Journal of Information and Data Management, 2026, 17:1, doi: 10.5753/jidm.2026.7427
This work is licensed under a Creative Commons Attribution 4.0 International License.

A Multi-Domain Scientific Paper Recommendation Dataset with Explicit Relevance Labeling

Ricardo A. Nascimento [**Instituto Federal de Goiás**| ***ricardomarcal02@hotmail.com***]
João F. dos Santos [**Instituto Federal de Goiás**| ***joaovitor.felipesantos@gmail.com***]
Adriano M. Camargo [**Instituto Federal de Goiás**| ***adrianocesar321@gmail.com***]
Sergio C. Canuto [**Instituto Federal de Goiás**| ***sergio.canuto@ifg.edu.br***]
Gustavo A. Costa [**Instituto Federal de Goiás**| ***gustavo.costa@ifg.edu.br***]
Daniel X. de Sousa [**Instituto Federal de Goiás**| ***daniel.sousa@ifg.edu.br***]

Federal Institute of Goiás, Campus Anápolis, Av. Pedro Ludovico, s/n, Residencial Reny Cury. 75131-457. Anápolis, GO, Brazil.

Received: 20 January 2026 • **Published:** 27 November 2026

Abstract

Scientific Paper Recommendation (SPRec) is hindered by a scarcity of public datasets that feature diverse research areas and explicit user feedback. Most benchmarks rely on implicit signals like authorship, which often fail to capture the evolving interests of researchers. Thusfore, this work introduces *Rec4SciBR*, a novel public dataset that addresses these gaps by providing rich textual descriptions of papers and projects across multiple domains, including explicit relevance labels from approximately two thousand researchers. Our study provides three primary contributions: (i) the release of *Rec4SciBR* as a representative, up-to-date resource for the SPRec community; (ii) a comprehensive benchmark comparing *Rec4SciBR* against established collections like DBLP and CiteULike across collaborative, content-based, and hybrid paradigms; and (iii) a critical evaluation of labeling strategies, demonstrating that relying solely on implicit labeling can severely degrade recommendation quality. Furthermore, we explore In-Context Learning (ICL) with Large Language Models, showing that models like GPT-4o-mini and Gemini 1.5 Flash can match traditional collaborative filtering performance without parameter fine-tuning. Our results highlight the importance of explicit feedback and rich content descriptions in building robust, personalized recommendation systems, positioning *Rec4SciBR* as a significant advancement for the SPRec ecosystem.

Keywords: Proceedings, Template, BCS, Content Repository, Indexing

1 Introduction

Recommender Systems (RecSys) have become a crucial tool across several domains, from entertainment to scientific research Wang and Yin [2025]; Wang *et al*. [2024]; Kanwal and Amjad [2024]. In the context of disclosing scientific knowledge and addressing the challenge of finding relevant papers in a huge volume of academic works, Scientific Paper Recommendation (SPRec) plays an essential role Bai *et al*. [2019]; Wang *et al*. [2024]; Ben Sassi [2025]; Flicke *et al*. [2025]. These systems primarily help researchers in identifying new works that align with their thematic interests and areas of expertise.

Despite the importance of these systems, there is a significant lack of publicly available and easily accessible datasets suitable for training SPRec models [Kreutz and Schenkel, 2022]. In practice, many datasets are restricted or private, or no longer accessible. Moreover, even the datasets that are available face important limitations. For instance, there is the issue of the knowledge areas diversity, where academic production showcases domain-specific behavioral patterns Montazerian *et al*. [2020]; Flicke *et al*. [2025]. Datasets constrained to specific areas (such as DBLP Sugiyama and Kan [2010] or PubMed[1]) fail to capture the diversity of researchers, necessarily introducing content bias. Other datasets remain outdated (such as CiteULike Wang *et al*. [2013]), and because they have been discontinued, they no longer reflect the most recent behavioral patterns of researchers, which may evolve for many reasons, including the changes in the technologies or even political conditions related to research funding Koltun and Hafner [2021].

Finally, the shortage of public SPRec's datasets that include explicit labels provided by the researchers themselves also represents a crucial limitation, both for parameter tuning and for reliable evaluation of effectiveness Kreutz and Schenkel [2022]. In practice, throughout an academic career, researchers' interests may change over time, creating a gap between the topics of their early on publications and those of their more recent works, a common phenomenon among senior researchers Liang *et al*. [2015]. In the same vein, there is the case of recent graduated PhDs, where the limited number of publications offers little information about their research interests.

In this context, indirect or implicit labeling of paper relevance – such as considering as relevant only the author's own articles Li *et al*. [2021] – may fail to capture a researcher's true interests, incorrectly assuming that these are restricted to the areas covered by their current publications. Likewise, approaches based on references Xie *et al*. [2021] or

[1] https://pubmed.ncbi.nlm.nih.gov/

citations Zhang *et al*. [2019] are also subject to this same bias. Ideally, such biases inherent to implicit labeling strategies can be mitigated through explicit labeling, in which researchers manually select the publications they consider most relevant to their research interests.

Given these limitations, this work mitigates such gaps by introducing a new dataset for scientific papers and projects recommendation systems, called *Rec4SciBR*. The *Rec4SciBR* dataset includes explicit relevant labels for papers and projects description, providing approximately two thousand of researchers, covering multiple areas of knowledge and incorporating up-to-date evaluations and content. The proposed dataset organizes scientific project descriptions, papers, and researcher profiles with information on authorship, coauthorship, including the explicit relevance labeling. All the information were characterized by considering the volume of interactions between users and scientific papers/projects, as well as the diversity of the descriptions of those items. This characterization is extended, as part of a comparative effort to the DBLP [Sugiyama and Kan, 2010] and CiteULike [Wang *et al*., 2013] collections.

In particular, by providing the scientific project data, *Rec4SciBR* can be used for evaluating other tasks, such as systems responsible for matching submissions to reviewers in peer-review processes, thereby increasing the alignment between reviewers' expertise and assigned works.

We also carefully evaluated the effectiveness of these datasets against state-of-the-art recommendation strategies, including collaborative filtering, content-based, and hybrid paradigms, as well as In-Context Learning (ICL) with Large Language Models (LLMs). Our results indicate that *Rec4SciBR* achieved superior performance with content-based filtering approaches, primarily due to the rich descriptions of objects, researchers, and papers. Furthermore, owing to these textual descriptions, *Rec4SciBR* also yielded impressive results using LLMs with ICL strategies, performing similarly to traditional collaborative filtering methods.

As the main claim of our dataset is the importance of explicit labeling, we also provide a comparison between explicit and implicit labels content, showing how far the researchers' interest change when considering implicit labeling. In particular, we demonstrate this comparison over similarity distance and content-filtering models, highlighting more clearly the advantages of explicit labeling in recommendation datasets.

