Datenbank Spektrum (2020) 20:181–187
https://doi.org/10.1007/s13222-020-00350-y

SCHWERPUNKTBEITRAG

Leveraging Arguments in User Reviews for Generating and Explaining Recommendations

Tim Donkers[1] · **Jürgen Ziegler**[1]

Received: 20 February 2020 / Accepted: 10 June 2020 / Published online: 1 July 2020
© The Author(s) 2020

Abstract
Review texts constitute a valuable source for making system-generated recommendations both more accurate and more transparent. Reviews typically contain statements providing argumentative support for a given item rating that can be exploited to explain the recommended items in a personalized manner. We propose a novel method called Aspect-based Transparent Memories (ATM) to model user preferences with respect to relevant aspects and compare them to item properties to predict ratings, and, by the same mechanism, explain why an item is recommended. The ATM architecture consists of two neural memories that can be viewed as arrays of slots for storing information about users and items. The first memory component encodes representations of sentences composed by the target user while the second holds an equivalent representation for the target item based on statements of other users. An offline evaluation was performed with three datasets, showing advantages over two baselines, the well-established Matrix Factorization technique and a recent competitive representative of neural attentional recommender techniques.

Keywords Recommender Systems · Explanations · Memory Networks

1 Introduction

Deciding which news articles to read, which product to buy, or which hotel to book has become an increasingly difficult task for web users due to the sheer amount of options available. In recent years, recommender systems (RS) have become well-established tools for alleviating the user's search and decision-making in such applications [25]. A recommendation issued by such a system can be considered a specific form of a claim, namely that the user will find the recommended item useful or pleasing. In contrast to classic argumentation theory, a recommendation claims neither general nor exclusive validity but is often personalized and may depend on local, temporal or other contextual factors. Recommendations typically do not aim at influencing a person's long-term beliefs, rather, they aim at supporting users in their decision-making in a specific interactive context such as an online shop, therefore also involving a strongly persuasive component.

Conventional recommender systems mostly function as black boxes and do not provide the user with explanations why a recommendation is given. This problem has stimulated considerable research into transparency and explainability of recommendations [30]. Explaining a recommendation aims at providing supportive evidence for the claimed suitability of the recommended item. The relation between a recommendation and its explanation can therefore be considered a specific form of argumentation, although very little research has thus far investigated explainability from this perspective [4, 22]. Since current RS mostly rely on quantitative approaches, explanations can usually not be derived from explicit system inferences but are mainly based on statistical concepts, depending on the recommendation approach taken. In the popular approach of Collaborative Filtering, for example, recommendations as well as explanations are based on item ratings given by users with similar preferences, following a form of *argumentum ad populum* scheme [7]. Content-based RS derive their recommendations from the similarity between a user's preferences and the (objective) properties of an object, enabling feature-based explanations (for a comparison of methods, see [6]), while hybrid systems apply a mixture of methods.

✉ Jürgen Ziegler
juergen.ziegler@uni-due.de

[1] Universität Duisburg-Essen, Forsthausweg 2, Duisburg, Germany

In addition to these basic approaches, user-generated content has increasingly been used for generating as well as for explaining recommendations, exploiting, for example, user-provided tags [19] or textual user reviews [32]. While textual feedback from other users has been shown to support and influence decision making [1], extracting item-related aspects and sentiments from reviews is still a challenging task. It is essential, however, for producing review-based explanations. A further challenge relates to the (plausible) assumption that the relevance of a particular review for the user's decision-making is both dependent on the user's own preferences and on the convincingness of the argumentation in the review. Determining the quality and convincingness of arguments in reviews, however, is a largely open research problem.

The ASSURE project, carried out in cooperation between the Interactive Systems Group (Prof. Jürgen Ziegler) and the Language Technology Lab (Prof. Torsten Zesch of the University of Duisburg-Essen, aims at leveraging review content for improving the accuracy of personalized recommendations as well as the quality of explanations, in particular by providing argumentative explanations. In this paper, we address the problem of explaining recommendations based on aspects extracted from reviews and present a novel neural architecture for modeling both user preferences and item-related aspects.

