Nova Base de Dados Brasileira para Sistemas de Recomendação de Artigos Científicos

João Vitor Felipe dos Santos[1], **Ricardo Marçal de Andrade Nascimento**[1], **Adriano César de Melo Camargo**[1], **Sergio Daniel Carvalho Canuto**[1], **Gustavo de Assis Costa**[1], **Daniel Xavier de Sousa**[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia de Goiás (IFG)
Goiás – Brasil

joaovitor.felipesantos@gmail.com, ricardomarcal02@hotmail.com, adrianocesar321@gmail.com, sergio.canuto@ifg.edu.br, gustavo.costa@ifg.edu.br, daniel.sousa@ifg.edu.br

Abstract. *This work presents a new dataset for Scientific Article Recommendation Systems (SARS). In addition to being scarce, many existing datasets in the SARS context rely solely on co-authorship relationships as a relevance criterion, overlooking the importance of explicit user evaluations. To address this issue, we propose a new dataset with explicitly defined relevance labels, comprising over 2,000 researchers, 30 areas of knowledge, and approximately 71,000 associated papers. The work includes a characterization and evaluation of the proposed dataset, alongside other widely used datasets in the literature.*

Resumo. *Este trabalho apresenta uma nova base de dados para Sistemas de Recomendação de Artigos Científicos (SRAC). Além de escassas, muitas das bases de dados no contexto SRAC utilizam apenas relações de coautoria como critério de relevância, negligenciando a importância das avaliações explícitas feitas por usuários. Para mitigar esse problema, propomos uma nova base de dados com rótulos de relevância explicitamente definidos, composta por mais de 2 mil pesquisadores, 30 áreas do conhecimento e aproximadamente 71 mil trabalhos associados. O trabalho inclui uma caracterização e avaliação da base proposta, juntamente com outras bases amplamente utilizadas na literatura.*

1. Introdução

Sistemas de Recomendação tornaram-se ferramentas indispensáveis em diversos domínios, do entretenimento à pesquisa científica [Sanchez-Lengeling et al. 2021]. No contexto da disseminação do conhecimento científico e do desafio de localizar publicações relevantes em meio ao volume crescente de trabalho acadêmicos, a especialidade dos Sistemas de Recomendação de Artigos Científicos (SRAC) desempenham um papel essencial [Price et al. 2022]. Esses sistemas auxiliam, sobretudo, pesquisadores na identificação de novos trabalhos que estejam alinhados com seus interesses temáticos e áreas de atuação.

Apesar da relevância desses sistemas, há uma escassez significativa de bases de dados públicas adequadas ao treinamento de modelos para SRAC [Kreutz and Schenkel 2022]. Na prática, muitas delas ou são de uso restrito/privado, ou não estão mais acessíveis. No entanto, mesmo as bases disponíveis enfrentam limitações importantes. Um exemplo é a pouca diversidade de áreas do

conhecimento. A produção acadêmcia apresenta padrões de comportamento específicos em cada área [Montazerian et al. 2020], e portanto, bases limitadas por área (como exemplo, DBLP e PubMed) desconsideram a diversidade de pesquisadores, o que necessariamente ocasiona viés de conteúdo. Outras bases se mantém desatualizadas (como exemplo, CiteULike), que por terem sido descontinuadas, não descrevem o padrão de comportamento mais recente dos pesquisadores, que pode evoluir devido à mudanças nas tecnologias vigentes, ou mesmo condições políticas relativas ao fomento da pesquisa [Koltun and Hafner 2021].

Por fim, a escassez de bases públicas para sistemas de recomendação de artigos científicos com rotulação explícita, feita pelos próprios pesquisadores, também representa uma limitação crucial tanto para o ajuste de parâmetros quanto para a avaliação fidedigna da eficácia dos modelos de recomendação [Kreutz and Schenkel 2022]. Na prática, ao longo da carreira acadêmica dos pesquisadores pode haver distanciamento de interesse entre os artigos publicados no início e os mais recentes, fenômeno comum entre pesquisadores sêniores [Liang et al. 2015]. Na mesma linha, no caso de recém-doutores, o número reduzido de publicações oferece pouca informação sobre seus interesses de pesquisa. Nesse contexto, a rotulação indireta ou implícita da relevância de publicações — por exemplo, considerando como relevantes apenas os próprios artigos do autor [Li et al. 2021] — pode ignorar os reais interesses do pesquisador, assumindo erroneamente que estes se limitam às áreas de suas publicações atuais. Da mesma forma, abordagens baseadas em referências [Xie et al. 2021] ou em citações [Zhang et al. 2019] também estão sujeitas a esse mesmo viés. Idealmente, tais vieses associados às estratégias de rotulação implícita podem ser mitigados por meio da rotulação explícita, feita pelos próprios pesquisadores, a partir da seleção manual das publicações que consideram mais relevantes para seus interesses de pesquisa.