This work is a follow-up to our preliminary introduction of *Rec4SciBR* [dos Santos *et al*., 2025]. In this version, we extend our research by providing several significant improvements: (i) a more detailed explanation and comparison of implicit versus explicit labels for scientific paper recommendation; (ii) an exploration of In-Context Learning (ICL) for recommendation tasks using Large Language Models (LLMs); and (iii) a more comprehensive analysis of the results initially presented in dos Santos *et al*. [2025]. Based on these new findings, we demonstrate that i) relying solely on implicit labeling can severely damage the quality of paper recommendations, and ii) we show that ICL with LLMs (which avoids the need for parameter fine-tuning) yields comparable performance with collaborative-filtering when using GPT-4o-mini and Gemini 1.5 Flash models.

In summary, in this work we provide three novel contributions:

- A new public dataset, allowing the scientific community to leverage its extensive content descriptions to support the development and evaluation of new SPRec models;
- A comprehensive comparison between three distinct and prominent scientific paper recommendation datasets—DBLP [Sugiyama and Kan, 2010], CiteULike [Wang *et al*., 2013], and *Rec4SciBR*—including effectiveness evaluation of content-based, collaborative filtering, hybrid, and ICL-based strategies;
- A thorough evaluation of labeling strategies, specifically examining the impact of explicit versus implicit labels on recommendation performance.

The remainder of this work is organized as follows. In Section 2, we provide a general overview of the context of Scientific Article Recommender Systems and existing datasets. In Section 3, we present a characterization of the proposed dataset, and in Section 4, a comparison using recommender system models. Finally, in Section 5, we present our conclusions.

2 Background and Related Works

The task in Scientific Paper Recommendation (SPRec) can be described as recommending scientific publications to a given researcher, exhibiting the idiosyncrasy of the academic environment Bulut *et al*. [2019]; Guo *et al*. [2020]. In particular, as highlighted by Alzoghbi *et al*. [2015], algorithms based on the textual content of users/items -— known as Content-Based Filtering Systems [Bai *et al*., 2019] —- are able to leverage the abundant textual content of papers by measuring textual similarity between items and researcher profiles. On the other hand, models based on exploring the similarity between researchers' preferences -— known as Collaborative Filtering Systems [Wang *et al*., 2022] —- may be compromised due to the scarcity or sparsity of data concerning researchers' preference evaluations.

In different contexts, such as movie and music recommendation systems, explicit user preference ratings play a central role in the development of models and in the evaluation of effectiveness, especially in hybrid models that combine Content-Based Filtering and Collaborative Filtering [Li and Zou, 2019]. However, most datasets for scientific paper recommendation lack records that establish this direct correlation between researchers and publications in terms of explicit ratings or relevance Kreutz and Schenkel [2022].

One of the few efforts to build public datasets with explicit relevance labels was the creation of the SPRD (*Scholarly Paper Recommendation Dataset*) [Sugiyama and Kan, 2010], based on the explicit labeling of papers of interest by 50 researchers in the field of Computer Science, which contains labels from 13 senior researchers. However, in addition to the limited number of researchers, access to this dataset is currently unavailable. Likewise, several correlated collections such as BibSonomy [Beel *et al*., 2014], DOCEAR [Beel *et al*., 2014], and PRSDataset [Guo *et al*., 2020] are also not accessible at the time of this research.

Among the datasets that are available, CiteULike [Wang *et al*., 2013] stands out as one of the most widely explored in the literature. Derived from the platform of the same name, which functioned as an academic social network aimed at discovering, organizing, and sharing bibliographic references, the dataset offered functionalities such as automatic tag recommendation, usage statistics, and personalized article suggestions. Despite the platform's discontinuation, data repositories reflecting the interactions of 5,551 users with articles published between 2004 and 2006 have been made publicly available[2].

Another widely used dataset in the field of Computer Science is DBLP [Ley, 2002]. Originally conceived as a Web experiment, DBLP evolved into a large-scale repository that currently contains more than 7 million publications and about 3 million authors, covering approximately 6,900 conferences and journals. The dataset organizes and provides detailed bibliographic metadata, including information about authorship, coauthorship, titles, venues, journals, publication dates, research areas, and the evolution of scientific communities. Due to its extensive availability of metadata, several works adopt DBLP for the development and evaluation of SPRec models [Ley, 2002].Due to the lack of explicit labeling data, the models evaluate this dataset considering implicit relevance labels.

On overall, the main datasets used in SPRec exhibit significant limitations. Outdated datasets, such as CiteULike that do not capture recent changes in researchers' interests. DBLP, although widely used, lacks diversity across knowledge areas and does not provide explicit relevance labels. Similar issues are observed in datasets such as OpenAlex, Web of Science, Scopus, and PubMed [Haupka *et al*., 2024], which, despite offering broad coverage of scientific publications and bibliographic metadata, do not include explicit relevance labels or access to personalized data (such as evaluation history, project descriptions, and researcher profiles). This absence makes it difficult to apply supervised and personalized recommendation techniques.

For those datasets without explicit labeling, the damage can be even more significant. The increasing emphasis on capturing recent user interests has become a central challenge in RecSys [Bao *et al*., 2025; Zhao *et al*., 2024], particularly in domains characterized by rapid knowledge expansion, such as scientific paper recommendation. For instance, the authors of Bao *et al*. [2025] argue that recent interactions play a decisive role in shaping high-quality recommendations, since long-term preference profiles often fail to represent a user's immediate research focus. In the context of SPRec scenario, implicit labeling can impose a bias by capturing long-term preferences and failing dramatically to reflect more recent interests (as we discussed in Section 4). Furthermore, to better address recent-interest adaptation, modern RecSys methods rely on dataset content to avoid additional training or fine-tuning and leveraging the ability of LLMs to learn directly from context, task know as In-Context Learning [Zhao *et al*., 2024].

Therefore, in order to provide an evaluation of recent interest for scientific researchers, the capability to explore In-Context Learning for SPRec, and given the shortage of accessible datasets for scientific paper recommendation tasks, we highlight the contribution of this work through the release and evaluation of the *Rec4SciBR* dataset.

3 *Rec4SciBR* – New Dataset for Paper Recommendation

In this section, we characterize the *Rec4SciBR* dataset, proposed for Scientific Paper Recommendation (SPRec) evaluation. We outline how it was built, its characteristics, limitations, and comparison with other datasets.

3.1 *Rec4SciBR* Construction Methodology

Rec4SciBR was built from a real-world scenario involving the recommendation of academic papers and scientific projects to more than 2,000 researchers and faculty members from a Brazilian institution. Through an extraction process, descriptions of articles and projects were extracted from the Lattes Platform[3]. Lattes Platform is a Brazilian curriculum repository widely used by nearly all research and educational institutions in Brazil. Thus, whenever a researcher publishes a new paper, they usually register the publication on the platform to be accessed by institution.