2 Goals & Challenges

Although explanations in RS can be discussed with respect to argumentation theory, this link has rarely been established in research. The most common forms of explanations, i.e. collaborative and feature-based, rely on statistical correlations found in the data and, thus, depict an abstract form of argumentation. While explanations based on textual feedback by other users more closely resemble how humans communicate with each other and usually provide deeper insights into an item's properties, principles of argumentation theory are generally not considered during their generation process. One reason for this is that manual as well as automated extraction of argumentative language patterns is still considered a challenging task [18]. But even if arguments were to be detected reliably, their application as explainable components in a RS framework is not trivially given.

One particular obstacle is the missing link between user preferences and argument relevance. Naturally, user opinions are multi-faceted and personal attitudes towards the different aspects of a product domain strongly contribute to their evaluation [32]. For instance, when choosing a movie to watch, the decision is presumably influenced by its genre, story, visuals, or by appearing actors etc. In addition, opinions about such aspects may be conflicting. Effective persuasion is, therefore, dependent on the consideration of the target audience's perspective with respect to specific aspect categories. The identification of aspects can, in this context, be described as a form of topic modeling in which a target entity, i.e. an item of the product domain, is linked with certain attributes towards which opinions can be expressed [21]. Being able to identify arguments per se, is, under this light, only a partial solution to the provision of relevant premises. Rather, the RS has to consider which pieces of available information are likely to be deemed as important by the target user and how this information can be represented in the larger context of a decision task. An analogy can easily be drawn from real-life: When friends give recommendations to each other, they usually accompany their claim by carefully selected reasons that are targeted at their vis-à-vis. Equivalently, in an automated setting, the alignment between argument and audience is crucial as well.

As a result, we define two prerequisites for the acquisition of personally relevant arguments: First, the identification of domain aspects representing the dimensions based on which arguments can be selected. Second, the derivation of a notion of how the target user evaluates these aspects. For instance, it would not suffice to identify *story* as a salient aspect of the movie domain. Instead, it is also important to assess which kinds of stories, i.e. the concrete aspect realization, the user prefers. In order to achieve such a level of distinction, a method to detect preferential relations between users and items has to be established. For the work at hand, we assume such personalized inferences can be derived from utterances that contain indicators of polarity, i.e. positive and negative sentiment.

While we focus on developing an architecture for modeling user preferences based on review data in our work presented here, in further research in the ASSURE project, we plan to address several critical challenges entailed by the integration of argumentation principles: Although sentiment analysis has provided successful techniques in practice, they only tell what opinions have been expressed, but not why these opinions are held in the first place. Consequently, there is no guarantee that the identified statements will be argumentative. It is not uncommon for users to only state that they liked or disliked a movie without giving any reasons why. In other cases, reviews might as well be descriptive only. For example, people may describe the *story* in great detail without adding any evaluative content. To make the problem even more complex, descriptive and evaluative components might not be adjacent in text, but be separated, for instance, by punctuation marks. Coreference resolution [27], argumentative zoning [29], or reasoning about entailment [31] are only some of the techniques that can play an important role to solve this problem. Otherwise,

extracted passages may be incomprehensible due to a lack of context [5].

Moreover, argumentation mining is not only concerned with the identification of individual claims and premises being made, but also with the derivation of relationships between them and how they work to together to support or undermine the overall message. The extraction of argument graphs is a powerful tool to provide users with rich explanations that shed light on an item's properties from various perspectives. Notably, the construction of argument graphs is not limited to a single review. Theoretically, one can assume a relation between the arguments being made in several reviews. Please refer to pertinent overview works to find several other current challenges of argumentation mining in general [e.g. 17].

While current argumentation mining techniques are still limited in solving the problems addressed above, some obstacles may be overcome in practice. For example, due to the large amount of available user reviews, it is not necessary to identify every single argument that could theoretically be found. It would rather be sufficient to identify a number of high quality arguments while dropping those argument candidates where the classifier is uncertain. The latter cases are often characterized by implicitness of premises and may, therefore, be hard to understand by users. Therefore, the extraction of unambiguous arguments, as indicated by, for example, discourse indicators such as *because*, might even be preferable.

