength of f_1. Moreover, the impact of the aspects on $\mathcal{P}^u_{\mathcal{I}}(f_3)$ should be null (given that all of f_3's aspects have neutral predicted rating) - thus the dialectical relations from $a_2, \ldots, a_n$

[10] This definition differs slightly from that in [54] as we neglect here inward relations for rated aspects, since their linked items do not have an effect on the aspects' predicted rating calculations.

[11] It is easy to see, by definition of $\mathcal{L}^u$, that if $\exists (i, a) \in \mathcal{L}^u$, then $r^u(i)$ is defined.

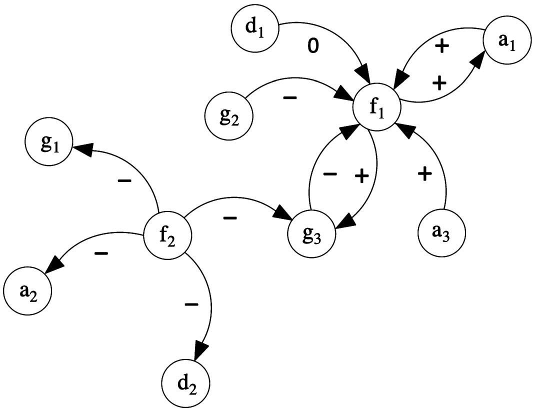


Fig. 3. A graphical representation of the TF corresponding to the directed A-I $\mathcal{F}^u$ for user u from A-I in Fig. 2. Here, '+' indicates 'support' ($\mathcal{L}^+$ in $\mathcal{F}^u$), '-' indicates 'attack' ($\mathcal{L}^-$ in $\mathcal{F}^u$) and '0' indicates 'neutraliser' ($\mathcal{L}^0$ in $\mathcal{F}^u$).

cannot be attacks. We use a neutralising relation that only dilutes the positive effect of a_1 so that the estimation of whether our user (dis)likes $a_2, \dots, a_n$ does not decrease or increase our estimation of whether the user (dis)likes f_3 nor does it necessarily decrease our estimation of whether the user (dis)likes f_1.

We now show theoretically that the behaviour of the argumentative scaffolding in relation to the predicted ratings of item-aspects is intuitive from an argumentation viewpoint (while remaining faithful to the RS). To do this, similarly to Section 3, we consider two A-Is ($\mathcal{F}^u = \langle \mathcal{I}, \mathcal{A}, \mathcal{T}, \mathcal{L}^u, \mathcal{U}, \mathcal{R} \rangle$ and $\mathcal{F}^{u\prime} = \langle \mathcal{I}', \mathcal{A}', \mathcal{T}', \mathcal{L}^{u\prime}, \mathcal{U}', \mathcal{R}' \rangle$, for $u \in \mathcal{U} \cap \mathcal{U}'$), with the former the starting point and the latter a modification of the former, and analyse the properties of their two corresponding TFs ($\langle \mathcal{X}, \mathcal{L}^-, \mathcal{L}^+, \mathcal{L}^0 \rangle$ and $\langle \mathcal{X}', \mathcal{L}^{-\prime}, \mathcal{L}^{+\prime}, \mathcal{L}^{0\prime} \rangle$, resp., assuming $\mathcal{X}' = \mathcal{X}$), taking $\mathcal{P}^u_\mathcal{X}$ to be a strength function $\sigma : \mathcal{X} \mapsto [-1, 1]$. We first prove that this σ is guaranteed to satisfy weak monotonicity, defined earlier, in a special case for the modification of ratings:

Proposition 3. *$\sigma = \mathcal{P}^u_\mathcal{X}$ satisfies the property of weak monotonicity at any (x, y) such that $\mathcal{R}'(u, x) = 0$.*

Proof. Here, for brevity, we refer to any quantity in the modified A-I $\mathcal{F}^{u\prime}$ with a prime $'$. Consider the case where $x \in \mathcal{L}^-(y)/x \in \mathcal{L}^+(y)/x \in \mathcal{L}^0(y)$. If $x \in \mathcal{I}$ and $y \in \mathcal{A}$ then by Definition 10 we know that $r^u(x) < 0/r^u(x) > 0/r^u(x) = 0$, resp. so setting $\mathcal{R}'(u, x) = 0$ gives $r^{u\prime}(x) > r^u(x)/r^{u\prime}(x) < r^u(x)/r^{u\prime}(x) = r^u(x)$, resp. and thus, by the fourth or fifth case of Definition 5, $\mathcal{P}^u_\mathcal{A}{}'(y) > \mathcal{P}^u_\mathcal{A}(y)/\mathcal{P}^u_\mathcal{A}{}'(y) < \mathcal{P}^u_\mathcal{A}(y)/\mathcal{P}^u_\mathcal{A}{}'(y) = \mathcal{P}^u_\mathcal{A}(y)$, resp. If $x \in \mathcal{A}$ and $y \in \mathcal{I}$ then by Definition 10 we know that $\mathcal{P}^u_\mathcal{A}(x) < 0/\mathcal{P}^u_\mathcal{A}(x) > 0/\mathcal{P}^u_\mathcal{A}(x) = 0$, resp., so setting $\mathcal{R}'(u, x) = 0$ gives $\mathcal{P}^u_\mathcal{A}{}'(x) > \mathcal{P}^u_\mathcal{A}(x)/\mathcal{P}^u_\mathcal{A}{}'(x) < \mathcal{P}^u_\mathcal{A}(x)/\mathcal{P}^u_\mathcal{A}{}'(x) = \mathcal{P}^u_\mathcal{A}(x)$, resp. and thus, by the second or third case of Definition 6, $\mathcal{P}^u_\mathcal{I}{}'(y) > \mathcal{P}^u_\mathcal{I}(y)/\mathcal{P}^u_\mathcal{I}{}'(y) < \mathcal{P}^u_\mathcal{I}(y)/\mathcal{P}^u_\mathcal{I}{}'(y) = \mathcal{P}^u_\mathcal{I}(y)$, resp. □

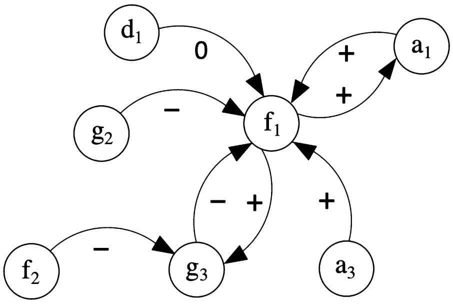
Thus, TFs may be deemed to represent the reasons for recommendations to users. In addition, the relations within this argumentative scaffolding may also highlight ways a user can modify the RS, intuitively driving the user's feedback in response to explanations drawn from the scaffolding (see Section 6) within interactions between the user and the RS. Consider, for example, the fact that attackers, supporters and neutralisers represent item-aspects which may be used as part of an explanation due to the distinct negative, positive or neutral, resp., effect they have on other item-aspects' predicted ratings. The following proposition characterises how they affect linked item-aspects (again, considering A-Is $\mathcal{F}^u$ and $\mathcal{F}^{u\prime}$ and corresponding TFs as mentioned earlier, before Proposition 3, and referring for brevity to any quantity in $\mathcal{F}^{u\prime}$ with a prime $'$, as in the proof of Proposition 3):