Diante dessas limitações, este trabalho propõe mitigar tais lacunas por meio da introdução de uma nova base de dados para sistemas de recomendação de artigos e projetos científicos, denominada *Rec4SciBR*. A *Rec4SciBR* possui rótulos de relevância de artigos e projetos científicos, além de contar com cerca de 2 (dois) mil pesquisadores brasileiros, abrangendo múltiplas áreas do conhecimento e incorporando avaliações e conteúdos atualizados. A base *Rec4SciBR* estrutura as descrições dos projetos, publicações e perfis dos pesquisadores com informações sobre autoria, coautoria e rotulação de relevância. As informações da base foram caracterizadas considerando o volume de interações entre usuários e publicações/projetos e a diversidade de descrição dos objetos armazenados. Essa caracterização é estendida, em um esforço comparativo, para as coleções DBLP[Sugiyama and Kan 2010] e CiteULike [Wang et al. 2013].

Os diferentes contextos de recomendação das coleções *Rec4SciBR*, DBLP[Sugiyama and Kan 2010] e CiteULike[Wang et al. 2013] foram criteriosamente comparados por meio da avaliação da eficácia de diferentes estratégias de recomendação. Para tal, foram considerados modelos de recomendação recentes de diferentes paradigmas, incluindo o método GraphRec [Rashed et al. 2019] baseado em grafos e filtragem colaborativa, a abordagens híbrida KGAT [Wang et al. 2019], técnicas baseadas em *self-attention* como SR-GNN [Wu et al. 2019], SASRec [Kang and McAuley 2018] e BERT4Rec [Sun et al. 2019], além de modelos baseados em conteúdo, como TF-IDF e BERT [Yang et al. 2020].

Os resultados obtidos indicam que a base de dados *Rec4SciBR* não apenas oferece um conjunto de informações bem estruturadas para experimentação acadêmica, como também se mostra promissora para a aplicação e avaliação de diferentes modelos de recomendação em cenários realistas. Em razão da presença de rótulos de relevância e da sequência de avaliações realizadas pelos usuários, a *Rec4SciBR* pode ainda ser utilizada para apoiar projetos no contexto de Large Language Models (LLMs), particularmente em aplicações baseadas em contexto (few-shots) [Lima et al. 2024] sobre pesquisadores brasileiros, sendo mais de 80% do seu conteúdo em Português. Quanto à aplicação dos modelos, observou-se que a *Rec4SciBR* apresentou melhor desempenho em abordagens híbridas e de filtragem baseada em conteúdo, atribuível, principalmente, à rica descrição dos objetos, pesquisadores e publicações.

Entendemos que a comunidade científica, em especial a comunidade brasileira, poderá se beneficiar com mais uma alternativa de base de dados pública, que possa contribuir na avaliação e construção de novos modelos. Ressaltamos ainda, que a contribuição apresentada não se trata de uma iniciativa pontual ou estática, mas sim de uma base de dados em constante atualização e expansão para incorporar novos dados, ampliando numero de usuários e suas avaliações.

O restante deste trabalho está organizado da seguinte forma. Na seção 2 descrevemos de forma mais geral o contexto de Sistemas de Recomendação de Artigos Científicos e bases de dados existentes. Na seção 3 apresentamos uma caracterização da base de dados proposta, e na seção 4 uma comparação com modelos de Sistemas de Recomendação. Por fim, na seção 5, descrevemos nossas conclusões.

🤗 **Base de Dados:** https://huggingface.co/datasets/Rec4SciBR/Rec4SciBR

2. Referencial Teórico

A tarefa de recomendação em Sistemas de Recomendação de Artigos Científicos (SRAC) pode ser descrita, de forma geral, como a recomendação de itens (publicações) para um determinado usuário/pesquisador [Bulut et al. 2019, Guo et al. 2020]. Modelagens para SRAC apresentam especificidades próprias do ambiente acadêmico. Em especial, conforme destacado por [Alzoghbi et al. 2015], modelos baseados no conteúdo textual de usuários/itens–denominados como Sistemas de Filtragem Baseada em Conteúdo [Bai et al. 2019]–são capazes de explorar o abundante conteúdo textual das publicações com a similaridade textual entre publicações e perfis de pesquisadores. Por outro lado, modelos baseados na exploração da similaridade entre as preferências de um usuário e as preferências dos demais usuários da base–denominados como Sistemas de Filtragem Colaborativa [Wang et al. 2022]–podem ser comprometidas devido à escassez ou esparsidade de dados referentes às avaliações de preferência dos pesquisadores.

Em contextos mais amplos, como em plataformas de recomendação de filmes e músicas, as avaliações explícitas de preferência dos usuários desempenham papel central no desenvolvimento de modelos e avaliação de efetividade das recomendações, especialmente em modelos híbridos que combinam a Filtragem Baseada em Conteúdo com Filtragem Colaborativa [Li and Zou 2019]. No entanto, a maioria das bases de dados para recomendação de artigos científicos carecem de registros que estabeleçam essa correlação direta entre usuários e publicações em termos de avaliação ou relevância explícita [Kreutz and Schenkel 2022].