3 Aspect-based Transparent Memories

As we have described in Sect. 2, the personalized extraction of polar structures is central to our purpose. Doing so requires the establishment of a user model that represents individual attitudes towards relevant aspects of the target domain. In this section, we introduce a novel method, which we call *Aspect-based Transparent Memories* (ATM), that models such multi-faceted user preferences and compares them to an item's properties in order to accurately predict numeric ratings while, at the same time, identifying candidate sentences to explain this prediction. Neural memory-based methods [cf. 10, 28] allow the externalization and structurization of possibly large amounts of knowledge. In our case, the memories are unique to a single person or item and control the process of encoding and decoding review data. Both steps, encoding (or *writing*) and decoding (or *reading*) are accompanied by mechanisms that are designed to impose transparency on the model.

The ATM architecture (Fig. 1) consists of two neural memories that can be viewed as arrays of slots for storing and thus memorizing information [10, 28]. The first memory component encodes representations of sentences

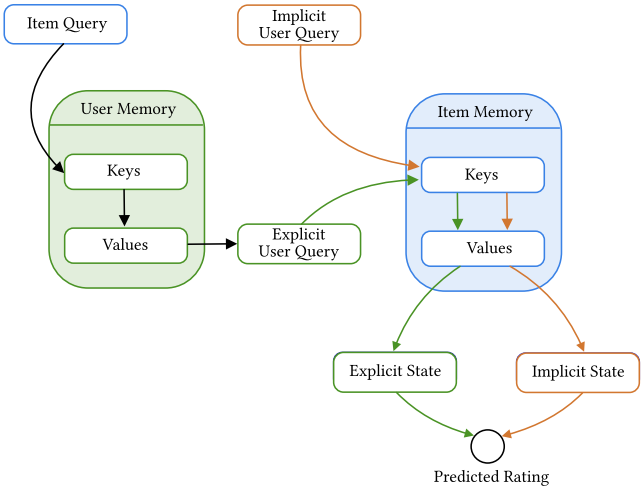

Fig. 1 Simplified schematic illustration of the proposed rating prediction pipeline including read operations for the neural memory components

composed by the target user. The second one is an equivalent variant for the target item and encompasses statements about the item by other users. Both memories are comprised of two subcomponents. First, aspect-based key vectors are used to perform the addressing operation, i.e. the selection of relevant memory locations. Keys are calculated by reconstructing sentences as a weighted combination of aspect embeddings such that the memory can be read in terms of topical overlap with the query vector. This model component is adapted from [11] and learns how to extract aspects in an unsupervised fashion while, at the same time, identifying the most salient of these learned aspects in each sentence. Conceptually, both aspect extraction as well as memory addressing can be described as a form of neural attention [2]. Value vectors, i.e. the content encoded into memory, depict the second component and contain the encoded sentence semantics. In our case, we encoded the sentences each with a bidirectional LSTM [12].

In order to predict the target rating, read operations extract elements from both user and item memory in a mixed-initiative fashion. The user memory is first queried by an item embedding, that is learned during the training process, to calculate the match between user preferences and item properties with respect to the appearing aspects. We call the result the *explicit* user state since it is derived from sentences put into writing as an active process by the user. This user state serves as a query to the aforementioned item memory. In other words, the explicit user interests are aligned with the opinions expressed by other users.

However, certain patterns found in rating behavior cannot be explained in terms of review content alone. For instance, a user may especially like *fantasy* movies, although they never mention this explicitly in any of their reviews. We assume that addressing the item memory only with the

explicit user state is insufficient. Therefore, we additionally train an *implicit* user representation that captures latent patterns similar to conventional collaborative filtering. This implicit state then serves as an additional query to the item memory. Both resulting vectors, explicit and implicit, can subsequently be combined to predict the target rating. A description of the architectural details and formalism used as well as extensions to the ATM architecture can be found in [5].

4 Explanations

Rendering recommendation models explainable has been recognized as a means to help users verify the underlying rationale by increasing transparency and accountability [e.g. 16]. Although retro-fit interpretable models have been proposed in the past [e.g. 24], we follow the line of argumentation that only model-intrinsic explanations allow faithful insights into the actual qualitative relationship between input features and recommendations [26]. Post-hoc explanations, on the other hand, cannot provide a sufficient level of certainty about their truthfulness as they usually only provide approximations of internal model states.

