Check for updates

An explainable content-based approach for recommender systems: a case study in journal recommendation for paper submission

Luis M. de Campos[1] · Juan M. Fernández-Luna[1] · Juan F. Huete[1]

© The Author(s) 2024

Abstract

Explainable artificial intelligence is becoming increasingly important in new artificial intelligence developments since it enables users to understand and consequently trust system output. In the field of recommender systems, explanation is necessary not only for such understanding and trust but also because if users understand why the system is making certain suggestions, they are more likely to consume the recommended product. This paper proposes a novel approach for explaining content-based recommender systems by specifically focusing on publication venue recommendation. In this problem, the authors of a new research paper receive recommendations about possible journals (or other publication venues) to which they could submit their article based on content similarity, while the recommender system simultaneously explains its decisions. The proposed explanation ecosystem is based on various elements that support the explanation (topics, related articles, relevant terms, etc.) and is fully integrated with the underlying recommendation model. The proposed method is evaluated through a user study in the biomedical field, where transparency, satisfaction, trust, and scrutability are assessed. The obtained results suggest that the proposed approach is effective and useful for explaining the output of the recommender system to users.

Luis M. de Campos, Juan M. Fernández-Luna and Juan F. Huete have contributed equally to this work.

This paper or a similar version is not currently under review by a journal or conference. This paper is void of plagiarism or self-plagiarism as defined by the Committee on Publication Ethics and Springer Guidelines.

✉ Juan M. Fernández-Luna
  jmfluna@decsai.ugr.es

  Luis M. de Campos
  lci@decsai.ugr.es

  Juan F. Huete
  jhg@decsai.ugr.es

[1] Departamento de Ciencias de la Computación e Inteligencia Artificial, Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación, CITIC-UGR, Universidad de Granada, C/ Periodista Daniel Saucedo Aranda, S/N, 18013 Granada, Spain

🖄 Springer

1 Introduction

Recommender systems (RS), a form of artificial intelligence (AI), offer personalized suggestions based on users' preferences and behaviors, widely utilized in online retail, streaming services, and social media platforms. Recognizing the significance of explainable AI both in a general context (Minh et al. 2022; Li et al. 2023; Barredo-Arrieta et al. 2020 and specifically within RS (Pavitha et al. 2022; Zhang and Chen 2020), the provision of explanations for recommended content serves multiple purposes. It not only enhances user trust but also improves transparency, allowing users to comprehend the reasoning behind recommendations, mitigates biases arising from data or algorithms, reinforces users' reliance on the system's judgments, and fosters increased engagement (Tintarev and Masthoff 2007; Zhang and Chen 2020). In this paper, we present an explanation approach tailored to enhance user understanding of RS recommendations in a venue recommendation problem, whereby a researcher with a recently written paper needs to decide to which journal the paper should be submitted.

When the RS is provided with the title and abstract of a new article, it suggests a group of appropriate venues for publication. The system uses a content-based RS (CBRS) where each journal included in the RS is represented by various textual subprofiles which group together every article published in the journal on the same topic. The input text is, therefore, matched against the journal subprofiles, and the user is presented with a sorted list of recommended journals in decreasing order according to their associated scores, which represent a type of aggregation of the subprofile scores.

Nevertheless, once the RS has recommended the most suitable journals according to the submitted title and abstract for the user to consider, users might not understand why certain suggestions were made and this could well result in loss of confidence in the RS. The proposed explanation approach is based on different explanation elements that generate information from different components from the CBRS in order to justify the recommendations (confidence in the recommendation, global topic word clouds, similar articles to the target one published in the same journal, specific topic word clouds covered by the journal and certain highlighted words from the target title and abstract). The proposed method is based only on content, i.e., the text of the articles published in their corresponding journals. Although other bibliometric information (such as impact factor, quartiles, co-authorship, etc.) might be used, the scope of this paper only focuses on text so as to measure the feasibility of this approach without external information.

This paper not only presents the design of this integral explanation scheme but also the results of a user study in order to determine how selected biomedical experts (the context of this evaluation) viewed the explanation tools presented and how useful they considered them to be.

The paper makes the following contributions:

- A novel approach for explaining CBRS which is totally integrated in the recommendation model particularized in the journal recommendation problem
- Verification of the transparency, satisfaction, trust, and scrutability of the proposal by means of a user study
- A comparison of the proposed method with those found in the state of the art

The remainder of the paper is organized as follows: Sect. 2 examines related work on explanations in recommender systems; Sect. 3 provides a brief overview of the specific system for recommending scientific journals for which we have developed explanation facilities; in Sect. 4, we describe different explanation elements developed; then, Sect. 5 presents the user study designed to evaluate the performance of the explainable recommender system from different perspectives; and finally, Sect. 6 outlines our conclusions and future lines of research.

2 State of the art

One of the major contributions dealing with explanation in RSs is the work by Tintarev and Masthoff (2007). They justify the implementation of good explanations since they can "help inspire user trust and loyalty, increase satisfaction, make it quicker and easier for users to find what they want, and persuade them to try or purchase a recommended item." Explainable RS (XRS) address the problem of why such items are recommended (Zhang and Chen 2020). Tintarev and Masthoff established seven benefits of the explanation: transparency (explain how the RS works), scrutability (users are able to express whether the RS is correct or not), trust (increase user confidence in the system), effectiveness (help the user make good choices), persuasiveness (make the user perform an action), efficiency (make decisions faster), and satisfaction (fulfillment of the user's needs). Two additional purposes might be education (learn something from the system) and debugging (identify problems in the RS) (Jannach et al. 2019). These elements are valid when designing RS explanation features and reflect the dimensions on which these should be based.

2.1 Classification of explainable recommender systems

The explanations provided by an RS could be classified into two models (Zhang and Chen 2020) depending on the interpretability of the explanations: while the model-intrinsic approach offers direct transparency for the RS decision, the model-agnostic one needs to create explanations after the decisions are taken. In this last case, the RS is a black box and explanation must be built on the top of it with all available resources. An example of this type is what Shmaryahu et al. (2020) call post hoc explanation. This is applied when the recommendation engine is based on a complex model with a low explainability level. A transparent model with a high explainability degree is then used to explain the recommendation. A second example can be found in another article (Iferroudjene et al. 2023), whereby the authors create an explanation context based on subgroup discovery on top of a top-n RS to identify active data for the recommendation. Moreover, Papadimitriou et al. (2012) established an alternative

categorization of the XRSs according to the resources used for explanation: human style, based on recommendations of similar users; item style, based on suggestions made to the user on similar items; and feature style, supported by features of an item which were previously considered interesting by a user. In addition, another classification of the explanation is presented in Radensky et al. (2022), where XRSs might explain a specific item recommendation (local), or in a model-based approach, global, which explains how recommendations are generally generated. Tintarev and Masthoff (2012) categorized the explanations as personalized or non-personalized according to whether the explanations are specific for each user or suit everyone. Finally, a recent categorization groups the explanation-based models into those where the recommendation model and the explanation function are separated, and the so-called recommendation-by-explanation models, where both actions are integrated (Rana et al. 2022). In these latter models, the user does not receive decreasingly sorted recommendations in terms of the confidence of the recommendation model, as it normally does, but rather explanations which are decreasingly sorted according to their quality.

2.2 Explanation styles

Regarding the explanation styles, Tintarev and Masthoff in Tintarev and Masthoff (2015) determine that there are different explanation styles according to the recommending model used to generate the recommendations: case, content, collaborative, demographic, knowledge, and utility based. Correspondingly, Nunes and Jannach (2017) identifies four types of explanation content: user preferences or user-provided or user-gathered input; inference process, extracted from the recommendation algorithm itself; background and complementary information such as additional information outside the user's context and based on their features (list of features and their advantages and disadvantages, and the most relevant explanation features).

Focusing on previously published explanation types, according to the taxonomy of explanations proposed by Nunes and Jannach (2017), explanations could be classified according to their presentation format into natural language-based explanations (e.g., predefined templates that are instantiated before explanation or structured language) and multimedia-based explanations (visualization in the form of graphs or other media formats).

2.3 Content-based explanations

Since the explanation approach presented in this paper is based on a content-based RS (CBRS), in this section we proceed to present a number of examples. In Verbert et al. (2013), an explanation approach is presented based on content and tag recommendation in the context of suggesting papers and talks from conferences. The first is supported by a user's profile from the text from the papers that the user has read, and the recommendation is performed by similarity computation. The second, however, uses tags that are assigned to conference talks. Once again, a matching is carried out between the tags of interest to the user and those from the talks. Similar users are also recommended according to profile closeness, and all the recommended

information is shown as a clustermap, combining recommended users, tags, talks, and papers. Cardoso et al. (2019) present IntersectionExplorer, a recommender system with explanation features in the context of conference paper recommendation. In this, they combine personal, social, and content relevance to perform the recommendation and enable multiple item sets from these relevance perspectives to be explored by means of a user interface. The recommender models applied are tag, bookmark, and bibliography based, with all of them using textual or tag content. Millecamp et al. (2019) designed an explanation for a CBRS, where profiles and items were represented by features. After computing the corresponding similarity, the explanation was given by a visual element that summarized the values of the user's features in terms of an interval, the exact values of the features for the recommended items, and a 5-point Likert scale to show the similarity of the recommended item with the user's profile. In Hernandez-Bocanegra and Ziegler (2020), an explanation is built on a model based on the user's profile and its latent features are matched against the latent features of the items and complemented with features representing sentiment information to be extracted from item reviews. The explanation elements were bar charts and tables to explain in terms of the features why the recommended items were selected.

In Sullivan et al. (2019), within the context of online news recommendation based on the user profiles storing topics, entities, and tags extracted from the read news articles, Sullivan et al. show new topics which might be of interest to the users and associated news. They use three explainability levels: visualization of the distribution of monthly read topics (user's reading behavior); visualization of that same distribution by the average user (contextualization with the community's behavior); and new recommended topics explained with the degree of relationships between the user's topics and new ones (exploitation of the user's profile). In Polleti and Cozman (2019), Polleti and Cozman proposed an explanation approach based on topic modeling to explain CBRS (this is, in fact, a model-agnostic method that could be used with any recommendation model): latent topics are extracted from the textual representation of the items and users' profiles. After obtaining the list of nearest items for a given user's profile, the common topics between a recommended item and the corresponding profile are then shown in order to justify the recommendation. In Pérez-Núñez et al. (2022), Pérez-Núñez et al. build a model to recommend and explain textual restaurant reviews written by other TripAdvisor users. Beginning with keywords from the textual reviews which are encoded using BoW, and given a user's information need, they use a classifier (multinomial logistic regression) to obtain recommendations. The explanation is a tag cloud generated by means of the common terms found in the user's query and the recommended items.

The paper Tsai and Brusilovsky (2019), in which the authors present an explainable hybrid recommended system in the context of conference recommendation, integrates several recommendation models based on content (keywords and topic similarities, respectively), social (co-authorship), and demographic information. They created five explanation elements, each generated from a given recommendation model (Venn diagrams containing tag clouds, topic similarities, co-authorship graphs, interest similarity, and geographic distance). Along the same lines, Louki et al. (2020) shows various explanation mechanisms for a hybrid RS. In terms of the content-based component, the user's profiles store tags and keywords, and tag and content-based similarities

between the profiles and items are incorporated into the hybrid recommendation. The general explanation comprises single explanations, each obtained from every model involved in the recommendation. For CBRS, therefore, they explain the recommendation of an item since it contains similar tags to those found in the profile and there are common tags between them.