Um dos poucos esforços na construção de bases de dados públicas com rotulação explícita de relevância foi realizado na construção da coleção SPRD (*Scholarly Paper Recommendation Dataset*), a partir da rotulação explícita de publicações de interesse de 50 pesquisadores da área de Ciência da Computação, e seu subconjunto, denominado SPRD_Senior, que contém a rotulação de 13 pesquisadores sêniores [Sugiyama and Kan 2010]. No entanto, além do número limitado de pesquisadores, o acesso a tal conjunto encontra-se indisponível. Da mesma forma, diversas coleções como BibSonomy [Beel et al. 2014], DOCEAR [Beel et al. 2014] e PRSDataset [Guo et al. 2020] não se encontram acessíveis até o momento desta pesquisa.

Das bases de dados disponíveis, a CiteULike [Wang et al. 2013] destaca-se como uma das mais amplamente exploradas na literatura. Derivada da plataforma homônima, que funcionava como uma rede social acadêmica voltada para a descoberta, organização e compartilhamento de referências bibliográficas, a base oferecia funcionalidades como recomendação automática de tags, estatísticas de uso e sugestões personalizadas de artigos. Apesar da descontinuidade da plataforma, repositórios de dados que refletem as interações de 5.551 usuários com artigos publicados entre 2004 à 2006 foi disponibilizado publicamente[1].

Outra base de dados amplamente utilizada na área da Ciência da Computação é a DBLP [Ley 2002]. Originalmente concebida como um experimento Web, a DBLP evoluiu para um repositório de grande escala, contendo atualmente mais de 7 milhões de publicações e cerca de 3 milhões de autores, abrangendo aproximadamente 6.900 conferências e periódicos. A base organiza e disponibiliza metadados bibliográficos detalhados, incluindo informações sobre autoria, coautoria, títulos, eventos, periódicos, datas de publicação, linhas de pesquisa e a evolução de comunidades científicas. Devido à grade disponibilidade de metadados, diversos trabalhos adotam a DBLP no desenvolvimento e avaliação de modelos para SRAC [Ley 2002].

Os principais conjuntos de dados utilizados em SRAC apresentam limitações significativas. Bases desatualizadas, como CiteULike, não capturam mudanças recentes nos interesses dos pesquisadores. Já a DBLP, embora amplamente utilizada, não possui diversidade de áreas nem oferece rótulos explícitos de relevância. Problemas semelhantes são observados em bases como OpenAlex, Web of Science, Scopus e PubMed [Haupka et al. 2024], que, apesar da cobertura de publicações científicas e metadados bibliográficos, não dispõem de rótulos explícitos de relevância nem o acesso a dados personalizados (como histórico de avaliação, descrição de projetos e caracterização dos pesquisadores) o que inviabiliza a aplicação de técnicas supervisionadas e personalizadas de recomendação. Dessa forma, considerando a escassez de bases de dados acessíveis para as tarefas de recomendação de artigos científicos, reforçamos a contribuição deste trabalho com a disponibilização e avaliação da base de dados *Rec4SciBR*.

3. Apresentação da Base de Dados *Rec4SciBR*

Nesta seção apresentamos a caracterização da base de dados *Rec4SciBR*, proposta para avaliação de modelos de Sistemas de Recomendação de Artigos Científicos (SRAC). Destacamos suas características, limitações e comparação com outras base de dados.

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

A base *Rec4SciBR* foi construída a partir de um cenário real, considerando a recomendação de artigos e projetos acadêmicos para mais de 2 mil pesquisadores e docentes de uma instituição brasileira de ensino e pesquisa. A descrição dos artigos e projetos foram obtidos a partir da Plataforma Lattes[2] com a coleta dos currículos, e as recomendações são feitas considerando a similaridade de conteúdo entre usuários e itens. Ou seja, pesquisadores recebem recomendações baseadas na similaridade dos termos mais frequentes em seus currículos Lattes (inclindo artigos, projetos e demais informações) em comparação com os trabalhos dos demais pesquisadores. Após recomendação os pesquisadores são convidados a avaliar cada item, fornecendo a avaliação explicita dos artigos e projetos. O pesquisador avalia cada recomendação atribuindo uma nota de 0 a 5.

Além dos rótulos, a base também inclui informações descritivas detalhadas dos usuários, artigos e projetos acadêmicos, considerando a disponibilidade pública dos currículos lattes. Considerando as informações dos pesquisadores, ressaltamos que não há nenhuma informação pessoal na base, como nome, endereço, contato ou intituição.