Proposition 4. *For any $x \in \mathcal{X}$ and $y \in (\mathcal{L}^-(x) \cup \mathcal{L}^+(x) \cup \mathcal{L}^0(x)) \cap (\mathcal{L}^{-\prime}(x) \cup \mathcal{L}^{+\prime}(x) \cup \mathcal{L}^{0\prime}(x))$:*

- *if $y \in \mathcal{A}$ and $\mathcal{P}^u_\mathcal{A}{}'(y) > \mathcal{P}^u_\mathcal{A}(y)$ then $\mathcal{P}^u_\mathcal{X}{}'(x) > \mathcal{P}^u_\mathcal{X}(x)$;*
- *if $y \in \mathcal{I}$ and $r^{u\prime}(y) > r^u(y)$ then $\mathcal{P}^u_\mathcal{X}{}'(x) > \mathcal{P}^u_\mathcal{X}(x)$;*
- *if $y \in \mathcal{A}$ and $\mathcal{P}^u_\mathcal{A}{}'(y) < \mathcal{P}^u_\mathcal{A}(y)$ then $\mathcal{P}^u_\mathcal{X}{}'(x) < \mathcal{P}^u_\mathcal{X}(x)$;*
- *if $y \in \mathcal{I}$ and $r^{u\prime}(y) < r^u(y)$ then $\mathcal{P}^u_\mathcal{X}{}'(x) < \mathcal{P}^u_\mathcal{X}(x)$.*

Proof. By inspection of Definitions 9 and 10 and Propositions 1 and 2. □

Intuitively, if we take an item-aspect's potential to reduce or increase other item-aspects' predicted ratings as an attacking or supporting strength, i.e. r^u for items (as given in Definition 10) and $\mathcal{P}^u_\mathcal{A}$ for aspects, adjustments to these attacking or supporting strengths may be characterised as *weakening* or *strengthening* attackers or supporters. In particular, we consider strengthening to be decreasing the r^u or $\mathcal{P}^u_\mathcal{A}$ of attackers (thus increasing their potential to reduce the strengths of



Item-Aspect	Predicted Rating
f$_1$	0.09
f$_2$	-1.00
a$_1$	0.13
a$_3$	1.00
d$_1$	0.00
g$_2$	-0.50
g$_3$	-0.38

Fig. 4. Example argumentative explanation for the recommendation of f_1 to u using the TF in Fig. 3 after cropping arguments without a path to f_1. Predicted ratings are also shown for reference.

arguments they attack) or increasing the r^u or $\mathcal{P}^u_{\mathcal{A}}$ of supporters (thus increasing their potential to increase the predicted ratings of arguments they support), and weakening to be the inverse. The property captured by Proposition 4 differs from the traditional properties for strength (i.e. gradual semantics), such as *reinforcement* [5] or *strict monotonicity* [12], where, for example, strengthening an attacker or a supporter always corresponds to increasing the attacked or supported, resp., argument's strength; this behaviour is not suitable here as strong attackers and strong supporters are at opposite ends of the same strength scale, i.e. $[-1, 1]$. We posit that our interpretation of weakening and strengthening makes sense from an argumentation viewpoint as, in TFs obtained from A-Is, an argument's semantic meaning is not fixed in that the argument may represent a user (strongly) liking its corresponding item-aspect if its strength is (extremely) positive or (strongly) disliking an item if the strength is (extremely) negative.

The next proposition shows that if a user profile remains constant, strengthening or weakening an argument can only affect another argument's predicted rating if there is a path from the former to the latter via the argumentative relations:

Proposition 5. *Let the profile π^u of user u be fixed. For any $x, y \in \mathcal{X}$, such that $r^{u\prime}(x) \neq r^u(x)$ if $x \in \mathcal{I}$ and $\mathcal{P}^u_{\mathcal{A}}{}'(x) \neq \mathcal{P}^u_{\mathcal{A}}(x)$ if $x \in \mathcal{A}$, $\mathcal{P}^u_{\mathcal{X}}{}'(y) \neq \mathcal{P}^u_{\mathcal{X}}(y)$ if and only if there exists a path $x_1, \ldots, x_n$ such that $x_1 = x$, $x_n = y$ and $x_{k-1} \in \mathcal{L}^-(x_k) \cup \mathcal{L}^+(x_k) \cup \mathcal{L}^0(x_k)$ for $k \in \{2, \ldots, n\}$.*

Proof. For any $(v, w) \in \mathcal{L}$, Definition 9 shows that if $(v, w) \in \mathcal{L}^u$ (and thus by Definition 10 $(v, w) \in \mathcal{L}^- \cup \mathcal{L}^+ \cup \mathcal{L}^0$), then $\mathcal{R}(u, w)$ is not defined and $r^u(v)$ is defined if $v \in \mathcal{I}$. It can thus be seen from Propositions 1 and 2 that if $(v, w) \in \mathcal{L}^- \cup \mathcal{L}^+ \cup \mathcal{L}^0$ and $r^{u\prime}(v) \neq r^u(v)$ or $\mathcal{P}^u_{\mathcal{A}}{}'(v) \neq \mathcal{P}^u_{\mathcal{A}}(v)$ then $\mathcal{P}^u_{\mathcal{X}}{}'(w) \neq \mathcal{P}^u_{\mathcal{X}}(w)$.

Let the default case for y be such that $\forall z \in \mathcal{L}^-(y) \cup \mathcal{L}^+(y) \cup \mathcal{L}^0(y)$, $r^{u\prime}(z) = r^u(z)$ if $z \in \mathcal{I}$ and $\mathcal{P}^u_{\mathcal{A}}{}'(z) \neq \mathcal{P}^u_{\mathcal{A}}(z)$ if $z \in \mathcal{A}$, i.e. nothing has changed and so by Definitions 5 and 6, $\mathcal{P}^u_{\mathcal{X}}{}'(y) = \mathcal{P}^u_{\mathcal{X}}(y)$.

If $x \in \mathcal{I}$ then the above logic extends to any $a \in \mathcal{A}$ such that $(x, a) \in \mathcal{L}^- \cup \mathcal{L}^+ \cup \mathcal{L}^0$, and then to any $i \in \mathcal{I}$ such that $(a, i) \in \mathcal{L}^- \cup \mathcal{L}^+ \cup \mathcal{L}^0$. If $x \in \mathcal{A}$ then the above logic extends to any $i \in \mathcal{I}$ such that $(x, i) \in \mathcal{L}^- \cup \mathcal{L}^+ \cup \mathcal{L}^0$. Thus the change is propagated if there is a path between x and y.

If y is such that $\mathcal{L}^-(y) \cup \mathcal{L}^+(y) \cup \mathcal{L}^0(y) = \emptyset$, by Definitions 9 and 10, $\mathcal{R}(u, y)$ is defined or $\forall i \in \mathcal{L}(y)$ and $\forall v \in \mathcal{U}$, $\mathcal{R}(v, y)$ is not defined, therefore $\mathcal{P}^u_{\mathcal{X}}{}'(y) = \mathcal{P}^u_{\mathcal{X}}(y)$. Thus the change is not propagated if there is no path between x and y.