In order to generate human-intelligible explanations, we propose to exploit the states of ATM's diverse attention components. Accessing attention values allows us to formulate two different types of explanations: The first one deals with the problem of representing the information space from the system's perspective. By indicating which aspects the system attends to and by clarifying how these aspects relate to concrete utterances, the target user becomes empowered to assess how the system evaluates and structures the input information. The second kind of explanation is rather concerned with extracting the reasons behind one concrete recommendation. Again, the attention components can be exploited to pick statements from other users that support this claim. In the following, we describe details concerning how to arrive at these types of explanations:

Aspect Extraction. In order to provide an overview of the information space, ATM can convey details about the average distribution of aspects in the whole data set (Fig. 2) or for a specific item (Fig. 3). The first step for this is to derive which aspects are deemed important by the system in the first place. The set of attended aspects can either be fixed a-priori or learned in unison with the remaining network parameters. Fixing them can be achieved by setting the aspect representation to the word embedding of the respective aspect term or by averaging the embedding vectors of several terms that together form a higher-order aspect. For instance, in the movie domain one such combination may consists of the embeddings of *story*, *storytelling*, *script* etc. Opposed to this, automatically identifying aspects can be achieved via extending the model cost function by an unsupervised loss that measures how well sentences can be reconstructed solely based on a combination of learned aspect embeddings [11]. For a given sentence, we can then derive the relative importance of each aspect. Consequently, averaging this importance rating over all sentences yields the overall distribution of occuring aspects in the data.

Personal Aspect Importance. Once salient aspects have been detected, ATM can utilize this information further to assess which aspects are assumed to be especially important to the target user (Fig. 2). As before, this information can be extracted by averaging the aspect weights; only this time the target sentences shall originate only from of the current user. Merely showing the distribution of personal aspect importance, however, doesn't yield sufficient transparency. Instead, we can additionally display exemplary sentences that strongly contributed to a particular aspect weight. Through this step, the user can better verify whether they agree with the assessment of their assumed aspect preferences.

Recommendation Explanation. The central explanatory component of ATM is a mechanism to communicate the reasoning behind one particular recommendation. As displayed in Fig. 1, ATM matches the user's explicit and implicit representations against statements formulated by other users. The resulting attention weights then indicate which sentences contain the largest overlap with the vector-

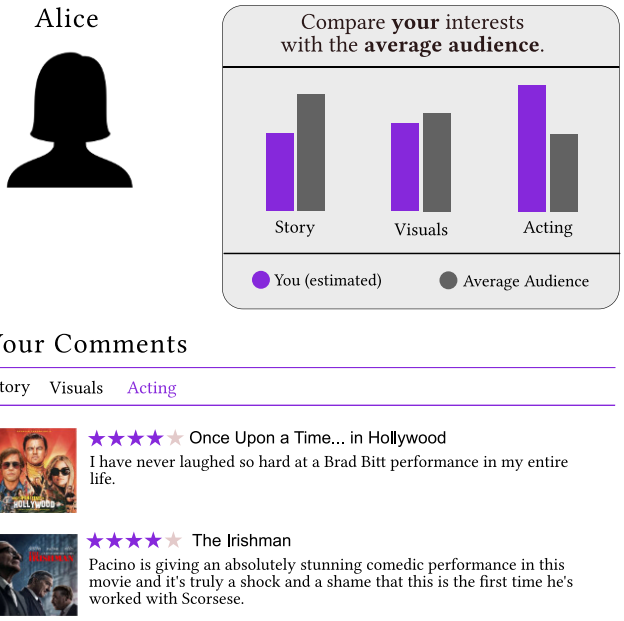

Fig. 2 Exemplary user profile that depicting (assumed) personal and average importance for three aspects as well as the target user's comments sorted by aspects