In conclusion, in CBRS-based explanation the available information usually used to explain the recommendation is the user's profile and comprises keywords, tags, or features. Additionally, latent topics extracted from the profiles and item texts are also considered. In terms of the explanation type, most approaches show in one way or another how the item covers the users' profiles.

2.4 Explanation elements

Going one step further, and independently of the previous explanation styles, a number of already published common types of explanation elements could be used to explain the recommendation: histograms, tables, or pie chart of ratings which show the rating distribution of similar users or ratings of similar products (Daher et al. 2017; Zhang and Chen 2020; Felfernig et al. 2021; Bilgic and Mooney 2015; Jannach et al. 2019; Pérez-Núñez et al. 2022); personalized or non-personalized tag clouds of the keywords that describe the recommended items (Gedikli et al. 2011; Chen 2013; Daher et al. 2017; Felfernig et al. 2021; Pérez-Núñez et al. 2022); common features or aspects between users and recommended items (Zhang and Chen 2020; Zhang et al. 2022; Vig et al. 2009; Millecamp et al. 2019; Pérez-Núñez et al. 2022) in the form of Venn diagrams (Tsai and Brusilovsky 2019); confidence percentage of the RS prediction (Daher et al. 2017); textual descriptions to show the reason for the recommendations (Daher et al. 2017); features detected in item descriptions and highlighted (Li et al. 2021), radar charts, which include the degree of various features for the recommended items (Daher et al. 2017; Zhang and Chen 2020; Felfernig et al. 2021; Tsai and Brusilovsky 2019); (elaborated) users' opinions on the recommended items (Zhang and Chen 2020; Hernandez-Bocanegra et al. 2020) and their graphical representation (Hernandez-Bocanegra and Ziegler 2020); generated natural language (Lully et al. 2018); keyword explanation, identifying the terms common to the recommended item and the user's profile (Bilgic and Mooney 2015); list of similar items with the corresponding user's rating and their impact on the recommendation (Bilgic and Mooney 2015) or similar items according to the user's preferences and their neighbors' (Shmaryahu et al. 2020); graphs, in the form of co-authorship (Tsai and Brusilovsky 2019), or bipartite ones containing users and items (Afchar et al. 2022) or items connected to their underlying topics (Polleti and Cozman 2019); or simply the formula to compute the relevance of a recommended item (Tsai and Brusilovsky 2019). As shown, there is a wide variety of explanation elements which depend on the available information and the recommending model.

2.5 Evaluation

The impact of the explanation, meanwhile, can be evaluated directly by means of a user study based on recruited users and a given task (Zhang and Chen 2020). The results of such a study are obtained by analyzing questionnaires completed by the users once they have finished the study. Two further methods have also been published (Vultureanu-Albisi and Badica 2022), one of which is online (real interaction with an RS) and the other is offline (where no users are considered in the evaluation). Another alternative is to directly measure the impact in the users' performance in a real environment, i.e., how accurate the decisions are, or how fast they are made (Jesus et al. 2021).

When designing an evaluation, since the designer must choose between some of the aims discussed at the beginning of this section as some might be mutually incompatible, any evaluation needs not only to identify the aim being investigated but also to employ suitable metrics (Tintarev and Masthoff 2012), mostly a questionnaire with the appropriate questions: for example, a user study is employed to evaluate persuasiveness and usefulness in Sato et al. (2019) where two specific questions were asked the users in a questionnaire (7-points Likert scale): For persuasiveness, *The explanation is convincing* and *The explanation triggers*; for usefulness, *The explanation is useful for choice* and *The explanation is easy to understand*. Persuasiveness is also evaluated in a user study in Louki et al. (2020), using the question *This explanation for the recommendation is convincing*. Effectiveness is mainly evaluated in Yao et al. (2022) with the question *Does the explanation help you decide whether you want to watch this recommended movie?*. Satisfaction is considered in Ferwerda et al. (2018) with the question *I found the programs that I chose to watch good*. In the study by Shulner-Tal et al. (2022), the System Causability Scale (based on system usability criteria) was utilized. Participants were instructed to assign scores to ten statements, and these scores were subsequently averaged for the comparison of different explanation styles.

3 Overview of the journal recommender system

In this section, we briefly describe the recommender system for which we have designed an explanation module (further details can be found in de Campos et al. 2022). This is a content-based system which is specifically devoted to recommending publication venues, and more specifically scientific journals, which might be suitable when attempting to publish a given article based on its content. Although the model may be used in any knowledge domain, the current version has been trained with a collection of articles in the biomedical domain extracted from PubMed/Scopus (Albusac et al. 2018).

Our system is based on a representation of each journal through a set of homogeneous thematic subprofiles, covering different topics considered within the journal. To achieve this, we begin with a document collection containing articles published in all the journals under consideration. Subsequently, we employ a text clustering algorithm,

specifically K-means,[1] to group these articles into K clusters[2] of thematically similar content. The clustering algorithm takes into account only the title, abstract, and keywords of each article. Each cluster, or more precisely, the terms within its constituent articles, represents a global topic within this collection.

We then analyze how each journal addresses each topic. For a given journal, we group together all its articles which belong to the same cluster, creating the corresponding journal subprofile. If no articles from a particular journal belong to a given cluster, that journal's subprofile for that cluster remains empty. The text of all these topically similar articles is then concatenated into a single macro-document, which represents the subprofile of the journal associated with the corresponding cluster/topic.

We, therefore, have a set of at most K documents associated with each journal. All of these documents from all the journals form a collection of subprofiles, which is subsequently indexed (we use the Lucene library[3]) to be used by an information retrieval system (IRS). We employ a language model with Jelinek–Mercer smoothing for this purpose.

Given a target article for which we want a journal recommendation for publication, we use its text as the query for the IRS. A list of the top-h subprofiles[4] is then retrieved, together with the corresponding scoring values $scr(j, t)$ indicating the relevance for the candidate article of those papers published in the journal j under the topic t. In other words, we compute a similarity degree between the candidate article and the way each journal j approaches each topic t. It is worth noting that this ranking shows a many-to-many relationship between topics and journals. While the same topic t might appear in various relevant journals, a given journal j may also encompass different topics relevant to the query.

To generate a list of recommended journals, we transform this subprofile ranking into a journal ranking using a fusion algorithm (de Campos et al. 2017). This algorithm aggregates the scores $scr(j, t)$ from all subprofiles retrieved from the same journal j, applying logarithmic penalization to account for their ranking position. This ensures that lower ranked subprofiles have a less influence on the final recommendation. The rationale is that if, for example, the candidate article primarily pertains to a topic t_1 but also touches on a topic t_2 to a lesser extent, a journal that covers both of these topics appears more relevant than a journal covering only one of them. The top n journals[5] from this newly generated ranking are then recommended to the user, as depicted in Fig. 1.

As we have seen, the recommendation process is based on information about topics (high level) as well as terms (low level) and their similarity to the query. However, the current system does not utilize any of the internal information it manages (such as what topics are discovered by the clustering algorithm, what articles and terms form the subprofiles associated with each journal, and what terms from these subprofiles match with those of the target article) to explain its recommendations. Instead, it

---

[1] alternatively, Latent Dirichlet Allocation could also be used, as seen in de Campos et al. (2023).

[2] with the parameter K denoting the number of clusters considered.

[3] https://lucene.apache.org/.

[4] h is a parameter, currently fixed to 40.

[5] n is another parameter, presently fixed to 10.

Title	School-Age Outcomes of Early Intervention for Preterm Infants and Their Parents: A Randomized Trial.
Abstract	To examine the child and parental outcomes at school age of a randomized controlled trial of a home-based early preventative care program for infants born very preterm and their caregivers. At term-equivalent age, 120 infants born at a gestational age of <30 weeks were randomly allocated to intervention (n = 61) or standard care (n = 59) groups. The intervention included 9 home visits over the first year of life focusing on infant development, parental mental health, and the parent-infant relationship. At 8 years' corrected age, children's cognitive, behavioral, and motor functioning and parental mental health were assessed. Analysis was by intention to treat. One hundred children, including 13 sets of twins, attended follow-up (85% follow-up of survivors). Children in the intervention group were less likely to have mathematics difficulties (odds ratio, 0.42; 95% confidence interval [CI], 0.18 to 0.98; P = .045) than children in the standard care group, but there was no evidence of an effect on other developmental outcomes. Parents in the intervention group reported fewer symptoms of depression (mean difference, -2.7; 95% CI, -4.0 to -1.4; P < .001) and had reduced odds for mild to severe depression (odds ratio, 0.14; 95% CI, 0.03 to 0.68; P = .0152) than parents in the standard care group. An early preventive care program for very preterm infants and their parents had minimal long-term effects on child neurodevelopmental outcomes at the 8-year follow-up, whereas primary caregivers in the intervention group reported less depression.
Keywords	None
Subjects	None

Recommended Journals

#1	Pediatrics
#2	BMC pediatrics
#3	Early human development
#4	Journal of paediatrics and child health
#5	The Cochrane database of systematic reviews
#6	Developmental medicine and child neurology
#7	JAMA
#8	BMJ Clinical research ed
#9	The Journal of pediatrics
#10	Archives of disease in childhood

Fig. 1 Text of the target article and list of recommended journals

simply provides an ordered list of recommended journals. Therefore, we attempt to examine and exploit these internal processes that the system carries out to generate meaningful explanations for its recommendations.

4 Description of the explanation elements

As stated in the previous section, once the user has entered the information about a target article (title, abstract and keywords), the system retrieves a ranked list of journals and there is the possibility of accessing different explanation elements. We have proposed various types of explanation, not only to test which of these are preferred by the users but also because, according to a number of previous studies (Louki et al. 2020; Papadimitriou et al. 2012; Tsai and Brusilovsky 2019), the combination of different types of explanations can be positive.

It should be noted that although the explanation elements considered are adapted to the specific type of application considered, namely recommending scientific journals,
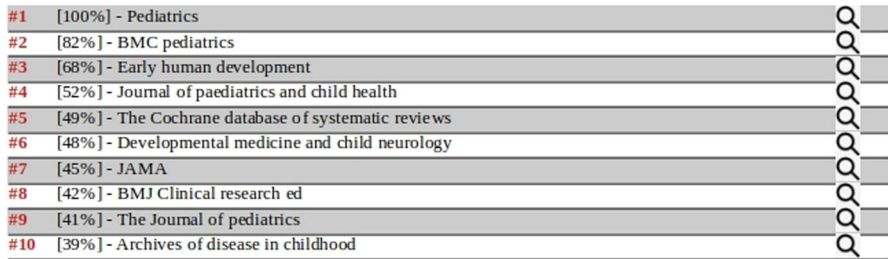
#1	[100%] - Pediatrics	🔍
#2	[82%] - BMC pediatrics	🔍
#3	[68%] - Early human development	🔍
#4	[52%] - Journal of paediatrics and child health	🔍
#5	[49%] - The Cochrane database of systematic reviews	🔍
#6	[48%] - Developmental medicine and child neurology	🔍
#7	[45%] - JAMA	🔍
#8	[42%] - BMJ Clinical research ed	🔍
#9	[41%] - The Journal of pediatrics	🔍
#10	[39%] - Archives of disease in childhood	🔍

Fig. 2 EE1: matching degrees between the recommended journals and the target article in Fig. 1. The given explanation text is: "TOP 10 RECOMMENDED JOURNALS AND THE RELATIVE CONFIDENCE OF THE SYSTEM IN EACH RECOMMENDATION (THE TOP JOURNAL ALWAYS RECEIVES 100% CONFIDENCE)"

as long as the base recommender system can be applied to other domains (such as, for example, expert finding), these elements can also be easily adapted.