No intuito de apresentar os dados de forma coerente com os algoritmos de sistemas de recomendação, apresentaremos a caracterização seguindo duas perspectivas. A dos sistemas de filtragem colaborativa e dos sistemas baseado em conteúdo. Na primeira, as conexões entre usuários e itens é a informação mais frequentemente utilizada para predizer os interesses dos usuários. Os algoritmos nesta categoria exploram diferentes extratégias para gerar uma boa representação das conexões entre usários e itens. Na segunda perspectiva, dos sistemas baseado em conteúdo, a descrição dos usuários e itens são utilizado com mais frequência. Neste caso, o objetivo é criar boas representações dos usuários e itens de forma que a comparação entre esses dois objetos tenham maior precisão. Naturalmente, os algoritmos ditos híbridos exploram essas duas informações.

Tabela 1: Resumo dos usuário e itens das bases avaliadas.

Dataset	Usuários	Itens	Conexões	Esparsidade	Área do Conhecimento
DBLP(AAN)	51645	30000	76699	0,995	Linguística Computacional
CiteULike	5551	16980	204986	0,783	Diversas
Rec4SciBR	2261	71895	19211	0,823	Diversas

3.1. Caracterização das Conexões entre usuários e itens

Apresentamos na Tabela 1 uma descrição numérica sobre usuários, itens e as conexões das bases de dados analisadas. Além da base *Rec4SciBR*, mostramos também as informações sobre CiteULike e *DBLP(AAN)*. No caso da DBLP, dado seu grande volume, mostramos somente uma fração (denominada *DBLP(AAN)*) que tem sido comumente usada em outros trabalhos [Sugiyama and Kan 2010]. A base *DBLP(AAN)* representa uma categoria de bases descritas por grandes repositótios (como OpenAlex, Web of Science, Scopus e PubMed[Haupka et al. 2024]), mas que não descrevem rótulos explícitos.[3]

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

[3]Embora a base *DBLP(AAN)* não seja diretamente comparável com a *Rec4SciBR*, por não ter rótulos explícitos, incluímos em nossas análises por ser amplamente utilizada na literatura, ampliando a completude do trabalho aqui proposto.

A Tabela 1 apresenta o número de usuários, itens (artigos e projetos no caso do *Rec4SciBR*), conexões, esparcidade e áreas do conhecimento. As conexões mostram o número de relações entre usuário e itens. Para as bases de dados *Rec4SciBR* e CiteULike essas conexões mostram o número de avaliações explicitas. Para o *DBLP(AAN)*, as conexões refletem a relação de autoria. Já a esparcidade mostra a ausência de conexões existentes entre usuários e itens. A base *Rec4SciBR* possui 2261 usuários e mais de 70 mil publicações, com esparcidade de 0,823.

Tabela 2: Resumo das conexões por usuário e por items.

	Métrica	***Rec4SciBR***	**CiteULike**	***DBLP(AAN)***
Conexões por Usuário	Número total de usuários com conexões	1.408	5.551	51.645
	Mínimo de conexões por usuário	1	10	1
	Máximo de conexões por usuário	178	403	233
	Média de conexões por usuário	6,75 (11,17)	36,90 (42,08)	1,49 (2,08)
Conexões por Item	Número total de publicações com conexões	7.710	16.980	30.000
	Mínimo de conexões por publicação	1	1	1
	Máximo de conexões por publicação	9	321	50
	Média de conexões por publicação	1,23 (0,60)	12,07 (11,85)	2,56 (1,80)

Já na Tabela 2 mostramos uma análise individualizada por usuário e itens. Neste caso, a base *Rec4SciBR* possui em media 6,75 avaliações por usuário (com desvio padrão de 11,17), com o mínimo de 1 e máximo de 178 avaliações. Em relação às publicações, são 7.710 publicações que receberam entre 1 e 9 avaliações, e uma média de 1,2 avaliações por publicação, com um desvio padrão de 0,60. Como descrito na Tabela 2, a base *DBLP(AAN)* possui mais de 51 mil usuários com conexões (relação de autoria).

Em linhas gerais, percebemos que os valores de conexão entre usuários e itens (tabelas 1 e 2) para a base *Rec4SciBR* são relativamente inferiores que a base CiteULike. Porém, 1.408 usuários possuem avaliações explícitas. Esses números demostram uma representatividade real que reflete uma instituição de mais de 2.000 pesquisadores, sendo então bastante útil na avaliação de modelos realistas. Ademais, como mostraremos a seguir, a base *Rec4SciBR* possui maior descrição individualizada dos usuários e itens.

Considerando a distribuição de rotulação explicita na base *Rec4SciBR*, verificamos que há uma grande predominância de notas altas, sendo 17.622 (91,73%) recebendo notas entre 4,0 e 5,0, 814 (4,23%) avaliações com notas entre 2,0 e 4,0, e apenas 775 avaliações com notas entre 0,0 e 2,0 (4,04%). Isso mostra que de forma geral, a base possui boa representatividade sobre o interesso dos usuários, destacando a qualidade da base de dados com o grande volume de rótulos relevantes.