Thus, following the standard Extract–Transform–Load (ETL) pipeline, we begin with the Extract phase, in which each XML (Extensible Markup Language) file contains a Curriculum Lattes records for a researcher. Each file includes structured information about publications, projects, and other academic activities. These elements are converted into an intermediate structured representation and stored in a JavaScript Object Notation (JSON) file.

In the Transform phase, the extracted textual data undergoes several preprocessing steps. First, stop-words of English, Portuguese and Spanish language are removed. Next, a general-word exclusion step is performed to remove commonly occurring academic terms that appear across many curricula but provide limited discriminatory power, e.g., "professor," "federal," and "institution." After this filtering process, the cleaned corpus is processed using TF–IDF weighting and different encoder models, better described in in Section 5.

In the Load phase, we build to the researchers a TF–IDF textual representation by concatenating the textual content extracted from their Curriculum Lattes entries. For scientific papers and research projects, the representation considers the title and the available descriptions. The resulting TF–IDF vectors are stored in a structured dataset that can be used for downstream tasks such as recommendation models.

To generate initial recommendations, we compute TF–IDF–based content similarity between users and items. In other words, researchers receive recommendations based on the similarity between the most relevant terms in their curricula (including publications, projects, and other academic information) and the textual descriptions of scientific works

[2] `https://github.com/js05212/citeulike-a`

[3] https://lattes.cnpq.br/

Table 1. Statistics of users and items .

Dataset	Users	Items	Interactions	Sparsity	Knowledge area
DBLP(AAN)	51.645	30.000	76.699	0,995	Computational Linguistics
CiteULike	5.551	16.980	204.986	0,783	Several
Rec4SciBR	2.261	71.895	19.211	0,823	Several

from other researchers. After receiving the recommendations, researchers are invited to evaluate each suggested item by providing explicit feedback on the recommended scientific articles and projects. Each recommendation is rated by the researcher on a 0–5 scale.

In addition to the labels, the dataset also includes rich textual features for users, papers, and scientific projects, extracted from publicly available information in Lattes CVs. We highlight that the dataset contains **no** personal or identifying information about researchers -— such as name, address, or contact details.

To align the presentation of the proposed dataset with common practices in Recommender Systems, we describe its structure from two complementary perspectives: collaborative filtering and content-based recommendation. In the collaborative filtering perspective, user–item interactions constitute the primary information for inferring user preferences. Algorithms in this class leverage various strategies to learn effective representations of these relationship patterns. In the content-based perspective, descriptive attributes of users and items play a crucial role. The objective is to construct high-quality representations that enable more accurate similarity comparisons between users and items.

3.2 Characterization of the Interactions between Users and Items

Table 1 depicts statistics of users, items, and user–item interactions of the datasets analyzed. In case of DBLP, due to its large size, we include only a commonly used subset (denoted *DBLP(AAN)*), which has been employed in prior works [Sugiyama and Kan, 2010; Bai *et al*., 2019]. The DBLP dataset represents a category of datasets derived from large scholarly repositories (such as OpenAlex, Web of Science, Scopus, and PubMed [Haupka *et al*., 2024]), but which do not contain explicit preference labels.[4]

Table 1 reports the number of users, items (scientific papers and projects, in the case of *Rec4SciBR*), interactions, sparsity, and coverage of knowledge areas. The interactions represent the number of relationships between users and items. For *Rec4SciBR* and CiteULike, these connections correspond to explicit ratings. For *DBLP(AAN)*, the connections reflect authorship relations – or implicit evaluation. Sparsity indicates the proportion of missing user–item interactions. The *Rec4SciBR* dataset contains 2,261 users and more than 70,000 publications, with a sparsity of 0.823.

Table 2 presents a user- and item-level analysis. In this case, the *Rec4SciBR* dataset contains an average of 6.75 interation per user (with a standard deviation of 11.17), with a minimum of 1 and a maximum of 178 iteration with different items. Regarding publications, 7,710 items received between 1 and 9 ratings, with an average of 1.2 ratings per publication and a standard deviation of 0.60. As shown in Table 2, the *DBLP(AAN)* dataset contains more than 51,000 users with connections (authorship relations).

Considering the distribution of explicit ratings in the *Rec4SciBR* dataset, we find a strong predominance of high scores: 17,622 ratings (91.73%) fall between 4.0 and 5.0, 814 ratings (4.23%) fall between 2.0 and 4.0, and only 775 ratings (4.04%) fall between 0.0 and 2.0. This indicates that the dataset provides a representative signal of user interests, highlighting its quality given the large volume of relevant labels.

Overall, we observe that the number of connections between users and items (Tables 1 and 2) in the *Rec4SciBR* dataset is relatively lower than in CiteULike. However, 1,408 users provided explicit ratings. These numbers reflect the real scenario of an institution with more than 2,000 researchers, making the dataset highly valuable for evaluating realistic recommendation models. Moreover, as we show in the following sections, the *Rec4SciBR* dataset offers a rich and more detailed descriptions of users and items.

3.3 Characterization – Description of users and items' content

Rec4SciBR provides the following information for scientific papers: venue, title, year of publication, language, type of production, and type of dissemination. Currently, the dataset includes over 31,532 journals among 57 distinct fields and 1,403 conferences. For academic projects content, the dataset provides the title and the description. For users, selected fields from the Lattes CV are included, such as the researcher's profile and area of expertise[5].

Regarding language distribution, 81.1% of the scientific publications are written in Portuguese and 17.3% in English. Academic projects are predominantly written in Portuguese. The remaining publications are in Spanish or other languages.

Table 3 details the characteristics of the two types of items considered in the recommendations: papers and projects, describing more than 65,000 papers and 6,000 research projects. In fact, we highlight that *Rec4SciBR* makes an important contribution by providing research projects along with explicit

[4]Although the *DBLP(AAN)* dataset is not directly comparable to *Rec4SciBR* due to the absence of explicit labels, we include it in our analysis because it is widely used in the literature, thereby increasing the completeness of this study.

[5]As we describe in next section, those field are crucial for ICL strategies.

Table 2. Summary of user's and items' interaction .

	Measure	*Rec4SciBR*	CiteULike	*DBLP(AAN)*
Relationships per users	Total quantity of users with relationship	1.408	5.551	51.645
	Min. relationships per user	1	10	1
	Max. relationships per user	178	403	233
	Average relatioships per user	6,75 (11,17)	36,90 (42,08)	1,49 (2,08)
Relationships per Item	Tota quantity of items with relationships	7.710	16.980	30.000
	Min. relationships per item	1	1	1
	Max. relationships per item	9	321	50
	Average relationships per item	1,23 (0,60)	12,07 (11,85)	2,56 (1,80)

Table 3. Statistics for Scientific Papers and Research Projects in *Rec4SciBR*.