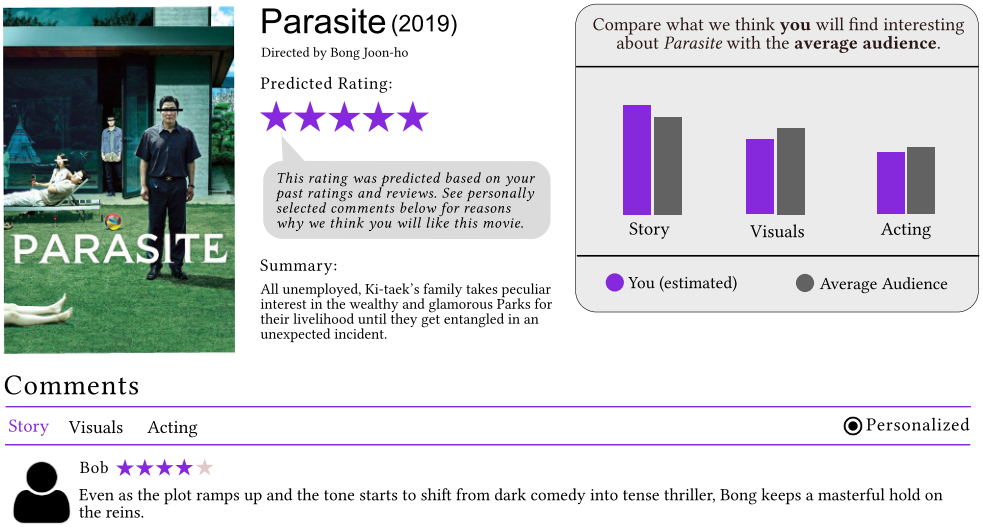

Fig. 3 Recommendation for the movie Parasite including the predicted rating, an overview of (assumed) aspect importance, and a personally selected comment that supports the predicted rating. Selection of comments can be personalized by toggling the respective radio button

ized user preferences. In other words, sentences with large attention weights are the best candidates to describe what properties of the target item the current user will probably like or dislike most (Fig. 3). Informally, the explanation process can be exemplified as follows: Let us assume a user has exhaustively dealt with storytelling in their past reviews. Concretely, they seem to like complex stories with a twist-ending a lot. ATM would then, based on concrete examples of these statement, derive a user representation that semantically reflects this preference. Now, if ATM were to generate recommendations for this user, it would find large overlaps between their preference representation and the embedding of sentences that also deal with twist-endings. As a result, not only will fitting movies receive larger overall scores, but individual sentences for this movie that contain concrete information about their ending would also be detected as salient. Please note that the same also applies for the implicit user representation. For a more detailed discussion of explaining recommendations with this process, please refer to our work presented in [5]. It also presents a user study aimed at evaluating the quality of the explanations generated.

Summarized, ATM can be seen as the first step towards a full-fledged argumentation-based explainable RS. It its current state, ATM is mostly concerned with detecting personally important information in review texts composed by other users. However, the extracted content is not yet presented in argumentative manner as no structural knowledge about arguments is represented in the model. This leads to several limitations that were already discussed in Sect. 2. Please note, however, that although such natural language explanations may eventually entail a causal structure, the underlying attention mechanism still only operates as a correlational statistical process. The resulting explanations, therefore, only express merely apparent causality. This is a phenomenon that has to be further investigated in future works via, for instance, the application of causal reasoning techniques [e.g. 8].

5 Evaluation

We have conducted experiments on three real-world datasets to demonstrate the effectiveness of ATM by comparing it to state-of-the-art RS. In the following, we present the datasets used, describe our experimental procedure, and introduce the baselines selected for comparison. Then, we evaluate and discuss performance.

Datasets. The *Yelp*[1] dataset is a large-scale dataset introduced in the context of the Yelp challenge. Since aspects only relate to specific domains, we filtered out all reviews for businesses not associated with the category *Restaurants*. *Kindle* is one category of the Amazon review dataset[2] containing reviews of e-books purchased from the Kindle store. *Movies* is another of the Amazon categories with movie and TV reviews. All datasets contain user reviews associated with a 5-star rating.

Procedure and Settings. In our experiments, we adopted the well-known Mean Squared Error (MSE) metric to evaluate recommender performance.

Reviews were first passed through a Stanford Core NLP Tokenizer [20] to obtain tokens which were then lowercased. Sentences were separated subject to the tokenizer result. Contractions were expanded and stopwords and punctuation were combined into a single token. We set a maxi-

[1] https://www.yelp.com/dataset/challenge.
[2] http://jmcauley.ucsd.edu/data/amazon/.

mum number of 30 words per sentence with a total of 150 sentences per user and item. Shorter sentences were padded accordingly. We used pretrained *fastText* embeddings [13] with dimensionality 300 for word embeddings.

