



INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE GOIÁS
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
CÂMPUS ANÁPOLIS
ESPECIALIZAÇÃO EM INTELIGÊNCIA ARTIFICIAL APLICADA

JOÃO VITOR FELIPE DOS SANTOS

IFG-Produz

**Análise e Caracterização de Uma Nova Base de
Dados Pública para Sistemas de Recomendação
de Artigos Científicos**

Anápolis, 2025.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE GOIÁS
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
CÂMPUS ANÁPOLIS
ESPECIALIZAÇÃO EM INTELIGÊNCIA ARTIFICIAL APLICADA

JOÃO VITOR FELIPE DOS SANTOS

IFG-Produz

Análise e Caracterização de Uma Nova Base de Dados Pública para Sistemas de Recomendação de Artigos Científicos

Monografia de Especialização apresentada à Coordenação do Curso de Especialização em Inteligência Artificial Aplicada como requisito parcial para obtenção do título de Especialista em Inteligência Artificial Aplicada.

Orientador: Prof. Dr. Daniel Xavier de Souza

Anápolis, 2025.

**TERMO DE AUTORIZAÇÃO PARA DISPONIBILIZAÇÃO
NO REPOSITÓRIO DIGITAL DO IFG - ReDi IFG**

Com base no disposto na Lei Federal nº 9.610/98, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia de Goiás, a disponibilizar gratuitamente o documento no Repositório Digital (ReDi IFG), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, em formato digital para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IFG.

Identificação da Produção Técnico-Científica

- | | |
|--|---|
| ☐ Tese | ☐ Artigo Científico |
| ☐ Dissertação | ☐ Capítulo de Livro |
| ☒ Monografia – Especialização | ☐ Livro |
| ☐ TCC - Graduação | ☐ Trabalho Apresentado em Evento |
| ☐ Produto Técnico e Educacional - Tipo: _____ | |

Nome Completo do Autor:

Matrícula:

Título do Trabalho:

Autorização - Marque uma das opções

1. ☒ Autorizo disponibilizar meu trabalho no Repositório Digital do IFG (acesso aberto);
2. ☐ Autorizo disponibilizar meu trabalho no Repositório Digital do IFG somente após a data ____/____/____ (Embargo);
3. ☐ Não autorizo disponibilizar meu trabalho no Repositório Digital do IFG (acesso restrito).

Ao indicar a opção **2 ou 3**, marque a justificativa:

- ☐ O documento está sujeito a registro de patente.
☐ O documento pode vir a ser publicado como livro, capítulo de livro ou artigo.
☐ Outra justificativa: _____

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O/A referido/a autor/a declara que:

- i. o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;
- ii. obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autor/a, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia de Goiás os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;
- iii. cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia de Goiás.

_____, Anápolis - GO, 25 / 03 / 2025
Local Data

Assinatura do Autor e/ou Detentor dos Direitos Autorais



INSTITUTO FEDERAL
Goiás

MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE GOIÁS
CÂMPUS ANÁPOLIS

ATA DE DEFESA PÚBLICA DE TRABALHO DE CONCLUSÃO DE CURSO

(Modalidade da Sessão: Webconferência)

No dia 11 do mês de fevereiro do ano de 2025, às 17 horas, no Programa de Pós-Graduação em Inteligência Artificial Aplicada, do Instituto Federal de Educação, Ciência e Tecnologia de Goiás (IFG) - Câmpus Anápolis, por meio de webconferência, deu-se a defesa do Trabalho de Conclusão de Curso (TCC) "**IFG-Produz: Análise e Caracterização de Uma Nova Base de Dados Pública para Sistemas de Recomendação de Artigos Científicos**", de autoria de **João Vitor Felipe dos Santos**, como requisito para a conclusão do Curso de Especialização em Inteligência Artificial Aplicada.

Sob a presidência do Orientador e Presidente da Banca, **Prof. Dr. Daniel Xavier de Souza** - IFG/Câmpus Anápolis, a Banca Examinadora teve como Avaliadores o **Prof. Dr. Otávio Calaça Xavier** - IFG/Câmpus Goiânia e o **Prof. Dr. Gustavo de Assis Costa** - IFG/Câmpus Jataí.

Em sessão pública, após a apresentação da pesquisa e dos seus resultados, assim como a defesa do Trabalho de Conclusão de Curso pelo discente, os integrantes da Banca Examinadora fizeram as suas arguições, considerações e avaliações. Depois de se reunir em sala separada para avaliação e deliberação, a Banca Examinadora retornou à sala de defesa pública para a proclamação do resultado. Assim, em conformidade com o Regulamento Geral dos Programas de Pós-Graduação *Lato Sensu* do Instituto Federal de Goiás (IFG), a Banca Examinadora manifestou-se pela **APROVAÇÃO** do Trabalho de Conclusão de Curso de **João Vitor Felipe dos Santos**.

Anápolis - GO, 11 de fevereiro de 2025.

Documento assinado eletronicamente por:

1. Prof. Dr. Daniel Xavier de Souza - IFG/Câmpus Anápolis
2. Prof. Dr. Otávio Calaça Xavier - IFG/Câmpus Goiânia
3. Prof. Dr. Gustavo de Assis Costa - IFG/Câmpus Jataí
4. João Vitor Felipe dos Santos - Discente

Documento assinado eletronicamente por:

- Otavio Calaca Xavier, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 12/02/2025 09:05:38.
- João Vitor Felipe dos Santos, 20222060180156 - Discente, em 11/02/2025 18:55:52.
- Gustavo de Assis Costa, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 11/02/2025 18:36:29.
- Daniel Xavier de Sousa, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 11/02/2025 18:33:16.

Este documento foi emitido pelo SUAP em 27/01/2025. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse <https://suap.ifg.edu.br/autenticar-documento/> e forneça os dados abaixo:

Código Verificador: 610819
Código de Autenticação: ee9ff017c0



Instituto Federal de Educação, Ciência e Tecnologia de Goiás
Avenida Pedro Ludovico, s/ nº, None, Reny Cury, ANÁPOLIS / GO, CEP 75131-457
(62) 3703-3359 (ramal: 3359)

Resumo

Título: IFG-Produz

Autor: João Vitor Felipe dos Santos

Orientador: Dr. Daniel Xavier de Souza

Este trabalho apresenta a caracterização e análise do IFG-Produz, uma nova base de dados pública disponibilizada para pesquisa em sistemas de recomendação de artigos científicos, visto que a escassez de bases de dados públicas e com informações relevantes para esse fim representa um desafio para a experimentação destes modelos. Assim, este estudo busca disponibilizar um conjunto de dados rico em informações, contendo interações entre usuários e publicações, rótulos de relevância e dados complementares, como autoria, conferências e periódicos. Foram realizados experimentos comparativos utilizando cinco modelos de recomendação: GraphRec, KGAT, SR-GNN, SASRec e BERT4Rec, além de abordagens baseadas em filtragem de conteúdo. Os resultados demonstram que o KGAT obteve melhor desempenho entre os modelos híbridos, enquanto técnicas baseadas em conteúdo, como TF-IDF e BERT, mostraram-se muito eficazes devido à representatividade dos títulos das publicações. Por outro lado, os modelos baseados em sequências apresentaram desempenho inferior, refletindo a ausência de um histórico temporal estruturado na base. As análises reforçam que o IFG-Produz é uma base de dados robusta e adequada para experimentação com diferentes técnicas de recomendação em diferentes cenários, permitindo a aplicação de algoritmos colaborativos, híbridos e baseados em conteúdo. Estudos futuros podem explorar conexões adicionais entre dados, como a relação entre palavras-chave e áreas de pesquisa, para ampliar a aplicabilidade do dataset em diferentes cenários de recomendação.

Palavras-chave

Sistemas de recomendação; IFProduz; Base de dados; Recomendação de artigos científicos; Modelos baseados em grafos; Inteligência artificial aplicada.

Abstract

Title: IFG-Produz: Analysis and Characterization of a New Public Dataset for Scientific Article Recommendation Systems

Author: João Vitor Felipe dos Santos

Advisor: Dr. Daniel Xavier de Souza

This study presents the analysis and characterization of IFG-Produz, a new public dataset designed for research in scientific article recommendation systems given that the scarcity of public datasets with relevant information poses a challenge for the experimentation of these models. Thus, this study aims to provide a dataset rich in information, including user-publication interactions, explicit relevance evaluations, and complementary metadata such as authorship, conferences, and journals. Experiments were conducted with five recommendation models—GraphRec, KGAT, SR-GNN, SASRec, and BERT4Rec—alongside content-based filtering approaches. Results indicate that KGAT achieved the best performance among hybrid models, while content-based methods, such as TF-IDF and BERT, proved highly effective due to the strong representativeness of publication titles. In contrast, sequential models underperformed, likely due to the lack of structured temporal history in the dataset. The findings reinforce that IFG-Produz is a robust and suitable dataset for experimentation with various recommendation techniques and scenarios, enabling the application of collaborative, hybrid, and content-based algorithms. Future studies may explore additional connections between data, such as the relationship between keywords and research areas, to expand the dataset’s applicability in different recommendation scenarios.

Keywords

Recommendation systems; IFProduz; Dataset; Scientific article recommendation; Graph-based models; Applied artificial intelligence

Lista de Figuras

Figura 1	– Diagrama estrutural da base de dados IFG-Produz.....	22
Figura 2	– Porcentagem de usuários por intervalo de interações (Comparativo).....	26
Figura 3	– Porcentagem de publicações por intervalo de interações (Comparativo).	28
Figura 4	– Quantidade de notas por frequência na base IFG-Produz.	29
Figura 5	– Word Cloud dos Títulos presentes na base AAN+DBLP	31
Figura 6	– Word Cloud dos Títulos presentes na base CiteULike.....	32
Figura 7	– Word Cloud dos Títulos presentes na base IFG-Produz.....	32
Figura 8	– Distribuição das publicações por idioma na base de dados IFG-Produz	33
Figura 9	– Palavras-chave mais frequentes por usuário na base IFG-Produz	34
Figura 10	– Palavras-chave mais relevantes por usuário na base IFG-Produz	35
Figura 11	– Relação de número de publicações por área de atuação (revistas)	36
Figura 12	– Proporção de tipos de publicação avaliadas na base IFG-Produz (à direita) e Média de notas por tipo de avaliação (à esquerda).	38
Figura 13	– Distribuição de anos de produção nas bases do IFG-Produz e AAN+DBLP.....	39
Figura 14	– Distribuição de publicações por usuário na base do IFG-Produz	40
Figura 15	– Word Cloud das Tags presentes na base CiteULike.....	41
Figura 16	– Distribuição de tags por publicação na base CiteULike	41
Figura 17	– Exemplo de estrutura da matriz de pesquisadores utilizada para alimentar o modelo GraphRec	45
Figura 18	– Exemplo de estrutura da matriz de publicações utilizada para alimentar o modelo GraphRec	46
Figura 19	– Arquitetura do modelo GraphRec.....	46
Figura 20	– Exemplo da estrutura de grafo alimentada no CKG para o ambiente IFG-Produz	49
Figura 21	– Arquitetura do modelo KGAT	51
Figura 22	– Arquitetura do modelo SR-GNN.....	53
Figura 23	– Exemplo de grafo de sessão e matriz de pesos geradas pelo modelo SR-GNN.....	54
Figura 24	– Arquitetura do modelo SASRec	56

Figura 25 – Arquitetura do modelo BERT4Rec	59
--	----

Lista de Tabelas

Tabela 1	– Avaliação de Disponibilidade de Datasets	20
Tabela 2	– Métricas principais das bases de dados utilizadas nos experimentos.....	24
Tabela 3	– Resumo das conexões por usuário das bases IFG-Produz, CiteULike e AAN+DBLP.....	25
Tabela 4	– Resumo das conexões por publicação nas bases IFG-Produz, CiteULike e AAN+DBLP.....	27
Tabela 5	– Resumo das avaliações relevantes na base IFG-Produz.....	28
Tabela 6	– Resumo das informações presentes na base IFG-Produz considerando as conexões de autoria.	30
Tabela 7	– Resumo das métricas das palavras-chave.	33
Tabela 8	– Resumo das métricas das revistas.....	35
Tabela 9	– Resumo das métricas das conferências.....	36
Tabela 10	– Resumo das métricas dos projetos de pesquisa.....	37
Tabela 11	– Resumo das métricas das publicações bibliográficas.	37
Tabela 12	– Resumo das palavras-chave na base CiteULike.	42
Tabela 13	– Hiperparâmetros utilizados para cada algoritmo nos diferentes bases de dados	65
Tabela 14	– Resultados dos algoritmos nos experimentos com as bases IFG-Produz, AAN+DBLP e CiteULike	68
Tabela 15	– Resultados dos algoritmos nos experimentos com a base IFG-Produz acrescidos de conexões adicionais de autoria	70
Tabela 16	– Resultados dos algoritmos no experimento com a base IFG-Produz acrescido de conexões adicionais das publicações	72

Lista de Abreviaturas e Siglas

AAN	<i>ACL Anthology Network</i>
BERT4Rec	<i>Bidirectional Encoder Representations from Transformers for Recommendation</i>
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CKG	<i>Collaborative Knowledge Graph</i>
DBLP	<i>Digital Bibliography & Library Project</i>
KGAT	<i>Knowledge Graph Attention Network</i>
RNN	<i>Recurrent Neural Network</i>
SASRec	<i>Self-Attentive Sequential Recommendation</i>
SR-GNN	<i>Session-based Recommendation with Graph Neural Networks</i>

Sumário

1	Introdução	12
2	Revisão Bibliográfica	15
2.1	Sistemas de Recomendação	15
2.1.1	Filtragem baseada em conteúdo	16
2.1.2	Filtragem colaborativa e Híbrida	16
2.2	Recomendação de artigos científicos.....	17
3	Caracterização de Bases de Dados para Recomendação de Artigos Científicos	19
3.1	Bases de Dados Disponibilizadas e Avaliadas	21
3.2	Caracterização Numérica	24
3.3	Avaliação dos Valores de Relevância (rótulos)	27
3.3.1	Variações da base IFG-Produt	29
3.4	Informações adicionais presentes nas bases	30
3.4.1	Títulos das publicações	30
3.4.2	Perfil dos usuários	33
3.4.3	Conferências e Revistas	34
3.4.4	Publicações e Projetos de Pesquisa	36
3.4.5	Ano de publicação.....	37
3.4.6	Autoria e Co-Autoria	39
3.4.7	Categorias e Tags	40
4	Metodologia	43
4.1	GraphRec	44
4.1.1	Descrição do Algoritmo	44
4.2	Knowledge Graph Attention Network	47
4.2.1	Dados de entrada	48
4.2.2	Modelo KGAT	49
4.3	Session-based Recommendation with Graph Neural Networks (SR-GNN)	52
4.4	Self-Attentive Sequential Recommendation (SASRec)	55
4.5	Bidirectional Encoder Representations from Transformers for Recommendation	58
5	Experimentação e Resultados	63
5.1	Dados utilizados e pré-processamento.....	63
5.2	Metodologia utilizada	64
5.3	Métricas de Avaliação e Treinamento dos Modelos	66

5.4	Resultados Gerais	67
5.4.1	Primeiro experimento: Análise geral dos modelos e bases de dados	67
5.4.2	Segundo experimento: IFG-Produz e conexões adicionais de autoria	70
5.4.3	Terceiro experimento: KGAT e conexões adicionais de publicações	72
5.5	Resumo das análises e conclusões	73
6	Conclusão	76
	Referências	78

Introdução

Nos últimos anos, os sistemas de recomendação tornaram-se ferramentas essenciais em diversos setores, desde o entretenimento até a pesquisa acadêmica (Sanchez-Lengeling *et al.*, 2021). No que diz respeito à propagação de conhecimento científico, a recomendação de artigos é fundamental para otimizar o acesso a informações, ajudando pesquisadores a identificar novas publicações de interesse dentro de seus campos acadêmicos. Contudo, a falta de bases de dados apropriadas para o desenvolvimento e avaliação desses sistemas ainda é um desafio (Kreutz e Schenkel, 2022).

Considerando esse contexto, o objetivo do presente estudo é contribuir com a proposição de uma nova base de dados para sistemas de recomendação e sua avaliação a partir dos resultados da aplicação em diversos modelos de recomendação, buscando analisar a influência da estrutura e conteúdo dos dados disponíveis na performance desses algoritmos.

Para tal, o estudo apresenta a base de dados IFG-Produz, um novo conjunto de dados públicos direcionados a sistemas de recomendação de artigos científicos. A base foi estruturada para fornecer informações descritivas sobre as publicações e perfis dos pesquisadores registrados, além de oferecer conexões entre usuários e publicações, incluindo informações sobre autoria, coautoria e rótulos de relevância baseados em avaliações prévias.

O trabalho também realiza uma análise descritiva e comparativa entre a base IFG-Produz e outros conjuntos de dados amplamente utilizados, como CiteULike (Wang, Chen e Li, 2013) e AAN/DBLP (Sugiyama e Kan, 2010), destacando suas principais características e diferenças estruturais. Para tal, avaliamos diversos aspectos, como o volume de interações entre usuários e publicações, a densidade das conexões, a representatividade dos dados disponíveis e a diversidade dos conteúdos armazenados em cada base de dados.

Essa análise permite a compreensão da distribuição das avaliações e interações dos usuários em cada base de dados, evidenciando padrões que podem impactar a performance de diferentes algoritmos de recomendação. Enquanto o CiteULike é

caracterizado por uma estrutura densa com múltiplas avaliações por publicação, a base AAN+DBLP apresenta um cenário mais esparsa, baseado em relações de autoria. Nesse contexto, a base de dados IFG-Produz se diferencia por fornecer não apenas avaliações explícitas dos usuários (rótulos), mas também um conjunto mais amplo de informações adicionais, como categorias temáticas, conferências, periódicos e informações descritivas do perfil de cada pesquisador.

Além desta análise descritiva, diferentes modelos de recomendação foram aplicados a estas bases de dados, permitindo uma avaliação inicial da viabilidade e potencial da base de dados disponibilizada para uso em diversos cenários de recomendação. Para isso, foram testados alguns modelos de recomendação baseados em grafos, com abordagens baseadas em filtragem colaborativa, abordagem híbrida e técnicas de *self-attention*, como GraphRec (Rashed, Grabocka e Schmidt-Thieme, 2019), KGAT (Wang *et al.*, 2019), SR-GNN (Wu *et al.*, 2019), SASRec (Kang e McAuley, 2018) e BERT4Rec (Sun *et al.*, 2019).

Os experimentos conduzidos avaliaram o desempenho desses algoritmos quanto a acurácia das sugestões, a habilidade de ordenar itens pertinentes e a robustez frente a cenários com alta esparsidade de dados. Adicionalmente, utilizamos métodos de filtragem de conteúdo já consolidados, tais como TF-IDF e BERT, para confirmar a representatividade das informações textuais na construção dos perfis dos usuários.

Esta análise experimental mostra que a base de dados sugerida não só proporciona um conjunto de informações bem estruturadas para experimentação acadêmica, mas também é promissora para a aplicação e experimentação de diferentes modelos de recomendação.

A conclusão de todas essas avaliações indica fortemente que a base de dados do IFG-Produz possui estrutura adequada para apoiar a implementação de várias metodologias de recomendação, sendo particularmente eficaz para modelos híbridos e filtragem baseada em conteúdo.

Sendo assim, este estudo busca contribuir para futuras pesquisas em sistemas de recomendação acadêmicos, disponibilizando uma nova base de dados robusta e bem estruturada, permitindo assim a experimentação de algoritmos mais complexos e possivelmente avanços na modelagem de algoritmos de recomendação mais precisos e eficazes.

O trabalho é dividido em capítulos, onde: o Capítulo 2 apresenta um breve referencial teórico; o Capítulo 3 apresenta as análises realizadas nas bases de dados apresentadas; o Capítulo 4 apresenta de forma detalhada os algoritmos utilizados na avaliação dessas bases de dados; o Capítulo 5 apresenta e discute os resultados obtidos; e, por fim, o Capítulo 6 resume essas análises e sugere futuros estudos para

aprofundar o entendimento da base de dados disponibilizada.

Revisão Bibliográfica

O avanço dos sistemas de recomendação tem exercido uma função essencial na customização da experiência do usuário nas mais diversas aplicações, desde plataformas de streaming até e-commerces. Estes sistemas de recomendação são concebidos com o objetivo de sugerir conteúdo relevante com base no comportamento e preferências dos usuários (Cazella, Nunes e Reategui, 2010).

Segundo Sanchez-Lengeling *et al.* (2021), outro aspecto muito relevante na análise desses sistemas é o contexto de aplicação, que influencia diretamente as características do modelo a ser adotado. Cada contexto apresenta desafios e necessidades específicas que moldam o processo de recomendação. Essa distinção ocorre porque tanto os atributos dos usuários quanto as propriedades dos itens variam conforme o cenário, exigindo abordagens personalizadas para capturar esses relacionamentos e otimizar as sugestões.

Considerando o objetivo deste trabalho – a proposição de uma nova base de dados para sistemas de recomendação e sua análise em diferentes modelos de recomendação – o embasamento teórico que é apresentado neste capítulo serve como suporte para a compreensão das técnicas e metodologias adotadas para a avaliação da base de dados.

2.1 Sistemas de Recomendação

Os sistemas de recomendação são modelos projetados com o objetivo de sugerir informações relevantes a usuários, com base em suas preferências, comportamentos anteriores ou nas preferências de outros usuários com perfis semelhantes. Como descrito por Cazella, Nunes e Reategui (2010), esses sistemas desempenham um papel crucial em serviços que envolvem a distribuição e comercialização de produtos ou conteúdo, estando presentes em diversas plataformas amplamente utilizadas nos dias de hoje.