	Measure	Quantity
Scientific Papers (SP)	Total of SP	65.210
	Average of SP per year	1552,61
	Oldest SP year	1.982
	Most recent SP year	2.024
Research // Projects (AP)	Total of AP	6.685
	Average project duration (years)	1,76
	Average number of characters in project descriptions	775,37

relevance assessments. This can be very helpful for systems that need to assign works to reviewers during peer-review evaluations, improving the likelihood that reviewers receive more closely aligned submissions.

A distinguishing feature of *Rec4SciBR* is the availability of textual content for both users and items (papers and projects). Table 4 provides a comparison across datasets, considering the number of terms used to describe papers and project titles. For this analysis, stop-words (in Portuguese and English) are removed from all datasets. As shown, *Rec4SciBR* contains a vocabulary nearly twice as large as those of *DBLP(AAN)* and CiteULike for both papers and projects -— approximately 31,000 and 34,000 terms, respectively. It also contains a greater average number of terms per item. These figures highlight the richness of the textual descriptions of our dataset and, as we will show in the next section, enable the construction of more informative representations for content-based models.

To better understand the textual content available for users, i.e. description of researcher profile and area of expertise, Table 5 presents statistics of area for *Rec4SciBR*. On average, each user has 60 words in their textual description, and the final dictionary has 62,138 unique terms. We do not include comparisons with the other datasets, as they do not provide textual descriptions for users.

Moreover, to better understand the user's content and item descriptions, we computed the most important terms using the TF–IDF representation. In Figure 1a, we identify the most frequent terms for groups of users. For example, the term "Education" is the most frequent term in the CV content of 275 researchers. Following that, administration and engineering appear as the second and third most frequent terms, respectively. In summary, the figure displays the 30 most frequent knowledge areas in the dataset, highlighting the diversity of expertise among the researchers.

A similar analysis is shown in Figure 1b, which presents the most frequent terms extracted from the names of journals cited in the Lattes CVs, which include terms such as Medicine, Interdisciplinary, and Engineering. Figure 1b further confirms the broad variety of topics represented in the proposed dataset.

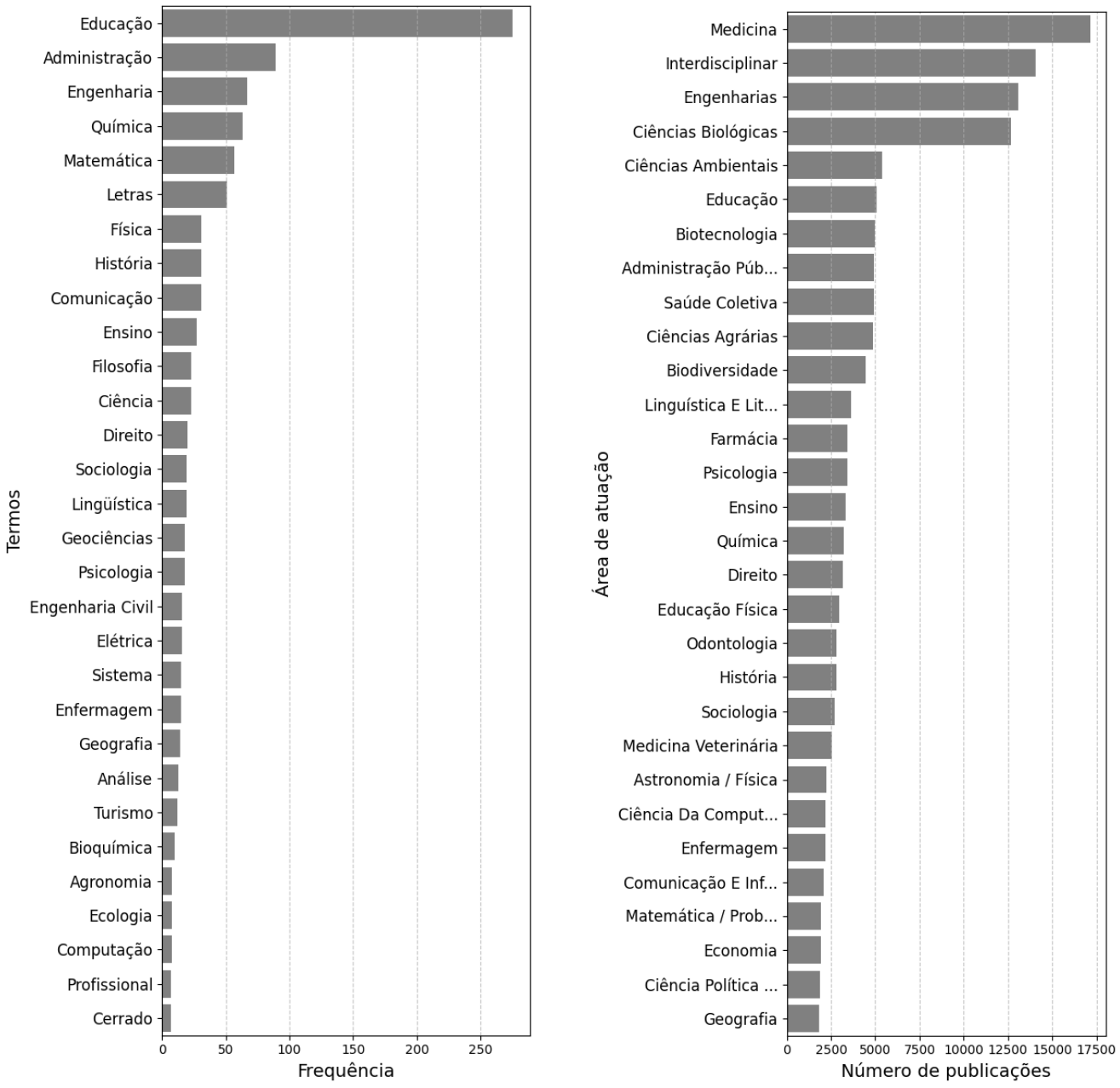
In summary, the *Rec4SciBR* dataset offers a rich description of both users and items, with strong potential for evaluating real-world Recommendation System models. This point is reinforced in the next section, where we describe its application on recommendation algorithms from literature, highlighting in particular the strong results obtained when leveraging the item and user representations.

4 Evaluations

We now present an in-depth analysis of *Rec4SciBR* from multiple perspectives, comparing it with other datasets as well as with algorithms commonly used in the recommender systems literature. This section is structured around three main objectives: i) analyzing the effectiveness of the datasets across different algorithms; ii) providing a detailed compar-

Table 4. Textual description of the items in the datasets.

Measure	***Rec4SciBR***		***DBLP(AAN)***	**CiteULike**
	Projects	Papers	Papers	Papers
Dictionary size	34.254	31.445	19.357	14.354
Average of words	58,7	8,64	6,8	6,1
Largest number of words	273	35	48	20



(a) Words most common for the user description – *Rec4SciBR*.