Therefore the proposition holds. □

This means that in order to explain a predicted rating for a given item-aspect, we may crop the TF corresponding to $\mathcal{F}^u$ so that its explanation is the sub-graph of the TF consisting of the item-aspects with a path to the explained item-aspect only. This cropping, along with careful selection of the information which constitutes the explanation (described in the next section), can alleviate any scalability issues which may arise for argumentation scaffolding with many arguments. For example, Fig. 4 shows the sub-graph of the graph in Fig. 3 that may be seen as a qualitative explanation for the recommendation f_1 to user u, indicating all item-aspects affecting the recommendation.

We will re-examine the characteristics of the argumentative scaffolding considered in this section when we show examples of argumentative explanations drawn from the argumentative scaffolding, in the next section.

6. Argumentative explanations

Within our RS, TFs drawn from A-Is form the basis for a variety of *argumentative explanations* for recommendations dictated by predicted ratings. These explanations are 'argumentative' because they provide a rationale for the recommendations using, as their main 'skeleton', sub-graphs of TFs (e.g. as in Fig. 4) providing content which can then be presented incrementally to users in different formats (e.g. as in Fig. 5, discussed later) to support different styles of interaction. The argumentative scaffolding (Proposition 4 in particular) also points to controlled forms of feedback from users during interactions. Thus, the use of our argumentative scaffolding affords great adaptability to our RS, firstly in the explanations'

Table 3
Example variations in explanation content for f_1, with argumentative artefacts in the linguistic explanations highlighted in bold. Note that many other variations are possible.

Requirements	Content	Linguistic Explanation
All supporters of f_1	a_1, a_3	*Catch Me If You Can was recommended* ***because*** *you like Leonardo DiCaprio and Tom Hanks.*
Strongest attacker and strongest supporter of f_1	a_1, d_1	*Catch Me If You Can was recommended* ***because*** *you like Leonardo DiCaprio,* ***despite the fact that*** *you dislike Biographies.*
A weak attacker of f_1 and its own attacker	g_3, f_2	*Catch Me If You Can was not recommended* ***because*** *it inferred that you don't like Dramas,* ***since*** *you disliked Moulin Rouge.*

customisability with regards to content and format and secondly in the way it allows modifications to be made to the RS via feedback mechanisms in user interactions. In the remainder of this section we discuss and illustrate different choices of *content* and *format* of argumentative explanations (Section 6.1) and different forms of feedback (Section 6.2).

6.1. Explanation customisation

We first consider the explanation content, i.e. the information for the rationale behind a recommendation which is delivered to the user: the requirements for identifying this content obviously vary depending on user and context. We posit that the subgraph of the TF identified in Proposition 5 provides an excellent source for this information, since it represents every item-aspect which may have had an effect on the recommendation. This means that explanations that faithfully represent *how* a recommendation was determined[12] may be drawn from this subgraph (as in Fig. 4).

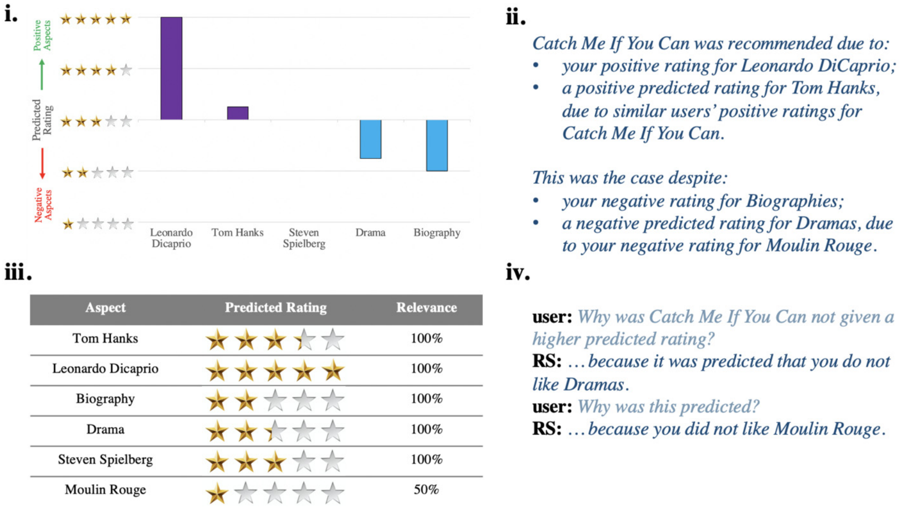
The content of an explanation may be selected from this subgraph depending on the user's requirements. For example, in a basic case where the user requests information on why an item was recommended, one straightforward way to provide an explanation is for the RS to determine the positive factors which led to this result, which, in our case, would be the supporters in the sub-graph of the TF. In the case of f_1 in the example in Fig. 4, this would correspond to the content column in the first row of Table 3, which in turn could be used to obtain a linguistic explanation as in the rightmost column in Table 3, using the conjunction *because* for the argumentative relation of support and utilising the full *width* of the supporters of f_1 in the TF. If a more balanced explanation for an item being recommended is required, the style of the explanation in the rightmost column in the second row of Table 3 may be more appropriate, where the strongest attacker and strongest supporter in (the sub-graph of) the TF are shown, again using appropriate conjunctions for the argumentative relations. However, this still uses only width in the TF and ignores reasons for and against used arguments. In our running example, consider the case where f_1 was *not* recommended; the third row of Table 3 shows how depth may be used to justify the RS's inference on the user's sentiment on *Dramas*. Here, the language represents the resolutely negative effects along this chain of reasoning.

We have provided a number of examples for selecting the content of explanations from TFs, but note that other methods could be useful, e.g. including neutralisers when the RS is explaining its uncertainty about an inference. Balog et al. [8] use templates to generate inferences of a user's sentiment on aspects in pairwise comparisons, e.g. *Catch Me If You Can was recommended* **because** *you like Biographies,* **especially** *those starring Leonardo DiCaprio*. Our argumentative scaffolding could support such explanations by comparing the aspects of a movie with their linked items, e.g. in our running example (to use a crude method) if all the items which are linked to both g_2 and a_3 are rated more highly than those linked to g_2 but not a_3, we may construct the same argumentative explanation.

Up to now we have only considered explanations of a linguistic format but other formats are possible, and the choice of the format is an important factor in how receptive a user is towards explanations [33]. The optimal format for an explanation varies significantly based on a range of factors including the application towards which the explanation is targeted and the goals of the explainee [49]. For example, a researcher testing an RS may prefer a graphical format which is true to the argumentative scaffolding itself, whereas a non-expert user may prefer a linguistic approach which gives the information in a natural, human-like manner. Argumentation frameworks themselves have been shown to be an effective way of supporting anthropomorphised *conversational* explanations, e.g. Cocarascu et al. [24] extract argumentation frameworks as explanations for review aggregations which are then used to facilitate conversations with users. Other forms of explanations which have been shown to be beneficial in RSs include tabular explanations, e.g. as in [70], where (paraphrased in our setting) the attacking and supporting item-aspects in an explanation may be represented in a table with other attributes shown, e.g. the item-aspect's strength and distance from the recommendation. Visual explanations in the form of charts have also been shown to perform well in studies on user preferences [38].