Ainda segundo Cazella, Nunes e Reategui (2010), esses sistemas geralmente são classificados, de acordo a abordagem adotada e sua finalidade, em três categorias principais: filtragem colaborativa, filtragem baseada em conteúdo e filtragem híbrida.

2.1.1 Filtragem baseada em conteúdo

Aggarwal (2016) descreve a abordagem de filtragem baseada em conteúdo como aquelas que concentram-se principalmente na similaridade entre os itens presentes no conteúdo disponível. Nesse caso, as recomendações são feitas com base na relação entre itens previamente consumidos ou avaliados por um usuário e novos itens disponíveis, sugerindo conteúdos semelhantes aos que aqueles usuários já consumiram no passado.

Esse tipo de abordagem é bastante utilizada e geralmente obtém um melhor desempenho em ambientes onde existem muitas informações representativas de atributos disponíveis, como palavras-chave ou outras informações textuais (Aggarwal, 2016).

Geralmente, os algoritmos baseados nessas abordagens trabalham em três etapas específicas. A primeira é o pré-processamento e extração de atributos, onde os atributos mais representativos dos itens de uma base de dados são extraídos e transformados em *embeddings* (Aggarwal, 2016).

O segundo passo envolve a construção de um modelo com o perfil dos usuários, utilizando as representações dos itens com os quais eles interagiram previamente. Por fim, a última etapa envolve a utilização do modelo aprendido para recomendar novos itens para esses usuários (Aggarwal, 2016).

2.1.2 Filtragem colaborativa e Híbrida

Já a filtragem colaborativa, segundo Cazella, Nunes e Reategui (2010), parte do princípio de que usuários com perfis semelhantes tendem a consumir conteúdos parecidos. Assim, ao analisar itens avaliados ou consumidos por um grupo de usuários, o sistema pode recomendar itens relevantes a outros usuários semelhantes.

Essa construção de perfis tende a acontecer principalmente criando representações dos usuários a partir de interações anteriores, utilizando rótulos de interação ou algum grau de relevância dessas interações. Dessa forma, essas abordagens diferem das baseadas em conteúdo por não exigirem o conhecimento do conteúdo dos itens avaliados (Cazella, Nunes e Reategui, 2010).

Ainda segundo Cazella, Nunes e Reategui (2010), a filtragem híbrida busca combinar os pontos fortes das duas abordagens anteriores. Ao considerar simultaneamente a similaridade entre perfis de usuários e a relação entre os itens

consumidos por cada indivíduo, essa abordagem visa aumentar a precisão das recomendações.

Além dessas categorias tradicionais, os avanços atingidos nos sistemas de recomendação têm introduzido novas abordagens, como, por exemplo, a filtragem baseada em sessões (Wu *et al.*, 2019), a filtragem demográfica e a filtragem baseada em conhecimento, o que amplia as possibilidades de personalização (Cazella, Nunes e Reategui, 2010).

2.2 Recomendação de artigos científicos.

Segundo Alzoghbi *et al.* (2015) e L, Liu e Pan (2021), o aumento significativo na quantidade e na velocidade de publicação de novos artigos científicos torna o processo de encontrar conteúdos relevantes um desafio para pesquisadores de diversas áreas. A busca por artigos relacionados a um tema, seja por palavras-chave ou com base em publicações anteriores de um autor, consome cada vez mais tempo dada a vasta quantidade de opções disponíveis.

Para enfrentar esse desafio, uma solução comum e promissora é a aplicação de sistemas de recomendação dentro das ferramentas de busca acadêmica e redes de conexão de pesquisadores (L, Liu e Pan, 2021). Embora essas abordagens compartilhem princípios com sistemas de recomendação utilizados em outros contextos, como streaming de conteúdo, a recomendação de artigos científicos apresenta especificidades únicas adaptadas ao ambiente acadêmico.

Segundo os autores, a grande quantidade de publicações científicas disponíveis nestas ferramentas de busca acadêmica impulsionou a adoção de sistemas de recomendação, especialmente aqueles baseados em filtragem colaborativa. No entanto, a alta esparsidade dos dados presentes nestas plataformas compromete o desempenho dos modelos.

Em outros casos, conforme descrito por Alzoghbi *et al.* (2015), são aplicados modelos baseados em conteúdo ou híbridos que utilizam o texto das publicações para realizar as recomendações. Porém, esses modelos também sofrem dentro do contexto de recomendação para artigos científicos, visto que muitas das informações mais representativas das publicações estão no texto completo, que raramente é disponibilizado pelas bases de dados.

Para mitigar esses problemas, muitos pesquisadores têm investido no desenvolvimento de metodologias e técnicas de coleta e introdução de informações adicionais auxiliares, capazes de enriquecer o processo de recomendação e melhorar os resultados desses modelos (L, Liu e Pan, 2021).

A correlação de relevância entre itens e usuários é um exemplo de informação adicional, já amplamente utilizada em sistemas de recomendação, como em plataformas de streaming, onde usuários frequentemente realizam avaliações ativas, facilitando a personalização das sugestões. No entanto, conforme apontado por Kreutz e Schenkel (2022), a maioria das bases de dados para recomendação de artigos científicos carece de registros que estabeleçam essa correlação direta entre usuários e publicações em termos de avaliação ou relevância, se baseando principalmente em autoria para gerar essa conexão.

Em alguns cenários, essa relação de relevância precisa ser inferida a partir de outras fontes de informação, como autoria, principalmente quando não há rótulos de avaliação disponíveis. Além disso, diversas abordagens se utilizam de informações e métodos adicionais, como a comparação semântica do conteúdo dos artigos, a análise de citações, o estudo de autoria e coautoria, além de características como o nível de experiência dos pesquisadores envolvidos. Essas abordagens visam superar as limitações dos dados disponíveis, permitindo a criação de sistemas mais eficazes e adaptados às necessidades específicas da pesquisa acadêmica (Kreutz e Schenkel, 2022).

Ainda segundo Kreutz e Schenkel (2022), outro desafio significativo no campo dos sistemas de recomendação de artigos científicos é a escassez de bases de dados públicas adequadas para análises e experimentações. Embora algumas bases, como o CiteULike (Wang, Chen e Li, 2013), sejam amplamente conhecidas e utilizadas, a maioria dos pesquisadores utiliza bases privadas, o que dificulta a replicação de experimentos, a validação de novas metodologias, a realização de análises comparativas e o desenvolvimento de melhorias efetivas.

Considerando os conceitos, metodologias e desafios apresentados, estabelecemos a base teórica tanto para a proposição da nova base de dados quanto para as análises realizadas com os modelos de recomendação baseados em grafos. No próximo capítulo, apresentaremos em detalhes a base de dados IFG-Produz, a qual buscamos disponibilizar através desta pesquisa, além das demais bases de dados utilizadas nas análises realizadas.

Caracterização de Bases de Dados para Recomendação de Artigos Científicos

Neste capítulo, apresentamos as bases de dados utilizadas no desenvolvimento desta pesquisa. Aqui aplicamos ênfase na disponibilização do IFG-Produz, uma nova base de dados criada para a aplicação em experimentos com recomendação de artigos científicos. Destacamos suas características, limitações e o processo de seleção, além de justificar sua relevância no contexto da recomendação. Também discutimos brevemente outras bases utilizadas para fins comparativos, como o CiteULike-a e o novo AAN+DBLP.

Apesar do crescente interesse da comunidade acadêmica, a acessibilidade a bases de dados de recomendação de artigos científicos ainda representa um grande desafio, principalmente no que se diz respeito a bases de dados que apresentam como rótulo as avaliações dos usuários.

Muitas das bases mencionadas em estudos anteriores não se encontram disponíveis nos repositórios originalmente indicados, observados durante o período de realização deste trabalho, dificultando sua utilização. Em especial, para a recomendação de artigos científicos, essa limitação reflete como um forte obstáculo na pesquisa sobre sistemas de recomendação, comprometendo a comparação entre abordagens e a validação de resultados.

Os autores Kreutz e Schenkel (2022) destacam esse problema ao analisar mais de 65 trabalhos no contexto de recomendação de artigos científicos, publicados entre janeiro de 2019 e outubro de 2021. Apenas nove desses estudos disponibilizaram bases de dados publicamente e que continham algum rótulo de relevância, das quais muitas já não estão mais acessíveis.

A Tabela 1 apresenta um levantamento detalhado sobre as bases de dados encontradas durante esta pesquisa, indicando sua disponibilidade pública e manutenção, incluindo as analisadas por Kreutz e Schenkel (2022).

As bases de dados foram avaliadas considerando sua acessibilidade e a atualização contínua. Casos de indisponibilidade incluem: i) links não mais acessíveis,

apesar de citados em publicações, ou ii) ausência de links ou informações que permitam acesso aos arquivos das bases de dados nas publicações em que foram citados.

Como observado, apenas quatro bases de recomendação de artigos científicos foram identificadas como disponíveis a partir dos meios disponibilizados pelos autores que as utilizaram.

Tabela 1 – *Avaliação de Disponibilidade de Datasets*

Dataset	Disponibilidade Pública em 12/2024	Manutenção em 12/2024
AAN original	Disponível	Sim
BBC	Indisponível	Não
BibSonomy	Indisponível	Sim
Citeulike-a	Disponível	Não
Citeulike-t	Disponível	Não
DBLP + Citations v1	Indisponível	Não
DBLP + Citations v11	Indisponível	Não
DOCEAR	Indisponível	Não
PRSDataset	Indisponível	Não
SPRD	Indisponível	Não
SPRD Senior	Indisponível	Não

Fonte: Desenvolvido pelo autor.

Além das bases de dados apresentadas em Kreutz e Schenkel (2022), encontramos também referências para as bases BibSonomy e DOCEAR, igualmente descritas na tabela 1.

A base BibSonomy é gerada a partir de snapshots gerados pela plataforma de mesmo nome, desde 2006, pelo Knowledge and Data Engineering Group da University of Kassel. Estes snapshots são disponibilizados de 6 em 6 meses por meio do site da universidade¹, porém o acesso desses arquivos requer o envio de um formulário e um termo de aceite via e-mail. Embora o formulário tenha sido devidamente enviado, nenhuma resposta foi recebida até a conclusão deste estudo.

Já a base DOCEAR (Beel *et al.*, 2017) não fornece instruções claras nem um link direto de acesso em sua publicação original, apesar de ser declarada de acesso

¹<https://www.kde.cs.uni-kassel.de/wp-content/uploads/bibsonomy/>

público. Tentativas de contato com os autores por e-mail foram realizadas, mas, até o momento da finalização deste trabalho, não houve retorno. O site da plataforma² também não indica nenhum meio de obter a base de dados, e informa que a mesma não é mantida atualmente.

Diante dessa escassez de bases de dados acessíveis para tarefas de recomendação de artigos científicos, reforçamos a contribuição deste estudo com a disponibilização e avaliação da base IFG-Produz, que fornece interações detalhadas e estruturadas para a recomendação de artigos científicos, que serão descritas em detalhes no subtópico a seguir.

Para garantir a consistência das nossas análises e comparação com a base de dados proposta, selecionamos outras duas bases de dados: CiteULike-a (Wang, Chen e Li, 2013) e AAN+DBLP (Sugiyama e Kan, 2010). Além da comparação dos dados presentes nas bases de dados, foi realizada uma análise do comportamento de cada um deles dentro de diferentes algoritmos de recomendação, os quais estarão descritos em mais detalhes no Capítulo 4.

3.1 Bases de Dados Disponibilizadas e Avaliadas

A base IFG-Produz foi coletada para este trabalho a partir de dados da plataforma de mesmo nome, disponibilizada pelo Instituto Federal de Goiás (IFG) para esta pesquisa.

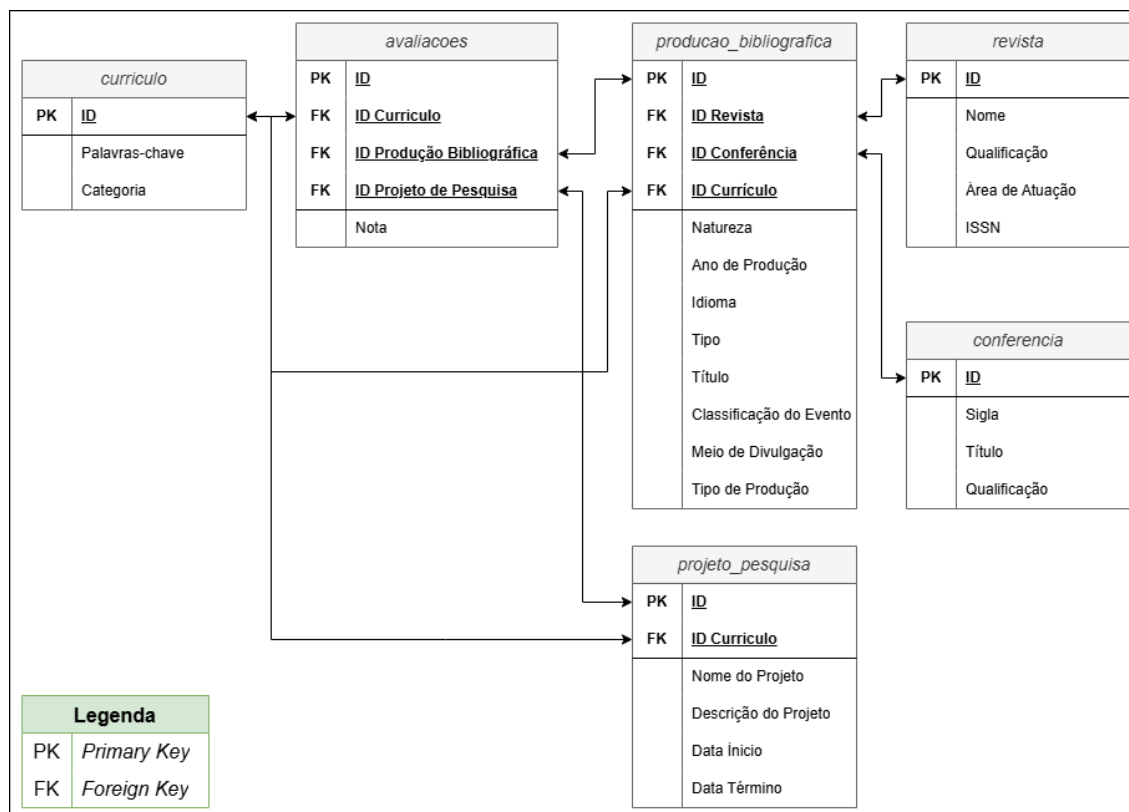
A plataforma foi desenvolvida com o propósito de conectar pesquisadores da instituição, facilitando a divulgação e avaliação dos trabalhos. Além disso, a plataforma é capaz de utilizar as avaliações e produções acadêmicas anteriores dos pesquisadores para recomendar novos conteúdos relevantes.

Além das publicações, a base também registra informações sobre projetos de pesquisa desenvolvidos, como ano de início e término, título e descrição. No contexto deste trabalho, tanto os artigos publicados quanto os projetos de pesquisa são caracterizados como publicações, e também tratados como itens.

Em adição a estas informações descritivas específicas de cada produção, a base inclui dados sobre conferências e periódicos científicos associados às publicações. No caso das conferências, são armazenados o título, a sigla e a avaliação Qualis registrada pela Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Para os periódicos, são registrados o nome, a qualificação, a área de atuação e o ISSN.

²<https://docear.org/>

Figura 1 – Diagrama estrutural da base de dados IFG-Produz.



Fonte: Desenvolvido pelo autor.

Por fim, a base também contém registros das avaliações (rótulos) realizadas pelos pesquisadores sobre as recomendações de publicações e projetos feitas pela plataforma. Ou seja, a plataforma IFG-Produz recomenda trabalhos aos pesquisadores com base em seus perfis de publicações e avaliações anteriores. Após a recomendação, o pesquisador pode avaliar a sugestão recebida com uma nota de 0 a 5.

A Figura 1 representa a estrutura geral dos dados presentes na base de dados, bem como as conexões entre cada uma das entidades presentes. A diversidade de informações armazenadas, acrescida dos rótulos de relevância, torna a base de dados IFG-Produz única e extremamente rica para exploração, especialmente no contexto de sistemas de recomendação.

A segunda base de dados avaliada em nosso trabalho é a CiteULike-a, disponibilizada por Wang, Chen e Li (2013), é amplamente utilizada na literatura sobre recomendação de artigos científicos devido à sua estrutura consistente e abrangente, como apresentado no trabalho de Kreutz e Schenkel (2022).

CiteULike foi uma plataforma de marcadores sociais para artigos científicos e uma rede social acadêmica que oferecia serviços gratuitos para descoberta, organização e compartilhamento de referências bibliográficas. Iniciada em 2004 pela

University of Manchester, foi bastante utilizada por pesquisadores, acadêmicos e estudantes, permitindo o armazenamento, marcação e compartilhamento de artigos acadêmicos, livros e outros materiais relevantes para suas áreas de interesse. Além disso, oferecia recursos como tags, estatísticas de uso e sugestões de artigos (Wang, Chen e Li, 2013).

Entretanto, de acordo com um comunicado da equipe de manutenção da plataforma, disponível em arquivo³, o CiteULike foi encerrado em março de 2019 devido a dificuldades financeiras e desafios na manutenção da base de dados, tornando inacessíveis os dados previamente disponibilizados pela fonte original.

Os autores Wang, Chen e Li (2013) disponibilizaram na plataforma Github⁴ uma versão da base, composta por dados coletados por Wang e Blei (2011), acrescidos de tags, baseados em snapshots do banco de dados original de 2011, denominada citeulike-a. Nesta pesquisa, tratamos CiteULike como este recorte da base analisado.

Essa base contém informações organizadas em três arquivos principais: um contendo as conexões entre usuários e publicações, onde a existência de uma conexão indica que o usuário marcou a publicação como relevante; outro com todas as tags ou palavras-chave; e um terceiro que associa as publicações às suas respectivas tags.

Dentro desse contexto, os usuários não possuem informações adicionais disponíveis, sendo identificados apenas por um ID. Já os artigos, além das tags, possuem título e resumo armazenados separadamente.

A terceira base de dados utilizada em nosso trabalho é baseada na base SPRD ((Sugiyama e Kan, 2010)), que foi disponibilizada publicamente pelos autores, mas encontrava-se indisponível nos links fornecidos no período de realização desta pesquisa. Entramos em contato com os autores e, a partir das orientações fornecidas por Sugiyama e Kan (2010) via e-mail, conseguimos organizar uma nova base, denominada aqui como AAN+DBLP. Essa base utiliza dados disponíveis na *ACL Anthology Network* (AAN) e na rede bibliográfica *Digital Bibliography & Library Project* (DBLP) .

A AAN é uma rede que conecta artigos, autores e citações na área de linguística computacional. Ela consiste em três sub-redes principais: relações de citação entre artigos, colaboração entre autores e citações de autores. Originalmente, a base continha 24.766 artigos e 124.857 citações. Como não possui uma métrica explícita de relevância ou interação direta entre usuários e itens, trabalhos que utilizam essa base frequentemente tratam as relações de citação, em substituição

³<https://web.archive.org/web/20190310145602/http://www.citeulike.org/news/>

⁴<https://github.com/js05212/citeulike-a>

aos rótulos de relevância. Por exemplo, Sugiyama e Kan (2010) definem os autores como usuários e utilizam a relação de autoria ou coautoria com um artigo como um vínculo a ser predito.

A base SPRD (Sugiyama e Kan, 2010), seguindo esse padrão, utilizou 597 artigos publicados na *ACL Anthology Network* entre 2000 e 2006. Os perfis dos usuários foram criados a partir dos autores desses artigos, com os trabalhos adicionais de cada autor recuperados pela ferramenta DBLP. Nesse contexto, a relevância de um item foi definida pela autoria ou coautoria: um trabalho publicado por um usuário foi considerado relevante para ele.

Baseados nesses procedimentos, replicamos o processo descrito, ampliando a quantidade de artigos para alcançar um volume de dados mais esparsos, adequado a um cenário real de recomendação. Conforme apresentado na Tabela 2, foram incluídos 30.000 artigos publicados na ACL entre 2006 e 2009. A base de dados resultante contém 51.645 usuários e 76.699 relações de autoria (ou coautoria) entre usuários e itens, oferecendo um panorama mais robusto de relevância dentro desse contexto.

Além disso, a base mantém os nomes dos pesquisadores como atributo adicional dos usuários e, para os artigos, preserva título, ano e revista de publicação, enriquecendo a base de dados com informações úteis para a recomendação.

Dataset	Usuários	Publicações	Conexões	Densidade	Área do Conhecimento
AAN + DBLP	51645	30000	76699	0.005	Linguística Computacional
CiteULike	5551	16980	204986	0.217	Diversas
IFG-Produz	2261	71895	19211	0.177	Diversas

Tabela 2 – *Métricas principais das bases de dados utilizadas nos experimentos*

Fonte: Desenvolvido pelo autor.

3.2 Caracterização Numérica

A descrição numérica das bases de dados analisadas pode ser encontrada na Tabela 2. Dentro desse contexto, o número total de usuários e itens representa a quantidade mantida nas bases de dados, mas não necessariamente a quantidade de usuários e publicações que possuem avaliações relevantes. Já as conexões indicam as relações entre usuários e publicações, sendo avaliações de itens recomendados no IFG-Produz, relevância no CiteULike e autoria no AAN+DBLP.

A densidade, também ilustrada na tabela, indica a relação entre as conexões presentes e o total de conexões possíveis, levando em conta apenas usuários e publicações que têm pelo menos uma interação documentada. Em outras palavras, um valor elevado sinaliza uma maior interação entre usuários e itens, enquanto um valor mais baixo indica uma estrutura de dados mais dispersa, com menos conexões em relação ao tamanho total da base.