Tabela 3: Números para as Publicações Científicas e Projetos Acadêmicos.

	Métrica	Valor
Publicações Científicas	Total de publicações bibliográficas	65.210
	Média de publicações por ano	1552,61
	Ano mais antigo de publicação	1.982
	Ano mais recente de publicação	2.024
Projetos Acadêmicos	Total de projetos de pesquisa	6.685
	Média de duração dos projetos (anos)	1,76
	Média de caracteres na descrição dos projetos	775,37

Tabela 4: Descrição textual dos itens das bases de dados.

Métrica	***Rec4SciBR***		***DBLP(AAN)***	**CiteULike**
	Projetos	Artigos	Artigos	Artigos
Total de termos únicos	34.254	31.445	19.357	14.354
Média de termos	58,7	8,64	6,8	6,1
Maior número de termos	273	35	48	20

3.2. Caracterização - Descrição de usuários e itens

A base *Rec4SciBR* disponibiliza as seguintes informações para os artigos científicos: nome do local de publicação, título do artigo, ano de produção, idioma, tipo da produção, meio de divulgação e área de atuação. Atualmente há cerca de 31.532 revistas em 57 distintas áreas de atuação e 1.403 conferências. Já para os projetos acadêmicos, são disponibilizadas: nome do projeto e descrição do projeto. Para os usuários, são disponibilizadas alguns campos do currículo Lates como: apresentação do pesquisador, área de atuação e descrição dos projetos.

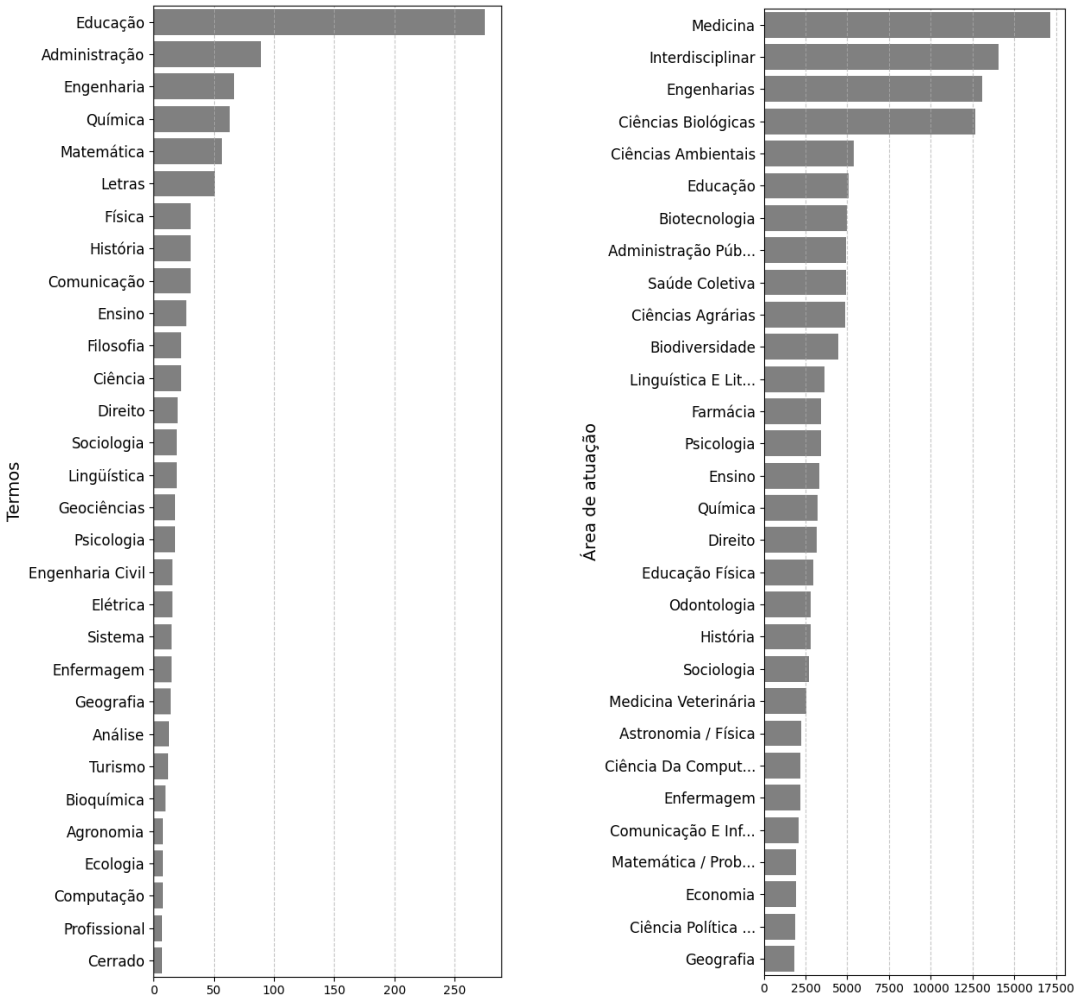
Considerando ainda a distribuição da língua, observa-se que as publicações científicas estão em 81,1% na língua portuguesa, e 17,3% em inglês. Já os projetos acadêmicos estão majoritariamente redigidos em português. O restante das publicações corresponde a materiais em espanhol e em outros idiomas. A Tabela 3 detalha as características dos dois tipos de itens empregados nas recomendações: publicações científicas e projetos acadêmicos. A base *Rec4SciBR* contém mais de 65 mil publicações e 6 mil projetos acadêmicos.