(b) Knowledge area most common for journals title – *Rec4SciBR*.

Figure 1. Analysis of most common terms for users and journals.

ison between implicit and explicit labeling on *Rec4SciBR*; and iii) showcasing the effectiveness of *Rec4SciBR* under an In-Context Learning (ICL) settings.

4.1 Methodology of Evaluation

We describe our results over a diverse variety of algorithms representing the main approaches in Recommender Systems, including: (i) Collaborative Filtering, with the models GraphRec [Rashed *et al.*, 2019], KGAT [Wang *et al.*, 2019], SR-GNN [Wu *et al.*, 2019][6], SASRec[Kang and McAuley, 2018][7], and BERT4Rec [Sun *et al.*, 2019][8]; (ii)

[6] https://github.com/CRIPAC-DIG/SR-GNN
[7] https://github.com/kang205/SASRec
[8] https://github.com/FeiSun/BERT4Rec

Table 5. Terms' Statistics regarding the users in *Rec4SciBR*.

Métrica	Valor
Dictionary size	22.138
Average of works per user	20
Largest size of words per user	1.287

Content-Based Filtering, using approaches such as TF-IDF[9] and BERT [Yang *et al*., 2020][10]; and (iii) Hybrid Filtering, represented by the KGAT model [Wang *et al*., 2019][11]. For all algorithms we are using the default parameter settings indicated in their respective original paper. This diversity of approaches aims to ensure a robust comparative analysis across different recommendation paradigms.

To ensure robustness and generalization of the results, 5-fold cross-validation was adopted, in which the dataset is divided into five subsets of users, and the evaluation is repeated five times, alternating the test subset and using the remaining ones for training, as is well established in the field [Said and Bellogín, 2014]. For the SR-GNN [Wu *et al*., 2019], SASRec [Kang and McAuley, 2018], and BERT4Rec [Sun *et al*., 2019] models -— since they are sequential algorithms -— we separated the two most recent papers associated with each user, assigning them to validation and testing, as described in Sun *et al*. [2019].

For the application of the BERT model, we used the SentenceBert (Multimodal) model[12] to construct new vector representations for users and items, and we compute similarity using the cosine function, following Yang *et al*. [2020].

The analysis of the models was conducted using metrics widely applied in the recommendation literature Avazpour *et al*. [2014], namely: Normalized Discounted Cumulative Gain (NDCG) and Mean Average Precision (MAP).

4.2 Collection and Processing Methodology

Regarding data pre-processing, a common step across all datasets was the removal of users with no interactions -— that is, users who neither rated nor produced any publications.

The datasets used by the SR-GNN [Wu *et al*., 2019] and SASRec [Kang and McAuley, 2018] models follow the treatment of the original papers: SR-GNN [Wu *et al*., 2019] discards user–item connections with fewer than five interactions, while SASRec removes users with fewer than three interactions. In contrast, the GraphRec [Rashed *et al*., 2019] and KGAT [Wang *et al*., 2019] models only restructure the data without altering its original content.

For the *Rec4SciBR* dataset applied in SR-GNN [Wu *et al*., 2019], it is necessary to binarize the ratings (originally ranging from 0 to 5), treating scores > 3 as relevant, otherwise irrelevant. In order to keep our task similar to other datasets, we only use papers data, avoiding the scientific projects.

For evaluating algorithms that rely on a sequential approach (SR-GNN [Wu *et al*., 2019], SASRec [Kang and McAuley, 2018] e BERT4Rec [Sun *et al*., 2019]), we recorded the order in which the labels were assigned, ensuring the preservation of each user's sequential interaction pattern.

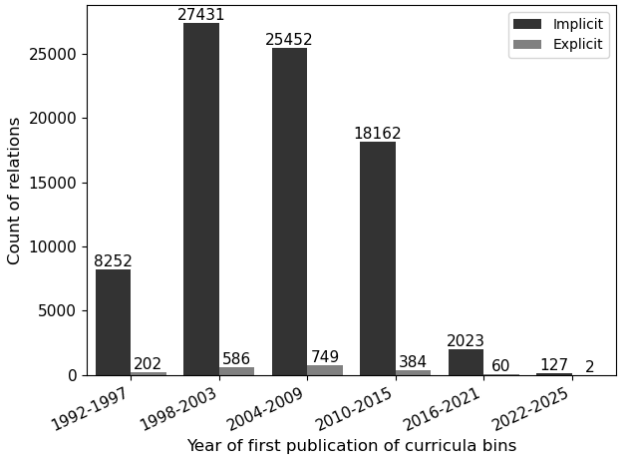


Figure 2. Counts of interaction by age (Implicit vs Explicit)

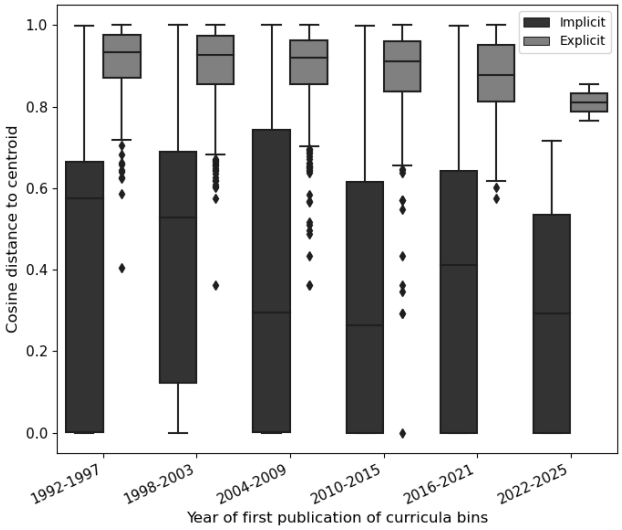


Figure 3. Distances by age bin (Implicit vs Explicit)

4.3 Effectiveness of the Experimental Results

We begin our analysis by showing the effectiveness of the datasets for the aforementioned algorithms. We use the original relevance relations of each dataset, that is, explicit and implicit labeling for both CiteULike and *Rec4SciBR* datasets, and implicit labeling by authorship for *DBLP(AAN)*. We also use the descriptions of each item and user as previously detailed.