Fig. 5 shows four alternative formats (in addition to the graphical format afforded by sub-graphs of TFs) of user explanation for the example from Fig. 4. Specifically, Fig. 5i shows a visual explanation in the form of charts exploiting the width in

[12] Note that not all RS explanations have this aim, e.g. see [50].



Catch Me If You Can was recommended due to:

- *your positive rating for Leonardo DiCaprio;*
- *a positive predicted rating for Tom Hanks, due to similar users' positive ratings for Catch Me If You Can.*

This was the case despite:

- *your negative rating for Biographies;*
- *a negative predicted rating for Dramas, due to your negative rating for Moulin Rouge.*

Aspect	Predicted Rating	Relevance
Tom Hanks	★★★½☆	100%
Leonardo Dicaprio	★★★★★	100%
Biography	★★☆☆☆	100%
Drama	★★½☆☆	100%
Steven Spielberg	★★★☆☆	100%
Moulin Rouge	★☆☆☆☆	50%

user: *Why was Catch Me If You Can not given a higher predicted rating?*
RS: *...because it was predicted that you do not like Dramas.*
user: *Why was this predicted?*
RS: *...because you did not like Moulin Rouge.*

Fig. 5. Possible visual (i), linguistic (ii), tabular (iii) and conversational (iv) explanations for f_1's predicted rating in our running example.

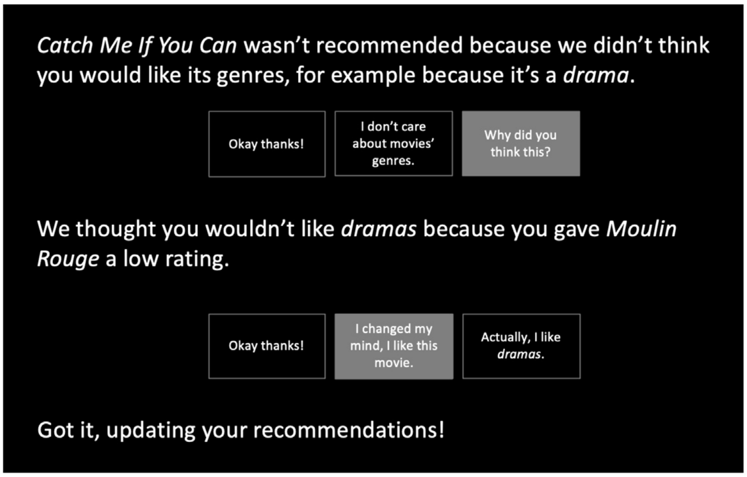
the TF, i.e. attacking and supporting aspects coloured by type and organised by their corresponding predicted ratings, thus giving the user a clear indication of each aspect's contribution to the predicted rating of the recommended item. Fig. 5ii, meanwhile, targets both depth and width in a linguistic format, which may be textual or spoken, e.g. by an AI assistant, depending on the requirements and preferences of the user. These explanations may be generated by templates or more complicated natural language generation processes, both employing the TF as the underlying knowledge base. Similar information is utilised in Fig. 5iii, which shows a tabular explanation similar to those of Vig et al. [70], where predicted ratings (translated to a 1-5 star scale) are shown alongside a *relevance* parameter, calculated here by inverting the distance from the recommended item. Finally, Fig. 5iv shows a conversational explanation, where the user has requested a counterfactual explanation as to why the item was not rated more highly. As the conversation progresses, the RS may step through the TF to formulate reasoning for its interactions, to which the user may respond with (possibly predetermined, as in [24]) responses. As with the linguistic explanations, conversational explanations may be textual or spoken. Note that other explanation formats are possible, in addition to those exemplified in Fig. 5, e.g. word clouds, as in [75], generated from supporters with word size weighted by strength.

6.2. Feedback

We now consider our RS's capability for explanation-driven feedback, regarding the way in which a user may interact with the explanation to provide the RS with more information. This is an important factor in that recommendations are highly unlikely to be perfect the first time and, even if they are, user preferences are dynamic and so in the ideal case an RS will adapt to their changes over time [21]. Our consideration here is whether and how the RS is able to elicit more information from the user via feedback mechanisms in these interactions.

Our explanations can leverage the argumentative reading of recommendations afforded by TFs to support feedback. For example, let us focus on explanations for a positive or negative predicted rating consisting of strong supporters or strong attackers, resp. In both cases, if the user disagrees with the predicted rating of the recommended item being so high or so low, resp., weakening the supporters or attackers, resp., will be guaranteed to adjust the predicted rating as desired, by Proposition 4. Meanwhile, if a user agrees with the contribution of an attacker or supporter, strengthening it will increase the effect it has. In the visual and tabular explanations in Fig. 5, it is easy to see how this intuitive behaviour allows simple indications of potential adjustments to the predicted ratings to be integrated into the explanation format such that their effect on the recommended item's predicted rating is clearly shown to the user. For example, a modifiable bar in the chart or selectable stars in the table for *Leonardo DiCaprio* could be shown along with an indication that any reduction in the predicted rating for *Leonardo DiCaprio* (thus the weakening of a supporter) would in turn reduce the predicted rating of Catch Me If You Can.

Other modifications supported by our RS, e.g. adjusting the user-specific constants or selecting a different set of similar users, could also be enacted by the argumentative explanations, e.g. if a user states that they care less/more about a particular type or that they do not consider the similar users' tastes to align with their own, resp. In the linguistic and conversational explanations, template-based interactions could be structured to include selectable user responses initiating desired modifications. For example, if a user initiates a conversational explanation with an indicated discrepancy, e.g. *I liked Catch Me If You Can, why didn't you recommend it to me?*, then the interaction with the user may be structured to include some of the possible modifications we have mentioned, e.g. as shown in Fig. 6. In the first interaction here, the user is told that the genres, particularly *Drama*, were the main reasons (possibly obtained by determining the type with the strongest



Catch Me If You Can wasn't recommended because we didn't think you would like its genres, for example because it's a *drama*.

Okay thanks!
I don't care about movies' genres.
Why did you think this?

We thought you wouldn't like *dramas* because you gave *Moulin Rouge* a low rating.

Okay thanks!
I changed my mind, I like this movie.
Actually, I like *dramas*.

Got it, updating your recommendations!

Fig. 6. An example conversational interaction driven by the argumentative explanations.

attackers) for this movie not being recommended. The user may then state they are satisfied with the explanation, reduce μ_{genre} (which may be guaranteed to increase *Catch Me If You Can*'s predicted rating due to the genres' negative effect on it) or ask for more reasons. In the illustration in the figure, the user selects the third option, and in the second interaction the attacker *Moulin Rouge* is highlighted as the negative reasoning. The user may then state that they are satisfied with the explanation or give a higher rating to *Moulin Rouge* or *Drama*, both of which may be guaranteed to increase *Catch Me If You Can*'s predicted rating by Proposition 4.

Less constrained approaches may also be taken within an iterative feedback process: if some of this (unconstrained) feedback leads to temporary unintended effects on other item-aspects' predicted ratings, further interactions will provide an opportunity for recalibration to adhere to users' preferences.