Um fator diferencial da *Rec4SciBR* é a disponibilidade de descrições textuais tanto para os usuários quanto para os itens (artigos e publicações). A Tabela 4 apresenta um comparativo entre as bases de dados, considerando a quantidade de termos utilizados para descrever os títulos das publicações e projetos. Para essa descrição, em todas as bases, foi realizada a remoção de *stopwords*. Como se observa na tabela, a base *Rec4SciBR* possui um dicionário de dados quase duas vezes maior que as base *DBLP(AAN)* e CiteULike, tanto para artigos como para projetos, cerca de 31 mil e 34 mil, respectivamente. Alé de possuir maior quantidade média de termos por item. Esses números evidenciam a quantidade de termos que descrevem a base, e como veremos na seção seguinte, permite construir melhor representação nos modelos baseado em conteúdo.

No intuito de entender melhor a descrição textual somente dos usuários, obtida a partir dos campos do currículo Lattes (apresentação do pesquisador, área de atuação

Tabela 5: Estatísticas dos termos associados a usuários da base *Rec4SciBR*.

Métrica	Valor
Total de termos únicos	22.138
Média de termos por usuário	20
Maior número de termos (1 usuário)	1.287



(a) Termos na primeira posição de importância para os usuários da base *Rec4SciBR*

(b) Áreas de atuação (por revista) com mais publicações científicas da base *Rec4SciBR*

Figura 1: Visualização da variedade de áreas para usuários e publicações.

e descrição dos projetos.), a Tabela 5 mostra as estatísticas para a base *Rec4SciBR*. Em média, são 60 palavras por usuário, e o dicionário resultante possui 62.138 palavras.

Para entender melhor o conteúdo da descrições de usuários e itens da base *Rec4SciBR*, contabilizamos os termos mais importante seguindo a representação TF-IDF. Na figura 1a identificamos os termos mais frequentes para grupos de usuários. Por exemplo, o termo "Educação" é o mais frequente no conteúdo dos currículos de 275 pesquisadores. Na sequência, vemos que administração e engenharia aparecem como segundo e

terceiro termos mais frequentes, respectivamente. Em suma, a figura mostar as 30 áreas do conhecimento mais frequentes da base de dados, destacando a pluralidade de conhecimento dos seus pesquisadores.

Mesma análise é feita na figura 1b, que mostra os termos mais frequentes para os nomes das revistas ou *journals* citados nos currículos Lattes. Ou seja, para as publicações em revistas, destacam-se os termos Medicina, Interdisciplinar e Engenharias. Vemos que a Figura 1b confirma a variedade de tópicos da base de dados proposta.

Em suma, a base de dados *Rec4SciBR* apresenta uma descrição rica de usuários e itens, com potencial para ser utilizada na avaliação de modelos reais de Sistemas de Recomendação. Este ponto é reforçado na próxima seção, na qual descrevemos seu uso por modelos de recomendação disponíveis na literatura, destacando, em especial, os bons resultados obtidos com a exploração das representações dos itens e dos usuários.

4. Avaliação em Sistemas de Recomendação

Essa seção apresenta uma análise da eficácia das base de dados analisadas, bem como as metodologias utilizadas para tal, considerando modelos de Sistemas de Recomendação frequentemente utilizados na literatura. O objetivo é analisar a performance da base *Rec4SciBR* em diferentes cenários, permitindo um indicativo inicial dos modelos que melhor exploram seu conteúdo.

4.1. Metodologia de Coleta e Processamento

A efetividade das estratégias de recomendação foi avaliada nas coleções *Rec4SciBR*, CiteULike [Wang et al. 2013] e *DBLP(AAN)* [Sugiyama and Kan 2010] previamente caracterizadas.

No que se refere ao pré-processamento dos dados, uma etapa comum a todas as bases foi a remoção de usuários sem interações, ou seja, que não avaliaram nem produziram publicações.

Os dados utilizados nos modelos SR-GNN e SASRec exigem critérios adicionais: o SR-GNN elimina conexões com menos de cinco interações, enquanto o SASRec remove usuários com menos de três. Já os modelos GraphRec e KGAT apenas reestruturam os dados sem modificar seu conteúdo original.

No caso da base *Rec4SciBR* aplicada no SR-GNN, foi necessário binarizar as avaliações (originalmente de 0 a 5), considerando nota > 3 como relevante. Além disso, tratamos tanto artigos quanto projetos de pesquisa como publicações.