We trained a variant of our proposed model with 10 aspects initialized with random values drawn from a Glorot uniform distribution [9]. The same distribution was used for randomly initializing the remaining parameters. Concerning the model-specific hyperparameters, we set $\lambda_D = 0.6$, $\lambda_s = 1.0$ and $\lambda_u = 1.0$. All of these values were selected via grid-search-like optimization.

Optimization was performed using Adam [14] and a learning rate of 0.001. We randomly split the data into training (80%), validation (10%), and test set (10%). The maximum number of epochs was set to 10. After training for one epoch with a batch size of 32, we calculated MSE on validation and test set. We report the results of the test set where the results of the validation set was lowest. All algorithms were implemented with Python using PyTorch [23].

Baselines. We compared our method against established recommendation models:

- *Matrix Factorization* (MF) [15] is one of the most popular collaborative filtering techniques.
- *Neural Attentional Rating Regression* (NARRE) [3] is a convolutional model that consists of two parallel attentive neural networks coupled by a final recommendation layer. The first network processes reviews of a target user in an attentive manner to derive a latent state. The second does the corresponding operation for the item side. Since we were mainly interested in comparing review-retrieval approaches with our aspect-based variant, we view NARRE, known to produce state-of-the-art recommendations, as a representative instance of the whole line of research.

The parameter values for NARRE were assigned subject to the evaluation in the original paper. The textual data made available to NARRE were chosen to match the total amount available to ATM.

Offline Performance. The results for the rating prediction task for ATM and the baselines are given in Table 1. As was to be expected, the two review-based approaches perform better than the conventional collaborative filtering model.

Furthermore, our proposed model outperforms NARRE on all considered datasets. One explanation for this is that its mixed-initiative approach allows ATM to more selectively distribute attention. While the integration between target user and item only happens in the later layers in NARRE, user and item side are interwoven in ATM right from the beginning. Additionally, breaking reviews down into sentences allows the model to distribute its attention on smaller semantic units. Finally, the integration of the review reconstruction pipeline strengthens the overall training signal.

Table 1 MSE for all baselines as well as the proposed ATM

	Yelp	Kindle	Movies
Baselines			
MF	1.379	0.739	1.488
NARRE	1.102	0.519	0.922
ATM	1.085	0.454	0.847

6 Conclusion

In this paper, we present an overview of the ASSURE project and the role of argumentation in recommender systems. We furthermore describe in more detail one of the solutions developed in the project: ATM is an approach to memorize user opinions on relevant item aspects found in raw review texts to derive multi-faceted user and item representations. We have shown that representing knowledge about multiple aspects in combination with external memories leads to more accurate recommendations. Offline experiments indicate that ATM outperforms other review-based models by at least a slight margin. The model can also serve as a basis for generating more informative explanations. These include the arrangement of review content with respect to aspect categories as well as the provision of personally selected user comments as decision support.

However, there is still room for improvement. Since ATM currently disregards any language structure beyond discriminating sentences, this may lead to explanations being detached from their original context which, in turn, impedes intelligibility. Consequently, the incorporation of deeper linguistic preprocessing appears necessary to improve the explanation performance. We are currently extending the approach by including representations of discourse markers and more complex argumentation mining techniques to reliably detect argumentative structures. Finally, we are also investigating means of formulating multi-perspective explanations based on supporting and attacking relations as derived from argument graphs generated from review data.