Uma densidade reduzida sugere que os usuários interagiram com um número limitado de publicações ou que as publicações receberam avaliações de um número limitado de usuários. Esse cenário pode representar um desafio para sistemas de recomendação baseados em filtragem colaborativa ou híbrida, pois a escassez de interações dificulta a construção de perfis representativos dos usuários (L, Liu e Pan, 2021). Com menos conexões disponíveis, torna-se mais difícil estabelecer similaridades entre usuários e itens, o que pode comprometer a eficácia das recomendações.

A distribuição das conexões por usuário também fornece *insights* relevantes. Nesta análise, consideramos apenas os usuários que possuem alguma conexão relevante com alguma das publicações. Como indicado na Figura 2, mais da metade dos usuários da base do IFG-Produz realizou entre uma e cinco avaliações, enquanto, segundo a Tabela 3, a média é de 6.75 avaliações por usuário, com um mínimo de 1 e máximo de 178 publicações avaliadas por um único pesquisador, e um desvio padrão de 11.17. Esses dados evidenciam um padrão de interação concentrado em poucos itens, o que pode impactar a modelagem dos perfis de usuários nos sistemas de recomendação.

Métrica	IFG-Produz	CiteULike	AAN+DBLP
Número total de usuários com conexões	1408	5551	51645
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)

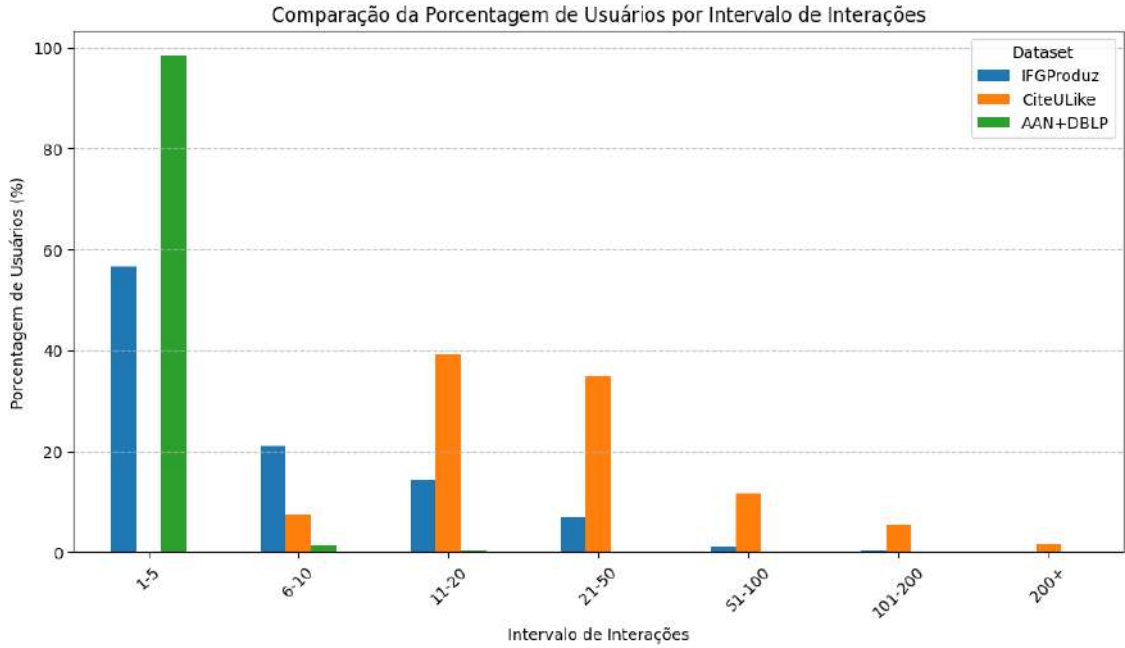
Tabela 3 – *Resumo das conexões por usuário das bases IFG-Produz, CiteULike e AAN+DBLP.*

Fonte: Desenvolvido pelo autor.

Já na base do CiteULike, ainda segundo o analisado na Tabela 3, cada usuário realizou no mínimo 10 avaliações, com uma média de aproximadamente 37 avaliações por usuário. A Figura 2 revela que cerca de 20% dos usuários realizaram mais de 50 avaliações.

Por fim, em relação aos autores incluídos na base de dados AAN+DBLP, a Tabela 3 e a Figura 2 mostram que cada autor possui pelo menos uma publicação,

Figura 2 – *Porcentagem de usuários por intervalo de interações (Comparativo).*



Fonte: Desenvolvido pelo autor.

com mais de 95% possuindo conexões com no máximo 5 publicações, resultando em uma média de aproximadamente 1,5 publicações por autor. Ou seja, ao considerar a autoria como métrica para indicar publicações relevantes para os autores, observa-se que esse número de conexões é relativamente menor do que a quantidade de avaliações realizadas por pesquisadores e usuários em outras bases de dados analisadas.

Já no caso do IFG-Produz, apesar de contar com menos usuários em comparação com as demais bases e apresentar uma densidade relativamente baixa, o *dataset* ainda demonstra boa representatividade dos pesquisadores ativos, com uma média de avaliações por usuário superior a base AAN+DBLP, por exemplo.

Novamente, é importante destacar que essa baixa densidade pode representar desafios para alguns algoritmos, especialmente os baseados em filtragem colaborativa (L, Liu e Pan, 2021). No entanto, essa característica também reflete de forma realista o padrão de interação em sistemas como o IFG-Produz, tornando-a uma fonte altamente relevante para avaliar a aplicabilidade dos modelos de recomendação em cenários práticos e menos idealizados.

Em relação às publicações, quando avaliamos a base do IFG-Produz, notamos mais sinais de sua maior esparsidade. Quase 100% das 7.710 publicações receberam entre uma e cinco avaliações, com o máximo registrado sendo nove avaliações para uma única publicação e uma média de 1.2 avaliações por publicação,

Métrica	IFG-Produz	CiteULike	AAN+DBLP
Número total de publicações com conexões	7710	16980	30000
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)

Tabela 4 – *Resumo das conexões por publicação nas bases IFG-Produz, CiteULike e AAN+DBLP.*
Fonte: Desenvolvido pelo autor.

com um desvio padrão de 0.60, como apresentado na Tabela 4. Vale destacar novamente que, para a análise, tanto de usuários quanto de publicações, a base foi filtrada para incluir apenas aqueles com pelo menos uma avaliação.

No caso do CiteULike, as conexões de publicações também mostram consistência, assim como as de usuários. De acordo com a Tabela 4, cada artigo recebeu pelo menos uma avaliação, com uma média de 12 avaliações por artigo. Assim, a estrutura bem definida, volume de interações e a diversidade de tópicos presentes na base CiteULike o tornam uma base sólida para experimentos em sistemas de recomendação.

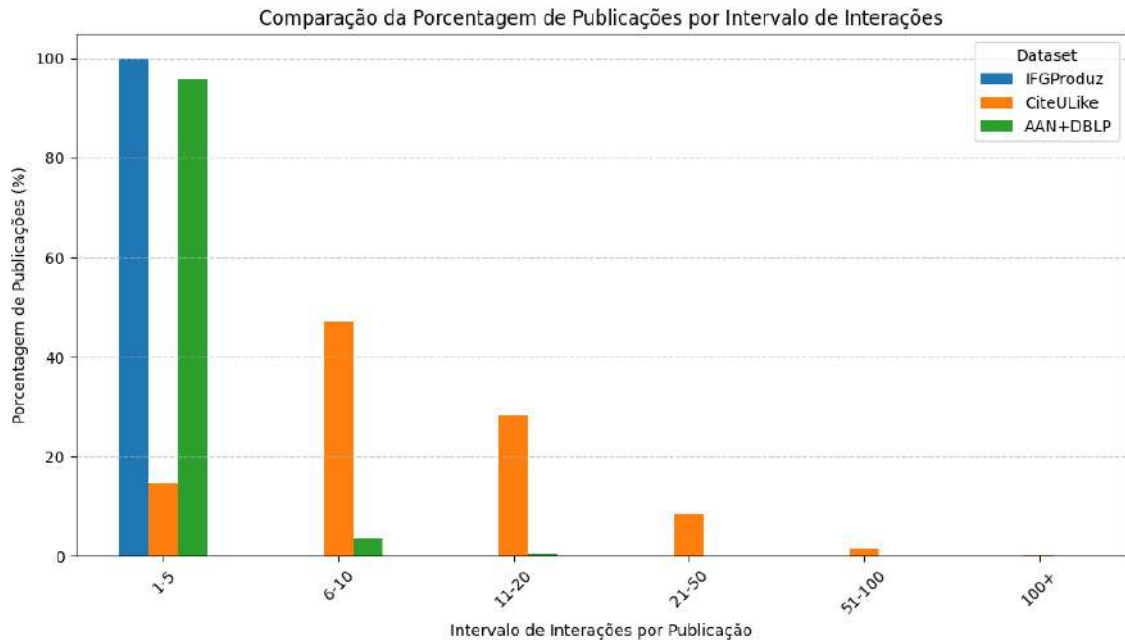
Já no AAN+DBLP, observamos novamente sinais da elevada esparsidade. Esse aspecto, aliado à menor quantidade de informações contextuais em relação a outras bases analisadas, representou desafios consideráveis para alguns modelos experimentados.

3.3 Avaliação dos Valores de Relevância (rótulos)

Como explicitado anteriormente, nas bases CiteULike e AAN+DBLP as relações entre usuários e itens são correspondentes a uma interação binária. No caso do CiteULike, essa relação indica que o usuário marcou a publicação como relevante, enquanto na AAN+DBLP indica que o usuário é autor ou coautor da publicação.

Já no caso do IFG-Produz, são armazenadas as interações de usuários com publicações de outros autores que foram recomendadas através da plataforma. Assim, é permitido aos usuários que avaliem as recomendações que foram feitas pelo sistema, avaliação esta caracterizada por uma nota que varia de 0 a 5, onde 0 caracteriza uma publicação não relevante e 5 caracteriza uma publicação muito relevante para o usuário.

No que se refere a estas avaliações, a distribuição das notas revela uma tendência significativa. A Figura 4 mostra que a maior parte das avaliações apresenta

Figura 3 – *Porcentagem de publicações por intervalo de interações (Comparativo).*

Fonte: Desenvolvido pelo autor.

Métrica	Valor
Número total de avaliações	19211
Média das notas de avaliação	4.69 (0.93)

Tabela 5 – *Resumo das avaliações relevantes na base IFG-Produz.*

Fonte: Desenvolvido pelo autor.

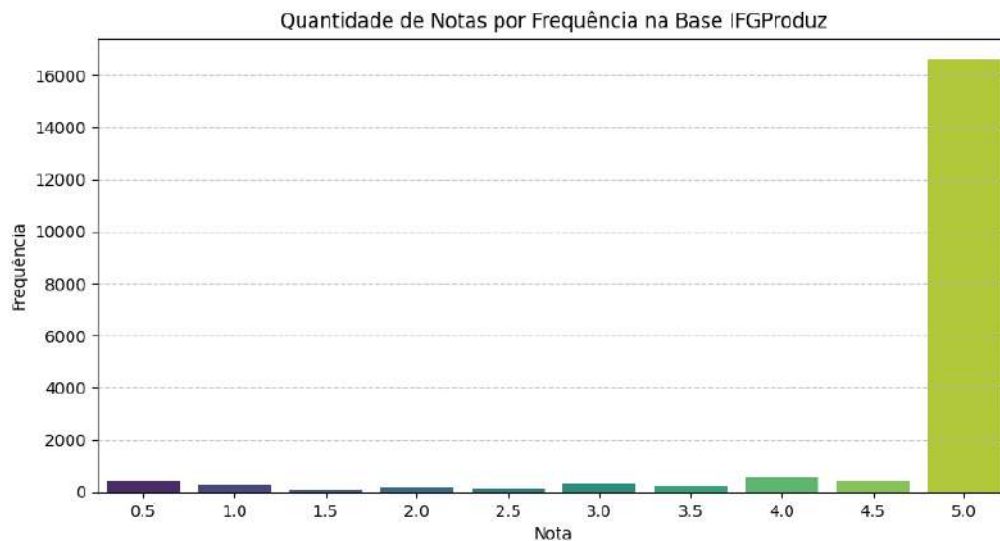
a nota 5.0, somando mais de 16.000 das 19.211 avaliações registradas. A Tabela 5 complementa essa análise, indicando que a média das notas é 4.7, com um desvio padrão de apenas 0.93, o que demonstra uma tendência de avaliações predominantemente positivas.

Essa informação pode indicar dois cenários distintos: o primeiro é que o método de recomendação atualmente utilizado na plataforma apresenta uma boa taxa de aceitação por parte dos usuários. O segundo sugere que os usuários tendem a não avaliar publicações recomendadas caso não despertem seu interesse.

De qualquer forma, esses dados fornecem um contexto favorável para a aplicação de diversos algoritmos de recomendação, especialmente aqueles que constroem o perfil do usuário com base nos itens avaliados positivamente.

Essa concentração, apesar de útil para definir o perfil de cada usuário, exige a inferência de avaliações negativas para balancear as entradas em alguns

Figura 4 – *Quantidade de notas por frequência na base IFG-Produt.*



Fonte: Desenvolvido pelo autor.

dos algoritmos de recomendação aplicados, como o GraphRec. Nesse caso, itens não avaliados são tratados como implicitamente avaliados de forma negativa, uma abordagem bastante utilizada, como apresentada por He *et al.* (2017), e descrita em mais detalhes na seção de experimentos.

3.3.1 Variações da base IFG-Produt

Um dos aspectos mais relevantes do IFG-Produt explorados durante os experimentos foi a incorporação de diferentes conexões entre usuários e itens para caracterizar relevância, enriquecendo assim a base de dados e possibilitando uma análise mais aprofundada.

Entre essas conexões, destacamos a possibilidade da utilização das conexões de autoria como critério adicional de relevância. Em nossos experimentos, ao incluir essas conexões juntamente com as avaliações relevantes, observamos um aumento significativo na quantidade de relações dentro do dataset, impactando diretamente as métricas gerais.

Sendo assim, apresentamos algumas das métricas mais relevantes para análise desse cenário na Tabela 6 a seguir.

Com base nas informações apresentadas na tabela, observa-se que a inclusão da autoria como critério de relevância resulta em um leve aumento na densidade do dataset, elevando o número total de conexões de 19.211 para 43.849. Além disso, a média de conexões por usuário cresce de 6,75 para 38,90, com um desvio padrão de 82,15 conexões. Apesar desse aumento significativo nas conexões por usuário, a

Métrica	IFG-Produz
Usuários	2261
Publicações	71895
Conexões	43849
Densidade	0.092
Número total de usuários com conexões	1408 usuários
Mínimo de conexões por usuário	1 conexão
Máximo de conexões por usuário	1043 conexões
Média de conexões por usuário	38.90 conexões (82.15)
Número total de publicações com ao menos uma conexão	7710 publicações
Mínimo de conexões por publicação	1 conexão
Máximo de conexões por publicação	9 conexões
Média de conexões por publicação	1.04 conexões (0.27)
Média das notas de avaliação	4.94 (0.93)

Tabela 6 – *Resumo das informações presentes na base IFG-Produz considerando as conexões de autoria.*
 Fonte: Desenvolvido pelo autor.

distribuição das conexões por item e a média das notas permanecem próximas aos valores obtidos quando consideradas apenas as avaliações.

Os impactos dessas mudanças nos resultados dos experimentos são discutidos mais adiante no Capítulo 5.

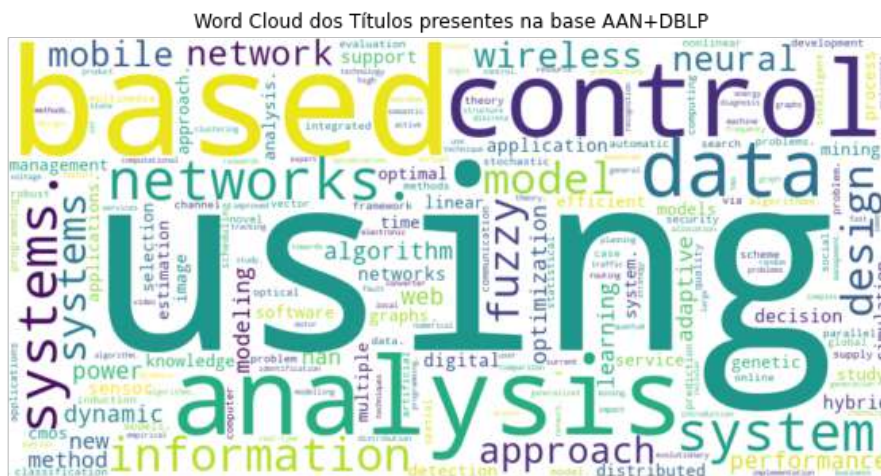
3.4 Informações adicionais presentes nas bases

Além das conexões entre usuários e publicações existentes nas bases de dados analisadas, cada base de dados possui algumas informações adicionais e características próprias, as quais detalharemos nessa seção.

3.4.1 Títulos das publicações

Em uma análise individual por base de dados, a Figura 5, que representa a nuvem de palavras (*word cloud*) dos títulos das publicações na base AAN+DBLP,

Figura 5 – *Word Cloud* dos *Títulos* presentes na base
AAN+DBLP



Fonte: Desenvolvido pelo autor.

evidencia que os títulos refletem a área de conhecimento predominante na AAN, conforme esperado: computação linguística e processamento de linguagem natural.

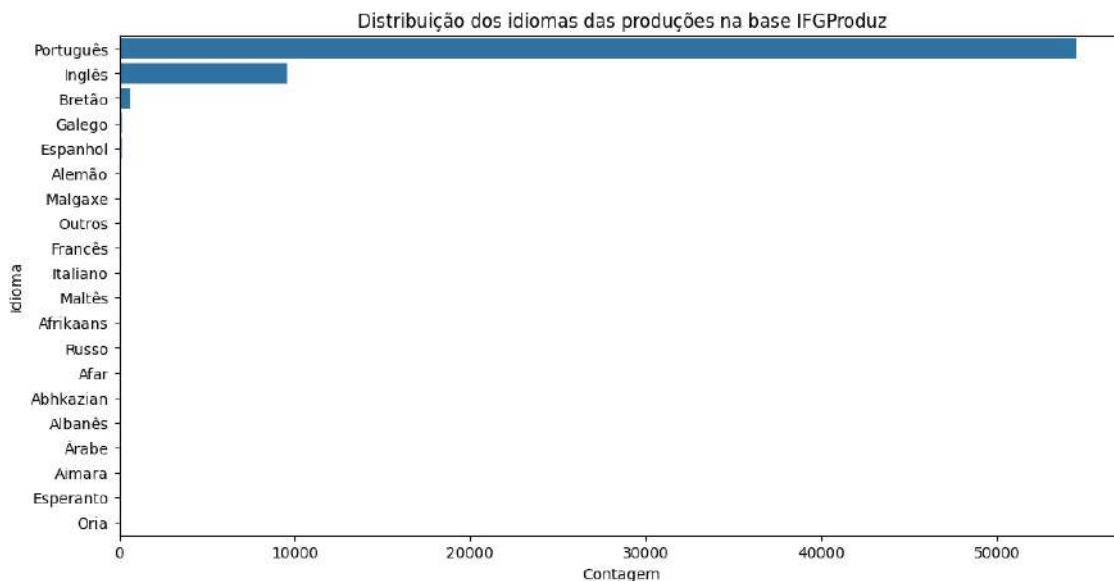
Da mesma forma, a *word cloud* dos títulos da base CiteULike, ilustrada na Figura 6, indica uma predominância de temas relacionados à bioinformática. Nem Wang, Chen e Li (2013) e nem Wang e Blei (2011), que originalmente geraram a base, oferecem uma descrição das áreas temáticas utilizadas para coletar as publicações, então não é possível oferecer uma comparação com essa análise e a intenção original dos autores.

Por outro lado, a *word cloud* da base IFG-Produz, representada na Figura 7, revela uma maior diversidade de temas, com forte presença de assuntos ligados à educação e ensino. Esse padrão reflete tanto a maior diversidade de áreas abordadas na base de dados quanto uma característica do perfil dos pesquisadores, que possuem forte vínculo com a docência e áreas relacionadas à educação.

Essa diversidade temática sugere que a base IFG-Produz pode ser utilizada em experimentos de recomendação de artigos científicos para diferentes domínios do conhecimento, permitindo a avaliação de modelos em cenários variados e ampliando sua aplicabilidade em contextos reais.

Outro diferencial relevante da base disponibilizada é a possibilidade de analisar algoritmos voltados para a língua portuguesa, um aspecto não observado nas demais bases analisadas nesta pesquisa. Os dados apresentados na Figura 8 reforçam essa característica, evidenciando que a maioria das publicações presentes na base está no idioma português.

Figura 8 – *Distribuição das publicações por idioma na base de dados IFG-Produz*



Fonte: Desenvolvido pelo autor.

3.4.2 Perfil dos usuários

Um dos principais diferenciais dessa base IFG-Produz é a riqueza de informações descritivas disponíveis tanto para os usuários quanto para as publicações. Além de fornecer uma base rica para a aplicação de modelos baseados em conteúdo, essas informações também possibilitam a modelagem de novas conexões.