Also, it is interesting to highlight that provided that the recommendation system is capable of offering a journal recommendation, an explanation for each one of them can also be generated as these explanations come from information that the RS uses to generate the recommendation. If the recommendation is bad, the explanation could be useful for the user to detect the inability of the RS for creating a quality recommendation.

In our case, we consider two levels of explanation: a global level that attempts to provide a general idea of why the entire set of journals was recommended, and a local level, where the explanations focus on each specific recommended journal.

4.1 Global explanation: why is this ranking obtained?

The two explanation elements (EE) considered in this level can be articulated in the following way.

4.1.1 EE1: ranking confidence

The original output of our recommender system was a list of journals, as depicted in Fig. 1. However, from the end user's perspective, it can be challenging to discern why a particular journal is ranked higher or lower than others and to what extent. Knowing such information can help users to make informed decisions regarding the most suitable journal for submitting the paper, ultimately enhancing the recommending experience.

Therefore, the first explanation element consists of simply showing a numerical score associated with every recommended journal, representing the matching degree between the target article and each journal, and reflecting system confidence in its recommendations (see Fig. 2, where in the caption of this figure we include the explanatory text given to the users). Opting for normalized scores (expressed as percentages) rather than raw scores enables consistent comparisons across various recommendations, irrespective of the retrieval algorithm or the specific target article. This choice enhances comparability and interpretation.

In particular, the used scores range from 0 to 100%, with higher values signifying the journal's greater suitability for publishing our article. They are obtained by dividing the raw score, a topic-based similarity degree, by the maximum score within the set of retrieved journals. Thus, from Fig. 2, it becomes evident that the topics covered in our article exhibit (approximately) twice the degree of similarity to the topics in the journal *Pediatrics* in comparison with the topics discussed in the journal *Developmental Medicine and Child Neurology*.

In previously published related work (Bilgic and Mooney 2015), the authors compute an influence degree of each item on the recommendation. This value is computed by removing the item from the training set and then recomputing the recommendation value for all the test items and finally considering the difference in scores with and without the item. Their experimental results show that the explanation based on this influence degree was very effective in comparison with the other two tested explanations. In Daher et al. (2017), they assign to each item the percentage of correct predictions in a system where the users can rate the prediction and feed the RS with that information. Although proved as useful, the problem of needing a sufficient number of predictions makes it not very operative. Our approach directly uses the (normalized) scores obtained from the CBRS to build a confidence percentage.

4.1.2 EE2: related topics

Topics play a crucial role in our system, as they enable more effective recommendations (de Campos et al. 2022). However, these topics were automatically derived from the entire corpus of published papers. As a consequence, there might be a gap between researchers' intuitions and the learned topics. In order to bridge this gap, another aspect of the explanation entails identifying the core themes discussed in the target article and scrutinizing how they are treated in the recommended journals. In order to rapidly and intuitively outline the essence of a topic, we selected the most significant terms (in a word cloud and/or a sorted list of terms), a widely employed technique in topic modeling (Aletras et al. 2017; Chi et al. 2019).

Continuing with our target article example (Fig. 1), the two word clouds in Fig. 3 illustrate how this paper can be approached from two distinct angles: one related to "`children health`" (Topic 56) and the other linked to "`controlled interventions studies`" (Topic 60). We also indicate in the caption the explanatory text presented to the user. Through the examination of these high-level topics, users can gain a more comprehensive understanding of the suitability of the recommended journals. Thus, in case these topics do not align with our expectations, we may have reservations about trusting the provided recommendations.

For the purpose of showing the related topics, two independent tasks must be undertaken. The first one is related to the approach used to determine the topic contents, and the second one involves establishing the number of relevant topics to display, ensuring it does not overwhelm the users.
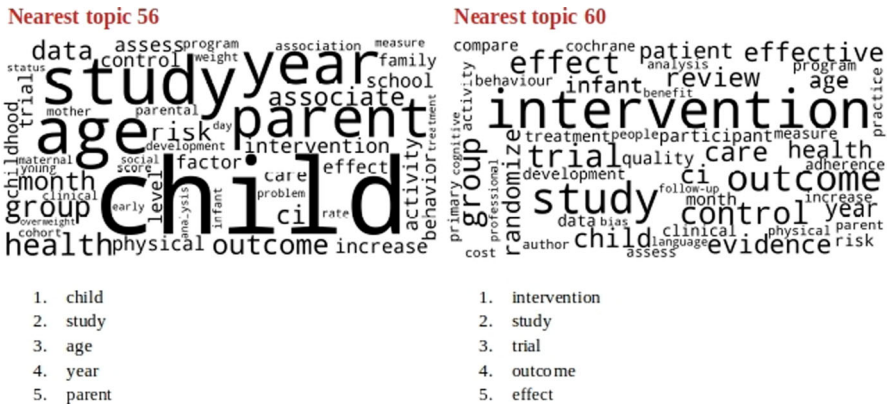
Nearest topic 56

1. child
2. study
3. age
4. year
5. parent

Nearest topic 60

1. intervention
2. study
3. trial
4. outcome
5. effect

Fig. 3 EE2: word clouds of related topics. The given explanation text is: "WITHIN THE RECOMMENDED JOURNALS, THESE two TOPICS ARE THE MORE RELEVANT TO YOUR RESEARCH. THE WORD CLOUDS SHOWCASE THE MOST COMMONLY USED WORDS UNDER EACH RESPECTIVE TOPIC." From these word clouds, researchers can see that papers focused on studies of `children health` as well as `controlled interventions` which have been published in the recommended journals are the most relevant to their research

Showing topics descriptions

In our case, topics has been learned from the entire collection of papers, grouping individual papers with a shared subject into the same cluster. In a broader sense, each topic is formed by the amalgamation of all the papers within the same cluster (i.e., all the journal subprofiles associated with this topic). For instance, Topic 56 and Topic 60 (in Fig. 3) are addressed in a total of 410 and 257 journals, respectively, with quite different scopes.

Thus, selecting the most significant terms from this amalgamation to describe a topic, i.e., utilizing the entire collection, might not contribute effectively to the understanding of our recommendations. This is because the resulting description can lead to a significant deviation from the context of our target research paper, making it challenging for the user to discern its relevance. In our running example, Topic 56 is predominantly characterized by terms related to `children asthma treatment`, as it is the prevalent health concern in a substantial number of papers published under this topic. Therefore, presenting such information would lead to confusion in understanding our explanations, as our paper is not related to asthma.

In order to overcome this problem, we specifically restrict the topic description to those journal subprofiles that are highly relevant to our paper. These subprofiles are those occupying the top positions in the ranking. This deliberate selection ensures that the obtained description is much more focused on the specificity of the target article, thereby enhancing its comprehensibility for the user.

We can say that the generated descriptions are personalized since they vary with the target research paper. Also, from a computational point of view, it is important to highlight the fact that we are referring to a relatively small number of journals discussing the topic, typically in the range of tens.

After choosing the journal subprofiles employed as the source text to build the word clouds, we provide a brief overview of the process used to select the most significant terms. Firstly, we utilize the spaCy library for entity recognition, identifying word groupings that exhibit strong relationships (in a general context, "White House" would be considered an entity) and treating them as a single unit. Any words not recognized as entities are considered separately. The resulting word stream is lemmatized, excluding the stop words, which are removed. Lemmatization is performed using the WordNetLemmatizer library in Python, taking into account their respective parts of speech in the sentence, determined using the NLTK POS tagger, also in Python. The resulting lemmas are then ready for counting their occurrences in the text stream to create the word clouds (with the top 50 lemmas being selected) or the list of terms (we show the top 15[6]).

Selecting the number of topics

Since it is rare for an article to be related to many high-level topics, only the most appropriate ones should be selected to explain the recommendation: sometimes, a single topic can fully explain the article, whereas other times, the inclusion of multiple topics can provide helpful explanations. However, for a given target article the topic ranking typically includes dozens of topics, being many of then marginally relevant.[7] As a consequence, an automated method for identifying the optimal number of topics is necessary.

For this purpose, we use a method which involves selecting the appropriate number of items from a set of available items. This methodology was proposed in (de Campos et al. 2021)[8] where the input is a set of m items of any type and the importance of each is represented by a weight $[w_1, \ldots, w_m]$; the output is the most important items.

In our case, the topics are the items and the weights are obtained by aggregating, for each topic t, the scores across all relevant journals (subprofiles) related to it, expressed as $w_t = \sum_j scr(j, t)$.

The methodology is based on ranking the topics in decreasing order of their weights and comparing, through a similarity or a distance measure, the probability distribution obtained by normalizing the topic weights with the vectors $(1, 0, 0, \ldots, 0)$, $(1, 1, 0, \ldots, 0), (1, 1, 1, \ldots, 0), \ldots, (1, 1, 1, \ldots, 1)$. Each vector represents the decision of selecting the first $k = 1, 2, \ldots$ topics with the greatest weight. The optimal decision is the one that optimizes the distance or similarity measure being used. We denote this function as *TopicSelection* $([w_{t_1}, \ldots, w_{t_m}], measure)$. It should be noted that the number of topics selected is not constant and relies on their individual weights. Many distance or similarity measures were considered in (de Campos et al. 2021), although all of these converged into only five selection strategies, ranging from the most restrictive Euclidean measure which always chooses the top ranked topic to

---

[6] although only 5 terms are displayed in Fig. 3, this is done to reduce its size.

[7] This is because many subprofiles exhibit some degree of similarity to the query which is composed by the title, abstract, and keywords of the target paper.

[8] In this case, it was applied to select the most representative topics that can be associated with a document which has been characterized by a probability distribution over the entire set of topics obtained from an LDA algorithm.

the Overlap measure that selects all the topics with nonzero weights. Intermediate strategies employ the Dice, Sorensen, or Cosine measure.

We conducted an offline experiment to determine the appropriate metric for this selection problem using a set of 32,864 test queries from our dataset (see Sect. 5.2). We counted the number of topics selected for each case and found that on average, the Overlap measure shows 11.96 topics, which is obviously excessive. Using the proposed methodology, we were able to select an average of 1.14, 1.69, and 2.85 topics using the Dice, Sorensen, and Cosine measures, respectively. The average number of topics selected using Dice was too restrictive and almost always (87% of the times), only one topic was displayed. The Cosine measure, on the other hand, often selected a high number of topics (24.4% of the times select four or more topics, with a maximum of 35 topics), which could be excessive. Sorensen measure represents a compromise (96.5% of the times shows one to three topics, with a maximum of eight different topics). Thus, given that papers can relate to multiple topics, but not an excessive number, we decided to use the Sorensen measure for our explanations, i.e., *TopicSelection*($[w_{t_1}, \ldots, w_{t_m}]$, Sorensen) is applied.