Para a avaliação do algortimos que utilizam uma abordagem sequencial, registramos a ordem em que os rótulos foram atribuídos, garantindo um registro do padrão sequencial dos usuários.

4.2. Metodologia de Avaliação

No que se refere aos modelos empregados, utilizou-se uma ampla variedade de algoritmos representativos das principais abordagens em sistemas de recomendação. As estratégias avaliadas incluem: **i) Filtragem Colaborativa**, com os modelos GraphRec [Rashed et al. 2019], KGAT [Wang et al. 2019], SR-GNN [Wu et al. 2019][4], SASRec

[4]https://github.com/CRIPAC-DIG/SR-GNN

[Kang and McAuley 2018][5], e BERT4Rec [Sun et al. 2019][6]; **ii) Filtragem Baseada em Conteúdo**, por meio de técnicas como TF-IDF[7] e BERT [Yang et al. 2020][8]; e **iii) Filtragem Híbrida**, representada pelo modelo KGAT [Wang et al. 2019][9]. Todas as estratégias foram aplicadas seguindo a parametrização padrão indicada em suas respectivas publicações originais.

Já com relação ao treinamento e execução dos modelos, os algoritmos GraphRec e KGAT foram treinados utilizando uma divisão padrão de 80% dos dados para treino, 10% para teste e 10% para validação, com amostragem aleatória após o pré-processamento. Já os modelos SR-GNN, SASRec e BERT4Rec, por serem algoritmos sequenciais, separamos as duas publicações mais recentemente associadas a cada usuário, atribuídas para validação e testes, conforme descrito em [Sun et al. 2019]. Para a aplicação do modelo BERT, utilizamos o modelo SentenceBert (Multimodal) [10] para construção da nova representação vetorial dos usuários e itens, e calculamos a similaridade pela função cosseno, seguindo [Yang et al. 2020].

Da mesma forma, os modelos de filtragem baseados em conteúdo (TF-IDF e BERT Multilíngue) constroem o perfil do usuário com todas as interações disponíveis, dispensando essa divisão.

Para garantir a robustez e generalização dos resultados, foi adotada validação cruzada em 5 partes (*5-fold cross-validation*), na qual o conjunto de dados é dividido em cinco subconjuntos de usuários, e a avaliação é repetida cinco vezes, alternando o subconjunto de teste e utilizando os demais para treinamento, conforme prática consolidada na área [Said and Bellogín 2014]. A análise dos modelos foi conduzida por meio de métricas amplamente aplicada na literatura de recomendação [Avazpour et al. 2014], a saber: Normalized Discounted Cumulative Gain (NDCG) e Mean Average Precision (MAP).

4.3. Resultados Gerais

Iniciamos nossa análise mostrando a efetividade das bases avaliadas para os modelos citados, necessariamente: GraphRec, KGAT, SR-GNN, SASRec, BERT4Rec, TF-IDF e BERT. Para esse experimento utilizamos as relações de relevância originais de cada base, ou seja, rotulação explicita para as base CiteULike e *Rec4SciBR*, e rotulação implícita por autoria para *DBLP(AAN)*. Também utilizamos as descrições de cada item e usuário descritos anteriormente.

Os resultados são descritos na Tabela 6. Como se observa, analisando a métrica NDCG10 para a base *Rec4SciBR*, algoritmos que exploram a filtragem baseada em conteúdo (TF-IDF e BERT), explorando as representações dos usuários e publicações, tiveram resultados mais eficazes que os algoritmos que exploram os modelos por filtragem colaborativa e filtragem híbrida. Padrão que se repetiu também para a base de dados

[5] https://github.com/kang205/SASRec

[6] https://github.com/FeiSun/BERT4Rec

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

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

[9] https://github.com/xiangwang1223/knowledge_graph_attention_network

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

Base	Métrica	Filt. Híbrida		Filt. Colaborativa			Filt. Conteúdo	
		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)	MAP	**0.779**	0.761	0.550	0.327	0.195	0.484	0.471
	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

Tabela 6: Resultados dos algoritmos nos experimentos com as bases *Rec4SciBR*, DBLP e CiteULike

CiteULike. Ou seja, a quantidade de termos únicos descrevendo os artigos e projetos (Tabela 4) parece contribuir para gerar a melhor representação entre usuários e artigos. Por outro lado, a baixa descrição das conexões entre usuários e itens (Tabela 2), justificam o resultado inferior para os modelos com Filtragem Colaborativa e Híbridos.

Inlusive, verificamos que para a base *Rec4SciBR*, o algoritmo TF-IDF teve resultado melhor que o BERT. Provavelmente, devida a diversidade de áreas da base, a separação de termos como vetores específicos possa ser mais apropriado que a a análise de contexto aplicado com o modelo BERT.