No que diz respeito aos usuários, o principal atributo descritivo são as palavras-chave associadas a cada um, extraídas de seu currículo Lattes e ordenadas por nível de importância. Essa é uma informação que também não é disponibilizada nos demais datasets analisados, onde não há nenhum atributo descritivo dos usuários. A tabela 7 apresenta algumas métricas relacionadas a essas palavras-chave:

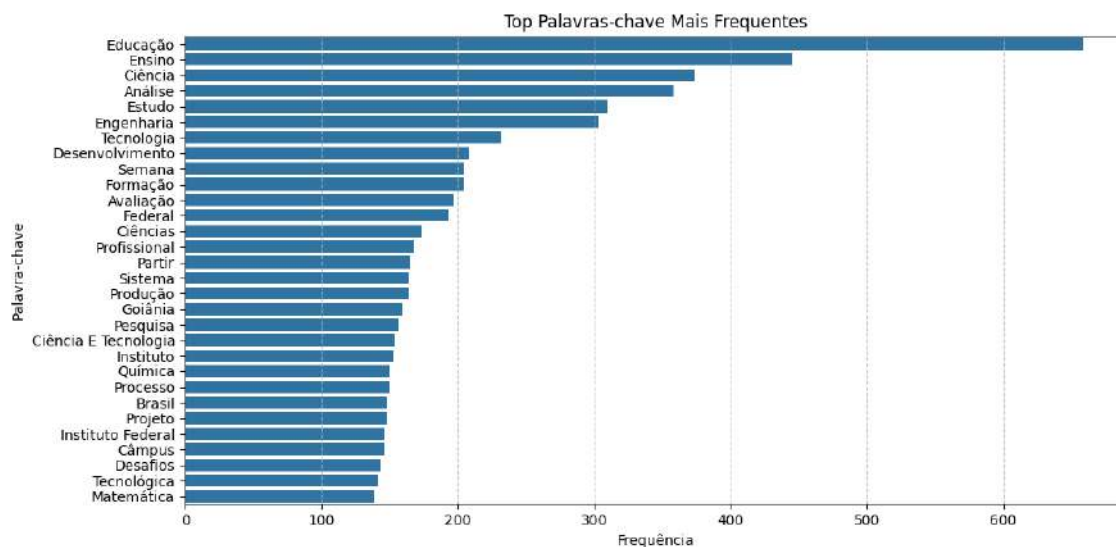
Métrica	Valor
Total de usuários na base	2261
Total de palavras-chave únicas	54150
Média de palavras-chave por usuário	50.38
Maior número de palavras-chave	1287 palavras

Tabela 7 – *Resumo das métricas das palavras-chave.*

Fonte: Desenvolvido pelo autor.

A partir da análise dessas métricas, é possível levantar um perfil dos

Figura 9 – *Palavras-chave mais frequentes por usuário na base IFG-Produz*



Fonte: Desenvolvido pelo autor.

usuários, considerando tanto as palavras-chave mais frequentes no geral, conforme ilustrado na Figura 9 quanto aquelas mais recorrentes como as mais importantes para cada usuário, representadas na figura 10.

Observamos aqui o mesmo padrão registrado ao examinar a *word cloud* dos títulos das publicações. Primeiramente, nota-se a presença de um caráter multidisciplinar nos dados. Em segundo lugar, nota-se a predominância de usuários com interesse voltado para a área educacional, sendo essa a palavra-chave mais frequente tanto na base como um todo quanto na primeira posição de relevância para os usuários.

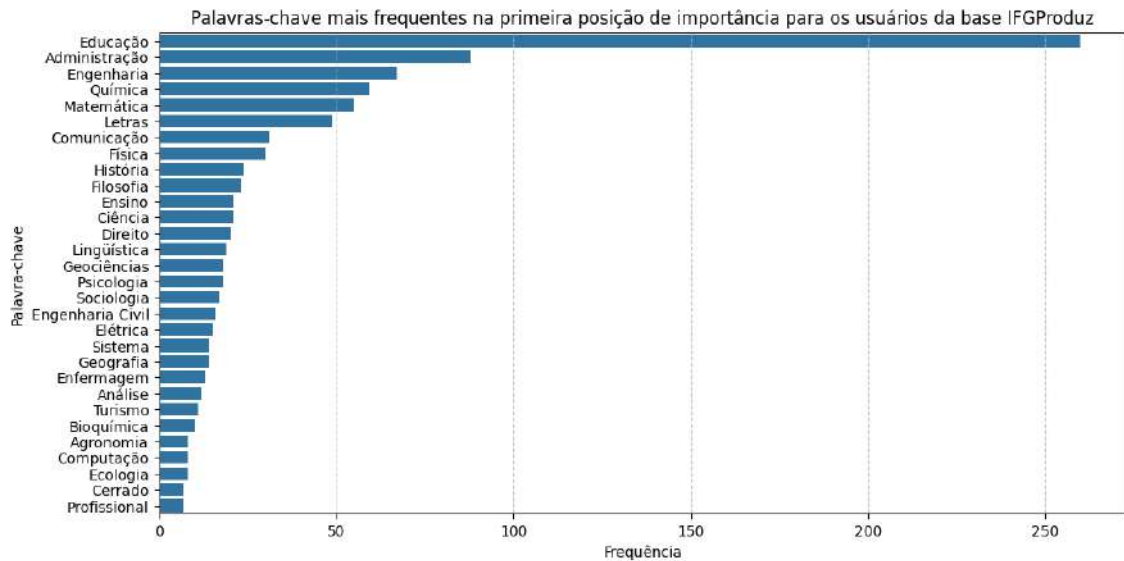
Mais uma vez, ressaltamos a contribuição da disponibilização destes dados, visto que essas informações sobre o perfil dos usuários podem auxiliar na construção de modelos de dados para recomendação mais robustos e precisos, o que não é percebido nas demais bases analisadas.

3.4.3 Conferências e Revistas

Além das informações descritivas dos usuários, a base IFG-Produz também contém detalhes estruturados sobre os itens, indo além dos atributos comuns a outras bases de dados, como idioma e tipo de publicação. A base armazena dados sobre as revistas e conferências associadas a cada publicação, proporcionando uma camada adicional de informação.

Embora o AAN+DBLP também registre informações sobre revistas, esses dados são armazenados apenas como texto corrido contendo o título da publicação.

Figura 10 – *Palavras-chave mais relevantes por usuário na base IFG-Produz*



Fonte: Desenvolvido pelo autor.

Apesar de representarem um recurso útil, exigem um alto nível de processamento para extração estruturada. Assim, essas bases não oferecem a mesma facilidade de acesso nem dados adicionais, como a área de atuação da revista, disponíveis no IFG-Produz.

No IFG-Produz, essas informações estão estruturadas em tabelas separadas, permitindo associações mais diretas entre publicações e seus respectivos veículos, uma maior facilidade de agrupamento desses veículos e o registro de atributos detalhados sobre as próprias revistas e conferências. Para as revistas, a base inclui nome, qualificação Qualis registrada pela Fundação CAPES, área de atuação e código ISSN. Algumas métricas sobre as revistas registradas podem ser observadas na Tabela 8.

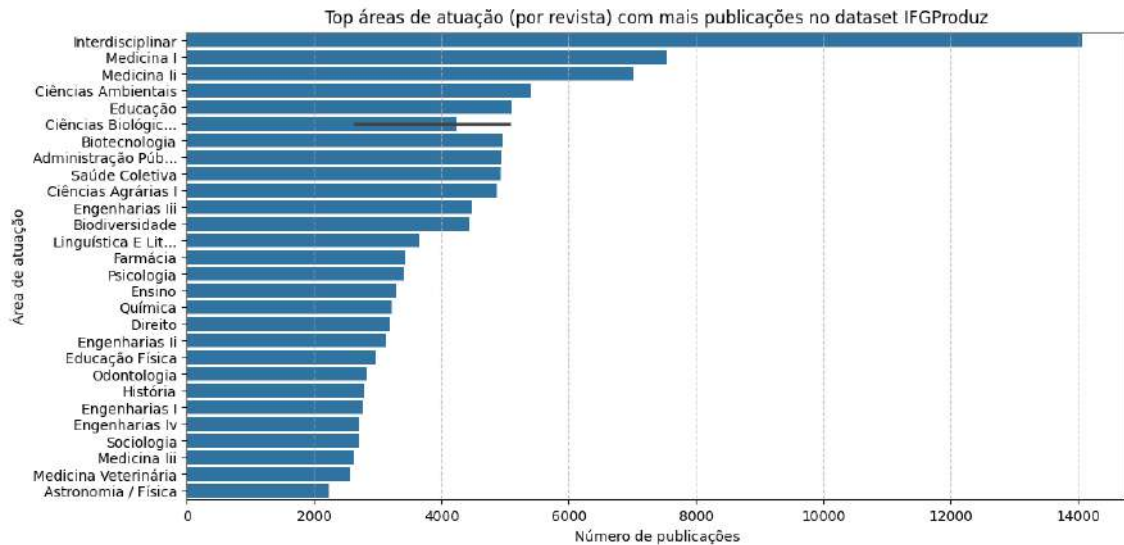
Métrica	Valor
Total de revistas únicas	158938
Total de áreas de atuação distintas	57
Total de ISSNs únicos	31531

Tabela 8 – *Resumo das métricas das revistas.*

Fonte: Desenvolvido pelo autor.

Outra análise relevante sobre as revistas registradas diz respeito às suas áreas de atuação. A Figura 11 apresenta a distribuição da quantidade de itens publicados em revistas, considerando as áreas mais frequentes. Essa análise reforça o

Figura 11 – *Relação de número de publicações por área de atuação (revistas)*



caráter interdisciplinar da base, evidenciando que revistas com abordagens interdisciplinares predominam nos registros. Esse aspecto pode ser um fator relevante para a recomendação, influenciando a diversidade dos itens sugeridos aos usuários.

Além das revistas, a base de dados também armazena informações sobre as conferências associadas às publicações, cujas métricas podem ser analisadas na Tabela 9. Para cada conferência, a base registra o título, a sigla e a nota Qualis atribuída pela Fundação CAPES.

Métrica	Valor
Total de conferências únicas	1415
Total de siglas distintas	1406

Tabela 9 – *Resumo das métricas das conferências.*

Fonte: Desenvolvido pelo autor.

Essas informações são extremamente valiosas para uma base voltada à área de recomendação, pois permitem a exploração de diferentes abordagens para a construção do perfil de um usuário. Além disso, oferecem diversas conexões adicionais, especialmente para metodologias baseadas em grafos, como as analisadas neste trabalho.

3.4.4 Publicações e Projetos de Pesquisa

Conforme detalhado anteriormente, as publicações analisadas neste estudo são categorizadas na base IFG-Produt como artigos publicados e projetos de

pesquisa. As Tabelas 10 e 11 apresentam dados relevantes sobre todas as publicações registradas.

Métrica	Valor
Total de projetos de pesquisa	6685
Média de duração dos projetos (anos)	1.76
Média de caracteres na descrição dos projetos	775.37

Tabela 10 – *Resumo das métricas dos projetos de pesquisa.*
Fonte: Desenvolvido pelo autor.

Métrica	Valor
Total de publicações bibliográficas	65210
Média de publicações por ano	1552.61
Ano mais antigo de publicação	1982
Ano mais recente de publicação	2024
Tipo de publicação mais comum	Capítulo de livro publicado

Tabela 11 – *Resumo das métricas das publicações bibliográficas.*
Fonte: Desenvolvido pelo autor.

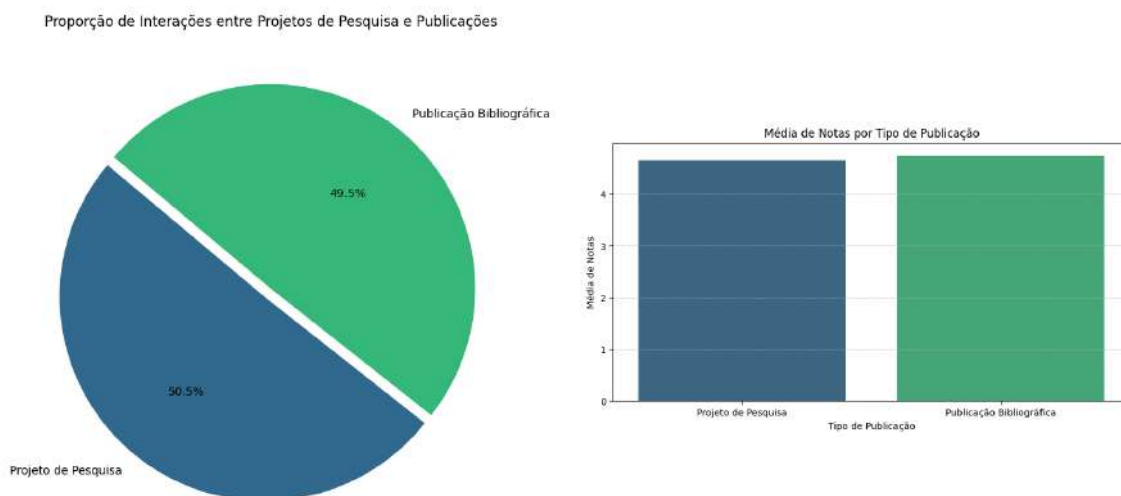
Apesar da base conter um número significativamente maior de artigos publicados, a análise dos gráficos na Figura 12 mostra que, ao considerarmos apenas os registros com avaliações relevantes, a amostra se torna mais balanceada entre os dois tipos de publicação. Esse equilíbrio pode ser um fator importante para futuras pesquisas que investiguem o impacto de características específicas de artigos e projetos de pesquisa em modelos de recomendação ou na construção do perfil dos usuários.

Para os experimentos realizados neste estudo, essas informações foram unificadas e tratadas genericamente como publicações, garantindo uma abordagem mais homogênea na análise dos algoritmos de recomendação.

3.4.5 Ano de publicação

Outra informação relevante a ser analisada nas bases IFG-Produz e AAN+DBLP é o ano de publicação dos itens registrados. Conforme ilustrado na

Figura 12 – *Proporção de tipos de publicação avaliadas na base IFG-Produz (à direita) e Média de notas por tipo de avaliação (à esquerda).*



Fonte: Desenvolvido pelo autor.

Figura 13, o IFG-Produz abrange publicações entre 1982 e 2024, com um pico de publicações em 2015.

Já para a construção do AAN+DBLP, foi considerado o intervalo entre 2006 e 2009, priorizando publicações mais recentes. Essa escolha visa manter a estrutura da base de dados alinhada à disponibilizada por Sugiyama e Kan (2010), como explicado anteriormente. Segundo Wang e Blei (2011), a base CiteULike-a foi coletada de artigos publicados entre 2004 e 2010, porém essa informação não é armazenada.

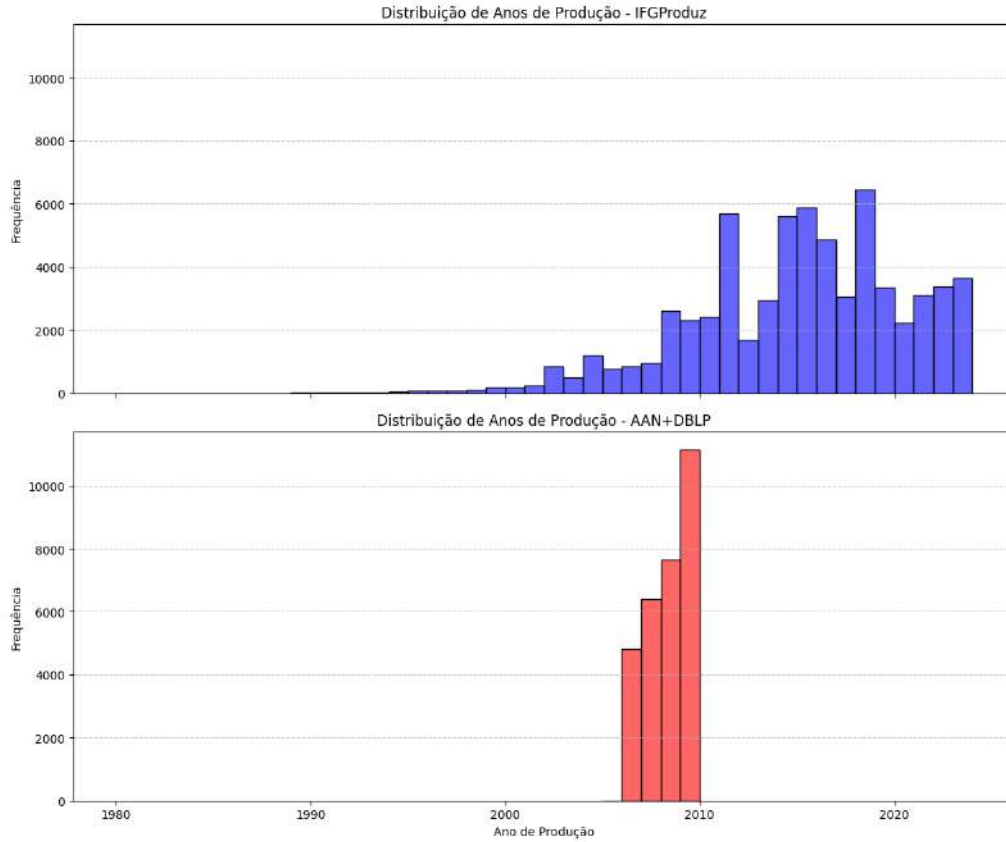
Embora a plataforma AAN mantenha registros desde 1965⁵, ampliar o intervalo temporal aumentaria significativamente o tamanho e a esparsidade da base de dados, dificultando a comparação com o IFG-Produz conforme proposto neste estudo.

Logo, a análise da distribuição temporal das publicações evidencia que o IFG-Produz possui um intervalo significativamente mais amplo em comparação com as demais bases, ao menos dentro do contexto dessa pesquisa. Enquanto o AAN+DBLP apresenta um período mais restrito e o CiteULike não contém essa informação, o IFG-Produz abrange um espectro mais extenso, proporcionando maior representatividade dos dados das publicações sem comprometer o tamanho da base.

A informação referente ao ano de publicação pode ser particularmente útil para possíveis estudos que busquem analisar o impacto do tempo de publicação na relevância dos itens recomendados. Além disso, essa informação pode permitir a avaliação de modelos ao longo de diferentes períodos históricos e contextos

⁵<https://aclanthology.org/events/coling-1965/>

Figura 13 – *Distribuição de anos de produção nas bases do IFG-Produz e AAN+DBLP*



Fonte: Desenvolvido pelo autor.

acadêmicos, garantindo uma análise mais robusta e aplicável a diversos cenários, sem prejudicar a performance dos experimentos.

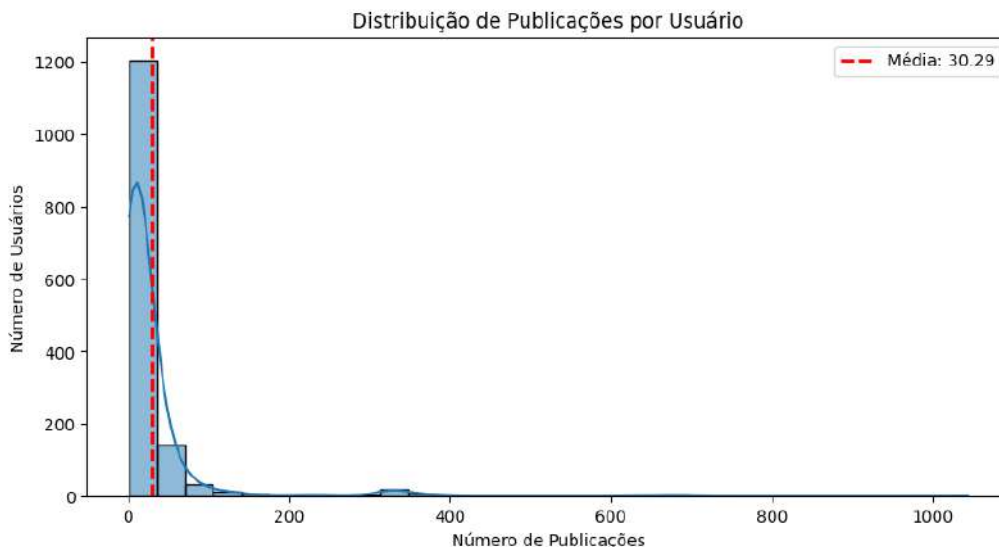
3.4.6 Autoria e Co-Autoria

Além da conexão entre usuários e itens a ser realizada a partir da avaliação de uma recomendação, existe ainda a conexão que está relacionada a autoria, visto que os usuários da plataforma também são pesquisadores. Como apresentado na figura 14, a média de publicações por usuário é de aproximadamente 30 publicações, sendo o mínimo uma publicação e o máximo sendo 1043 publicações para um único usuário, nesse caso considerando inclusive as publicações que não foram avaliadas.

Isso permite algumas análises adicionais, onde conseguimos adicionar mais conexões a partir dessa informação de autoria. Considerando os dados com essas conexões adicionais, continuamos com os 1127 usuários e 7146 itens, porém agora com um total de 43.850 conexões, sendo 24.889 conexões de autoria.

Todas essas informações adicionais descritas nos subtópicos anteriores representam um grande diferencial da base IFG-Produz, onde muitas não são encontradas

Figura 14 – *Distribuição de publicações por usuário na base do IFG-Produz*



Fonte: Desenvolvido pelo autor.

em nenhuma das outras bases analisadas para esta pesquisa. Inicialmente, essas informações sugerem uma base extremamente sólida não apenas para a aplicação de modelos baseados em conteúdo, mas também para a modelagem de novas conexões a serem utilizadas em modelos de filtragem colaborativa. Dessa forma, esses atributos adicionais podem ser incorporados à estrutura do grafo utilizada nesses modelos, possivelmente enriquecendo suas representações.

3.4.7 Categorias e Tags

A base CiteULike também possui uma informação adicional interessante, que são as tags. A base possui um total de 16.980 tags, utilizadas como termos-chave para descrever os artigos.