To conclude this section, we relate our approach to those in the literature. As previously mentioned, a number of different tag cloud approaches have previously been described and these include (Gedikli et al. 2011), where the authors present a basic approach where a tag cloud is created directly from the words in the item description and their number of occurrences. In addition, they show a second alternative, called the personalized tag cloud, where the users express their interests in the tags and the tag-based diagram shows how the items cover such interests. The authors of this paper founded the use of tag cloud in the fact that "explanations based on tag clouds are not only well accepted by the users but can also help to improve the efficiency and effectiveness of the explanation process." An aggregated view of several users in the context of group recommendation in tag cloud form is presented in Felfernig et al. (2021), where the graphic contained not only the tags but also the indication of which users preferred each tag, resulting in a very interesting way of explaining. Pérez-Núñez et al. (2022) propose a tag cloud where the importance of the tags that characterize a product is not the typical term frequency but is learned by means of a deep learning process. In Chen (2013), based on a collaborative filtering RS where the items contain a textual representation, the tag clouds are colored according to terms found in the user's positive, negative, or neutral reviews. The suggested method can enhance the acceptance rate of recommendations and enhance user satisfaction.

The word clouds for our second explanation element are generated in a totally different way to these approaches. Firstly, we find ourselves faced with a CBRS context; secondly, we could say that the recommendations are personalized, but not in the way proposed in Gedikli et al. (2011) but in terms of the user's input (the article submitted to the RS) as the recommendations are first generated from the most relevant subprofiles in the ranking by means of a process of entity detection and then from a word frequency count to determine word size.

Titles of the most similar articles published in this journal (EE3)

| #1 | [● ○ ○] (1) Long-term benefits of home-based preventive care for preterm infants: a randomized trial. |
| #2 | [● ○ ○] (2) Preventive care at home for very preterm infants improves infant and caregiver outcomes at 2 years. |
| #3 | [● ○ ○] (7) Five-year follow-up of harms and benefits of behavioral infant sleep intervention: randomized trial. |

## Titles of the most similar articles published in this journal (EE3)

#1	[● ○ ○] (5) The effect of in-hospital developmental care on neonatal morbidity, growth and development of preterm Taiwanese infants: a randomized controlled trial.
#2	[○] (19) Early communication in preterm infants following intervention in the NICU.
#3	[○ ●] (34) Family functioning, burden and parenting stress 2 years after very preterm birth.

Fig. 4 EE3 Explanation text: "LIST OF ARTICLES PUBLISHED FOR THE RECOMMENDED JOURNAL, WHICH ARE SIMILAR TO THE TARGET ARTICLE. THE COLORS OF THE TRAFFIC LIGHTS REPRESENT THE DEGREE OF SIMILARITY WITH RESPECT TO THIS TARGET ARTICLE (GREEN VERY SIMILAR; ORANGE, MEDIUM SIMILARITY; RED, LOW SIMILARITY). This figure shows EE3 explanations for the first (top) and third (bottom) recommended journals, *Pediatrics* and *Early Human Development*, respectively, which are similar to the target article in Fig. 1

4.2 Local explanations: why is a particular journal recommended?

The system also offers more specific explanations about each of the recommended journals. By clicking on the magnifying glass to the right of the journal title (see Fig. 2), a detailed explanation for each journal can be found.

4.2.1 EE3: journal-related articles

The first specific explanation (Element 3) for a journal is a list of up to three articles published in the journal which are most similar to the target article (Fig. 4 shows the list associated with the first and third recommended journals for the target article used in Fig. 1).

Each article has associated a traffic light, where the colors means the degree of similarity with respect to the target article: green, very similar; orange, medium similarity; and red, low similarity.

In this way, users can observe how their target articles are similar in content to other articles previously published in the journal, and this provides a different, more specific reason for the journal recommendation.

In order to find the set of related articles, we used an auxiliary index to index each paper separately rather than topic-based subprofiles. We submitted the same query (title, abstract and keywords) and obtained a ranked list of related articles using the same similarity measure, i.e., the Jelineck–Mercer language model.

For each recommended journal, we decided to show up to three articles (the most relevant ones published in the journal), but only if they belong to the 50 most similar ones. This is because we believe that articles beyond this threshold may not be sufficiently similar to the target paper. It is worth noting that as a result of this restriction,

it is possible that zero, one, or two articles may be displayed for a given journal. This may be an indication of a possibly inappropriate journal recommendation.

In order to help users quickly identify the relevance of the displayed articles, we have color-coded them: articles with a high relevance (ranked in the top 10) are displayed in green, those with a medium relevance (ranked between 11 and 25) are displayed in orange, and those with a low relevance (ranked below 25) are displayed in red. The position of each article in the ranking is also shown.

This explanation element is related to what Bilgic and Mooney (2015) call the Influence Style Explanation: in the context of the explanation, they show the training items with the greatest impact on the recommendation and their corresponding user's ratings, as well as a score representing their influence on the recommendation. It is a class of similar items to the one recommended but based on users' previous experience. Shmaryahu et al. (2020) also justify the recommendation showing some similar and previously rated items to the recommended one, although this technique is not the best one for the users in their study. In our case, this explanation element does not offer previously rated items but similar articles to the target one published in the recommended journal, based exclusively on content similarity. Moreover, for this explanation element 3, the list of similar items does not have any direct impact on the recommendation and is only a mere explanation in contrast with how the similar items are used in Bilgic and Mooney (2015) and Shmaryahu et al. (2020).

4.2.2 EE4: journal-related topics

Another explanation element, specific for each recommended journal, consists of a set of word clouds that depict the primary topics and their coverage within the journal. While we employ a similar representation technique as in EE2, utilizing word clouds in an abstract manner, there are several distinctions. Firstly, we exclusively consider relevant subprofiles from the specific journal as input so that the topics selected (the most representatives for this journal) can be different from those in the global explanation EE2. Moreover, in cases where the same topic is selected, variations may emerge in how that topic is specifically addressed within the journal. This distinction is exemplified in Fig. 5, displaying the word cloud representation of Topic 56 in *Developmental Medicine and Child Neurology*, the sixth recommended journal. Globally, Topic 56 was associated with `child health research` in EE2 (see Fig. 3). However, within the recommended journal, its focus is narrowed to issues related to `motor function and cerebral palsy in children`. Such insights can be valuable for users in determining the potential relevance of this journal for publishing their research.

Another difference lies in the methodology for determining the number of topics presented to the user. To address this, we employ the same topic selection algorithm as previously discussed. However, in this scenario, the weights are determined by the raw scores $w_{tj} = scr(j, t)$, which quantify the relevance of each subprofile in a specific journal j. Regarding our selection strategy, tailored to a single journal focus, follows a more restrictive approach by limiting the number of chosen topics. Particularly, we base this selection on the Dice measure, which, in practical terms, results in the inclusion of few topics, typically one or two. More formally, if the

Topic 56

School-Age Outcomes of Early Intervention for Preterm Infants and Their Parents: A Randomized Trial.

To examine the child and parental outcomes at school age of a randomized controlled trial of a home-based early preventative care program for infants born very preterm and their caregivers. At term-equivalent age, 120 infants born at a gestational age of <30 weeks were randomly allocated to intervention (n = 61) or standard care (n = 59) groups. The intervention included 9 home visits over the first year of life focusing on infant development, parental mental health, and the parent-infant relationship. At 8 years' corrected age, children's cognitive, behavioral, and motor functioning and parental mental health were assessed. Analysis was by intention to treat. One hundred children, including 13 sets of twins, attended follow-up (85% follow-up of survivors). Children in the intervention group were less likely to have mathematics difficulties (odds ratio, 0.42; 95% confidence interval [CI], 0.18 to 0.98; P = .045) than children in the standard care group, but there was no evidence of an effect on other developmental outcomes. Parents in the intervention group reported fewer symptoms of depression (mean difference, -2.7; 95% CI, -4.0 to -1.4; P < .001) and had reduced odds for mild to severe depression (odds ratio, 0.14; 95% CI, 0.03 to 0.68; P = .0152) than parents in the standard care group. An early preventive care program for very preterm infants and their parents had minimal long-term effects on child neurodevelopmental outcomes at the 8-year follow-up, whereas primary caregivers in the intervention group reported less depression.

Fig. 5 EE4 and EE5 Explanation text: "YOUR RESEARCH IS LINKED TO Developmental Medicine and Child Neurology JOURNAL THROUGH TOPIC 56. THE TERMS HIGHLIGHTED (RED/BOLD FACE) IN YOUR SUBMISSION ARE THE ONES THAT PLAYED A SIGNIFICANT ROLE IN DETERMINING THIS ASSOCIATION. THE WORD CLOUD SHOWCASES THE MOST FREQUENTLY USED TERMS IN THOSE PAPERS THAT, UNDER THIS TOPIC, HAVE BEEN PUBLISHED IN THIS JOURNAL"

topics related to the candidate journal j in the output ranking are $t_h, \ldots, t_k$ we use the *TopicSelection*$([w_{t_h j}, \ldots, w_{t_k j}],$ Dice$)$.

4.2.3 EE5: related terms

The final explanation element associated with a given journal also relates to the set of selected topics, but in this case, we aim to concentrate our attention on those significant terms in the target article that contribute the most to the relevance of each selected subprofile. In order to do so (see Fig. 5, right-hand side), the title and the abstract of the target article are displayed with a selection of terms marked in color in order to illustrate the coincidences. In this way, the user can see which article terms are mainly responsible for the journal recommendation and decide whether the system is focusing on the most important ideas or rather on less important ones. This information can help the user decide whether to accept the recommendation or not.

In order to accomplish this, we first isolate the contribution of each term to the subprofile score $scr(j, t)$ of the selected topic t for the given journal j, whereby terms that do not match the subprofile have a zero contribution. Following de Campos et al. (2018), we then sort the terms in decreasing order of their contribution and select the ones that achieve at least 90% of the full contribution. We use the cosine measure to compute the similarity between the contributions of both the set of selected terms and the full set of terms. Finally, these selected terms are highlighted in a different color (red) so that they stand out. Additionally, in order to focus the user's attention on the most significant phrase in the paper, i.e., the one that most contributes to the final score, we also decided to highlight this phrase (in bold).

Continuing with our ongoing example, it is worth noting that when researchers consider both EE4 and EE5, they may decide against submitting their paper to *Developmental Medicine and Child Neurology* journal. This decision can be driven by the observation that the journal's scope, under Topic 56, appears to be only loosely aligned to the content of this particular research (mainly generic terms related to children, parents, infants, etc. are highlighted and by themselves can explain the 90% of the resulting score). Making such decision is crucial to prevent publishing the paper in an unsuitable journal which could potentially restrict the visibility and impact of our research.