The results are presented in Table 6. As expetected and analyzing the NDCG@10 metric for *Rec4SciBR*, algorithms that rely on content-based filtering (TF-IDF and BERT) leveraging user and paper representations, achieve a more effective results than algorithms based on collaborative filtering and hybrid approaches. This pattern also appears for

[9] `https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html`

[10] `https://tfhub.dev/google/universal-sentence-encoder-Multilingual/3`

[11] `https://github.com/xiangwang1223/knowledge_graph_attention_network`

[12] https://www.kaggle.com/models/google/universal-sentence-encoder/tensorFlow2/multilingual/2?tfhub-redirect=true

Dataset	Measure	Hybrid		Collaborative Filtering			Content Filtering	
		GraphRec	KGAT	SRGNN	SASRec	BERT4Rec	TF-IDF	BERT
Rec4SciBR	NDCG@10	0.647	0.738	0.235	0.292	0.121	**0.885**	0.766
	MAP	0.507	**0.645**	0.242	0.262	0.184	0.511	0.154
DBLP(AAN)	NDCG@10	**0.838**	0.824	0.546	0.343	0.187	0.512	0.507
	MAP	**0.779**	0.761	0.550	0.327	0.195	0.484	0.471
CiteULike	NDCG@10	0.785	0.843	0.150	0.731	0.422	**0.860**	0.609
	MAP	0.700	**0.788**	0.121	0.677	0.382	0.188	0.541

Table 6. Effective evaluation to *Rec4SciBR*, DBLP, and CiteULike datasets

CiteULike. In other words, the large number of unique terms (Table 4) seems to contribute to generating better representations among users and articles. On the other hand, the limited description of the interactions between users and items (Table 2) explains the inferior results of Collaborative Filtering and Hybrid models. We remark that these good results for content-based in SPRec is also observed in Dhanda and Verma [2016].

We also observed that, for the *Rec4SciBR* dataset, the TF-IDF algorithm performs better than BERT. This is likely due to the diversity of areas in the dataset, where representing terms as specific vectors may be more appropriate than the context-based analysis performed by BERT.

The results for the *DBLP(AAN)* dataset are also quite interesting. Although being sparse, the dataset contains a total number of interactions per user and per item (Table 2) greater rather than the other datasets. Consequently, this explains the better performance exploiting the user–item connections -— those based on Collaborative Filtering and Hybrid methods. At the same time, since the dataset offers limited descriptions of its objects, algorithms using content-based approaches yield lower results for *DBLP(AAN)* .

4.4 Explicit and Implicit Labeling

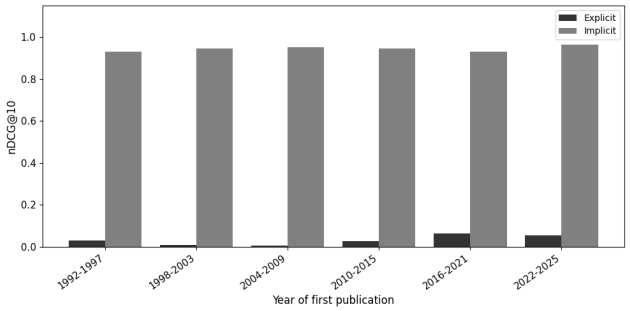


Figure 4. nDCG@10 evaluation for implicit training and explicit and implicit test set, using TF-IDF as ranking.

In our work, we claim that explicit labeling increases data quality by reducing two types of bias: i) a bias toward older papers, typically associated with senior researchers who may no longer be interested in early publications, and ii) a sparsity bias related to junior researchers, who are at the beginning of their careers and therefore have fewer published papers. Thus, we leverage the fact that *Rec4SciBR* contains both explicit and implicit labels to clearly support this claim,

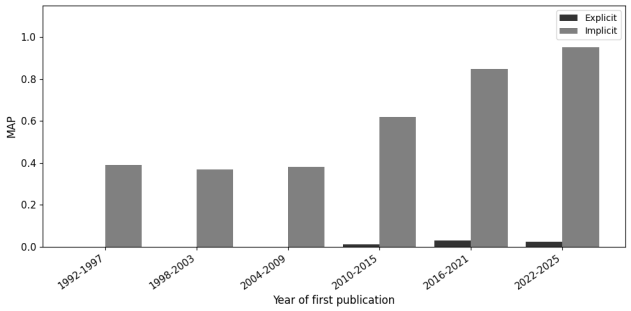


Figure 5. MAP evaluation for implicit training and explicit and implicit test set, using TF-IDF as ranking.

describing the differences between these two types of labels in terms of content and effectiveness.

We begin by presenting Figure 2, showing the distribution of explicit and implicit labels of *Rec4SciBR* when joining researchers by the year they obtained their PhD, highlighting senior and junior researchers. As can be observed, researchers from earlier years (e.g., 1998 and 2003) exhibit substantially more implicit data than those from more recent years (e.g., 2022 and 2025). As expected, the number of implicit labels is significantly higher than the number of explicit ones, since explicit labeling (approximately 2,000 labels in total) depends on active action from researchers. We also observe that junior researchers have fewer explicit labels, as this group is smaller in institution that we have obtained the data.

To further analyze the content differences between explicit and implicit labels, Figure 3 compares the similarity across the same PhD-year ranges shown in Figure 2. Thus, the figure presents box plots of cosine similarity values where explicit and implicit labels are compared against the same vector, a centroid computed solely from implicit labels. Overall, the figure highlights a substantial divergence between explicit and implicit similarities. For example, researchers from the 2004 and 2009 years show implicit label similarities clustered around 0.3 of cosine similarity, whereas explicit labels reach values above 0.9, indicating a substantial difference of user interests representation, and confirming that in many cases the most recent interest of researcher can be different from early publications.

To demonstrate the impact of these differences on content-based recommendation models, Figures 4 and 5, nDCG@10 and MAP, respectively, illustrate the effectiveness of TF-IDF-based representations. In this experiment, we use implicit labels as training data to model researchers' interests

Measure	GraphRec	KGAT	TF-IDF	Zero-shot		Few-shot	
				GPT 4o-mini	Gemini 2.5-FL	GPT 4o-mini	Gemini 2.5-FL
nDCG@10	0.589 [0.57, 0.61]	0.672 [0.66, 0.69]	0.887 [0.87, 0.91]	0.606 [0.56, 0.65]	0.627 [0.59, 0.67]	0.686 [0.65, 0.72]	0.663 [0.63, 0.70]
MAP	0.463 [0.44, 0.48]	0.556 [0.53, 0.58]	0.539 [0.49, 0.58]	0.526 [0.48, 0.57]	0.558 [0.51, 0.60]	0.590 [0.55, 0.63]	0.572 [0.53, 0.61]

Table 7. Effective evaluation comparing baselines to LLM Recommendation.

Zero-shot Prompt
How will user with description "*{description}*" rate this paper: "*{title}*"? (1 being lowest and 10 being highest). Attention! Just give me back the exact number as a result, and you don't need a lot of text.

(a) Zero-shot prompt configuration, considering only the user description and the candidate paper title.

Few-shot Prompt
Here is user rating history: *{History of 3 interactions, each in the format: "Title, Rating;"}* Based on above rating history, how will user with description "*{description}*" rate this paper: "*{title}*"? (1 being lowest and 10 being highest). Attention! Just give me back the exact number as a result, and you don't need a lot of text.