Segundo Wang, Chen e Li (2013), essa versão da base foi construída a partir da base coletada por Wang e Blei (2011), apenas adicionando as tags. Em nenhum dos dois trabalhos é citado um critério de seleção temática das publicações.

A análise das tags indica que os tópicos abordados pelos artigos são amplamente diversificados, como ilustrado na nuvem de palavras da Figura 15, porém bastante relacionados a *softwares* e *bioinformática*. Além disso, cada artigo apresenta ao menos uma tag, conforme a Tabela 12, sendo 357 o máximo de tags registradas por artigo, com uma média de aproximadamente 18 tags por item.

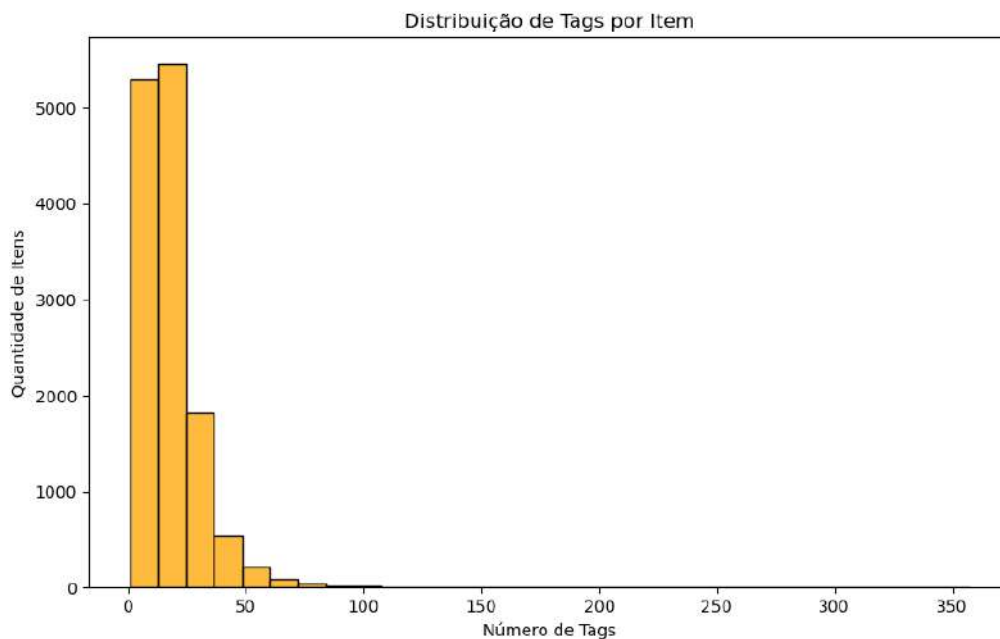
Já a Figura 16 evidencia que a grande maioria dos artigos possui até 50 tags, o que contribui para uma boa representatividade dos itens por meio dessas palavras-chave.

Figura 15 – *Word Cloud das Tags presentes na base CiteU-Like*



Fonte: Desenvolvido pelo autor.

Figura 16 – *Distribuição de tags por publicação na base CiteULike*



Fonte: Desenvolvido pelo autor.

Métrica	Valor
Número total de tags	46388
Mínimo de tags por publicação	1
Máximo de tags por publicação	357
Média de tags por publicação	17.7

Tabela 12 – *Resumo das palavras-chave na base CiteULike.*

Fonte: Desenvolvido pelo autor.

Esses dados evidenciam a boa representatividade das publicações na base CiteULike, tornando-a uma referência sólida para as análises desta pesquisa. Sua estrutura permite uma comparação adequada com a base IFG-Produz, possibilitando uma avaliação mais abrangente do desempenho dos algoritmos de recomendação em diferentes contextos.

Com base nessas análises, concluímos que a base IFG-Produz apresenta características, dados e uma estrutura similar a bases de dados amplamente utilizadas em algoritmos de recomendação científica, como o CiteULike, permitindo assim uma comparação válida em nossos experimentos. Além disso, destacamos a riqueza de informações adicionais oferecidas pelo IFG-Produz, que não são contempladas por nenhuma outra base de dados analisada nesta pesquisa, tornando-a uma alternativa relevante para experimentos em recomendação acadêmica.

A versão completa da base IFG-Produz utilizada neste estudo está disponível na plataforma Kaggle⁶, juntamente com uma descrição detalhada dos dados. O acesso ao público está disponível a partir da data de publicação desta pesquisa.

É importante destacar que, embora este trabalho apresente semelhanças com a proposta de Camargo (2022), que também caracteriza a base IFG-Produz, a presente pesquisa se diferencia por sua abordagem mais aprofundada. Além de uma revisão de literatura mais extensa, realizamos experimentos com uma variação maior de algoritmos e de preparação dos dados. Além disso, disponibilizamos a base de forma pública, proporcionando uma contribuição ainda mais ampla para a comunidade acadêmica.

No próximo capítulo, apresentaremos uma análise detalhada dos algoritmos utilizados nos experimentos, explorando o comportamento do IFG-Produz em diferentes modelos de recomendação.

⁶kaggle.com/datasets/2e9085673224c6a9c70dc0d3adf08c3de38bc560b9262cfefb06b5596d79cf90

Metodologia

Considerando a apresentação das bases de dados no capítulo anterior, descrevemos aqui alguns algoritmos de recomendação encontrados na literatura. Nosso objetivo é comparar o desempenho desses algoritmos em cada base de dados, analisando suas características e particularidades dos modelos em relação a cada base de dados, em especial ao IFG-Produz.

Para a execução desses modelos, buscamos algoritmos de estado da arte que apresentassem bom desempenho em ambientes de recomendação baseados em grafos e em redes self-attention, permitindo explorar diferentes configurações das conexões presentes nas bases de dados.

O principal objetivo de utilizarmos algoritmos baseados em grafos para a análise é a capacidade oferecida pelos mesmos de modelar e conectar os dados de maneira flexível e estruturada, permitindo a investigação de diversas abordagens para o uso do dataset. Além disso, priorizamos algoritmos que oferecessem mecanismos para otimizar a performance tanto em cenários com maior disponibilidade de informações quanto em ambientes com dados mais esparsos. Essas escolhas visam enriquecer a análise e avaliar o impacto da quantidade de informações na performance final, considerando que a base IFG-Produz contém diversas informações adicionais que geralmente não estão presentes em outras bases de dados de recomendação.

No total, foram implementadas cinco abordagens, que aplicam diferentes técnicas de pré-processamento dos dados, diferentes técnicas de embeddings e modelos de predição. Todas as abordagens apresentaram ótimos resultados em seus trabalhos originais em comparação com métodos tradicionais de recomendação.

Cabe pontuar aqui um dos obstáculos no desenvolvimento da presente pesquisa, que foi a adaptação dos algoritmos fornecidos pelos autores para o contexto do presente trabalho. Muitas das abordagens não fornecem uma descrição clara da formatação necessária para os dados, e nem acesso direto às predições pelo algoritmo após a execução, dificultando a implementação destes algoritmos em diferentes bases de dados e a aplicação das métricas de forma padronizada.

Dessa forma, apresentamos na sequência os modelos utilizados e os expli-

camos já considerando as adaptações para o contexto de recomendação de artigos científicos, e considerando as bases de dados avaliadas em nosso trabalho.

4.1 GraphRec

Uma das primeiras abordagens para recomendação que foi analisada e utilizada na construção deste trabalho é a proposta por Rashed, Grabocka e Schmidt-Thieme (2019), nomeada pelos autores de GraphRec.

O GraphRec é um modelo de recomendação relativamente simples, adaptável a diferentes contextos e que seja capaz de trabalhar de forma efetiva com uma pouca quantidade de informação e igualmente com uma grande quantidade de informação adicional.

Segundo os autores, sistemas de recomendação baseados em matrizes de fatoração costumam ter uma excelente performance, porém exigem estruturas mais complexas para conseguirem abstrair informações adicionais ao contexto da recomendação, e muitas das vezes não conseguem generalizar de volta a cenários simples com informações escassas após serem modificados.

Para a implementação do modelo GraphRec, foi utilizada a implementação realizada por Camargo (2022) como base, visto que o autor realizou um trabalho similar ao comparar algumas técnicas de recomendação baseadas em conteúdo com a abordagem do GraphRec, também utilizando a base de dados do IFG-Produz como uma das fontes, e disponibilizou sua implementação publicamente.

4.1.1 Descrição do Algoritmo

Em resumo, o modelo proposto por Rashed, Grabocka e Schmidt-Thieme (2019) é composto de dois componentes principais, representados na figura 19. O primeiro componente é responsável por extrair vetores de atributos latentes tanto dos pesquisadores quanto das publicações, mapeando-os para um espaço compartilhado, e os alimentar em uma rede neural com ativação não-linear.

Para alimentar esse primeiro componente, utilizam-se duas matrizes distintas: uma baseada em pesquisadores e outra em publicações. A matriz de pesquisadores é construída com base nas conexões entre pesquisadores e publicações (ou avaliações), considerando o grau de relevância atribuído a cada interação. Esse grau de relevância pode variar dependendo da base de dados específica utilizada. (Rashed, Grabocka e Schmidt-Thieme, 2019)

Na base de dados do IFG-Produz, por exemplo, a relevância é representada por um valor numérico de um a cinco, que corresponde ao grau de importância que

Figura 17 – *Exemplo de estrutura da matriz de pesquisadores utilizada para alimentar o modelo GraphRec*

**Matriz de conexão entre
pesquisadores e publicações**

	$\bar{l}_1$	$\bar{l}_2$	$\bar{l}_3$
u_1	5		1
u_2	3	2	
u_3			4

Fonte: Produzida pelo autor.

um pesquisador atribui a uma publicação lida. A Figura 17 ilustra essa relação, destacando que as interseções entre pesquisadores e publicações contêm valores numéricos apenas quando existe uma interação relevante entre eles, indicando assim o grau de relevância dessa conexão.

Rashed, Grabocka e Schmidt-Thieme (2019) descrevem que, devido às possíveis diferentes escalas destes valores, todos eles são normalizados entre 0 e 1, onde 0 representa uma interação vazia, e 1 uma interação com avaliação de valor máximo.

A matriz de publicações, por sua vez, é inicialmente construída como uma matriz quadrada, representando as interações entre as próprias publicações. (Rashed, Grabocka e Schmidt-Thieme, 2019). Posteriormente, essa matriz é enriquecida com a adição de colunas que representam atributos suplementares de cada item, por meio de concatenação, ampliando a representação dos dados disponíveis, como ilustrado na figura 18.

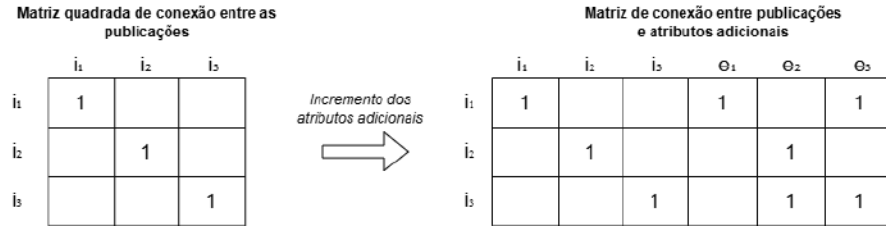
Neste caso, como as interações não representam uma avaliação, mas sim a presença de um atributo específico em uma publicação, os valores das conexões são binários: 0 indica a ausência de conexão e 1 indica uma conexão verdadeira.

A partir das técnicas de fatoração de matrizes descritas por Rashed, Grabocka e Schmidt-Thieme (2019), são extraídos vetores de embeddings que representam os atributos de pesquisadores e publicações contidos nessas matrizes.

O segundo componente é um conjunto adicional de parâmetros baseados em grafos, ativado condicionalmente conforme os parâmetros de execução do algoritmo. Os autores sugerem a ativação deste componente em situações onde há pouca informação inicial ou adicional sobre pesquisadores e publicações.

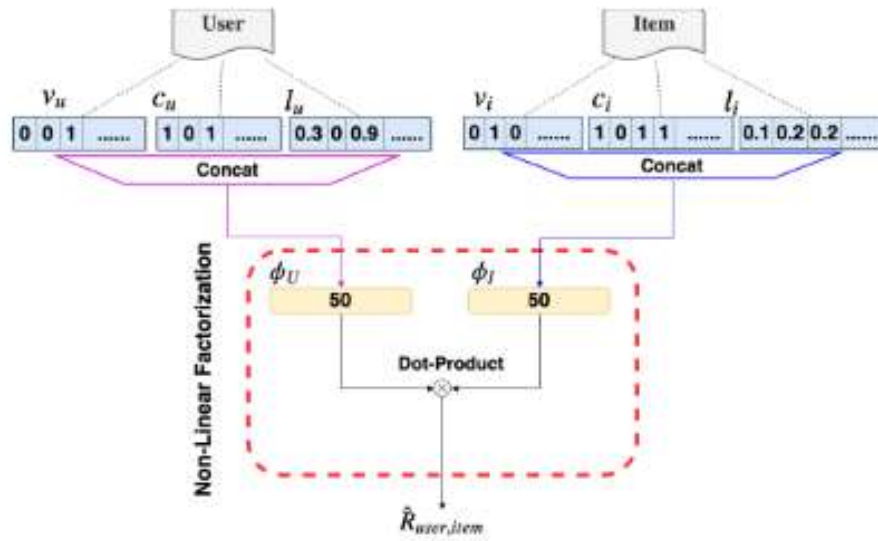
Quando ativado, o modelo, durante a execução do primeiro componente, gera novos vetores com base nas matrizes de interação já existentes. Esses vetores são criados utilizando conceitos de matrizes laplacianas e fornecem informação adicional

Figura 18 – *Exemplo de estrutura da matriz de publicações utilizada para alimentar o modelo GraphRec*



Fonte: Produzida pelo autor.

Figura 19 – *Arquitetura do modelo GraphRec*



Fonte: (Rashed, Grabocka e Schmidt-Thieme, 2019).

ao modelo de predição (Rashed, Grabocka e Schmidt-Thieme, 2019).

De acordo com Kipf e Welling (2017), a decomposição da matriz laplaciana de um grafo e o uso de seus vetores como atributos podem melhorar a precisão das predições, pois esses vetores encapsulam propriedades fundamentais do grafo de interação. Essa mesma abordagem é empregada por Rashed, Grabocka e Schmidt-Thieme (2019), onde, ao decompor a matriz laplaciana, extraem-se atributos genéricos do grafo, que são então concatenados aos embeddings originais obtidos no primeiro componente.

Na Figura 19, esse conjunto adicional de parâmetros é representado por l_i e l_u , indicando vetores de parâmetros adicionais gerados pela decomposição da matriz e concatenados aos vetores originais.

Dessa forma, o modelo proposto por (Rashed, Grabocka e Schmidt-Thieme, 2019) permite dois modos de funcionamento. Em contextos onde há informações suficientes para predições confiáveis, utilizam-se apenas os vetores gerados pelas matrizes de interação pesquisador-publicação e publicação-atributos. No entanto,

quando essas informações são insuficientes, o modelo pode gerar novos vetores a partir de atributos do grafo, que são concatenados aos vetores originais, enriquecendo a representação de cada entidade.

No caso da aplicação do algoritmo na presente pesquisa, foram feitos experimentos com os dois modelos de funcionamento. Porém, como não apresentaram uma diferença estatística relevante, mantivemos os resultados obtidos pelo método que utiliza as informações adicionais.

Na base IFG-Produz, foram utilizados os tipos de publicação como atributos adicionais dos pesquisadores e publicações. Para a base CiteULike, as tags, que representam palavras-chave das publicações, foram escolhidas, com um limite estabelecido no número de tags por artigo. Por fim, para a base construída com dados da AAN e DBLP, o atributo selecionado foi o ano de publicação.

Os vetores gerados a partir dos componentes descritos são então utilizados para alimentar uma rede neural com funções de ativação não-linear. Conforme descrito pelos autores, esses vetores não são empregados como parâmetros da rede, como ocorre na maioria dos métodos de fatoração de matrizes, mas sim como valores de ativação nas camadas da rede neural (Rashed, Grabocka e Schmidt-Thieme, 2019).

Os autores relatam que o modelo alcançou resultados superiores aos de alguns métodos de referência, mesmo sem utilizar atributos externos, indicando que atributos genéricos de grafos podem servir como substitutos eficazes para atributos adicionais quando estes não estão disponíveis (Rashed, Grabocka e Schmidt-Thieme, 2019).

4.2 Knowledge Graph Attention Network

O modelo *Knowledge Graph Attention Network* (KGAT) é uma abordagem desenvolvida por Wang *et al.* (2019). De acordo com os autores, e alinhado com os objetivos do nosso trabalho, os sistemas de filtragem colaborativa sofrem pela impossibilidade de incorporar informações adicionais sobre as entidades e o contexto da aplicação.

Dessa forma, o principal objetivo do modelo proposto é incorporar informações às estruturas de relação usuários-itens já existentes, construindo uma estrutura baseada em grafos do conhecimento capaz de combinar essas informações. O objetivo é melhorar o resultado das recomendações, considerando um maior detalhamento e explicando o motivo de, por exemplo, dois usuários gostarem de um certo item. Essa estrutura de grafo construída foi nomeada pelos autores de *Collaborative Knowledge Graph* (CKG) . (Wang *et al.*, 2019)

A escolha do KGAT para análise neste trabalho se justifica, pois: i) é considerado um dos algoritmos estado da arte em Sistemas de Recomendação utilizando GNNs, como descrito por Gao *et al.* (2023), e também já experimentado dentro do contexto de recomendação de artigos científicos por Tang, Liu e Qian (2021) e ii) o mesmo propõe avaliar não somente as informações que conectam usuário e item, mas também descrições adicionais do lado do item que justificam a relação com o usuário.

Na sequência, descrevemos o funcionamento do algoritmo KGAT adaptando-o para o contexto de recomendação de artigos científicos, mais especificamente dentro do ambiente da base de dados do IFG-Produz. Iniciaremos pela preparação dos dados e a construção do CKG, a estrutura considerada como entrada do modelo KGAT.

4.2.1 Dados de entrada

Como mencionado anteriormente, o CKG ou "Grafo de Conhecimento de Conteúdo" representa um grafo que integra tanto as conexões entre pesquisadores e publicações já existentes quanto as interações entre publicações e seus atributos adicionais (Wang *et al.*, 2019).

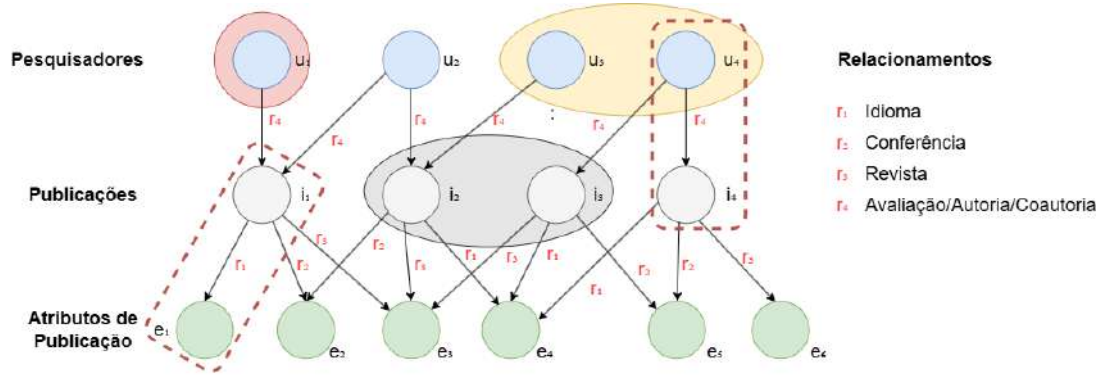
Em outras palavras, em vez de utilizar apenas as relações diretas entre pesquisadores e publicações, adaptaremos o KGAT para considerar também os atributos adicionais das publicações para avaliar sua relevância para um pesquisador, como os dados de revista, conferência e área do conhecimento.

Para tal, a estrutura de CKG gerada utiliza os dois grafos na disposição dos dados existentes: o grafo bipartido, formado pelas conexões diretas entre pesquisadores e publicações, e o grafo com novas conexões triplas, estruturadas como 'sujeito-propriedade-objeto' (Wang *et al.*, 2019).

No contexto deste estudo, exemplos dessas triplas incluem 'pesquisador-avaliação-publicação', 'pesquisador-autor-publicação', e no caso de atributos adicionais para os itens, um exemplo seria 'publicação-conferência-AAAI'. Na figura 20, podemos ver essas triplas destacadas nos retângulos pontilhados. Uma delas indica que o pesquisador u_4 possui uma conexão de autoria com a publicação i_4 , caracterizando assim uma tripla. Da mesma forma, a publicação i_1 possui uma conexão de área de pesquisa com o atributo e_1 , formando outra tripla.

Os nodos nesse contexto representam os pesquisadores, publicações e seus atributos, enquanto as arestas indicam o tipo de conexão entre eles. Essas descrições, mostradas na Figura 20, ilustram como o uso do CKG possibilita a identificação de relações complexas que, em geral, são ignoradas por modelos tradicionais.

Figura 20 – Exemplo da estrutura de grafo alimentada no CKG para o ambiente IFG-Produt



Fonte: Produzida pelo autor.

Na imagem, o usuário u_1 , destacado em vermelho, representa o usuário para o qual se deseja gerar uma recomendação. O pesquisador u_1 é autor de i_1 e não possui conexão direta com os trabalhos i_2 e i_3 . Porém, esses dois trabalhos compartilham atributos de tipo e meio de publicação com i_1 , uma relação que não seria capturada por um modelo tradicional e que pode ser relevante para a recomendação.