In Li et al. (2021), in the context of an RS service, Li et al. detect contexts (places, dates, companion, etc.) and contextual features from reviews of these services (hotels, attractions, venues, etc.) and highlight in the text reviews those relevant for the users according to their profiles. According to their experimental results, this highlighting is very useful for users because it provides a personalized explanation and allows them to identify relevant features of the explanation. The underlying idea of explanation element 5 is similar, but in this case, we only work with plain words and do not refer to any kind of underlying concept or feature.

4.3 Categorizing the explanation elements

Finally, once we have described different explanation elements proposed in this paper, we categorize them according to the different types of explanations and RSs described in Sect. 2.

Our proposed explanation is basically a model-intrinsic approach since the explanation comes from the recommender model.

In terms of the explanation resources, which are classified as human, item, or feature style, we can conclude that our approach mainly corresponds to the last type since we consider keywords from the article texts and these are used as a kind of journal feature. Since the underlying RS is content based, the recommendation is made based on the similarity between the target article and the journal profiles. However, explanation EE3, which consists of the list of similar articles published in the recommended journal, could be considered as a type of item style explanation.

Our proposal is distinctly local, as it provides an individualized explanation for each recommended item. Additionally, it is non-personalized, given that, despite adapting results to the target paper, we do not consider any information (such as interests or preferences) pertaining to the individual seeking the recommendation. The only information supplied by the user is the title and abstract of the target article. Since the recommendation and explanation are not integrated into our recommender model and the explanations are attached to the recommended items, it is not a recommendation-by-explanation one.

Focusing on the explanation content, we could say that our approach combines the user input (the text provided by the user), the information obtained in the inference process (the recommendation itself), and feature-based information, since keywords are also used to perform the explanation. Moreover, in our case, the explanation style is a combination of content-based and case-based information, named accordingly

in Tintarev and Masthoff (2015), or item style explanation as in Papadimitriou et al. (2012). The reason for the first is obvious since we are working with a content-based RS, and for the second since similar items to the target are presented for a given recommended journal.

Finally, we would like to emphasize that our system consistently provides accurate explanations for all generated recommendations, as these explanations rely directly on the internal information managed by the RS. The persuasiveness of these explanations is a matter that a user study could unveil. In any case, this may depend on the inherent quality of the recommendations. If the recommendations are not appropriate, it is conceivable that the explanations may lack conviction, for example, featuring topics or highlighting terms that are not central to the target article. However, this could be construed positively, offering users reasons to reject such recommendations.

5 Design of the user study

5.1 Objective

The main objective of the user study described in this section is to determine whether the previously presented designed explanations are deemed to be suitable by biomedical experts and help to understand and accept the RS recommendations.

In line with the explanation benefits outlined by Tintarev and Masthoff Tintarev and Masthoff (2007), in this study we focus on transparency, satisfaction, trust and scrutability of the explanations, with the further inclusion of quality and novelty of the recommendations. With these aims in mind, we designed the questionnaires to gather user feedback after interaction with the XRS. We do not consider effectiveness, persuasiveness and efficiency since our XRS is not an online, real system enabling actions to be performed once the explanations have been received.

5.2 Dataset

The test collection used in the experimentation is called PMSC-UGR (Albusac et al. 2018) and has been created by the authors from PubMed and Scopus. It contains the title, abstract, keywords, citations, and authors of papers published between the years 2007 and 2016 from 1002 different journals. The articles published in the first nine years (a total of 276,679 papers) were used to build the journal subprofiles and feed the RS. The remaining articles from the year 2016 (a total of 32,864) configure the test set from where the articles for showing the recommendations and their explanations were selected in the user study.

5.3 Participants

In order to perform a user study of our recommender system and its explanation facilities, we recruited a number of researchers (mostly university professors) who were responsible for a large number of biomedical publications. Of these, seventeen (10

women and 7 men) completed the user study and evaluated a total of 68 submission recommendations, each one implying the analysis of the 10 suggested journals, which have therefore been considered valid for the result analysis. The distribution according to their expertise is as follows: medical doctors (6), nurse (1), biostaticians (one a researcher from a public hospital) (4), librarian from an hospital library (1), biochemists (2), and computer scientists with expertise in bio-informatics (3).

5.4 Protocol

The objective of the user study had two folds: to assess the quality of both the explanations provided for the recommendations and the recommendations themselves. With this objective in mind, researchers were invited to imagine the following scenario: They had recently written an article and were currently deliberating on the choice of a journal by employing the RS. They were instructed to make the ultimate decision by not solely relying on the suggested journal list but also by scrutinizing the explanations accompanying each recommendation.

To simulate this scenario, the participants were instructed to select some articles from a pool of already published ones.[9] Below, we detail how this pool of articles has been created and provide the motivations behind it. Now, our focus shifts to detailing how participants reviewed the recommended journals and their respective explanations. For the sake of simplicity, we have omitted some details of the protocol; however, a comprehensive description of the protocol provided to the users for conducting the evaluation is presented in "Appendix Appendix A:."

Specifically, upon selecting a target article (by clicking a magnifying glass), users were presented with a ranking of 10 journals recommended by the system for that particular article. At this stage, they are not observing the explanation but only a recommendation, the original output for the RS, so the participants can evaluate the list of suggested journals without any explanation. After clicking in the explanation link, the same ranking is offered completed with their corresponding confidence percentages (EE1) and several word clouds of the most related topics (EE2). The user then might click, in turn, on the explanation for each suggested journal to obtain its specific explanations (EE3, EE4, and EE5).

After observing all the explanation for such target article, the participant is requested to complete a specific questionnaire (therefore, there are as many filled questionnaires as target articles evaluated). Once all the selected articles had been analyzed, participants were instructed to finalize the process by completing a global questionnaire at the end.

The questions from these two questionnaires are shown in Tables 1 and 2. Except for a few questions where an item was selected, all the questions involved scaling responses on a 7-point Likert scale (1—strongly disagree, 2—disagree, 3—somewhat disagree, 4—neither agree nor disagree, 5—somewhat agree, 6—agree, 7—strongly agree). In addition, the global questionnaire contained a question for the users to freely comment on the evaluation.

---

[9] All of them with a more recent publication date than the articles used to build the system.

Table 1 Specific questionnaire for each target article and each user

SQ1: The list of recommended journals, without information about the explanation, seems reasonable to me

SQ2: Once I know the explanation, I understand why the system suggests such a recommendation, independently on its correctness

SQ3: Once I know the explanation, I consider that the recommended journals make sense

SQ4: The explanation helps me to make the decision of whether accepting or not the journals suggested by the system

SQ5: I am satisfied with how the recommender systems works in this specific case

Table 2 Global questionnaire for each user

GQ1: The list of recommendations seems to be reasonable

GQ2: The systems recommends new (unknown) journals where the paper could be submitted

GQ3: The explanation has proved to be helpful to understand what are the foundations of the system to recommend journals

GQ4: The explanation helps me to better decide about the suitability or not of a journal

GQ5: My trust in the recommender system is higher once I know how it explains the recommendations

GQ6: When some of the recommended journals are no suitable for the target article, I understand the reasons through the explanation

GQ7: The explanation helps me to know the words from the target article that cause the recommendation of such journals

GQ8: In general, I understand the explanation

GQ9: What explaining element was easier to understand? (select one option)

GQ10: What explaining element was more difficult to understand? (select one option)

GQ11: What explaining elements were more helpful to understand the explanation? (select one option)

GQ12: In case this recommender system were publicly available, I would use it to find potential journals where I could submit my articles

GQ13: Have you previously used other journal recommender systems?

GQ14: I shall recommend the use of the system to my colleagues

According to the selected aims of this user study which were outlined at the beginning of Sect. 5.1, Table 3 provides details about which questions deal with each purpose.

Creating the pool of articles

In addition to the global objective of determining the quality of the explanations, we want to evaluate whether the perceived quality depends on the quality of the recommendations. With this aim, the articles subject to evaluation by participants in the study were chosen from the test set and categorized based on the system's objective

Table 3 Classification of questions in the specific (SQ) and global (GQ) questionnaires

Aim	Questions
Questions to evaluate different explanations aims	
Transparency	SQ2, GQ3, GQ6, GQ7, GQ8
Satisfaction	SQ5, GQ12, GQ14
Trust	SQ3, GQ5
Scrutability	SQ4, GQ4
Preferred types of explanations	GQ9, GQ10, GQ11
Questions to evaluate recommendations	
Quality	SQ1, GQ1
Novelty	GQ2

performance.[10] The classification comprises three categories: Category A includes articles for which the first journal recommended by the system aligns with the actual journal of publication; Category B encompasses articles where the actual journal is among the ten recommended by the system; and Category C consists of articles where the actual journal does not match any of the recommended journals. Importantly, participants were deliberately kept unaware of the actual journal of publication to prevent bias in their assessments. This categorization of articles can be useful for investigating potential correlations between user evaluations of explanations and the system's ability to accurately identify the journal of publication. The objective is to discern whether such correlations exist within the evaluated articles.

Also, we want to evaluate whether the opinions of the participants are influenced by their expertise in the target research field. To fulfill this purpose, we introduced a fourth Category (D), comprising articles authored by the users participating in the system evaluation, which were directly provided by them. For these articles, the actual journal where the paper was published has been excluded from the lists of recommended journals to mitigate potential bias in users' opinions stemming from its position in the ranking. This fact was explicitly communicated to the participants. The inclusion of Category (D) is justified by our assumption that assessing explanations for papers they are intimately familiar with would result in higher-quality evaluations, given their in-depth knowledge of the content and topics covered.

The distribution of articles in the pool by category is as follows: A(5), B(14), C(9), D(23). Although these categories were not disclosed to the participants, they were directed to guarantee that each participant selects at least one article from each category. However, responses were not received for all the selected papers. Finally, a total of 68 recommendations and explanations were evaluated by the 17 experts participating in the evaluation. The number of articles evaluated within Categories A, B, C and D is 15, 16, 14 and 23, respectively.

---

[10] This information was intentionally withheld from the users.
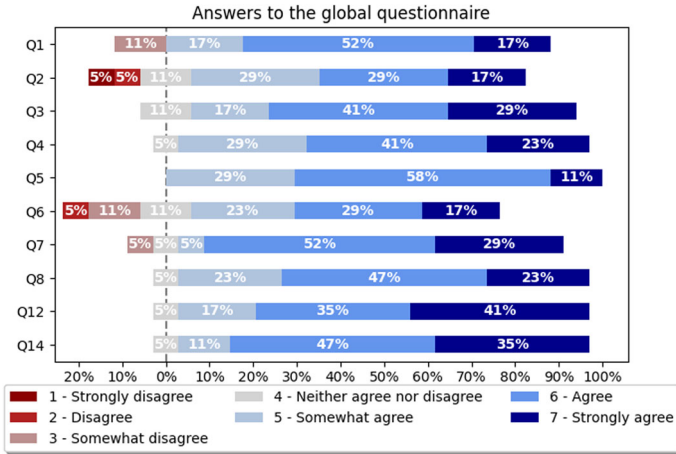
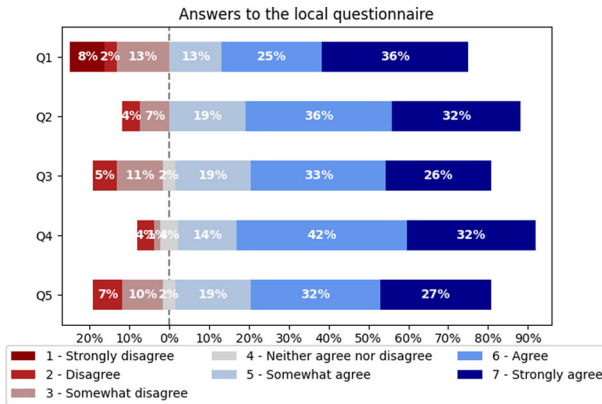**Fig. 6** Results of the global questionnaire

Fig. 7 Results of the specific questionnaire

5.5 Results

In this section, the results of the evaluation are presented as well as the insights from their analysis.