(b) Few-shot prompt configuration, incorporating historical user-item interactions and candidate paper title.

Table 8. Prompts designed for LLM-based Recommendation.

and evaluate two test sets: explicit labels and the remaining implicit labels. As expected, implicit data exhibit higher similarity to the training set than explicit labels. Consequently, we observe that recommendation models relying solely on item content are strongly biased when trained and evaluated on implicit data, which can lead to an overestimation of model performance.

4.5 Evaluation ICL strategies with *Rec4SciBR*

An important requirement in current recommendation tasks is the ability to capture users' recent interests Bao *et al*. [2025]; Zhao *et al*. [2024]. In this context, In-Context Learning (ICL) provides an effective strategy by enabling LLMs to adapt at inference time using only the user's most recent information as contextual examples, without requiring retraining or fine-tuning processing. However, effectiveness using ICL relies on providing relevant contextual information that allows LLM to perform better predictions. Therefore, in this section, we take advantage of users and items content in *Rec4SciBR* to evaluate the effectiveness in LLM predictions.

With this goal, we follow the methodology described in Liu *et al*. [2023], where the authors evaluate ChatGPT's performance across several recommendation tasks, including rating prediction, sequential recommendation, direct recommendation, explanation generation, and review summarization. Among these tasks, rating prediction is the most closely aligned with the algorithms evaluated in our work, as it estimates the numerical score a user would assign to an unseen item based on past user–item interactions. The authors demonstrate that rating prediction achieves highly competitive results, providing strong evidence for the effectiveness of LLM-based recommendation strategies.

Thus, we design specific prompts for both zero-shot and few-shot settings, as illustrated in Figure 8. Our objective is to prompt the model to output a discrete rating in the range $[1, 5]$, given a user description and the title of a candidate item. In the few-shot setting, the prompt is augmented with a small set of in-context demonstrations that are distinct from the items used for evaluation. Specifically, for each user, three positively rated items are selected as in-context examples, while the remaining highly rated items are held out for testing, subject to a fixed evaluation budget. The LLM-generated ratings are then used to induce a ranking over the held-out positive items and 49 randomly sampled negative candidate items.

The effectiveness results in terms of NDCG@10 and MAP, along with their corresponding confidence intervals, are reported in Table 7. From these results, we observe that LLM-based approaches do not surpass the content-based (TF–IDF) baseline, although they achieve performance comparable to collaborative-filtering methods. For instance, GPT-4o-mini obtains statistically similar results to the KGAT method, with an NDCG@10 of approximately 0.686. In the zero-shot setting, Gemini exhibits the best performance, whereas in the few-shot setting, GPT-4o-mini achieves the highest results. Overall, these findings provide evidence that leveraging LLMs on the *Rec4SciBR* dataset can yield highly competitive recommendation performance.

5 Conclusion

This work presents the *Rec4SciBR* dataset, a new collection designed for the development and evaluation of Scientific Paper Recommendation Systems. The proposal aims to address important gaps observed in widely used datasets in the literature, such as the absence of explicit relevance labels and limited thematic diversity.

Characterization of the dataset shows that *Rec4SciBR* is promising for evaluating recommendation models within the

factual context of research institutions. The experimental results demonstrate that, due to the rich textual descriptions of users and items, content-based models achieved impressive outcomes; specifically, TF-IDF and BERT reached superior performance, highlighting the dataset's value for practical applications. Our results also emphasize the importance of explicit labeling over implicit labeling, illustrating how researcher interests evolve over time. Finally, we demonstrate that by leveraging the textual content of *Rec4SciBR*, it is possible to achieve highly relevant results using ICL strategies, for instance, matching the effectiveness of collaborative filtering approaches.

Regarding the analysis of limitations, the diversity of knowledge areas poses significant challenges in identifying similarities and relevant patterns. Particularly when recommending content across highly distinct domains and assessing their representativeness of different quantities. To mitigate this issue, we will continue to expand the proposed dataset by including more researchers and additional article recommendations. Furthermore, we plan to explore models that weight the representation of each domain to better account for cross-domain variations.

In summary, *Rec4SciBR* represents a significant advancement for the SPRec ecosystem, offering a public, up-to-date dataset that reflects the diversity of scientific research. For future work, we plan to expand the dataset with new users and evaluations, as well as integrate richer semantic attributes and heterogeneous graphs to enhance hybrid and deep learning–based models.

Acknowledgements

This work is supported by the Instituto Federal de Educação, Ciência e Tecnologia de Goiás (IFG), by CNPq (443011/2023-0) e Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR) (408490/2024-1).

References

Alzoghbi, A., Arrascue Ayala, V. A., Fischer, P. M., and Lausen, G. (2015). Pubrec: Recommending Publications Based on Publicly Available Meta-data. *2015 CEUR Workshop Proceedings*, 1458:11–18.

Avazpour, I., Pitakrat, T., Grunske, L., and Grundy, J. (2014). Dimensions and Metrics for Evaluating Recommendation Systems. In Robillard, M. P., Maalej, W., Walker, R. J., and Zimmermann, T., editors, *Recommendation Systems in Software Engineering*, pages 245–273, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X., and Xia, F. (2019). Scientific Paper Recommendation: A Survey. *IEEE Access*, 7:9324–9339. DOI: 10.1109/ACCESS.2018.2890388.

Bao, K., Yan, M., Zhang, Y., Zhang, J., Wang, W., Feng, F., and He, X. (2025). Customizing in-context learning for dynamic interest adaption in LLM-based recommendation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14278–14291, Vienna, Austria. Association for Computational Linguistics. DOI: 10.18653/v1/2025.findings-acl.735.

Beel, J., Langer, S., Gipp, B., and Nürnberger, A. (2014). The Architecture and Datasets of Docear's Research Paper Recommender System. *D-Lib Magazine*, 20(11/12). DOI: 10.1045/november14-beel.

Ben Sassi, I. (2025). ScientiaRec: a Scientific Article Recommendation System with LLM-Driven Feature Extraction. In *EARL 2025 - Workshop on Evaluating and Applying Recommendation Systems with Large Language Models at RecSys '25*, Prague, Czech Republic.

Bulut, B., Kaya, B., and Kaya, M. (2019). A Paper Recommendation System Based on User Interest and Citations. In *Proceedings of the 2019 1st International Informatics and Software Engineering Conference (UBMYK)*, pages 1–5. DOI: 10.1109/UBMYK48245.2019.8965533.