7. User studies

We now present the results from two user studies examining the qualities of argumentative explanations as judged by humans, specifically with regards to varying the explanations' content and format. Our aim with these user studies is to explore whether the variety of contents and formats that our RS encompasses is worthwhile, and to inform deployment of our methodology to build applications. Given this focus, we leave as future work a comparison with explanations offered by other methods from the literature (including by other hybrid RSs and by content-based RSs). We also leave as future work user studies on explanations' feedback, since feedback warrants a full investigation in itself, requiring full deployment of our RS and data collection on recommendations' take-up.

In order to perform these user studies, we used crowdsourcing to elicit ratings from, recommend movies to and generate explanations for participants, before asking them to rate the varying explanations that we provided. This is an inherently imperfect process since it is difficult to measure how users evaluate explanations of recommendations without consuming the items that are recommended to them, though many works have somewhat remedied this issue, e.g. [15], where instead of asking users to read a book they have been recommended, they are asked to read a summary describing it. There is also evidence that in the particular domain of movies, users are able to estimate the value of items reasonably well even without consuming them [46].

Our first user study concerns the *content* of an explanation. We used Amazon's *Mechanical Turk* to conduct an experiment where 76 participants were asked to rate a maximum of 70 movies on a scale of 1-5 stars (translated to our $[-1, 1]$ scale, straightforwardly, using the formula $y = ((x - 1)/2) - 1$), along with the option to state that they had not seen the movie. Once participants had rated ten films positively (> 3 stars) or after rating all films, we calculated recommendations using our RS for the participants who had rated at least 2 movies positively (to avoid cases where there was inadequate information for the RS to generate reasonable recommendations). Participants were shown three recommendations, i.e. movies they had not seen with the highest predicted rating, along with an explanation for each. As mentioned, we wished to vary explanation content in this study, and so we varied the width and depth of the TF utilised in the explanations (as discussed in Section 6.1).

We delivered the explanations in a linguistic, textual format throughout and we did not permit any feedback from participants. We thus had:

- a *baseline* explanation which did not utilise the argumentative scaffolding: *users similar to you gave high ratings to this film*; we chose this baseline as it has a similar content to explanations found in the literature (e.g. those of Herlocker et al. [38]) and is thus likely familiar to users;

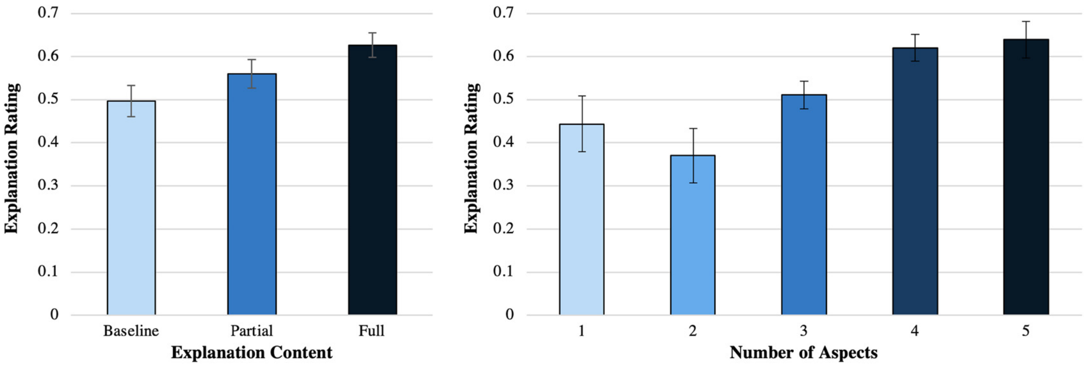


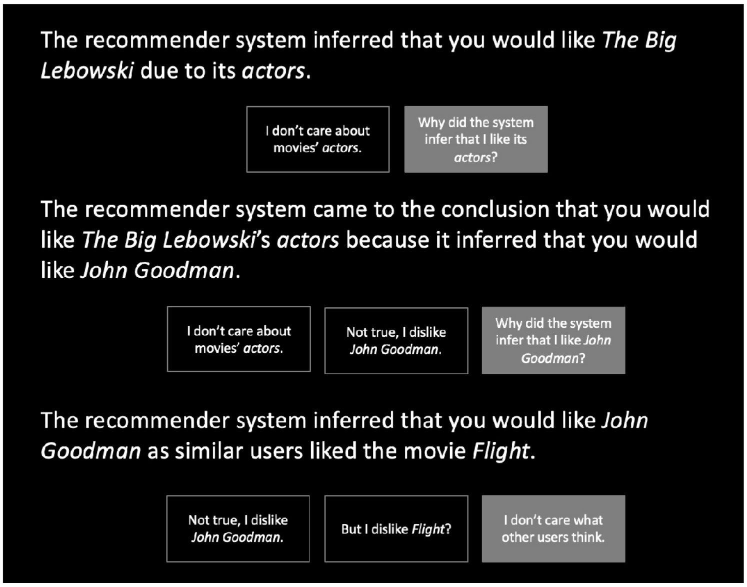
Fig. 7. Results from the first user study showing the mean explanation rating from users by explanation format and number of aspects with standard error indicated.

- one *partial* explanation which utilised width by mentioning between 1 and 5 of the recommended item's linked aspects which had the highest predicted ratings, i.e. the recommended item's strongest supporters, e.g. *you like films starring Anjelica Huston, films starring Owen Wilson, films directed by Wes Anderson, comedy films and Drama films*; and
- one *full* explanation which utilised width and depth by taking the partial explanations and adding these aspects' linked items which had the highest predicted ratings, i.e. their strongest supporters, e.g. *you like films directed by Ridley Scott such as Gladiator, Action films such as Mission: Impossible II and Drama films such as The Big Kahuna*.

The number of examples depended on the number of supporters that were available in the argumentative scaffolding. For each explanation, participants were prompted with the message 'what follows is an explanation for our recommendation, please rate this explanation', followed by the explanation and a scale (represented as [0,1] here) allowing the participants to provide their rating, from *very poor* (represented as 0 here) to *great* (represented as 1 here), i.e. the users saw only *very poor* and *great*, not the numbers. The order of the three explanations (baseline, partial, full) was randomly determined for each participant. Fig. 7 shows that participants generally preferred full compared to partial explanations and partial compared to baseline explanations, in agreement with previous research [13]. This also aligns with the findings of Kulesza et al. [43] that users value completeness in explanations, even at the expense of soundness (though our argumentative explanations are necessarily sound due to their faithfulness to the RS). Similarly, we find that the more aspects present, the higher the explanation rating. This visual inspection is supported by an analysis of variance, that shows a significant effect of explanation type ($F(2, 216) = 4.31$, $p = .014$) and number of aspects ($F(1, 216) = 18.53$, $p < .001$) but no effect of order ($F(1, 216) = 0.13$, $p = .720$) and no significant interactions ($p > .140$). Interestingly, although there is a general preference for more information, since explanations with 4 or 5 aspects received significantly higher ratings than explanations with 1, 2 or 3 aspects, there might be evidence for a non-linear effect, since we observe a very small, non-significant difference between explanations with 4 vs. explanations with 5 aspects. Of course, this saturation effect requires further validation and experimentation with explanations that contain more aspects than used here; for example, it would be interesting to consider whether a threshold for the number of aspects included may exist towards higher user satisfaction. We leave this investigation to future work. A final point we can take from the experiment overall was that, despite the general trend towards explanations with more information, users' explanatory preferences clearly varied among the three types and over the number of aspects, showing the value of argumentative scaffolding's ability to support numerous forms of argumentative explanation.