Funding Open Access funding provided by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are

included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Askalidis G, Malthouse EC (2016) The value of online customer reviews. In: Proceedings of the 10th ACM Conference on Recommender Systems, ACM
2. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate (arXiv preprint arXiv:14090473)
3. Chen C, Zhang M, Liu Y, Ma S (2018) Neural attentional rating regression with review-level explanations. In: Proceedings of the 2018 World Wide Web Conference on World Wide Web, pp 1583–1592 (International World Wide Web Conferences Steering Committee)
4. Chesnevar CI, Maguitman AG, González MP (2009) Empowering recommendation technologies through argumentation. In: Rahwan I, Simari GR (eds) Argumentation in artificial intelligence. Springer, Heidelberg, Berlin, New York, pp 403–422
5. Donkers T, Kleemann T, Ziegler J (2020) Explaining recommendations by means of aspect-based transparent memories. In: Proceedings of the 25th International Conference on Intelligent User Interfaces, pp 166–176
6. Gedikli F, Jannach D, Ge M (2014) How should I explain? A comparison of different explanation types for recommender systems. Int J Hum Comput Stud 72(4):367–382
7. Gena C, Grillo P, Lieto A, Mattutino C, Vernero F (2019) When personalization is not an option: an in-the-wild study on persuasive news recommendation. Information 10(10):300
8. Ghazimatin A, Balalau O, Roy RS, Weikum G (2019) PRINCE: Provider-side interpretability with counterfactual explanations in recommender systems (arXiv preprint arXiv:191108378)
9. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp 249–256
10. Graves A, Wayne G, Danihelka I (2014) Neural turing machines (arXiv preprint arXiv:14105401)
11. He R, Lee WS, Ng HT, Dahlmeier D (2017) An unsupervised neural attention model for aspect extraction. In: Long Papers. Proceedings of the 55th annual meeting of the association for computational linguistics, vol 1. In, Vancouver, Canada, pp 388–397
12. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
13. Joulin A, Grave E, Bojanowski P, Mikolov T (2016) Bag of tricks for efficient text classification (arXiv preprint arXiv:160701759)
14. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization (arXiv preprint arXiv:14126980)
15. Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. Computer 8:30–37
16. Kunkel J, Donkers T, Michael L, Barbu CM, Ziegler J (2019) Let me explain: impact of personal and impersonal explanations on trust in recommender systems. CHI Conference on Human Factors in Computing Systems Proceedings, CHI 2019, ACM, New York, NY, USA
17. Lawrence J, Reed C (2020) Argument mining: a survey. Comput Linguist 45(4):765–818
18. Lippi M, Torroni P (2016) Argumentation mining: state of the art and emerging trends. ACM Trans Internet Technol 16(2):10
19. Loepp B, Donkers T, Kleemann T, Ziegler J (2019) Interactive recommending with Tag-enhanced Matrix Factorization (TagMF). Int J Hum Comput Stud 121:21–41
20. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp 55–60
21. McAuley J, Leskovec J, Jurafsky D (2012) Learning attitudes and attributes from multi-aspect reviews. In: 2012 IEEE 12th International Conference on Data Mining, IEEE, pp 1020–1025
22. Naveed S, Donkers T, Ziegler J (2018) Argumentation-based explanations in recommender systems: conceptual framework and empirical results. In: Adjunct publication of the 26th Conference on User Modeling, Adaptation and Personalization, ACM UMAP '18, Singapore, pp 293–298
23. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A (2017) Automatic differentiation in pytorch
24. Ribeiro MT, Singh S, Guestrin C (2016) Why should i trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 1135–1144
25. Ricci F, Rokach L, Shapira B (2015) Recommender systems: introduction and challenges. In: Recommender systems handbook. Springer, Heidelberg, Berlin, New York, pp 1–34
26. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206
27. Soon WM, Ng HT, Lim DCY (2001) A machine learning approach to coreference resolution of noun phrases. Comput Linguist 27(4):521–544
28. Sukhbaatar S, Weston J, Fergus R (2015) End-to-end memory networks. In: Advances in neural information processing systems, pp 2440–2448
29. Teufel S, Siddharthan A, Batchelor C (2009) Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, vol 3. Association for Computational Linguistics, Singapore, pp 1493–1502
30. Tintarev N, Masthoff J (2011) Designing and evaluating explanations for recommender systems. In: Recommender systems handbook. Springer, Heidelberg, Berlin, New York, pp 479–510
31. Zanzotto FM, Pennacchiotti M, Moschitti A (2009) A machine learning approach to textual entailment recognition. Nat Lang Eng 15(4):551–582
32. Zhang Y, Lai G, Zhang M, Zhang Y, Liu Y, Ma S (2014) Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In: Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval, ACM, pp 83–92