Os círculos amarelo e cinza indicam grupos de usuários e itens relevantes para a recomendação, que não seriam capturados de forma explícita por modelos tradicionais. Isso ocorre porque esses itens e usuários não possuem uma conexão direta com o usuário alvo; em vez disso, estão ligados a atributos adicionais compartilhados pelos itens, o que enriquece o processo de recomendação (Wang *et al.*, 2019).

A estrutura CKG construída é então utilizada para alimentar o modelo KGAT.

4.2.2 Modelo KGAT

O modelo KGAT é composto de três principais componentes: camada de embeddings, camadas de propagação e a camada de predição, conforme ilustra a Figura 21.

A camada de embeddings tem a função de parametrizar a estrutura do CKG, preservando o formato do grafo de conhecimento enquanto aprende representações vetoriais das entidades e de suas relações (Wang *et al.*, 2019). Para isso, é utilizada a metodologia *TransR* Yan *et al.* (2019), que permite gerar embeddings para cada entidade e relação, levando em conta a estrutura e as conexões presentes no CKG.

A metodologia *TransR* é uma técnica amplamente utilizada em aprendizado de máquina para representar grafos de conhecimento, projetando dois vetores em relação a um espaço vetorial r . Dada uma tripla (h, r, t) em um grafo de

conhecimento, onde h é a entidade de origem (*head*), neste caso pesquisadores, t é a entidade de destino (*tail*), publicações, e r é a relação entre elas, o objetivo do TransR é modelar essas relações por meio de uma operação de translação no espaço vetorial r , gerando embeddings correspondentes. Ao representar as triplas no espaço vetorial, o modelo aprende a parametrizar as entidades e relações de modo a respeitar a estrutura e as interações específicas entre elas, determinadas por cada tripla (Wang *et al.*, 2019).

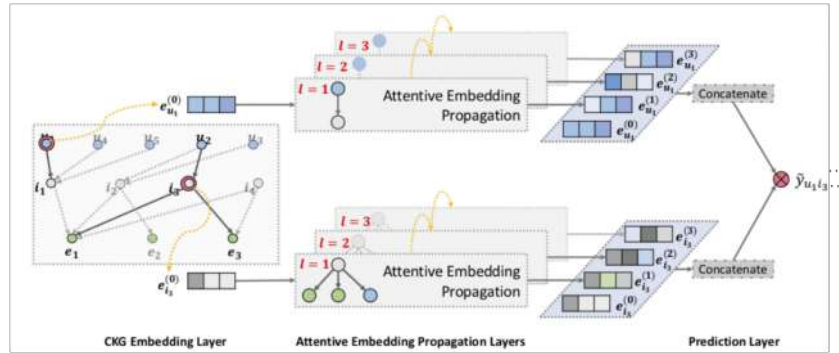
A saída da camada de embeddings são dois vetores de embeddings: um deles representando os pesquisadores, a partir das interações diretas entre pesquisadores e publicações, e o outro representando as publicações, a partir das conexões entre as publicações e seus atributos, relações essas que foram mapeadas nas triplas anteriormente. Essa camada de embeddings é representada na figura 21 como '*CKG Embedding Layer*'. Os vetores de saída desta etapa são usados na etapa de propagação, nomeada na figura como *Attentive Embedding Propagation Layers*.

Na etapa de propagação, o KGAT emprega camadas de propagação de embeddings attention para capturar informações contextuais e relacionais (Wang *et al.*, 2019). Cada camada de propagação contém três componentes principais: i) propagação de informação, que transmite dados entre entidades conectadas por múltiplas triplas; ii) atenção baseada em conhecimento, que calcula a importância das conexões com base na proximidade das entidades no espaço relacional; e iii) agregação de informação, que combina a representação original da entidade com as representações de sua rede de conexões, utilizando diferentes agregadores. (Wang *et al.*, 2019).

Dessa forma, os autores são capazes de utilizar a arquitetura de GNN, propagando recursivamente os embeddings, porém aproveitando a ideia das redes attention, aplicando pesos a essa propagação e permitindo destacar a importância de cada uma delas. (Wang *et al.*, 2019)

Por fim, Wang *et al.* (2019) descrevem o agrupamento dessas camadas em um único conjunto, permitindo a exploração de conexões em um nível ainda mais alto. As representações aprendidas em cada uma das camadas são propagadas, tornando-se parte da representação inicial da iteração da camada seguinte, a camada de predição. Esse processo é ilustrado na figura 21, representado pelas setas pontilhadas amarelas conectando cada um dos *Attentive Embedding Propagation Layers*.

Como descrito anteriormente, esse processo de propagação é aplicado para dois tipos de vetores, um que representa os pesquisadores e suas conexões com as publicações, e outro que representa as publicações e suas conexões com seus atributos.

Figura 21 – *Arquitetura do modelo KGAT*

Fonte: (Wang *et al.*, 2019).

Após o fim do processamento de todas as camadas de propagação, o modelo possui dois conjuntos com diversos vetores de embeddings, sendo um vetor para cada camada de propagação que foi aplicada. Todo este processo de propagação permite enriquecer os embeddings obtidos na primeira etapa, além de controlar a força de propagação a partir do ajuste da quantidade de camadas de propagação aplicadas. (Wang *et al.*, 2019)

O modelo então aplica um mecanismo de agregação de camadas para combinar esses conjuntos de representações em um único vetor, a partir de uma operação de concatenação, obtendo assim um vetor para pesquisadores e outro para publicações.

Por fim, na camada de predição, é calculado o produto interno dos vetores de usuários e itens para prever sua pontuação de relevância entre a publicação e o pesquisador Wang *et al.* (2019), operação esta representada pela fórmula:

$$\hat{y}(u, i) = \mathbf{e}_u^\top \mathbf{e}_i; \quad (1)$$

onde $\mathbf{e}_u$ e $\mathbf{e}_i$ são os vetores de embeddings para o pesquisador u e a publicação i , respectivamente, e $\hat{y}(u, i)$. Essa pontuação reflete a probabilidade de uma correspondência entre um pesquisador e uma publicação, com base nas informações contextuais e relacionais capturadas ao longo do processo.

Como resultado da aplicação deste modelo, (Wang *et al.*, 2019) descrevem resultados consistentes, muito próximos e, por vezes, melhores que os obtidos por outras metodologias já consolidadas. Dentro do contexto da presente pesquisa, os resultados também se apresentam positivos, e serão apresentados com mais detalhes no capítulo 5

4.3 Session-based Recommendation with Graph Neural Networks (SR-GNN)

Outra abordagem utilizada na construção deste trabalho é *Session-based Recommendation with Graph Neural Networks* (SR-GNN) Wu *et al.* (2019). O algoritmo se apoia nos conceitos de recomendação baseada em sessão juntamente com estruturas baseadas em GNN e mecanismos attention para obter uma melhor performance de recomendação.

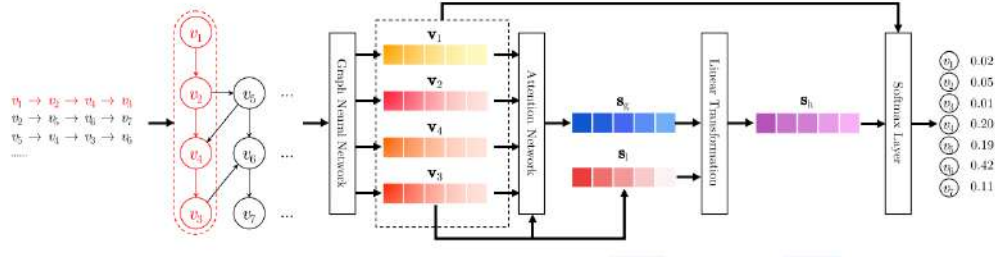
A abordagem de Wu *et al.* (2019) foi selecionada para análise neste trabalho por combinar redes neurais baseadas em grafos com uma estratégia diferente das demais analisadas: a recomendação baseada em sessões. Essa escolha permite explorar diferentes abordagens que utilizam estruturas de grafos e avaliar os resultados em configurações distintas das bases de dados.

No contexto desta pesquisa, a recomendação baseada em sessões é um método cujo objetivo é gerar sugestões para um pesquisador com base em seu comportamento durante uma única sessão de interação. Em outras palavras, em situações onde os pesquisadores acessam um sistema em sessões separadas por períodos de tempo, a abordagem busca prever qual publicação o pesquisador provavelmente acessará em seguida.

Essa previsão é feita exclusivamente a partir dos dados sequenciais da sessão atual, sem depender do histórico completo ou do perfil de preferências de longo prazo do pesquisador. Para cada sessão, o modelo calcula a probabilidade de acesso a todas as publicações disponíveis, classificando-as e identificando os candidatos mais prováveis para recomendação (Wu *et al.*, 2019).

Conforme descrito por Wu *et al.* (2019), os sistemas de recomendação baseados em sessões geralmente enfrentam algumas limitações, como a dificuldade em estimar representações precisas de pesquisadores devido às variações de comportamento entre diferentes sessões. Além disso, esses sistemas frequentemente lidam com um grande número de sessões anônimas, o que torna ainda mais desafiador construir representações confiáveis para cada pesquisador em contextos específicos. (Wu *et al.*, 2019)

Diante dessas limitações, os autores Wu *et al.* (2019) propõem um modelo com o objetivo principal de gerar representações mais precisas para as publicações, visando melhorar o desempenho dos algoritmos que utilizam esses vetores como entrada. Em vez de assumir que cada sessão pode ser representada por um vetor latente generalizado das preferências do usuário, os autores optam por modelar diretamente a sessão com base em todas as publicações presentes nela. Essa abordagem busca capturar de forma mais fiel as interações contextuais específicas

Figura 22 – Arquitetura do modelo SR-GNN

Fonte: (Wu *et al.*, 2019).

de cada sessão. (Wu *et al.*, 2019)

A primeira etapa consiste em representar cada sessão de cada usuário presente na base de dados como um grafo direcionado. No contexto de recomendação de artigos científicos, cada nó do grafo corresponde a uma publicação, enquanto as arestas conectam publicações avaliadas sequencialmente pelo pesquisador durante a sessão.

Um aspecto importante na construção desse grafo é o tratamento de itens que aparecem repetidamente na sequência de acessos. Cada conexão é ponderada com base em um peso normalizado, calculado como a frequência de ocorrência do item dividido pelo grau de saída do nó correspondente. Isso garante que itens acessados múltiplas vezes tenham suas conexões ajustadas para refletir essa repetição.

A Figura 23, disponibilizada pelos autores, apresenta um exemplo desse grafo, incluindo as conexões direcionadas e a matriz de pesos associada. Na ilustração, é possível observar como as conexões originadas de publicações repetidas possuem pesos menores em comparação às arestas de publicações acessadas apenas uma vez.

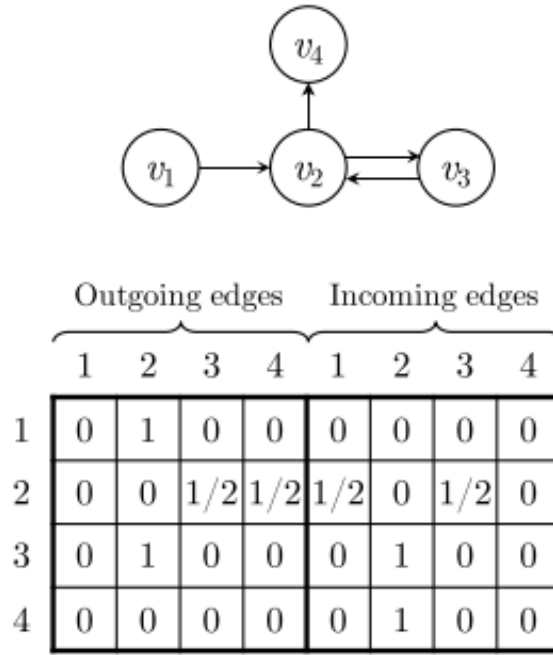
Por exemplo, a aresta que sai (*outgoing edge*) de V_2 para V_4 possui um peso de $\frac{1}{2}$, pois o item V_2 foi acessado duas vezes na sequência, enquanto a aresta que sai de V_3 para V_2 possui peso 1, visto que o item V_3 foi acessado apenas uma vez.

Segundo Wu *et al.* (2019), todos estes grafos de sessões construídos são alimentados em uma GNN que utiliza mecanismos de *gating* para aprender os vetores *embeddings* de cada uma das publicações destas sessões. Esse processo pode ser observado na figura 22, que ilustra os grafos de sessão como entrada de uma GNN, e como saída, os vetores que representam cada um dos nós dos grafos.

Tendo em mãos os vetores que representam todas as publicações, é necessário gerar os vetores de *embedding* que representam cada sessão. Para isso, Wu *et al.* (2019) descrevem uma estratégia que busca combinar a preferência a longo prazo do pesquisador com os interesses do mesmo dentro daquela sessão específica.

Para isso, primeiramente, para uma sessão s se determina um vetor local,

Figura 23 – Exemplo de grafo de sessão e matriz de pesos geradas pelo modelo SR-GNN



Fonte: (Wu *et al.*, 2019).

denominado s_l , que representa basicamente a última publicação acessada dentro de uma sessão. Ou seja, para uma sessão $s = [v_{s,1}, v_{s,2}, \dots, v_{s,n}]$, sendo a última publicação acessada v_n , o vetor local pode ser definido basicamente como $s_l = v_n$. (Wu *et al.*, 2019)

Na Figura 22, podemos observar que o vetor do nó v_3 , que foi o último acessado no exemplo para a sessão s , se torna o vetor s_l . Para a representação a longo prazo dos interesses do pesquisador, é gerado um vetor *embedding* global para a sessão, denominado s_g . Esse vetor é gerado a partir da agregação de todos os vetores de nós daquela sessão. Nessa etapa, um mecanismo *soft-attention* é utilizado para priorizar informações de diferentes níveis de importância dos nós. (Wu *et al.*, 2019).

Para obter a representação da sessão como um vetor *embedding*, é criado um vetor híbrido denominado s_h , que basicamente combina s_l e s_g em um único vetor por meio de uma transformação linear sobre a concatenação desses *embeddings*. (Wu *et al.*, 2019). Assim, ao fim dessa etapa, cada sessão é representada como um vetor que é capaz de representar a preferência global do usuário e a preferência durante a sessão.

Para realizar a predição, o *score* de cada publicação candidata é calculado pela multiplicação entre o *embedding* da publicação e o vetor s_h gerado para a sessão. Em seguida, aplica-se uma função *softmax* ao *score*, gerando a saída do modelo: um

valor que representa a probabilidade de a publicação ser a próxima a ser acessada pelo pesquisador naquela sessão (Wu *et al.*, 2019).

Esse processo é repetido para todas as publicações candidatas, resultando em um conjunto de probabilidades que permite ranqueá-las e recomendar as publicações com maior probabilidade de acesso.

Para viabilizar os experimentos com as bases de dados deste trabalho no modelo SR-GNN, adaptamos as estruturas das bases de dados para simular um conjunto de sessões. Nessa adaptação, todas as publicações avaliadas por um pesquisador foram agrupadas em uma única sessão, com a ordem das publicações disposta de forma aleatória.

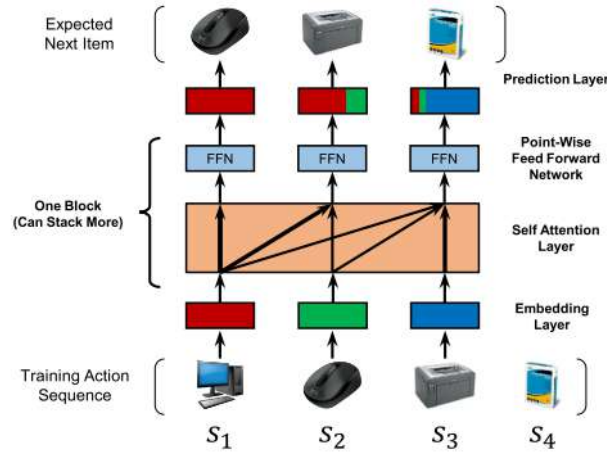
4.4 Self-Attentive Sequential Recommendation (SASRec)

O *Self-Attentive Sequential Recommendation* (SASRec) , proposto por (Kang e McAuley, 2018), é outro algoritmo analisado e implementado na realização deste trabalho. O modelo tem como objetivo realizar recomendações observando o contexto a longo prazo, assim como nos modelos *Recurrent Neural Network* (RNN) , porém utilizando apenas uma pequena quantidade de dados e de passos.

Segundo os autores, sistemas de recomendação sequenciais performam bem em casos onde os dados são esparsos, mas têm problemas em contextos onde os dados são mais complexos e com mais conexões. Isso ocorre justamente por serem modelos que assumem que a próxima ação é condicionada apenas por poucas conexões anteriores. (Kang e McAuley, 2018)

Por outro lado, os modelos baseados em RNN são capazes de abstrair e performar bem em ambientes com dados complexos, porém exigem grandes quantidades de dados e iterações antes de conseguir realizar as predições. Nesse caso, todas as conexões de determinada entidade são sumarizadas e utilizadas para prever a próxima conexão daquela entidade(Kang e McAuley, 2018). Dessa forma, como explicitado anteriormente, o principal objetivo de (Kang e McAuley, 2018) é unir as duas abordagens, sendo capazes de realizar predições efetivamente observando todo o contexto, porém com menos ações e em ambientes com dados menos densos.

O modelo foi selecionado para análise neste trabalho por apresentar uma abordagem alinhada às demais, utilizando informações estruturadas em grafos para recomendação, além de explorar a aplicação de informações adicionais e a complexidade do algoritmo.

Figura 24 – *Arquitetura do modelo SASRec*

Fonte: (Kang e McAuley, 2018).

Ademais, os autores reportam resultados superiores em relação a metodologias tradicionais e disponibilizaram os algoritmos publicamente, o que facilitou a realização dos experimentos.

O modelo proposto pelos autores e ilustrado na Figura 24 utiliza uma arquitetura composta por uma camada de embedding, camadas de self-attention e uma camada de predição, cuja entrada é uma sequência de publicações nas quais houve interações com o pesquisador, e a saída é uma versão "deslocada" dessa sequência, prevendo a próxima publicação a ser consumida. (Kang e McAuley, 2018)

Para fins de descrição, no contexto da recomendação sequencial abordado pelos autores, a entrada dos dados do algoritmo consiste em uma sequência de ações realizadas pelo usuário. Já no contexto da presente pesquisa, voltada para a recomendação de artigos científicos, essas ações representam conexões entre um pesquisador e uma publicação, sejam elas de avaliação ou autoria.

A camada de embeddings recebe essa sequência de avaliações associada a um pesquisador e transforma os elementos da sequência em vetores de dimensões fixas. Assim, uma sequência de avaliações alimentada no algoritmo é convertida em uma sequência de publicações $s = (s_1, s_2, \dots, s_n)$, onde n o tamanho máximo da sequência processada pelo algoritmo, definido previamente durante a configuração do modelo. (Kang e McAuley, 2018).

De acordo com Kang e McAuley (2018), se o tamanho da sequência de publicações exceder o limite n , o algoritmo considera apenas as n publicações mais recentes. Por outro lado, caso a sequência tenha um número de elementos inferior ao valor máximo, o algoritmo aplica um *padding* com o valor 0 até que a sequência atinja o tamanho n .

Com essa sequência ajustada, os autores constroem uma matriz de embed-

dings para cada elemento da sequência, representando cada publicação. Além disso, é adicionado um embedding posicional a cada matriz, a fim de preservar a ordem dos itens no contexto, já que o modelo de *self-attention* utilizado não considera, de forma nativa, informações sobre a posição dos elementos. Essa inclusão é essencial para capturar a ordem dos itens, um aspecto necessário no modelo de recomendação sequencial proposto. (Kang e McAuley, 2018)

Exemplificando, o item da sequência de entrada dos dados s_1 , após o primeiro passo, obtem-se uma matriz M_{s_1} . Essa matriz é então adicionada a matriz de embedding posicional P_1 , indicando sua posição na sequência. Por fim, a saída da camada de embeddings é o conjunto das matrizes de todas as n publicações da sequência de avaliações alimentada, representada pelo embedding $\hat{E}$ na equação 2, disponibilizada por Kang e McAuley (2018):

$$\hat{E} = \begin{bmatrix} M_{s_1} + P_1 \\ M_{s_2} + P_2 \\ \vdots \\ M_{s_n} + P_n \end{bmatrix} \quad (2)$$

Assim como no modelo SR-GNN proposto por Wu *et al.* (2019), onde o perfil de um pesquisador era representado pelos embeddings das publicações que compunham seu conjunto de interações, o modelo SASRec também permite induzir o embedding de um pesquisador a partir dos embeddings das publicações avaliadas em sua sequência, o que se encaixa no contexto dos dados presentes nas bases de dados avaliadas.

Com a estrutura de entrada $\hat{E}$ definida, a próxima camada, denominada de *self-attention*, calcula as interações ponderadas entre os itens da sequência, permitindo que o modelo capture relações significativas entre todos os elementos. (Kang e McAuley, 2018)

Conforme descrito pelos autores, a operação de *self-attention* utiliza matrizes lineares para transformar o embedding $\hat{E}$ obtido na etapa anterior em três matrizes de projeção: *query*, *key* e *value*. Essas matrizes são então processadas por uma camada de atenção, possibilitando que o modelo identifique não apenas relações diretas entre itens, mas também interações assimétricas, ampliando sua flexibilidade e precisão.

Após o bloco de *self-attention*, uma camada de rede *Point-Wise Feed-Forward* é aplicada. Segundo (Kang e McAuley, 2018), essa camada introduz não linearidade no modelo, permitindo que ele considere interações entre diferentes

dimensões latentes e aumente sua capacidade de representação.