In our second user study, we investigate user take-up on various *formats* for explanations. Specifically, we consider three forms of *interactive explanations* (IEs), which are: *IE1*, of a *tabular* format; *IE2*, of a *linguistic* format; and *IE3*, of a *conversational* format. These are along the lines of formats (ii)-(iv) in Fig. 5 (we left the evaluation of take-up of the *visual* format of explanations, i.e. format (i) in Fig. 5, to future work). Note that IE1 contains somewhat less content than IE2 and IE3, since it utilises only width, and not depth, of the underpinning argumentative scaffolding.

We once again used *Mechanical Turk* to conduct our experiment, where we asked 75 participants for ratings on 70 movies each before generating explanations for recommendations for the participants who had rated at least five movies including at least three positive and/or three negative ratings. Each participant was then presented with three positive recommendations and/or (resp.) three negative recommendations. If a participant showed disagreement with the recommendation, we offered two types of IE for that recommendation, asking the participants which of the two IEs they preferred to give pairwise comparisons between explanation formats. We had 51 such occurrences, as 23 participants did not rate enough movies or indicated that they agreed with the recommendations (and one participant gave nonsensical responses). These pairwise comparisons were IE1 vs. IE2 (i.e. comparing tabular and textual formats) and IE2 vs. IE3 (i.e. comparing textual and conversational formats). We also counterbalanced the order of presentation, and so showed: 17 participants IE1 and then IE2, 15 participants IE2 and then IE1, 9 participants IE2 and then IE3 and 10 participants IE3 and then IE2. An example IE3 adapted from the second user study is shown in Fig. 8 (showing how content and collaborative filtering can both be channeled within the interaction with the user).



The recommender system inferred that you would like *The Big Lebowski* due to its *actors*.

I don't care about movies' *actors*.
Why did the system infer that I like its *actors*?

The recommender system came to the conclusion that you would like *The Big Lebowski*'s *actors* because it inferred that you would like *John Goodman*.

I don't care about movies' *actors*.
Not true, I dislike *John Goodman*.
Why did the system infer that I like *John Goodman*?

The recommender system inferred that you would like *John Goodman* as similar users liked the movie *Flight*.

Not true, I dislike *John Goodman*.
But I dislike *Flight*?
I don't care what other users think.

Fig. 8. An example conversational interaction from the second user study.

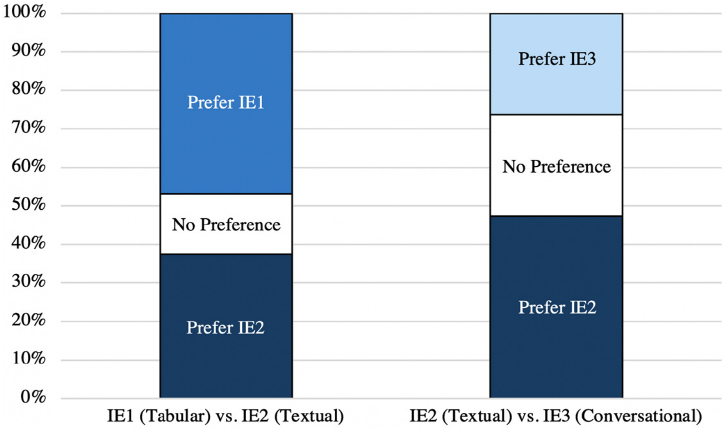


Fig. 9. Results from the second user study showing the participants' diverse preferences in explanation format in the pairwise comparisons. (The order in which explanations were delivered to the participants had little effect on these results and so this was ignored.)

The overall results from our second user study are shown in Fig. 9, demonstrating a slight preference for IE1 over IE2, i.e. tabular format over textual (despite the tabular IE having slightly less information), and a preference for IE2 over IE3, i.e. textual format over conversational. The sample size was fairly small but feedback from the participants regarding the latter comparison referred to a preference for having all the information up front, statically. However, the most clear finding was that the participants' preferences varied significantly regarding explanation format, highlighted by the fact that few participants opted for the 'no preference' option. This demonstrates the importance of the argumentative scaffolding, which naturally supports a host of different explanation formats, allowing diverse user preferences to be accounted for.

8. Conclusions

We have proposed a novel RS, built on argumentative scaffolding. The method is explainable, in a flexible way, while also giving reasonable results when compared with existing (non-explainable) RSs in the literature (which we have proven empirically): it is encouraging to see that effectiveness (in the form of precision) does not need to be sacrificed to support explainability. Our RS is also adaptive in two important ways. The argumentative scaffolding firstly allows the extraction of a diverse repertoire of explanations, varying in their characteristics of content and format, and so allows the customisation to the explanatory preferences of users, which we show in a user study both differ across users and show a general inclination towards more information. The second adaptive feature of our method is the ability to support feedback mechanisms, given the link between the argumentative scaffolding/explanations and the predicted ratings driving the recommendations, proven here via theoretical analysis. We illustrate with a number of examples how the former capability may be used to address the issue of an unachievable one-size-fits-all explanation for humans, and how the latter helps to achieve the perpetual goal for RSs of aligning recommendations with user preferences, particularly because they are dynamic and thus change over time.

This paper opens a number of potentially fruitful avenues for future work. We would firstly like to undertake a comprehensive analysis of the explanation content, varying not just width and depth in the explanations but also the method for selecting arguments, e.g. considering attackers to generate counterfactual explanations or pairwise comparisons of arguments as in [8]. Further user studies on the explanation format would also be a very interesting research direction, since the most suitable format would likely not only depend on the individual user but also the chosen application. An investigation aiming to ascertain the most effective forms of feedback mechanisms (not covered in the experiments described above) would require a much lengthier and more complex experiment, e.g. to ensure statistical significance in rating accuracy improvements after feedback had been provided (again, see [8]). The experimental process could also be further enhanced, e.g. by asking participants how much they would pay to watch a certain (unseen) movie, after having seen different types of explanations. Further improvements to the RS itself could also be targeted including generating the user-specific constants systematically and optimally, e.g. by learning on bootstrapping, or by allowing different forms of aspects to be considered, e.g. those of a continuous nature. Moreover, it would be interesting to consider hierarchical forms of A-Is, where aspects may admit sub-aspects and the argumentation scaffolding would be, as a result, more complex. Another important line of work would be the deployment of the RS in other contexts, e.g. in recommendation in e-commerce or music streaming, where A-Is seem to be perfectly suitable, as does the interactive nature of our explanations. These alternative contexts, e.g. music streaming, may be more amenable to implicit feedback, as discussed in Section 2. It would also be interesting to investigate how the RS's context affects the preferred explanatory characteristics, e.g. is the requirement for more information more prevalent in an e-commerce setting? The study of the integration of mechanisms for implicit feedback, e.g. users stopping movies early as negative ratings, into our RS is also left as future work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was partially funded by the UK Human-Like Computing EPSRC Network of Excellence EP/R022291/1. 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. The authors are grateful to all members of the Computational Logic and Argumentation group at Imperial College London for useful feedback on the set up for the user studies in Section 7, and to Ben Wilkins and Jinfeng Zhong for pointing out some typographical errors in [54], rectified here. Finally, we would like to thank the AI Journal's reviewers and the editor for their suggestions, which we believe led us to significantly improve this paper.