Figures 6 and 7 show the charts with the answers to the global and specific questionnaires, respectively. In these figures, the stacked bar charts display the distribution (proportion) of user responses, color-coded to represent different answer categories. In order to facilitate interpretation, each bar is aligned around the neutral answer (neither agree nor disagree) which is marked by a vertical line. This alignment serves as a reference point, allowing for easy identification of negative responses (ranging from strongly disagree to somewhat disagree) and positive responses (ranging from somewhat agree to strongly agree).

Table 4 shows the averages and standard deviations of the scores for the questions

Table 4 Average and standard deviation results for the specific questionnaire on a scale from 1 to 7

	SQ1	SQ2	SQ3	SQ4	SQ5
Average	5.28	5.74	5.43	5.87	5.43
s.d	1.98	1.36	1.51	1.22	1.55

Table 5 Average and standard deviation results for the global questionnaire on a scale from 1 to 7

	GQ1	GQ2	GQ3	GQ4	GQ5	GQ6	GQ7	GQ8	GQ12	GQ14
Average	5.65	5.12	5.88	5.82	5.82	5.12	5.94	5.88	6.12	6.12
s.d	1.13	1.60	0.96	0.86	0.62	1.45	1.06	0.83	0.90	0.83

Table 6 Number of positive, negative, and neutral answers for each question in the specific questionnaire

	SQ1	SQ2	SQ3	SQ4	SQ5
Positive	51	60	54	61	54
Negative	17	8	12	4	12
Neutral	0	0	2	3	2

Table 7 Number of positive, negative, and neutral answers for each question in the global questionnaire

	GQ1	GQ2	GQ3	GQ4	GQ5	GQ6	GQ7	GQ8	GQ12	GQ14
Positive	15	13	15	16	17	12	15	16	16	16
Negative	2	2	0	0	0	3	1	0	0	0
Neutral	0	2	2	1	0	2	1	1	1	1

in the specific questionnaire. Table 5 displays the same information for the questions in the global questionnaire with graded responses.

We can see in these tables that, on average, the evaluation of the explanation capabilities of the system (as well as its recommendations) is positive: the average scores are always between 5 and 6 (between somewhat agree and agree). In particular, the questions which scored close to or greater than 6 are: SQ2, GQ3, GQ7 and GQ8 (relating to transparency); SQ4 and GQ4 (relating to scrutability); GQ5 (relating to trust); and GQ12 and GQ14 (relating to satisfaction). The lowest scored question is GQ2 (5.12), relating to the novelty of the recommendations, although this could be explained by the fact that our evaluators are all experienced researchers in the biomedical field.

In Tables 6 and 7, we also display, for the specific and global questionnaires, respectively, the number of positive, negative, and neutral answers for each question, where we consider a score of 5, 6, or 7 to be positive, a score of 1, 2, or 3 to be negative, and a score of 4 to be neutral. From these data, using a one-sample test of proportions for each question, we have found that the proportion of positive answers is significantly greater than the proportion of negative answers in every case (with p-values that are mostly very low). This confirms that our system is positively evaluated by the users.

Focusing on the specific questionnaire outlined in Table 6, it is evident the positive impact of explanations when comparing the responses to SQ1 with those of other

Table 8 Average results for the specific questionnaire focusing on the subjective opinions of the participants, i.e., when users judge the raw recommendations as positive (positive SQ1) or negative (negative SQ1)

Class	SQ1	SQ2	SQ3	SQ4	SQ5
Positive SQ1	6.31	6.22	6.04	6.23	6.07
Negative SQ1	2.18	4.29	3.58	4.76	3.47

questions. Thus, while the recommendations, without the presence of explanations, were initially received positively (51), the counts for other specific questions, which consider the impact of the explanation, increased. Conversely, the opposite trend is observed for instances of negative responses to SQ1 (17). These findings confirm the positive role that explanations play in users' perception of the RS.

Digging deeper, our goal is to investigate whether users' subjective opinions on the suggested journals have an impact on their perception of the utility of the explanation. For this purpose, in Table 8, we also present how users perceived the explanations in two distinct situations: one where the user does not like the suggestions (articles with a negative value for SQ1) and another where the user agrees with them (articles with a positive value for SQ1).

First, let us focus on negative SQ1 answers: Without explanation, the participants gave an average rating of 2.18. However, after evaluating the explanations, their perception of how the RS works (SQ5) increased significantly to an average value of 3.47. [11] Furthermore, it is noteworthy to mention that the Pearson correlation score between SQ1 and SQ5 is -0.673. This score indicates that as SQ1 decreases, there is a greater improvement in user perception after the explanation. This shift can be attributed to the fact that explanations aid in comprehending the RS output (SQ2, averaging 4.29) and considering the sense behind some of the recommendations (SQ3, averaging 3.58). Notably, these explanations prove valuable in their decision-making process (SQ4, averaging 4.76). These are all observations indicating the usefulness of explanations for users.

Conversely, when participants agreed with the provided recommendations (positive SQ1 answers), despite finding the explanations helpful (with an average rating greater than 6 in all cases), there was a marginal decline in their overall perception. The post-explanation SQ5 score equaled 6.07, in contrast to the pre-explanation score of SQ1, which was 6.31,[12] being in this case positively correlated (Pearson correlation between SQ1 and SQ5 equals to 0.72). One potential interpretation of this scenario is that participants might utilize explanations to deduce that some of the recommended journals lack coherence, leading to a decline in SQ3 compared to SQ1.

These findings suggest that when users agree with the recommendations, the subsequent explanations may appear irrelevant. On the other hand, when users do not like the recommendation, the explanations prove to be beneficial. Similar outcomes have been observed in a different domain as is the design of explanations in recruitment

---

[11] Results of the paired-t test indicated that there is a significant medium difference between before and after explanations, $p = 0.031$.

[12] Results of the paired-t test indicated that there is a significant small difference between before and after explanations, $p = 0.009$.

Table 9 Average results for the specific questionnaire for each class of article on a scale from 1 to 7: the number of articles evaluated within Categories A, B, C, and D is 15, 16, 14, and 23, respectively

Class	SQ1	SQ2	SQ3	SQ4	SQ5
A	5.87	6.13	6.00	6.27	6.00
B	4.81	5.81	5.25	5.56	5.19
C	5.00	5.57	5.14	5.43	4,93
D	5.39	5.52	5.35	6.09	5.52

Table 10 Frequency of answers evaluating different explanation elements as easy (GQ9), difficult (GQ10), and helpful (GQ11) and previous experience with journal recommender systems (GQ13)

	GQ9	GQ10	GQ11	GQ13	
EE1	5	3	8	Yes	2
EE2	2	3	0	No	15
EE3	6	0	5		
EE4	0	8	1		
EE5	4	3	3		

decision support systems (Shulner-Tal et al. 2022). Therefore, we want to highlight the contribution of the explanation in those situations where the recommended journals do not align with the user's criteria, which can be perceived as a weakness for a RS. This perception of inefficacy can lead to decreased trust in the system, dissatisfaction with the user experience, and a reluctance to rely on the recommendations offered by the RS. The results obtained clearly indicate that explanations enhance user satisfaction, helping in their decision-making process. This assertion is corroborated by responses to GQ1, which participants answered after evaluating recommendations for all their selected papers. In this instance, 88.23% of participants found the recommendations to be reasonable, in contrast with the 75% positive responses obtained for SQ1.

In the preceding discussion, we examined the acquired results in relation to subjective user opinions. Now, our attention shifts to the objective quality of the system measured by its capability of suggesting the actual journal. It is worth noting that there may be additional relevant journals, different from the actual journal, where the paper could be published.[13] However, valuable insights can be gleaned from this analysis. Specifically, in Table 9, we break down the average scores for the questions in the specific questionnaire into the four categories of articles considered. It appears that there is a clear positive correlation between the opinions of the users concerning the quality of the system (recommendations and explanations) and the objective results obtained. The trends suggest that the better the system objectively performs for an article (Categories C, B, A), the higher users score subjective system performance for this article. For Category D, we obtained intermediate results (mainly between the results for Categories A and B).

In terms of the questions which involved selecting an item and which mainly individually analyzed various explanation elements, Table 10 provides information about

---

[13] This can be reflected by the positive outcomes achieved for category C, where users still find value in the list of recommended journals even when the system is unable to retrieve the actual journal.

which explanation elements are the easiest and which are the most difficult to understand and also which are the most helpful. We can also see that most participants had no previous experience with this type of recommender system.

There are three explanation elements (EE1, EE3, and EE5) that stand out in terms of understandability and usefulness. The easiest to understand is the list of related articles, followed by the numerical scores and the highlighted terms in the title and abstract. The same EEs are also considered the most helpful, reversing in this case the order of scores and related articles. These results seem to tally with the findings of previous work (Carenini and Moore 2000; Ehsan et al. 2019; Shmaryahu et al. 2020) which concur that the preference is for short, easy-to-understand explanations. On the other hand, the users considered the two types of word cloud to be less useful (probably because they have not been understood properly, especially EE4). In this sense, it seems that users need to become familiar with the provided explanation elements to minimize their cognitive effort, aligning with the results in Gedikli et al. (2014).

In order to conclude this section, we examine different explanation objectives as outlined in Table 3. Generally speaking, users have reported a positive perception of the *system's quality* (GQ1, SQ1). This tallies well with the findings of our offline evaluation (de Campos et al. 2022), where the system successfully included the correct journal among the top-10 recommended journals in 71.21% of the cases. Additionally, the users have reported *satisfaction* with the given explanation (GQ12, GQ14), with an average rating of 6.12. It is of note that this satisfaction is particularly strong when the top-ranked suggestion aligns with the "correct" journal (see Table 9, Category A, SQ5 column), leading to easily understandable explanations. However, when a recommendation might not tally with the user's opinions (see Table 9, Category C, SQ5 column), understanding the explanation becomes more challenging.

The explanations clearly enhance the *transparency* of the recommendations with an average rating of 5.9 (GQ3, GQ7, GQ8). This perception of transparency seems to be closely tied to the quality of the recommendations themselves (see Table 9, Category C, SQ2 column, and also GQ6 in Table 5). Moreover, these explanations help users reason about the correctness of the recommendation system and enable them to make decisions, as reflected by an average rating of 5.85 for *scrutability* (SQ4, GQ4). More significantly, when users need to make decisions for their own papers (see Table 9, Category D, SQ4), the importance of explanations becomes even more pronounced. This scenario represents the real-world use of the RS and emphasizes the crucial role that explanations play since users rely heavily on them to guide their decision-making regarding the choice of journal for submission.