Embora o mecanismo de *self-attention* seja eficiente ao considerar todos os itens anteriores na sequência, ele opera de forma linear. Por isso, a camada *Point-Wise Feed-Forward* é essencial para capturar interações mais complexas entre dimensões latentes, complementando o aprendizado da *self-attention*.

O conjunto formado pela camada de *self-attention* e a camada *Point-Wise Feed-Forward* constitui um bloco, que pode ser visto destacado na Figura 24. Esses blocos podem ser empilhados para permitir a modelagem de transições mais complexas entre itens. No entanto, os autores destacam que o aumento da rede pode resultar em problemas como *overfitting*, instabilidade no treinamento e um aumento significativo no tempo necessário para o aprendizado.

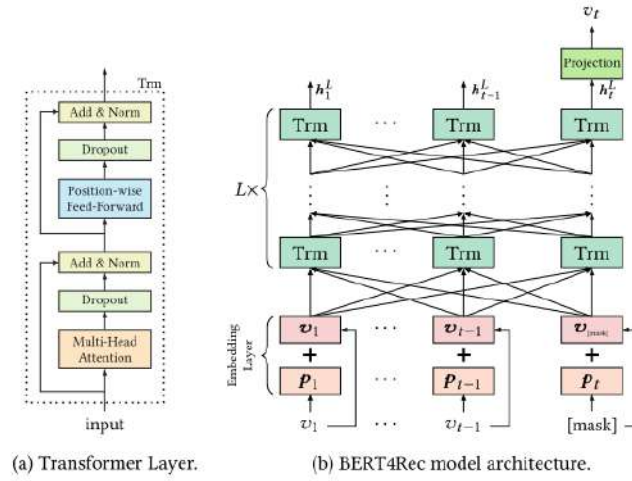
Por fim, a camada de predição utiliza um modelo de fatoração matricial (MF) para calcular a relevância do próximo item com base nas informações extraídas pelos blocos de self-attention descritos anteriormente. Para cada publicação, é obtido um *score*, que é utilizado para ranquear as publicações candidatas e indicar qual seria a próxima publicação da sequência avaliada pelo pesquisador.

De acordo com os próprios autores, o trabalho foi capaz de construir um modelo eficiente para recomendações sequenciais, aproveitando os mecanismos de *self-attention* para capturar dependências complexas dentro dos dados de entrada, mesmo em contextos nos quais modelos RNN não apresentavam bom desempenho. Os resultados demonstraram uma melhora significativa na performance em relação a outros modelos lineares, tornando este trabalho de grande relevância para a presente pesquisa.

4.5 Bidirectional Encoder Representations from Transformers for Recommendation

O modelo *Bidirectional Encoder Representations from Transformers for Recommendation* (BERT4Rec), proposto por Sun *et al.* (2019), baseia-se no mesmo princípio apresentado por Kang e McAuley (2018) no modelo SASRec: as metodologias sequenciais convencionais não representam de forma eficaz as preferências dos usuários. A proposta de Sun *et al.* (2019) busca aprimorar os modelos unidirecionais utilizando um modelo bidirecional, juntamente com mecanismos de attention para melhorar a capacidade de predição.

Os autores compartilham da visão de Kang e McAuley (2018) sobre o potencial dos modelos baseados em attention para recomendação sequencial, reconhecendo o SASRec como próximo de sua abordagem. No entanto, o principal diferencial do

Figura 25 – Arquitetura do modelo BERT4Rec

Fonte: (Sun *et al.*, 2019).

BERT4Rec está na aplicação de um modelo bidirecional para gerar as representações das preferências dos usuários.

Detalhando um pouco mais o modelo e aplicando-o ao contexto de recomendação de artigos científicos, a proposta de Sun *et al.* (2019) consiste em um sistema capaz de aprender representações do comportamento de um pesquisador em relação a uma sequência de publicações. Isso é realizado por meio da aplicação de um modelo baseado em self-attention, que empilha múltiplas camadas de *transformers* bidirecionais, como ilustrado no item *b* da Figura 25.

Em cada camada, o modelo revisa iterativamente a representação de cada publicação, permitindo que esta troque informações com todas as demais publicações da sequência. Essa abordagem difere fundamentalmente dos métodos baseados em RNN, que transmitem informações de forma sequencial, passo a passo, restringindo a troca de informações apenas às publicações anteriores na sequência.

Ao possibilitar interações globais entre os itens da sequência, o modelo BERT4Rec se mostra mais eficiente na captura de relações complexas, particularmente em cenários como o de recomendação de artigos científicos, onde as conexões entre essas publicações podem ser menos lineares e mais diversificadas.

O BERT4Rec recebe como entrada uma sequência de publicações que representam os itens com os quais um pesquisador interagiu ou avaliou. A primeira etapa do processo descrito por Sun *et al.* (2019) consiste na aplicação de uma camada de embedding, responsável por transformar cada publicação da sequência em um vetor de tamanho fixo N , capaz de capturar suas características latentes. Adicionalmente, é incorporado um embedding posicional, já que o mecanismo de attention não considera, de forma nativa, informações sobre a posição dos elementos

na sequência.

Essa inclusão dos embeddings posicionais, similar ao modelo SASRec, traz uma limitação quanto ao tamanho máximo da sequência que pode ser processada pelo algoritmo. Em casos onde a sequência de publicações excede esse limite N , apenas os N últimos itens são considerados.

Dessa forma, a saída dessa camada de embeddings é composta por vetores que representam as publicações da sequência, enriquecidos com informações posicionais relativas aos demais itens. Nesse ponto, a publicação que será alvo da predição é adicionada à sequência, como representado no item b da Figura 25 pelo item P_t . Para integrá-lo ao modelo, aplica-se um *token* especial denominado *mask*, indicado na figura como $v(mask)$. Este elemento mascarado é processado junto com os demais itens nas camadas subsequentes.

As representações resultantes dessa etapa alimentam um conjunto de camadas transformer, cuja estrutura foi descrita no item a da Figura 25. Cada camada transformer recebe como entrada vetores que representam as publicações da sequência e computa representações ocultas mais ricas e informativas, transformando os vetores iniciais em versões refinadas. Cada uma dessas camadas é composta por duas subcamadas principais.

A primeira é a *Multi-Head Self-Attention*, responsável por identificar dependências entre as publicações, independentemente de suas posições na sequência original. Aqui, diversas funções de attention são aplicadas em paralelo, permitindo a modelagem de relações complexas entre os itens. As representações geradas são então combinadas e encaminhadas para as camadas subsequentes.

A segunda, denominada *Position-wise Feed-Forward Network*, introduz não linearidade ao modelo, possibilitando interações entre as diferentes dimensões das representações. Sua funcionalidade é equivalente à camada *Point-Wise Feed-Forward* descrita no modelo SASRec por Kang e McAuley (2018).

Na etapa seguinte, o BERT4Rec calcula iterativamente as representações ocultas de cada publicação utilizando as camadas transformer. Conforme ilustrado no item b da Figura 25, essas camadas são organizadas em uma estrutura matricial, onde as linhas correspondem às camadas (L) e as colunas representam os vetores das publicações na sequência (t).

Cada camada L aplica as funções *transformer* simultaneamente, calculando a função de *attention* para todos os itens da sequência. A saída de cada uma dessas operações alimenta todos os *transformers* da próxima camada L , promovendo um fluxo contínuo de informações ao longo das camadas empilhadas entre todos os itens da sequência.

Essa arquitetura viabiliza a bidirecionalidade proposta pelos autores, per-

mitindo que a representação de uma publicação influencie não apenas a representação das publicações subsequentes, mas também as publicações anteriores Sun *et al.* (2019).

Após essa troca de informações, o modelo obtém as representações finais para todas as publicações da sequência de entrada. As publicações previamente mascaradas pelo *token mask* são então previstas por meio de uma rede *feed-forward* de duas camadas com ativação GELU, que utiliza as representações ocultas como entrada.

O treinamento do BERT4Rec é baseado na tarefa de preenchimento de lacunas (Cloze Task), conforme definida por W. (1953). Nesse processo, uma proporção dos itens é mascarada, e o modelo é treinado para prever os itens ocultos com base no contexto fornecido pelos itens adjacentes. Segundo Sun *et al.* (2019), esse método pode ainda gerar um número maior de amostras de treinamento em comparação com a predição sequencial tradicional, permitindo o treinamento de um modelo de representação mais poderoso.

Durante a fase de teste, o modelo é ajustado para prever apenas a próxima publicação com base no histórico de comportamento do pesquisador. Isso é feito aplicando o *token mask* ao final da sequência de interações, alinhando o treinamento ao objetivo final de recomendação.

Por meio desse modelo, os autores obtiveram uma performance superior em comparação a diversos algoritmos, incluindo o SASRec, ao avaliarem diferentes bases de dados.

Assim, resumindo os modelos apresentados, temos: O GraphRec, que combina métodos de fatoração de matrizes com redes neurais para produzir embeddings de usuários e itens, possibilitando a identificação de relações latentes em grafos; O KGAT, que emprega um Grafo de Conhecimento Colaborativo (CKG) e técnicas *attention* para disseminar embeddings entre entidades interligadas, melhorando as recomendações ao incluir informações extras; O SR-GNN, que utiliza redes neurais em grafos para sugerir itens com base em sessões, simulando interações sequenciais entre os elementos; O SASRec, que emprega técnicas de *self-attention* para identificar padrões temporais nas interações dos usuários, dando prioridade à ordem de consumo; e finalmente, o BERT4Rec, que utiliza um modelo bidirecional fundamentado em *transformers* para identificar dependências globais entre os itens analisados, possibilitando previsões mais acuradas sobre as futuras interações.

Concluimos, portanto, que os modelos apresentados possuem bons resultados em seus trabalhos de origem, além de abranger diferentes modelos e estratégias de recomendação, o que nos permite a realização de uma análise exploratória ampla e descritiva das bases de dados escolhidas.

Assim, apresentados estes modelos, seguimos para o capítulo de experimentação, onde serão descritos os experimentos realizados utilizando os algoritmos, além dos resultados obtidos.

Experimentação e Resultados

Este capítulo apresenta a metodologia experimental adotada na pesquisa e os resultados obtidos com a implementação dos algoritmos de recomendação descritos no capítulo 4 nas bases de dados avaliadas no capítulo 3. O objetivo é analisar a performance da base de dados IFG-Produz em diferentes cenários e conjuntos de informações, comparando-a com outras bases existentes. Assim, buscamos validar a efetividade da base IFG-Produz para futuras análises e aplicações em algoritmos de recomendação.

Inicialmente, descrevemos o ambiente experimental, incluindo os procedimentos de pré-processamento das bases e as configurações iniciais dos experimentos. Por fim, apresentamos os resultados obtidos para todas as bases de dados e experimentos, acompanhados de uma breve análise.

5.1 Dados utilizados e pré-processamento

Os experimentos utilizaram as bases CiteULike Wang, Chen e Li (2013), AAN+DBLP Sugiyama e Kan (2010) e IFG-Produz, conforme descrito no Capítulo 3. Cada base de dados foi submetida a um pré-processamento específico para garantir compatibilidade com os algoritmos aplicados.

Uma modificação comum aos três casos é a remoção de usuários sem interações com as publicações, ou seja, aqueles que não avaliam nem produzem nenhuma publicação.

Além disso, alguns dos algoritmos exigem pré-processamentos específicos. O SR-GNN, por exemplo, remove conexões com menos de cinco instâncias de interação, enquanto o SASRec exclui usuários com menos de três interações. Já os modelos GraphRec e KGAT apenas ajustam a estrutura dos dados sem modificar suas informações originais.

Para o SR-GNN, especificamente na base IFG-Produz, ajustamos os valores de avaliação, originalmente entre 0 e 5. Como o algoritmo opera apenas com conexões

binárias, estabelecemos a nota 3 como ponto de corte, ou seja, publicações com nota maior que 3 são consideradas relevantes, enquanto as demais são descartadas.

Importante ressaltar que, para a base IFG-Produz, tratamos tanto projetos de pesquisa quanto artigos publicados como publicações. Nos experimentos que utilizam informações adicionais, consideramos apenas as informações adicionais dos artigos, pois os projetos de pesquisa não possuem esses atributos.

Outro ponto importante é que, para a aplicação de algoritmos baseados em sessão ou sequência, como SR-GNN, SASRec e BERT4Rec, as bases de dados não contêm informações sequenciais. Assim, geramos sequências aleatórias a partir dos itens relevantes de cada usuário. Nesses casos, perde-se uma informação relevante presente na base do IFG-Produz, que é o grau de relevância que um pesquisador atribui a cada publicação avaliada.

5.2 Metodologia utilizada

Os códigos-fonte utilizados nos experimentos foram obtidos a partir dos repositórios disponibilizados pelos artigos analisados. O código do GraphRec pode ser acessado no repositório do GitHub¹, disponibilizado por Rashed, Grabocka e Schmidt-Thieme (2019). O código-fonte do KGAT está disponível no repositório², fornecido por Wang *et al.* (2019). A implementação do SRGNN pode ser encontrada no repositório do GitHub³, disponibilizado por Wu *et al.* (2019). O código do SASRec está disponível no repositório⁴, conforme apresentado por Kang e McAuley (2018). Por fim, o BERT4Rec pode ser encontrado no seguinte repositório⁵, disponibilizado por Sun *et al.* (2019).

Para garantir a reprodutibilidade dos experimentos, os hiperparâmetros utilizados estão detalhados na Tabela 13. A seleção dos valores variou conforme o algoritmo. Para GraphRec e KGAT, utilizamos o método *grid search*, explorando os intervalos de valores sugeridos por Rashed, Grabocka e Schmidt-Thieme (2019) e Wang *et al.* (2019). Já para SR-GNN, SASRec e BERT4Rec, adotamos os parâmetros padrão definidos em suas implementações originais. Embora tenhamos realizado alguns testes manuais variando esses parâmetros, os melhores resultados ainda foram obtidos com os valores originais.

¹<https://github.com/ahmedrashed-ml/GraphRec>

²https://github.com/xiangwang1223/knowledge_graph_attention_network

³<https://github.com/CRIPAC-DIG/SR-GNN>

⁴<https://github.com/kang205/SASRec>

⁵<https://github.com/FeiSun/BERT4Rec>

Algoritmo	IFG-Produs	CiteULike	AAN+DBLP
GraphRec	LR: 0.001	LR: 0.001	LR: 0.0002
	λ_u : 0.08, λ_i : 0.02	λ_u : 0.05, λ_i : 0.1	λ_u : 0.1, λ_i : 0.01
	Emb: 50	Emb: 100	Emb: 100
	Épocas: 550	Épocas: 200	Épocas: 400
	Batch: 1000	Batch: 500	Batch: 1000
KGAT	LR: 0.001	LR: 0.005	LR: 0.001
	L2: 1e-5	L2: 1e-5	L2: 1e-2
	Emb: 1024	Emb: 1024	Emb: 1024
	Épocas: 100	Épocas: 100	Épocas: 100
	Node Drop: 0.1	Node Drop: 0.1	Node Drop: 0.1
	Mess Drop: 0.1	Mess Drop: 0.1	Mess Drop: 0.1
SR-GNN	LR: 0.001		
	L2: 1e-5		
	Épocas: 100		
	Batch: 100		
SASRec	LR: 0.001		
	Dropout: 0.5		
	Épocas: 200		
	Batch: 128		
BERT4Rec	LR: 1e-4		
	Dropout: 0.2		
	Batch: 256		

Tabela 13 – *Hiperparâmetros utilizados para cada algoritmo nos diferentes bases de dados*

Em relação às especificidades de cada algoritmo, no caso do GraphRec, realizamos experimentos com e sem o mecanismo adicional que utiliza informações da

matriz para melhorar a performance em cenários com menos informações contextuais, descrito no Capítulo 4. Embora tenha havido uma pequena redução no erro médio com esse mecanismo ativado, a diferença não foi estatisticamente relevante. Por isso, optamos por considerar apenas os resultados com essa funcionalidade ativada.

Além dos cinco modelos apresentados no Capítulo 4, todos os experimentos, com exceção do último, também incluem dois algoritmos de filtragem baseados em conteúdo: TF-IDF⁶ e o modelo zero-shot do BERT multilíngue⁷. Esses modelos, amplamente utilizados para representação de texto por meio de embeddings, constroem o perfil do usuário a partir dos itens considerados relevantes no cenário e calculam a similaridade desse perfil com os demais itens.

5.3 Métricas de Avaliação e Treinamento dos Modelos

Para avaliar o desempenho dos algoritmos, utilizamos três métricas: Normalized Discounted Cumulative Gain (NDCG)Camargo (2022), que mede a qualidade da ordenação dos itens recomendados; Root Mean Squared Error (RMSE)Avazpour *et al.* (2014), que calcula o erro médio quadrático entre as predições dos modelos e os valores reais; e Mean Average Precision (MAP)Camargo (2022), que pondera a precisão média das recomendações considerando o ordenamento dos itens relevantes. Para o NDCG, utilizamos os 10 primeiros usuários ranqueados pelos algoritmos.

Essas métricas foram escolhidas por serem amplamente utilizadas Avazpour *et al.* (2014) e por possibilitarem uma análise equilibrada entre a qualidade da ordenação e precisão das recomendações.

Além disso, normalizamos o erro médio para a base IFG-Produz nos modelos GraphRec, TF-IDF e BERT, pois esses algoritmos preveem valores dentro do intervalo original de avaliações (1 a 5). Aplicamos uma normalização simples para ajustar os valores ao intervalo de 0 a 1, garantindo uma análise comparável ao restante dos algoritmos. O mesmo foi feito para o KGAT, cujas predições variam entre -1 e 1, necessitando de normalização correspondente.

Os algoritmos GraphRec, KGAT e SR-GNN 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. Em todos os experimentos,

⁶https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

⁷<https://tfhub.dev/google/universal-sentence-encoder-Multilingual/3>

consideramos a média dos resultados em 5 *folds*, com um teste pareado a 95% de confiança para verificar a real significância das métricas.

Já os modelos SASRec e BERT4Rec, baseados em self-attention, utilizam todos os dados para treino, teste e validação. 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.

5.4 Resultados Gerais

Para esta pesquisa, realizamos três grupos principais de experimentos, que aqui são divididos e apresentados em subtópicos, onde observamos os resultados obtidos e algumas análises.

5.4.1 Primeiro experimento: Análise geral dos modelos e bases de dados

O primeiro experimento envolve os cinco algoritmos analisados — GraphRec, KGAT, SR-GNN, SASRec e BERT4Rec — aplicados às três bases de dados estudadas.

Utilizamos as relações de relevância originais de cada base (relevância para o CiteULike, autoria para o AAN+DBLP e avaliação para o IFG-Produz), além de algumas relações adicionais disponíveis, como tags no CiteULike e os atributos tipo de produção, idioma, classificação do evento, natureza e meio de divulgação no IFG-Produz. Essas informações correspondem às conexões primárias das publicações em cada base, ou seja, atributos diretamente vinculados aos itens, e foram aplicadas aos algoritmos que suportam essas conexões, como o GraphRec e KGAT.

Considerando os algoritmos e as bases de dados descritos nos capítulos anteriores, executamos os experimentos com todos os algoritmos e todas as bases de dados. Os resultados e as métricas gerais obtidas se encontram na Tabela 14.

Analisando os resultados por base de dados e considerando os melhores modelos avaliados, no CiteULike, o KGAT apresentou o melhor desempenho com NDCG@10 de 0.8434 e MAP de 0.7878, superando o GraphRec, que obteve 0.7849 e 0.6998, respectivamente, além de um erro médio maior.

No AAN + DBLP, observamos um padrão semelhante, porém GraphRec e KGAT apresentaram resultados muito similares, configurando um empate estatístico com NDCG@10 de 0.83 e MAP entre 0.77 e 0.76, embora o GraphRec tenha exibido um erro médio maior.

Métrica	GraphRec	KGAT	SRGNN	SASRec	BERT4Rec	TF-IDF	BERT
IFG-Produz							
NDCG@10	0.6471	0.7378	0.2351	0.2920	0.1210	0.8848	0.7657
MAP	0.5066	0.6448	0.2416	0.2619	0.1837	0.5110	0.1541
RMSE	0.6950	0.5920	0.9308	0.9763	—	0.5160	0.6344
AAN+DBLP							
NDCG@10	0.8378	0.8237	0.5463	0.3247	0.1871	0.5124	0.5067
MAP	0.7791	0.7612	0.5503	0.3267	0.1947	0.4841	0.4712
RMSE	0.6995	0.3724	0.7131	0.8832	—	0.1737	0.3349
CiteULike							
NDCG@10	0.7849	0.8434	0.1504	0.7310	0.4220	0.8604	0.6086
MAP	0.6998	0.7878	0.1209	0.6767	0.3820	0.1878	0.5410
RMSE	0.7068	0.4679	0.9069	0.7317	—	0.5307	0.6838

Tabela 14 – *Resultados dos algoritmos nos experimentos com as bases IFG-Produz, AAN+DBLP e CiteULike*

Fonte: Desenvolvido pelo autor.

Por fim, no IFG-Produz, o KGAT novamente se destacou como o melhor algoritmo para métricas de ordenação, com NDCG@10 de 0.7378 e MAP de 0.6448, considerando apenas esses modelos de filtragem colaborativa e híbridos.

Os resultados destes dois modelos demonstram que ambos apresentam desempenho positivo e um padrão similar em todas as bases. O KGAT se destacou ao atingir os melhores resultados entre os dois, além de se aproximar mais dos valores reais de predição. Já o GraphRec exibiu um desempenho equilibrado e positivo, especialmente no AAN+DBLP.