Dhanda, M. and Verma, V. (2016). Recommender system for academic literature with incremental dataset. *Procedia Computer Science*, 89:483–491. Twelfth International Conference on Communication Networks, ICCN 2016, August 19– 21, 2016, Bangalore, India Twelfth International Conference on Data Mining and Warehousing, ICDMW 2016, August 19-21, 2016, Bangalore, India Twelfth International Conference on Image and Signal Processing, ICISP 2016, August 19-21, 2016, Bangalore, India. DOI: https://doi.org/10.1016/j.procs.2016.06.109.

dos Santos, J. V. F., de Andrade Nascimento, R. M., de Melo Camargo, A. C., Canuto, S. D. C., de Assis Costa, G., and de Sousa, D. X. (2025). Nova base de dados brasileira para sistemas de recomendação de artigos científicos. Brazil. Brazilian Symposium on Databases.

Flicke, M., Angrabeit, G., Iyengar, M., Protsenko, V., Shakun, Illia andCicvaric, J., Kargi, B., He, H., Schuler, Lukas andScholz, L., Agnihotri, K., Cao, Y., and Geiger, A. (2025). Scholar inbox: Personalized paper recommendations for scientists. In Mishra, P., Muresan, S., and Yu, T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 307–317, Vienna, Austria. Association for Computational Linguistics. DOI: 10.18653/v1/2025.acl-demo.30.

Guo, G., Chen, B., Zhang, X., Liu, Z., Dong, Z., and He, X. (2020). Leveraging Title-Abstract Attentive Semantics for Paper Recommendation. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020)*, pages 67–74. AAAI Press. DOI: 10.1609/aaai.v34i01.5335.

Haupka, N., Culbert, J. H., Schniedermann, A., Jahn, N., and Mayr, P. (2024). Analysis of the Publication and Document Types in OpenAlex, Web of Science, Scopus, Pubmed and Semantic Scholar.

Kang, W.-C. and McAuley, J. (2018). Self-Attentive Sequential Recommendation. `https://arxiv.org/abs/1808.09781`.

Kanwal, T. and Amjad, T. (2024). Research paper recommendation system based on multiple features from citation network. *Scientometrics*, 129:5493–5531. DOI: 10.1007/s11192-024-05109-w.

Koltun, V. and Hafner, D. (2021). The h-Index Is No Longer

an Effective Correlate of Scientific Reputation. *PLOS ONE*, 16(6):1–16. DOI: 10.1371/journal.pone.0253397.

Kreutz, C. K. and Schenkel, R. (2022). Scientific Paper Recommendation Systems: A Literature Review of Recent Publications. *International Journal on Digital Libraries*, 23:335–369.

Ley, M. (2002). The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In Laender, A. H. F. and Oliveira, A. L., editors, *String Processing and Information Retrieval*, pages 1–10, Berlin, Heidelberg. Springer Berlin Heidelberg.

Li, Y., Wang, R., Nan, G., Li, D., and Li, M. (2021). A Personalized Paper Recommendation Method Considering Diverse User Preferences. *Decision Support Systems*, 146. DOI: 10.1016/j.dss.2021.113546.

Li, Z. and Zou, X. (2019). A Review on Personalized Academic Paper Recommendation. *Computer and Information Science*, 12:33. DOI: 10.5539/cis.v12n1p33.

Liang, W., Lu, Z., Jin, Q., Xiong, Y., and Wu, M. (2015). Modeling of Research Topic Evolution Associated with Social Networks of Researchers. In *Proceedings of the 2015 IEEE 12th International Conference on Ubiquitous Intelligence and Computing, 2015 IEEE 12th International Conference on Autonomic and Trusted Computing, and 2015 IEEE 15th International Conference on Scalable Computing and Communications (UIC-ATC-ScalCom)*, pages 1169–1174. DOI: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.213.

Liu, J., Liu, C., Zhou, P., Lv, R., Zhou, K., and Zhang, Y. (2023). Is chatgpt a good recommender? a preliminary study.

Montazerian, M., Zanotto, E. D., and Eckert, H. (2020). Prolificacy and Visibility versus Reputation in the Hard Sciences. *Scientometrics*, 123(1):207–221. DOI: 10.1007/s11192-020-03369-w.

Rashed, A., Grabocka, J., and Schmidt-Thieme, L. (2019). Attribute-Aware Non-Linear Co-Embeddings of Graph Features. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19)*, pages 314–321. Association for Computing Machinery. DOI: 10.1145/3298689.3346999.

Said, A. and Bellogín, A. (2014). Comparative Recommender System Evaluation: Benchmarking Recommendation Frameworks. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*, pages 129–136, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2645710.2645746.

Sugiyama, K. and Kan, M.-Y. (2010). Scholarly Paper Recommendation via User's Recent Research Interests. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, pages 29–38, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/1816123.1816129.

Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., and Jiang, P. (2019). BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. `https://arxiv.org/abs/1904.06690`.

Wang, H., Chen, B., and Li, W.-J. (2013). Collaborative Topic Regression with Social Regularization for Tag Recommendation. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*.

Wang, L., Du, W., and Chen, Z. (2024). Multi-feature-enhanced academic paper recommendation model with knowledge graph. *Applied Sciences*, 14(12). DOI: 10.3390/app14125022.

Wang, S. and Yin, Z. (2025). Discourse-aware scientific paper recommendation via qa-style summarization and multi-level contrastive learning. *ArXiv*, abs/2511.03330.

Wang, W., Tang, T., Xia, F., Gong, Z., Chen, Z., and Liu, H. (2022). Collaborative Filtering With Network Representation Learning for Citation Recommendation. *IEEE Transactions on Big Data*, 8(5):1233–1246. DOI: 10.1109/TBDATA.2020.3034976.

Wang, X., He, X., Cao, Y., Liu, M., and Chua, T.-S. (2019). KGAT: Knowledge Graph Attention Network for Recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 950–958. Association for Computing Machinery.

Wu, S., Tang, Y., Zhu, Y., Wang, L., Xie, X., and Tan, T. (2019). Session-Based Recommendation with Graph Neural Networks. `https://arxiv.org/pdf/1811.00855`.

Xie, Y., Sun, Y., and Bertino, E. (2021). Learning Domain Semantics and Cross-Domain Correlations for Paper Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*, pages 706–715, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3404835.3462975.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Hernandez Abrego, G., Yuan, S., Tar, C., Sung, Y.-h., Strope, B., and Kurzweil, R. (2020). Multilingual Universal Sentence Encoder for Semantic Retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-demos.12.

Zhang, Y., Wang, M., Gottwalt, F., Saberi, M., and Chang, E. (2019). Ranking Scientific Articles Based on Bibliometric Networks with a Weighting Scheme. *Journal of Informetrics*, 13(2):616–634. DOI: 10.1016/j.joi.2019.03.013.

Zhao, Z., Fan, W., Li, J., Liu, Y., Mei, X., Wang, Y., Wen, Z., Wang, F., Zhao, X., Tang, J., and Li, Q. (2024). Recommender systems in the era of large language models (llms). *IEEE Trans. on Knowl. and Data Eng.*, 36(11):6889–6907. DOI: 10.1109/TKDE.2024.3392335.