Based on these findings, we can conclude that providing explanations helps users to *trust* (GQ5) the decisions made by the RS and increases their confidence in its reliability. However, as expected, the quality of the recommendations plays a crucial role in shaping this trust and confidence.

Limitations of this study   A possible limitation of this study is the relatively small number of users involved in the evaluation of our explanation proposal. It is important to note that the individuals conducting the study must possess a comprehensive understanding of the subject matter and be familiar with the journals suitable for the publication of the work. The problem is that in the biomedical context where this

study was conducted, it is very difficult to find available senior researchers who could serve as volunteers. Nevertheless, despite this drawback, we believe that it is possible to conclude that the explanation approach presented in this paper has been widely accepted as a useful tool for making users aware of the recommendation reasons.

A second limitation arises as a consequence of the aforementioned one, given that the final number of evaluated papers is relatively low. Although we contend that the quantity of evaluated papers is adequate for deriving meaningful results, a larger number of evaluated papers would have enabled more robust and conclusive findings. However, the strength of the obtained conclusions is reinforced by the fact that all participants are senior researchers.

Lastly, during the evaluation, we presented the explanation elements to users as if they were interacting with an operational recommender system. We organized the explanations into two stages (EE1–EE2 initially, followed by EE3–EE4–EE5), where each group of explanation elements could be considered a compact explanation. Within each group, the explanation elements could be displayed in a different order to participants, aiming to avoid potential biases or counterbalancing issues influenced by the presentation order of the explanation elements.

6 Conclusions and future works

In this paper, we have presented a novel explanation approach for complementing the output of a journal CBRS and describing the reasons why the RS suggested such journals for publishing the paper provided by the user.

The elements of the explanation (tag clouds, RS scores, similar papers, abstract highlighting), although not new in themselves, are totally adapted to the current problem and to the RS model, offering an example of a model-intrinsic explanation and combining global and local information extracted from the model.

In order to validate the explanation approach in terms of a number of aims (transparency, satisfaction, trust and scrutability), we designed and implemented a user study in the biomedical domain, where the users, after interacting with the explainable RS, had to complete questionnaires to record their opinions about the process.

From the questionnaire results, we might conclude that, overall, the explanation provided proved useful for understanding why the RS suggested a particular journal, not only as separate elements but as a whole, offering a combined explanation which was accepted by most users involved in the study. This finding, although observed in other contexts, is also documented in a wide range of publications, such as Gedikli et al. (2014), Tintarev and Masthoff (2015), Millecamp et al. (2019), Hernandez-Bocanegra et al. (2020) and Zhang and Chen (2020). Also, the study underscores the significant contribution of explanations in scenarios where the recommended journals do not align with the user's criteria. Similar patterns have been identified in a distinct domain (Shulner-Tal et al. 2022).

In addition, we found a positive correlation between the objective performance of the RS and user satisfaction.

Regarding the EEs, it seems that all of them (with the exception of the tag clouds) are easy to understand and offer an added value for interpreting the RS output. However,

the two EEs based on tag clouds are more difficult to understand and must therefore be redefined in further research.

In terms of future work, following the suggestions of various users in the observation fields of the questionnaires, we aim to improve the interpretability of the tag clouds as it was sometimes difficult to understand the general concepts or topics that these intended to express. One solution could be to complement the use of tag clouds with a natural language description generated with a large language model to describe the topics integrated in them. Additionally, a number of users claimed that the terms contained were too general and this could in turn lead to confusion and result in an unclear or not useful tag cloud interpretation. We would consequently redefine the selection policy of the tag clouds in an attempt to find the most expressive ones.

Similarly, several users proposed enhancing the explanation by incorporating specific features of the recommended journals, such as the impact factor or other bibliometric indicators. This addition aims to enrich the explanation with supplementary information, potentially improving the decision-making process.

In this context, integrating the journal impact factor (or any other relevant measure) could be accomplished through an online reranking of the initial recommendations, considering the scientometric index of the journals. The decision to include its effect could be delegated to the users, who could utilize a slider, for instance, to adjust the balance between the two extremes (ranging from 100% content to 100% impact factor). Other bibliometric indicators, such as quartiles, can be easily integrated into the recommendation ranking by grouping journals accordingly. The same approach could be applied to the type of journal access; results could be organized by open access and others.

In a previous work, de Campos et al. (2022), the author's previous publications were also incorporated into the RS model as an additional factor which also impacts the journal recommendation process. Another interesting line of future research, therefore, would be to incorporate this into the explanation process, with the fact of having previously published in the recommended journals being another key element for explaining them.

Meanwhile, this explanation scheme will be exported out of the biomedical field to cover any type of journal domain and this is a straightforward process. In this way, this kind of explanation, which has proved useful in this domain, would be successfully exported to others, thereby widening the range of researchers who could benefit from the explanation facilities provided.

Finally, we plan to adapt the recommendation system to the problem of academic expert finding (for example, given a scientific text, to recommend researchers whose expertise would be useful for collaborations in the topics covered by the provided text) and to modify the explanation approach to deal with the singularities of this problem.

Appendix A: Description of the protocol of the user study

Before commencing the evaluation, users were instructed to review the study's guidelines to familiarize themselves with various EEs, understand their interpretation, grasp the flow of the explanation, and learn how to conduct assessments.

To initiate this process, we provided users with the following introductory background regarding the issue of venue recommendation:

Background

Given a paper ready to be submitted to a journal or conference, venue in general, the problem of venue recommendation attempts to automatically identify suitable publication venues for such paper and suggest them to the author. This is a difficult problem for various reasons: firstly, because there is a huge number of possible publication venues, and secondly, because even within a single specific research domain, there are thousands of publications. It is not, therefore, easy for researchers to be aware of every academic venue that would suit their domain of interest. The problem is further exacerbated by the increasing number of papers which contain multidisciplinary research and by the dynamic change in the scope of certain venues. The situation is also more difficult for new inexperienced researchers and for experienced researchers who move to new research areas.

In a previous research, we developed a journal recommender system that tries to ease this task by suggesting suitable journals that the author could consider for publication.

In this new research, we are conducting a user study to check whether explanation features designed for the journal recommender system are useful for understanding the recommendation and increasing the user's confidence on it.

The second part is focused on the aims of the evaluation:

Aims

At this point, it is important to remark that we are trying to evaluate separately the quality of the recommendations offered by the system (i.e., whether the proposed journals seem you appropriate to publish the given article) and the quality of the explanations offered by the system to justify its recommendations. Our main goal is to evaluate the quality of the explanation facilities of the system, although knowing your opinion about the quality of the recommendations is also important. It should be noticed that it is possible that the system makes correct recommendations and at the same time the explanations of these recommendation are useful, but it can also be the case that the recommendations are okay but the explanations are poor and also that the recommendations are not appropriate but the explanations of these bad recommendations make sense (and even can be useful to decide not to accept the recommendations). Both bad recommendations and bad explanations are also possible.

This system, in its current state, represents a first step within the journal selection process by recommending journals that are suitable to publish a given paper based solely on thematic content. In this sense, when you are deciding whether the journal recommendations offered by the system are good or not, the decision should be made based exclusively on the thematic content of the articles published in the recommended journal(s) compared with that of the target article, without taking into account other factors concerning the journals, as for example impact factors, editorial boards or duration and hardness of the evaluation process.

Next, users were provided with a comprehensive description of all the EEs they would encounter in the evaluation. Additionally, instructive examples in the form of graphics, as illustrated in Figs. 2, 3, 4, and 5 from Sect. 4 of the paper, accompanied the descriptions. These examples aimed to illustrate specific elements and guide users on how to interpret them, serving as explanations for the recommended journals. The objective was to acquaint users with the EEs and enhance their understanding of how to interpret them in the context of explanations:

Explanation Elements

[EE1]—In the list of recommended journals, attached to each journal there is a percentage that reflects the confidence of the system on the recommended journal, i.e., showing decreasing degrees of matching between the target article and each of the journals, as estimated by the system. These scores explain the ranking of recommended journals provided by the RS as they give information about the certainty of the RS on each recommended journal.

[EE2]—An association of the target article to one or several thematic/topical areas which are commonly covered by the recommended journals. These thematic areas are represented in the form of word clouds. The words represented in them are the most important ones for those topics. These word clouds explain the main topics of the recommended journals in relation to the target article and the user could use this information to assess whether the topics of the target article are related to one or several topics extracted from all the journals in the ranking. If not, it could be reasonable to think that the recommendation might not be totally useful as the journals do not publish papers in the target article's topics. This EE does not depend on a specific recommended journal.

[EE3]—A list of the up to three most similar articles published in a selected journal from the ranking (local to a chosen journal). Each similar article has associated a traffic light highlighting the green, orange or red lights according to the degree of similarity with the target article (green, very similar; orange, medium similarity; red, slightly similar). Three very similar papers to the one to be submitted (all in green color) would explain the fact that this journal has already published papers with closely related content so it would be a good option to submit the target article. Otherwise, if they are in red, they are less related and therefore the journal may be less appropriate than others, although it still could be a viable option.

[EE4]—One or several word clouds which reflect topics commonly covered by the articles published in a selected journal (local to a chosen journal). If the journal topics shown are similar to the topics covered by the target article, it could be concluded that the target article would fit very well in the scope of the journal.

[EE5]—For each of the shown word clouds/thematic areas (EE4), the abstract of the target article with those words in common with this area colored, in order to illustrate the coincidences. The more words colored in red in the abstract more coincidences with the words forming the topic, so more appropriate the journal could be for publishing.

Finally, the users could read a text with the flow of the evaluation:

Evaluation Process

To begin with the evaluation, a list of preloaded target articles are shown grouped in four groups. The user is asked to evaluate several of them from each group (at least one from each category). After clicking in the corresponding target article link (magnifying glass), a first page is shown containing the information of the target article (title, abstract, keywords and concepts) and the recommendation in the form of a ranking of 10 journals. This is not part of the explanation, only the recommendation itself. The user should click on a link presented in this page for the explanation. After this action, the global explanations for that target article and their 10 recommended journals, i.e., EE1 and EE2, are shown. In the EE1 ranking, the user may click on the link of each recommended journal for receiving its local explanation, giving way to explanations EE3, EE4 and EE5 (specific of that journal).

Once the evaluation of a target article is finished, a questionnaire is requested to be filled, containing specific questions related to this last evaluation (link found at the right top corner of the web page of the explanations). Please, be aware that this questionnaire must be completed for each target article selected.

This process could be repeated for each target article as many times as desired by the user.

After iterating this process for several target articles, and interacted with the recommendations and their explanations, a final questionnaire with questions related to the whole evaluation process is asked to be filled, which is found at the left main menu.

At this point the evaluation is finished.