References

[1] B. Abdollahi, O. Nasraoui, Using explainability for constrained matrix factorization, in: Proceedings of the 11th ACM Conference on Recommender Systems, RecSys, 2017, pp. 79–83.
[2] C.C. Aggarwal, Recommender Systems - The Textbook, Springer, 2016.
[3] M. Aliannejadi, H. Zamani, F. Crestani, W.B. Croft, 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, 2019, pp. 475–484.
[4] N.S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, Am. Stat. 46 (1992) 175–185.
[5] L. Amgoud, J. Ben-Naim, D. Doder, S. Vesic, Acceptability semantics for weighted argumentation frameworks, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI, 2017, pp. 56–62.
[6] K. Atkinson, P. Baroni, M. Giacomin, A. Hunter, H. Prakken, C. Reed, G.R. Simari, M. Thimm, S. Villata, Towards artificial argumentation, AI Mag. 38 (2017) 25–36.
[7] K. Balog, F. Radlinski, Measuring recommendation explanation quality: the conflicting goals of explanations, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 329–338.
[8] K. Balog, F. Radlinski, S. Arakelyan, 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, 2019, pp. 265–274.
[9] P. Baroni, G. Comini, A. Rago, F. Toni, Abstract games of argumentation strategy and game-theoretical argument strength, in: PRIMA: Principles and Practice of Multi-Agent Systems - 20th International Conference, 2017, pp. 403–419.
[10] P. Baroni, D. Gabbay, M. Giacomin, L. van der Torre (Eds.), Handbook of Formal Argumentation, College Publications, 2018.
[11] P. Baroni, A. Rago, F. Toni, How many properties do we need for gradual argumentation?, in: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018, pp. 1736–1743.
[12] P. Baroni, A. Rago, F. Toni, From fine-grained properties to broad principles for gradual argumentation: a principled spectrum, Int. J. Approx. Reason. 105 (2019) 252–286.
[13] C. Bechlivanidis, D.A. Lagnado, J.C. Zemla, S. Sloman, Concreteness and abstraction in everyday explanation, Psychon. Bull. Rev. (2017) 1–14.
[14] P. Bedi, P.B. Vashisth, Argumentation-enabled interest-based personalised recommender system, J. Exp. Theor. Artif. Intell. 27 (2015) 199–226.
[15] M. Bilgic, R.J. Mooney, Explaining recommendations: satisfaction vs. promotion, in: Beyond Personalization Workshop, IUI, vol. 5, 2005, p. 153.

[16] D. Billsus, M.J. Pazzani, Learning collaborative information filters, in: Proceedings of the 15th International Conference on Machine Learning ICML, 1998, pp. 46–54.
[17] C.E. Briguez, M.C. Budán, C.A.D. Deagustini, A.G. Maguitman, M. Capobianco, G.R. Simari, Argument-based mixed recommenders and their application to movie suggestion, Expert Syst. Appl. 41 (2014) 6467–6482.
[18] R.D. Burke, Hybrid recommender systems: survey and experiments, User Model. User-Adapt. Interact. 12 (2002) 331–370.
[19] R.D. Burke, Hybrid systems for personalized recommendations, in: Intelligent Techniques for Web Personalization, IJCAI Workshop, ITWP, 2003, pp. 133–152.
[20] C. Cayrol, M.-C. Lagasquie-Schiex, On the acceptability of arguments in bipolar argumentation frameworks, in: Symbolic and Quantitative Approaches to Reasoning with Uncertainty, 8th European Conference, ECSQARU, 2005, pp. 378–389.
[21] X. Chen, Y. Zhang, Z. Qin, Dynamic explainable recommendation based on neural attentive models, in: The 33rd AAAI Conference on Artificial Intelligence, 2019, pp. 53–60.
[22] C.I. Chesñevar, A.G. Maguitman, M.P. González, Empowering recommendation technologies through argumentation, in: Argumentation in Artificial Intelligence, Springer, 2009, pp. 403–422.
[23] K. Christakopoulou, F. Radlinski, K. Hofmann, Towards conversational recommender systems, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 815–824.
[24] O. Cocarascu, A. Rago, F. Toni, Extracting dialogical explanations for review aggregations with argumentative dialogical agents, in: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS, 2019, pp. 1261–1269.
[25] F. Costa, S. Ouyang, P. Dolog, A. Lawlor, Automatic generation of natural language explanations, in: Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion, 2018, pp. 57:1–57:2.
[26] M. Czarkowski, A scrutable adaptive hypertext, Ph.D. thesis, University of Sydney, Australia, 2006, http://hdl.handle.net/2123/10206.
[27] M.F. Dacrema, P. Cremonesi, D. Jannach, 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, pp. 101–109.
[28] J. Dalton, V. Ajayi, R. Main, Vote goat: conversational movie recommendation, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018, pp. 1285–1288.
[29] M. Du, N. Liu, X. Hu, Techniques for interpretable machine learning, Commun. ACM 63 (2020) 68–77.
[30] P.M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artif. Intell. 77 (1995) 321–357.
[31] D.M. Gabbay, Logical foundations for bipolar and tripolar argumentation networks: preliminary results, J. Log. Comput. 26 (2016) 247–292.
[32] A.J. García, G.R. Simari, Defeasible logic programming: an argumentative approach, Theory Pract. Log. Program. 4 (2004) 95–138.
[33] F. Gedikli, D. Jannach, M. Ge, How should I explain? A comparison of different explanation types for recommender systems, Int. J. Hum.-Comput. Stud. 72 (2014) 367–382.
[34] T. George, S. Merugu, A scalable collaborative filtering framework based on co-clustering, in: Proceedings of the 5th IEEE International Conference on Data Mining ICDM, 2005, pp. 625–628.
[35] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2019) 93:1–93:42.
[36] F.M. Harper, J.A. Konstan, The movielens datasets: history and context, ACM Trans. Interact. Intell. Syst. 5 (2015) 19:1–19:19.
[37] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T. Chua, Neural collaborative filtering, in: Proceedings of the 26th International Conference on World Wide Web, WWW, 2017, pp. 173–182.
[38] J.L. Herlocker, J.A. Konstan, J. Riedl, Explaining collaborative filtering recommendations, in: Proceeding on the ACM Conference on Computer Supported Cooperative Work, CSCW, 2000, pp. 241–250.
[39] N. Hug, Surprise, a Python library for recommender systems, http://surpriselib.com, 2017.
[40] A. Ignatiev, Towards trustable explainable AI, in: Proceedings of the 29th International Joint Conference on Artificial Intelligence, IJCAI, 2020, pp. 5154–5158.
[41] D. Kim, B. Suh, Enhancing vaes for collaborative filtering: flexible priors & gating mechanisms, in: Proceedings of the 13th ACM Conference on Recommender Systems, RecSys, 2019, pp. 403–407.
[42] Y. Koren, R.M. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, IEEE Comput. 42 (2009) 30–37.
[43] T. Kulesza, S. Stumpf, M.M. Burnett, S. Yang, I. Kwan, W. Wong, Too much, too little, or just right? Ways explanations impact end users' mental models, in: IEEE Symposium on Visual Languages and Human Centric Computing, 2013, pp. 3–10.
[44] J. Kunkel, T. Donkers, L. Michael, C. Barbu, J. Ziegler, Let me explain: impact of personal and impersonal explanations on trust in recommender systems, in: Conference on Human Factors in Computing Systems, CHI, 2019, p. 487.
[45] D. Lemire, A. Maclachlan, Slope one predictors for online rating-based collaborative filtering, CoRR, arXiv:cs/0702144, 2007.
[46] B. Loepp, T. Donkers, T. Kleemann, J. Ziegler, Impact of item consumption on assessment of recommendations in user studies, in: Proceedings of the 12th ACM Conference on Recommender Systems, RecSys, 2018, pp. 49–53.
[47] X. Luo, M. Zhou, Y. Xia, Q. Zhu, An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems, IEEE Trans. Ind. Inform. 10 (2014) 1273–1284.
[48] J. McInerney, B. Lacker, S. Hansen, K. Higley, H. Bouchard, A. Gruson, R. Mehrotra, Explore, exploit, and explain: personalizing explainable recommendations with bandits, in: Proceedings of the 12th ACM Conference on Recommender Systems, RecSys, 2018, pp. 31–39.
[49] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267 (2019) 1–38.
[50] C. Musto, F. Narducci, P. Lops, M. de Gemmis, G. Semeraro, Linked open data-based explanations for transparent recommender systems, Int. J. Hum.-Comput. Stud. 121 (2019) 93–107.
[51] S. Naveed, T. Donkers, J. Ziegler, 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, pp. 293–298.
[52] F. Radlinski, N. Craswell, A theoretical framework for conversational search, in: Conference on Conference Human Information Interaction and Retrieval, CHIIR, 2017, pp. 117–126.
[53] A. Rago, O. Cocarascu, C. Bechlivanidis, F. Toni, Argumentation as a framework for interactive explanations for recommendations, in: 17th International Conference on Principles of Knowledge Representation and Reasoning, KR, 2020.
[54] A. Rago, O. Cocarascu, F. Toni, Argumentation-based recommendations: fantastic explanations and how to find them, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI, 2018, pp. 1949–1955.
[55] P. Resnick, H.R. Varian, Recommender systems, Commun. ACM 40 (1997) 56–58.
[56] P. Rodríguez, S. Heras, J. Palanca, J.M. Poveda, N.D. Duque, V. Julián, An educational recommender system based on argumentation theory, AI Commun. 30 (2017) 19–36.
[57] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (2019) 206.
[58] T. Schnabel, P.N. Bennett, T. Joachims, Shaping feedback data in recommender systems with interventions based on information foraging theory, in: Proceedings of the 12th ACM International Conference on Web Search and Data Mining, WSDM, 2019, pp. 546–554.