Os resultados da base *DBLP(AAN)* são também bastante interessantes. Embora a mesma possua maior esparsidade, a base possui um número total de conexões por usuário e por item (Tabela 2) superior comparado as outras bases. Consequentemente isso justifica os bons resultados que exploram as conexões entre usuários e itens – com Filtragem Colaborativa e Híbrida. Ao mesmo tempo, como a base possui pouca descrição dos seus objetos, os algoritmos que utilizam abordagem baseada em conteúdo apresentam resultados menores.

5. Conclusão

Este trabalho apresenta a base de dados *Rec4SciBR*, uma nova coleção voltada para o desenvolvimento e avaliação de Sistemas de Recomendação de Artigos Científicos (SRAC), com foco na realidade da produção científica brasileira. A proposta visa suprir lacunas importantes observadas nas bases de dados amplamente utilizadas na literatura, como a ausência de rótulos de relevância explícitos e a limitada diversidade temática.

Os resultados experimentais demonstraram que a *Rec4SciBR* é promissora para a avaliação de modelos de recomendação baseados em conteúdo, especialmente em razão da riqueza das descrições textuais dos usuários e itens. Modelos como TF-IDF e BERT apresentaram desempenho superior, evidenciando o valor da base para aplicações práticas.

Como análise de limitações, identificamos que a predominância da língua portuguesa pode reduzir a eficácia de modelos de linguagem amplamente treinados em inglês, exigindo adaptação ou fine-tuning para garantir uma representação semântica adequada dos dados. Ainda, a diversidade de áreas do conhecimento impõe desafios adicionais na identificação de similaridades e padrões relevantes, especialmente quando se busca recomendar conteúdos entre domínios muito distintos. Para mitigar o problema desse último tópico, seguiremos trabalhando para ampliação da base proposta, includindo novos pes-

quisadores e recomendações de novos artigos.

Em suma, a *Rec4SciBR* representa um avanço importante para o ecossistema de SRAC no Brasil, oferecendo uma fonte de dados pública, atualizada e representativa da diversidade da pesquisa nacional. Como perspectivas futuras, sugerimos a ampliação da base com novos usuários e avaliações, bem como a integração de atributos semânticos mais ricos e grafos heterogêneos para potencializar modelos híbridos e baseados em aprendizado profundo.

Agradecimentos

Este trabalho é apoiado pelo Instituto Federal de Educação, Ciência e Tecnologia de Goiás (IFG), pelo 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). .

Referências

[Alzoghbi et al. 2015] 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 et al. 2014] 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 et al. 2019] Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X., and Xia, F. (2019). Scientific Paper Recommendation: A Survey. *IEEE Access*, 7:9324–9339.

[Beel et al. 2014] 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).

[Bulut et al. 2019] 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.

[Guo et al. 2020] 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.

[Haupka et al. 2024] 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 and McAuley 2018] Kang, W.-C. and McAuley, J. (2018). Self-Attentive Sequential Recommendation. `https://arxiv.org/abs/1808.09781`.

[Koltun and Hafner 2021] Koltun, V. and Hafner, D. (2021). The h-Index Is No Longer an Effective Correlate of Scientific Reputation. *PLOS ONE*, 16(6):1–16.

[Kreutz and Schenkel 2022] 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 2002] 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 et al. 2021] 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.

[Li and Zou 2019] Li, Z. and Zou, X. (2019). A Review on Personalized Academic Paper Recommendation. *Computer and Information Science*, 12:33.

[Liang et al. 2015] 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.

[Lima et al. 2024] Lima, M., Silva, E., and da Silva, A. (2024). Um Estudo sobre o Uso de Modelos de Linguagem Abertos na Tarefa de Recomendação de Próximo Item. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 510–522, Porto Alegre, RS, Brasil. Sociedade Brasileira de Computação.

[Montazerian et al. 2020] Montazerian, M., Zanotto, E. D., and Eckert, H. (2020). Prolificacy and Visibility versus Reputation in the Hard Sciences. *Scientometrics*, 123(1):207–221.

[Price et al. 2022] Price, R., Skopec, M., Mackenzie, S., Nijhoff, C., Harrison, R., Seabrook, G., and Harris, M. (2022). A Novel Data Solution to Inform Curriculum Decolonisation: The Case of the Imperial College London Masters of Public Health. *Scientometrics*, 127(2):1021–1037.

[Rashed et al. 2019] 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.

[Said and Bellogín 2014] 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.

[Sanchez-Lengeling et al. 2021] Sanchez-Lengeling, B., Reif, E., Pearce, A., and Wiltschko, A. B. (2021). A Gentle Introduction to Graph Neural Networks. *Distill*.

[Sugiyama and Kan 2010] 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.

[Sun et al. 2019] 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 et al. 2013] 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 et al. 2022] 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.

[Wang et al. 2019] 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 et al. 2019] 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 et al. 2021] 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.

[Yang et al. 2020] 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.

[Zhang et al. 2019] 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.