When engaging with the RS and its explanations, to serve as a reminder and aid in understanding the significance of each EE, accompanying descriptions were provided. These descriptions can be found in the captions of Figs. 2, 3, 4, and 5, corresponding to Sect. 4 of the paper.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Afchar, D., Melchiorre, A., Schedl, M., Hennequin, R., Epure, E., Moussallam, M.: Explainability in music recommender systems. AI Mag. **43**, 190–208 (2022)

Albusac, C., de Campos, L.M., Fernández-Luna, J.M., Huete, J.F.: PMSC-UGR: A test collection for expert recommendation based on PubMed and Scopus. In: Advances in Artificial Intelligence. CAEPIA 2018, LNAI 11160, pp. 34–43 (2018)

Aletras, N., Baldwin, T., Lau, J.H., Stevenson, M.: Evaluating topic representations for exploring document collections. J. Am. Soc. Inf. Sci. **68**, 154–167 (2017)

Barredo-Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020)

Bilgic, M., Mooney, R.: Explaining recommendations: Satisfaction vs. promotion. In: Proceedings of the Beyond Personalization Workshop in Conjunction with International Conference on Intelligent User Interfaces (IUI'05) (2015)

Cardoso, B., Sedrakyan, G., Gutiérrez, F., Parra, D., Brusilovsky, P., Verbert, K.: IntersectionExplorer, a multi-perspective approach for exploring recommendations. Int. J. Hum. Comput. Stud. **121**, 73–92 (2019)

Carenini, G., Moore, J.: An empirical study of the influence of argument conciseness on argument effectiveness. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, pp. 150–157 (2000)

Chen, W., Hsu, W., Lee, M.L.: Tagcloud-based explanation with feedback for recommender systems. In: Proceedings of SIGIR Conference (2013)

Chi, J., Ouyang, J., Li, C., Dong, X., Li, X., Wang, X.: Topic representation: finding more representative words in topic models. Pattern Recogn. Lett. **123**, 53–60 (2019)

Daher, J.B., Brun, A., Boyer, A.: A review on explanations in recommender systems. Hal Open Science. https://hal.science/hal-01836639 (2017)

de Campos, L.M., Fernández-Luna, J.M., Huete, J.F.: Committee-based profiles for politician finding. Int. J. Uncertain. Fuzziness Knowl. Based Syst. **25**(Suppl. 2), 21–36 (2017)

de Campos, L.M., Fernández-Luna, J.M., Huete, J.F.: On the selection of the correct number of terms for profile construction: theoretical and empirical analysis. Inf. Sci. **430–431**, 142–162 (2018)

de Campos, L.M., Fernández-Luna, J.M., Huete, J.F., Redondo-Expósito, L.: LDA-based term profiles for expert finding in a political setting. J. Intell. Inf. Syst. **56**(3), 529–559 (2021)

de Campos, L.M., Fernández-Luna, J.M., Huete, J.F.: Publication venue recommendation using profiles based on clustering. IEEE Access **10**, 106886–106896 (2022)

de Campos, L.M., Fernández-Luna, J.M., Huete, J.F.: Use of topical and temporal profiles and their hybridisation for content-based recommendation. User Model. User-Adap. Interact. **33**(4), 911–937 (2023)

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., Riedl, M.O.: Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 263–274 (2019)

Felfernig, A., Tintarev, N., Tran, T., Stettinger, M.: Designing explanations for group recommender systems. arXiv:2102.12413 (2021)

Ferwerda, B., Swelsen, K., Yang, E.: Explaining content-based recommendations. https://www.bruceferwerda.com/docs/Ferwerda_DigitalTVGuide.pdf (2018)

Gedikli, F., Ge, M., Jannach, D.: Understanding recommendations by reading the clouds. In: Lecture Notes in Business Information Processing, vol. 85. Springer (2011)

Gedikli, F., Jannach, D., Ge, M.: How should I explain? A comparison of different explanation types for recommender systems. Int. J. Hum. Comput. Stud. **72**, 367–382 (2014)

Hernandez-Bocanegra, D.C., Ziegler, J.: Explaining review-based recommendations: effects of profile transparency, presentation style and user characteristics. i-com **19**(3), 181–200 (2020)

Hernandez-Bocanegra, D.C., Donkers, T., Ziegler, J.: Effects of argumentative explanation types on the perception of review-based recommendations. In: Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct) (2020)

Iferroudjene, M., Lonjarret, C., Robardet, C., Plantevit, M., Atzmueller, M.: Methods for explaining Top-N recommendations through subgroup discovery. Data Min. Knowl. Disc. **37**, 833–872 (2023)

Jannach, D., Jugovac, M., Nunes, I.: Explanations and user control in recommender systems. In: Personalized Human–Computer Interaction, pp. 133–158. De Gruyter Oldenbourg (2019)

Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., Pedro, Gama, J.: How can I choose an explainer? An application-grounded evaluation of post-hoc explanations. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 805-815 (2021)

Li, L., Chen, L., Dong, R.: CAESAR: context-aware explanation based on supervised attention for service recommendations. J. Intell. Inf. Syst. **57**, 147–170 (2021)

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., Zhou, B.: Trustworthy AI: from principles to practices. ACM Comput. Surv. **55**(9), 177 (2023)

Louki, P., Schaffer, J., Pjara, J., O'Donovan, J., Getoor, L.: Generating and understanding personalized explanations in hybrid recommender systems. ACM Trans. Interact. Intell. Syst. **10**(4), 31 (2020)

Lully, V., Laublet, P., Stankovic, M., Radulovic, F.: Enhancing explanations in recommender systems with knowledge graphs. Procedia Comput. Sci. **137**, 211–222 (2018)

Millecamp, M., Naveed, S., Verbert, K., Ziegler, J.: To explain or not to explain: the effects of personal characteristics when explaining feature-based recommendations in different domains. In: IntRS '19: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (2019)

Minh, D., Wang, H.X., Li, Y.F., Nguyen, T.N.: Explainable artificial intelligence: a comprehensive review. Artif. Intell. Rev. **55**, 3503–3568 (2022)

Nunes, I., Jannach, D.: A systematic review and taxonomy of explanation in decision support and recommender systems. User Model. User-Adap. Interact. **27**(3–5), 393–444 (2017)

Papadimitriou, A., Symeonidis, P., Manolopoulos, Y.: A generalized taxonomy of explanation styles for traditional and social recommender systems. Data Min. Knowl. Disc. **24**, 555–583 (2012)

Pavitha, N., Pungliya, V., Raut, A., Bhonsle, R., Purohit, A., Patel, A., Shashidhar, R.: Movie recommendation and sentiment analysis using machine learning. Glob. Transit. Proc. **3**(1), 279–284 (2022)

Pérez-Núñez, P., Díez, J., Bahamonde, A., Luaces, O.: Text-based recommender system with explanatory capabilities. ResearchSquare. https://doi.org/10.21203/rs.3.rs-1536768/v2 (2022)

Polleti, G.P., Cozman, F.G.: Explaining content-based recommendations with topic models. In: 8th Brazilian Conference on Intelligent Systems, pp. 800–805 (2019)

Radensky, M., Downey, D., Lo, K., Popovic, Z., Weld, D.: Exploring the role of local and global explanations in recommender systems. In: CHI's Extended Abstract. arXiv:2109.13301 (2022)

Rana, A., D'Addio, R.M., Manzato, M.G., Bridge, D.: Extended recommendation-by-explanation. User Model. User-Adap. Interact. **32**, 91–131 (2022)

Sato, M., Nagatani, K., Sonoda, T., Zhang, Q., Ohkuma, T.: Context style explanation for recommender systems. J. Inf. Process. **27**, 720–729 (2019)

Shmaryahu, D., Shami, G., Shapira, B.: Post-hoc explanations for complex model recommendations using simple methods. In: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems. CEURS, 2682:paper3 (2020)

Shulner-Tal, A., Kuflik, T., Kliger, D.: Fairness, explainability and in-between: understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system. Ethics Inf. Technol. **24**, 2 (2022)

Sullivan, E., Bountouridis, D., Harambam, J., Najafian, S., Loecherbach, F., Makhortykh, M., Kelen, D., Wilkinson, D., Graus, D., Tintarev, N.: Reading news with a purpose: explaining user profiles for self-actualization. In: Proceedings of 27th Conference on User Modeling, Adaptation and Personalization Adjunct (UMAP' 19 Adjunct) (2019)

Tintarev, N., Masthoff, J.: A survey of explanations in recommender systems. In: IEEE 23rd International Conference on Data Engineering Workshop, Istanbul, Turkey, 2007, pp. 801–810 (2007)

Tintarev, N., Masthoff, J.: Evaluating the effectiveness of explanations for recommender systems. User Model. User-Adap. Interact. **22**, 399–439 (2012)

Tintarev, N., Masthoff, J.: Explaining recommendations: design and evaluation. In: Recommender Systems Handbook, chapter 10, pp. 353–382. Springer (2015)

Tsai, C., Brusilovsky, P.: Explaining recommendation in an interactive hybrid social recommender. In: IUI '19: Proceedings of the 24th International Conference on Intelligent User Interfaces March, pp. 391–396 (2019)

Verbert, K., Parra, D., Brusilovsky, P., Duval, E.: Visualizing recommendations to support exploration, transparency and controllability. In: Proceedings of the 2013 International Conference on Intelligent User Interfaces (2013)

Vig, J., Sen, S., Riedl, J.: Tagsplanations: explaining recommendations using tags. In: Proceedings of Intelligent Users Interface Conference. pp. 47–56 (2009)

Vultureanu-Albisi, A., Badica, C.: A survey on the effects of adding explanation to recommender systems. Concurr. Comput. **34**(20), e6834 (2022)

Yao, Y., Wang, C., Li, H.: Counterfactually evaluating explanations in recommender systems. arXiv:2203.01310 (2022)

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

Zhang, Z., Chen, L., Jiang, T., Li, Y., Li, L.: Effects of feature-based explanation and its output modality on user satisfaction with service recommender systems. Front. Big Data **5**, 897381 (2022)

Luis M. De Campos received the B.Sc. degree in mathematics, in 1984. He completed his Ph.D. Thesis in 1988, researching on fuzzy measures and integrals and became a Lecturer in computer science at the University of Granada, Spain, in 1991. He is currently a Professor with the Department of Computer Science and Artificial Intelligence, University of Granada. His current research interests include information retrieval, recommender systems, Bayesian networks, and machine learning.

Juan M. Fernández-Luna received the degree in computer science from the University of Granada, Spain, in 1994, and the Ph.D. degree from this same university in 2001, working on a thesis in which several retrieval models based on Bayesian networks for information retrieval were designed. He is currently a Professor at the Computer Science Department, University of Granada. His main research interests include XML retrieval, working in collaboration with Juan F. Huete and Luis M. de Campos in XML personalization, collaborative IR, recommender systems and learning to rank, areas in which they have published papers in prestigious journals and international conferences and edited special issues.

Juan F. Huete received the Ph.D. degree in 1995, researching on the uncertainty treatment in artificial intelligence under the formalism of Bayesian networks. He is currently a Professor at the Department of Computer Science and Artificial Intelligence, University of Granada. His research interests include information retrieval, and designing retrieval models based on these graphical models. He is also working in the recommender system field and other fields like collaborative IR or learning to rank. He has been a co-editor of special issues about Bayesian networks and information retrieval, teaching and learning IR, and personalization.