[59] S. Seo, J. Huang, H. Yang, Y. Liu, Interpretable convolutional neural networks with dual local and global attention for review rating prediction, in: Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys, 2017, pp. 297–305.

[60] A. Sepliarskaia, J. Kiseleva, F. Radlinski, M. de Rijke, Preference elicitation as an optimization problem, in: Proceedings of the 12th ACM Conference on Recommender Systems, RecSys, 2018, pp. 172–180.

[61] T. Silveira, M. Zhang, X. Lin, Y. Liu, S. Ma, How good your recommender system is? A survey on evaluations in recommendation, Int. J. Mach. Learn. Cybern. 10 (2019) 813–831.

[62] K. Sokol, P.A. Flach, Explainability fact sheets: a framework for systematic assessment of explainable approaches, in: FAT* '20: Conference on Fairness, Accountability, and Transparency, 2020, pp. 56–67.

[63] Y. Sun, Y. Zhang, Conversational recommender system, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR, 2018, pp. 235–244.

[64] J.C. Teze, S. Gottifredi, A.J. García, G.R. Simari, Improving argumentation-based recommender systems through context-adaptable selection criteria, Expert Syst. Appl. 42 (2015) 8243–8258.

[65] N. Tintarev, J. Masthoff, A survey of explanations in recommender systems, in: Proceedings of the 23rd International Conference on Data Engineering Workshops, ICDE, 2007, pp. 801–810.

[66] N. Tintarev, J. Masthoff, Explaining recommendations: design and evaluation, in: Recommender Systems Handbook, 2015, pp. 353–382.

[67] S.E. Toulmin, The Uses of Argument, Cambridge University Press, 1958.

[68] A. Töscher, M. Jahrer, R.M. Bell, The BigChaos Solution to the Netflix Grand Prize, 2009.

[69] L. van Velsen, T. van der Geest, R. Klaassen, M.F. Steehouder, User-centered evaluation of adaptive and adaptable systems: a literature review, Knowl. Eng. Rev. 23 (2008) 261–281.

[70] J. Vig, S. Sen, J. Riedl, Tagsplanations: explaining recommendations using tags, in: Proceedings of the 14th International Conference on Intelligent User Interfaces, IUI, 2009, pp. 47–56.

[71] M.G. Vozalis, K.G. Margaritis, Applying SVD on generalized item-based filtering, Int. J. Comput. Sci. Appl. 3 (2006) 27–51.

[72] B. Walek, V. Fojtik, A hybrid recommender system for recommending relevant movies using an expert system, Expert Syst. Appl. 158 (2020) 113452.

[73] H. Wang, N. Wang, D. Yeung, Collaborative deep learning for recommender systems, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 1235–1244.

[74] X. Wang, X. He, Y. Cao, M. Liu, T. Chua, KGAT: knowledge graph attention network for recommendation, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD, 2019, pp. 950–958.

[75] Y. Wu, M. Ester, FLAME: a probabilistic model combining aspect based opinion mining and collaborative filtering, in: Proceedings of the 8th ACM International Conference on Web Search and Data Mining, WSDM, 2015, pp. 199–208.

[76] Y. Xian, Z. Fu, S. Muthukrishnan, G. de Melo, Y. Zhang, Reinforcement knowledge graph reasoning for explainable recommendation, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 285–294.

[77] J. Zhang, X. Shi, S. Zhao, I. King, STAR-GCN: stacked and reconstructed graph convolutional networks for recommender systems, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI, 2019, pp. 4264–4270.

[78] Y. Zhang, X. Chen, Explainable recommendation: a survey and new perspectives, Found. Trends Inf. Retr. 14 (2020) 1–101.

[79] Y. Zhang, X. Chen, Q. Ai, L. Yang, W.B. Croft, Towards conversational search and recommendation: system ask, user respond, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM, 2018, pp. 177–186.

[80] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, S. Ma, Explicit factor models for explainable recommendation based on phrase-level sentiment analysis, in: The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2014, pp. 83–92.

[81] Q. Zhao, F.M. Harper, G. Adomavicius, J.A. Konstan, Explicit or implicit feedback? Engagement or satisfaction?: a field experiment on machine-learning-based recommender systems, in: Proceedings of the 33rd Annual ACM Symposium on Applied Computing, SAC, 2018, pp. 1331–1340.

[82] L. Zheng, V. Noroozi, P.S. Yu, Joint deep modeling of users and items using reviews for recommendation, in: Proceedings of the 10th ACM International Conference on Web Search and Data Mining, WSDM, 2017, pp. 425–434.