Por outro lado, os algoritmos SR-GNN, SASRec e BERT4Rec tiveram o menor desempenho nas três bases. No IFG-Produz, todos apresentaram baixa performance, sendo o BERT4Rec o de pior resultado. É possível atribuir o baixo desempenho desses modelos principalmente à estrutura dos dados. Como explicado anteriormente, nenhuma das bases possuía informações organizadas de forma sequencial nem registros da data em que as conexões foram realizadas. Especialmente no caso do IFG-Produz, perdemos informações importantes que eram os pesos de cada interação. Com sequências geradas aleatoriamente e informações faltantes, era esperado que esses algoritmos apresentassem um desempenho inferior aos demais.

Interessante observar, porém, que para a base AAN+DBLP, o SR-GNN teve um desempenho um pouco melhor dentre estes, superando os modelos baseados em conteúdo. Atribuímos o bom desempenho do SR-GNN a dois fatores. Primeiro, os dados dessa base organizados resultam em diversas sequências de tamanho menor, com informações similares para cada usuário. Conforme descrito por Wu *et al.* (2019) e Kang e McAuley (2018), esses modelos sequenciais e baseados em RNN apresentam dificuldades em lidar com listas extensas de itens e se saem melhor quando há muitas sessões com interações similares.

Embora nenhuma das bases possua múltiplas sequências por usuário, no AAN+DBLP as sequências são, em média, menores que nas outras bases, o que pode explicar o desempenho mais equilibrado do SR-GNN nesses experimentos. Além de maiores, as sequências de itens nas bases CiteULike e IFG-Produz tendem a ser um pouco mais diversificadas, também contribuindo para o resultado observado.

Já no CiteULike, os algoritmos SASRec e BERT4Rec tiveram um desempenho positivo e equilibrado, que atribuímos à capacidade dos modelos baseados em *self-attention* de agrupar itens similares em categorias, conforme descrito por Kang e McAuley (2018). Como essa base apresentou a definição dos itens em algumas poucas categorias bem definidas e similares, organizados por tags, os modelos conseguiram identificar essas relações nas sequências dos usuários, mesmo que fossem relativamente mais longas do que as do AAN+DBLP e IFG-Produz.

Nestes três algoritmos, também podemos atribuir o resultado mais baixo aos poucos experimentos onde foram realizados o ajuste de parâmetros, principalmente no BERT4Rec, que deveria em teoria obter um resultado superior ao obtido pelo SASRec.

Por fim, analisamos os algoritmos de filtragem baseados em conteúdo, TF-IDF e BERT Multilíngue. Esses modelos tiveram uma ótima performance no IFG-Produz e no CiteULike, mas um desempenho um pouco inferior no AAN+DBLP. Associamos esse bom resultado à boa representatividade e similaridade das publicações presentes nos perfis dos usuários.

Durante a análise das bases de dados, observamos que o IFG-Produz e o CiteULike possuem mais publicações por usuário, além de títulos de publicações mais representativos do perfil individual de cada usuário. Dessa forma, os modelos baseados em conteúdo conseguem obter maior precisão nessas bases, mesmo utilizando apenas os títulos como fonte de informação.

Assim, nestes primeiros experimentos, observamos que o KGAT apresentou um dos melhores desempenhos entre os algoritmos de filtragem colaborativa e híbridos, provavelmente devido à sua capacidade de incorporar nativamente atributos adicionais das publicações e demais entidades, proporcionando uma representação

mais rica de usuários e itens, além de uma melhor capacidade de abstração das conexões existentes em todos os cenários.

Além disso, os algoritmos de filtragem baseadas em conteúdo demonstraram um desempenho geral superior, refletindo a boa representatividade dos usuários a partir dos itens relevantes, especialmente nas bases IFG-Produz e CiteULike.

5.4.2 Segundo experimento: IFG-Produz e conexões adicionais de autoria

O segundo experimento foca exclusivamente no IFG-Produz e avalia todos os modelos anteriores, mas com um aumento na quantidade de conexões dos usuários com os itens. Aqui, acrescentamos as conexões de autoria às publicações relevantes de um usuário, além das conexões originais baseadas em avaliação, visto que, assim como na base AAN+DBLP, a autoria pode ser considerada uma relação relevante.

Esse experimento busca aumentar a representatividade do perfil dos usuários sem recorrer a informações adicionais dos itens além daquelas do primeiro experimento, permitindo avaliar o impacto dessas conexões na performance dos algoritmos. Os resultados desse experimento estão apresentados na Tabela 15.

IFG-Produz com conexões adicionais de usuários (autoria)							
Métrica	GraphRec	KGAT	SRGNN	SASRec	BERT4Rec	TF-IDF	BERT
NDCG@10	0.6174	0.6800	0.1871	0.0163	0.0136	0.9201	0.8205
MAP	0.4736	0.5665	0.1533	0.0179	0.0112	0.3133	0.1582
RMSE	0.6844	0.6702	0.9633	0.9957	—	0.3797	0.6251

Tabela 15 – *Resultados dos algoritmos nos experimentos com a base IFG-Produz acrescidos de conexões adicionais de autoria*

Fonte: Desenvolvido pelo autor.

Em uma análise inicial, observamos um decréscimo na performance de todos os algoritmos de filtragem colaborativa, híbrida e baseados em self-attention, enquanto as metodologias baseadas em conteúdo apresentaram melhorias mais evidentes e estatisticamente relevantes.

Uma das principais mudanças ao considerar as conexões de autoria como relevantes na base foi o aumento da esparsidade da base de dados. Como cada usuário possui, em média, 30 publicações (Capítulo 3), mas cada publicação está associada a apenas um usuário, ampliamos a sequência de itens relevantes por usuário, mas mantivemos um número relativamente baixo de conexões por item.

O impacto dessa mudança nos modelos baseados em sessões e self-attention já foi discutido, onde apontamos a partir do referencial que sequências únicas, diversificadas e mais longas tendem a prejudicar o desempenho desses algoritmos, o que podemos perceber em uma grande extensão nesse experimento.

Para os modelos GraphRec e KGAT, a redução na performance pode ser explicada por um fator evidenciado por Rashed, Grabocka e Schmidt-Thieme (2019), que destacam que bases menos esparsas tendem a favorecer os algoritmos de filtragem colaborativa. Ou seja, quanto menor a esparsidade, melhor a representatividade nesses modelos. Assim, ao aumentar a esparsidade, observamos uma queda na performance.

Em ambos os casos, esse efeito pode também estar relacionado ao aumento da diversidade e quantidade dos dados adicionados com essas conexões de usuário, o que impacta a estrutura original da base.

Além disso, não foi realizada uma curadoria temática ou temporal para avaliar as publicações do usuário. Publicações mais antigas podem conter temáticas menos relevantes e distintas das publicações mais recentes e das últimas publicações avaliadas por um pesquisador. Além disso, é possível que os pesquisadores possuam diversas áreas de pesquisa em andamento, mas podem buscar recomendações em áreas específicas.

Interessante observar que, por si só, a autoria e coautoria tendem a ser uma boa métrica representativa de relevância, como observado nos resultados obtidos para a base AAN+DBLP, mas combinadas com as avaliações, causam uma mudança na estrutura das informações que tende a piorar essa performance.

Por outro lado, observamos um aumento considerável na performance dos algoritmos de filtragem baseados em conteúdo, onde houve um aumento no NDGC@10 e uma diminuição do MAP e erro médio. Atribuímos esse resultado justamente ao aumento na quantidade de itens por usuário e à diversidade desses dados, uma vez que as conexões entre usuários e itens não afetam diretamente o desempenho desses modelos. Com mais itens relevantes atribuídos a cada usuário, a representatividade de seu perfil melhora, o que aumenta a assertividade dos modelos TF-IDF e BERT Multilíngue. Dessa forma, o modelo recomenda as publicações relevantes em uma posição mais alta do rank, mas erra um pouco mais em relação às demais.

Esse é um resultado que reforça a qualidade e riqueza informacional dos títulos das publicações na base IFG-Produz, demonstrando que esses atributos são bem estruturados e úteis para outros sistemas de recomendação.

5.4.3 Terceiro experimento: KGAT e conexões adicionais de publicações

Por fim, o último experimento envolveu a execução do KGAT, que apresentou o melhor desempenho entre os modelos baseados em grafos analisados, aplicado a base IFG-Produz com a adição de conexões adicionais relacionadas aos itens. Essas conexões incluem dados de conferência e revistas, como as conexões destas com as publicações e as áreas de atuação das mesmas. Os resultados desse experimento estão apresentados na Tabela 16.

IFG-Produz com conexões adicionais de publicações	
Métrica	KGAT
NDCG@10	0.7739
MAP	0.7082
RMSE	0.5027

Tabela 16 – *Resultados dos algoritmos no experimento com a base IFG-Produz acrescido de conexões adicionais das publicações*

Fonte: Desenvolvido pelo autor.

Podemos perceber um leve aumento na performance, com o NDCG@10 atingindo 0.7739, em comparação com 0.7378 no experimento inicial. No entanto, ao considerar as métricas de comparação pareadas, não podemos afirmar que essa melhoria seja estatisticamente significativa.

Embora as conexões adicionais das publicações tenham ampliado o grafo de conhecimento utilizado pelo KGAT, os resultados indicam que essa expansão não aprimora significativamente a representatividade dos dados, ao contrário do que esperávamos.

É possível que os usuários permaneçam agrupados pelas mesmas informações gerais, sugerindo que essas novas conexões específicas não acrescentam distinções tão relevantes para a personalização das recomendações.

Uma possível explicação para isso é a grande quantidade de projetos de pesquisa na base de avaliações, que não possuem informações de revista ou conferência, reduzindo a representatividade dessas conexões e limitando seu impacto na recomendação.

Além disso, sabemos que o algoritmo desenvolvido por Wang *et al.* (2019) atribui um peso maior às conexões diretas entre usuários e publicações do que às

conexões adicionais dos itens. Embora essas conexões adicionais sejam importantes, seu impacto é menos expressivo, o que sugere que para influenciar significativamente a performance do modelo, essas conexões precisam ser altamente representativas.

Ou seja, ao adicionar apenas essas conexões adicionais dos itens, não conseguimos um volume de informações relevantes o suficiente para melhorar significativamente a performance do algoritmo, ao menos no nível testado.

Esse resultado é reforçado pelos experimentos anteriores, que indicam uma maior representatividade nos itens avaliados diretamente pelo usuário do que em suas conexões com outros usuários semelhantes. Isso é suportado ainda pelo bom desempenho dos algoritmos baseados em filtragem de conteúdo, que dependem exclusivamente das características dos itens para gerar recomendações.

5.5 Resumo das análises e conclusões

Os resultados obtidos nos apresentam um cenário otimista em relação à aplicação do IFG-Produz para a implementação de diferentes modelos de recomendação, principalmente em relação àqueles baseados em conteúdo.

Para os modelos de filtragem colaborativa e híbrida, com um foco na utilização de estruturas de grafo, como o GraphRec e o KGAT, notamos que o IFG-Produz obteve um resultado estatisticamente inferior às demais bases de dados, com NDCG@10 variando aproximadamente entre 0.65 e 0.73 para essas abordagens, enquanto para as demais bases esse valor foi entre 0.79 e 0.84. Podemos atribuir esse fenômeno ao menor número de usuários e conexões existentes no IFG-Produz. O CiteULike possui mais usuários e uma média de conexões por usuário maior, enquanto o AAN+DBLP conta com uma menor média de conexões por usuário, mas compensa com um número significativamente maior de usuários.

Já os algoritmos voltados para predição sequencial, como SR-GNN e SAS-Rec, em geral, não apresentaram um bom desempenho, resultando em baixos valores de NDCG e MAP e um alto erro médio. Esse comportamento era esperado, pois nenhuma das bases de dados analisadas registra a ordem temporal das interações dos usuários com os itens, nem as divide em sessões, prejudicando o desempenho desses modelos que dependem de sequências. Além disso, esses modelos não consideram o peso das conexões, que é uma informação muito importante no caso do IFG-Produz. A única exceção a esse baixo resultado é a da base CiteULike para o modelo SAS-Rec, que obteve um bom NDCG@10 (0.7310), porém com alto erro médio (0.7317), resultado que atribuímos ao escopo limitado de categorias das publicações presentes na base de dados.

Já no caso dos algoritmos com filtragem baseada em conteúdo, também obtivemos resultados positivos para todas as bases de dados, porém percebemos um resultado superior para o IFG-Produz em relação às demais. Atribuímos esse bom resultado à representatividade dos títulos presentes nas demais bases de dados, que se limitam a áreas do conhecimento específicas e geram um menor impacto na representatividade do perfil dos usuários. Isso é refletido ainda mais no segundo experimento, onde ampliamos a quantidade de conexões e, consequentemente, publicações por usuários. Nesse experimento, as análises realizadas com TF-IDF e BERT-Multilíngue demonstram uma ótima performance para o IFG-Produz, que atingiu um NDCG@10 de 0.9201 no experimento com o TF-IDF e 0.8205 para o BERT. Isso indica justamente que as informações descritivas presentes no IFG-Produz são capazes de representar de forma precisa o perfil dos usuários que as avaliam. Já as análises dos demais modelos nesse experimento demonstram uma diminuição na performance com relação aos experimentos anteriores, indicando que as conexões de autoria, nesse contexto somadas às conexões de avaliação, aumentam a esparsidade dos dados e introduzem uma diversidade de informação que prejudica a performance dos mesmos.

Um último experimento, realizado utilizando o KGAT com conexões adicionais das publicações no IFG-Produz, não demonstrou nenhuma melhora estatística de performance com relação aos demais experimentos com esse modelo e base de dados, mantendo um NDCG@10 próximo de 0.77. Esse fenômeno sugere que as novas conexões específicas escolhidas para o experimento (revistas e conferências) não acrescentam distinções tão relevantes para a personalização das recomendações dentro do IFG-Produz, ou que, neste cenário específico, podem ter sua representatividade dissolvida dentre os projetos de pesquisa existentes na base, que não possuem essas informações adicionais.

Em uma visão geral, podemos perceber que os resultados obtidos nos experimentos com o IFG-Produz se alinham com resultados nas demais bases de dados, principalmente o CiteULike. Percebemos um resultado positivo nos modelos com resultado baseado em conteúdo e modelos de filtragem híbrida e colaborativa, e um pior resultado nos modelos de recomendação sequenciais e baseados em *self-attention*.

Esses resultados apresentam fortes indícios de que experimentos realizados com a base IFG-Produz resultam em uma análise condizente com a realidade dos cenários de aplicação da recomendação de artigos acadêmicos. Além disso, a quantidade de informações existentes na base disponibilizada permite análises mais aprofundadas em relação às que foram realizadas na presente pesquisa, explorando novas conexões e distribuição dos dados em diferentes modelos de recomendação.

Um possível estudo futuro envolve a aplicação do KGAT (ou outros modelos de filtragem colaborativa) em um grafo ainda mais complexo, incorporando não apenas as publicações relevantes e autorias dos usuários, mas também informações como palavras-chave do perfil dos usuários, áreas de aplicação das revistas e outros atributos. Isso permitiria analisar o comportamento do modelo diante de uma estrutura mais rica em conexões e verificar se tais informações, em conjunto, impactam significativamente seu desempenho.

Além disso, outra possibilidade promissora seria a utilização de técnicas de *data augmentation* para coletar e integrar mais dados do perfil Lattes dos pesquisadores, de onde essas informações já são atualmente coletadas. Pesquisas nessas abordagens podem incluir a extração automática de novas palavras-chave, temas de pesquisa emergentes e conexões indiretas entre pesquisadores, enriquecendo ainda mais a representação existente e potencialmente melhorando a qualidade das recomendações geradas pelo modelo.

Conclusão

Sistemas de recomendação de artigos científicos são fundamentais para otimizar o acesso a informações pertinentes por parte de pesquisadores. Porém, a escassez de bases de dados apropriadas para a análise e experimentação desses sistemas dificulta a experimentação em diferentes cenários e com diferentes modelos.

Por isso, neste trabalho, apresentamos e disponibilizamos ao público uma nova base de dados voltada para o meio de sistemas de recomendação de artigos científicos, o IFG-Produz. Apresentamos uma descrição dos dados presentes nesta base, realizamos algumas análises com essas informações, apresentando padrões relevantes para esses sistemas, e comparamos os dados e estruturas com mais duas bases de dados, a CiteULike e a AAN+DBLP.

A análise dos dados presentes no IFG-Produz nos mostra que o mesmo inclui informações descritivas inovadoras sobre o perfil dos usuários, como palavras-chave ranqueadas por importância, e sobre as publicações, como dados descritivos de revistas e conferências. Além disso, outras informações comuns às outras bases de dados, como autoria e coautoria ou rótulos das avaliações de relevância, também estão presentes. Essas características apontam que a base IFG-Produz permite a construção de perfis consistentes de usuários e publicações. Além disso, a quantidade elevada de informações adicionais permite a aplicação de modelos mais complexos capazes de abstrair e explorá-las de maneira eficiente.

Já os experimentos práticos demonstraram que a base de dados apresenta ótimos resultados na aplicação em modelos de filtragem baseados em conteúdo, resultados positivos, porém inferiores às demais bases de dados nos modelos baseados em filtragem colaborativa e híbrida. Também foi demonstrado que aumentar a quantidade de conexões adicionando as conexões de autoria dos usuários na base IFG-Produz não melhora sua performance nos algoritmos baseados em grafo, mas melhora o ranqueamento de itens relevantes nos modelos de filtragem baseados em conteúdo.

Atribuímos esse padrão de comportamento, no geral, à boa representatividade dos usuários presentes da base de dados do IFG-Produz, tanto por meio de

suas avaliações de recomendações quanto por suas conexões, mas principalmente por meio dos títulos e termos descritivos das publicações interagidas.

Contudo, este estudo representa apenas um primeiro passo na exploração das possibilidades oferecidas pelo IFG-Produz. Em particular, sugerimos futuros trabalhos que realizem a investigação de filtragem colaborativa com diferentes conexões adicionais dos usuários e publicações, bem como no aprimoramento das abordagens baseadas em conteúdo, aproveitando a diversidade de informações disponíveis na base que aqui foram apresentadas.

Considerando a estrutura geral e construção do dataset, outra abordagem interessante seria explorar a possibilidade da utilização de técnicas de *data augmentation* para coletar e integrar mais dados do perfil Lattes dos pesquisadores aos dados já existentes. O enriquecimento das representações apresenta um bom potencial de melhorar a qualidade das recomendações geradas pelo modelo.

Além disso, novas análises do impacto de diferentes configurações de dados e hiperparâmetros na performance dos algoritmos podem fornecer "insights" valiosos para a otimização dos sistemas de recomendação nessa base de dados.

Dessa forma, espera-se que este trabalho e a base de dados disponibilizada sirvam como base para futuras pesquisas, incentivando a comunidade a explorar e aprimorar os sistemas de recomendação acadêmicos, contribuindo para uma disseminação mais eficiente e personalizada do conhecimento científico.

Referências

AGGARWAL, Charu C. Content-based recommender systems. *In: _____. **Recommender Systems: The Textbook***. Cham: Springer International Publishing, 2016. p. 139–166. ISBN 978-3-319-29659-3. Disponível em: https://doi.org/10.1007/978-3-319-29659-3_4. 16

ALZOGHBI, Anas; AYALA, Victor Anthony Arrascue; FISCHER, Peter M; LAUSEN, Georg. **Pubrec: Recommending publications based on publicly available meta-data**. 2015. Disponível em: https://opus.bibliothek.uni-augsburg.de/opus4/frontdoor/deliver/index/docId/59481/file/D01_CRC69_Alzoghbi.pdf. Acesso em: 19 dez 2024. 17

AVAZPOUR, Iman; PITAKRAT, Teerat; GRUNSKE, Lars; GRUNDY, John. Dimensions and metrics for evaluating recommendation systems. *In: _____. **Recommendation Systems in Software Engineering***. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. p. 245–273. ISBN 978-3-642-45135-5. Disponível em: https://doi.org/10.1007/978-3-642-45135-5_10. 66

BEEL, Joeran; CAREVIC, Zeljko; SCHAIBLE, Johann; NEUSCH, Gabor. **RARD: The Related-Article Recommendation Dataset**. 2017. Disponível em: <https://arxiv.org/abs/1706.03428>. 20

CAMARGO, Adriano C. M. **Categorização e Avaliação de Algoritmos de Sistemas de Recomendação na base de dados IFG-Produz**. 2022. Disponível em: <https://repositorio.ifg.edu.br/handle/prefix/1869>. Acesso em: 10 out 2024. 42, 44, 66

CAZELLA, Sílvio César; NUNES, MASN; REATEGUI, Eliseo. A ciência da opinião: Estado da arte em sistemas de recomendação. **André Ponce de Leon F. de Carvalho; Tomasz Kowaltowski.(Org.). Jornada de Atualização de Informática-JAI**, p. 161–216, 2010. 15, 16, 17

GAO, Chen; ZHENG, Yu; LI, Nian; LI, Yinfeng; QIN, Yingrong; PIAO, Jinghua; QUAN, Yuhao; CHANG, Jianxin; JIN, Depeng; HE, Xiangnan; LI, Yong. A survey

of graph neural networks for recommender systems: Challenges, methods, and directions. **ACM Trans. Recomm. Syst.**, Association for Computing Machinery, New York, NY, USA, v. 1, n. 1, mar. 2023. Disponível em: <https://doi.org/10.1145/3568022>. 48

HE, Xiangnan; LIAO, Lizi; ZHANG, Hanwang; NIE, Liqiang; HU, Xia; CHUA, Tat-Seng. **Neural Collaborative Filtering**. 2017. Disponível em: <https://arxiv.org/abs/1708.05031>. 29

KANG, W.; MCAULEY, J. **Self-Attentive Sequential Recommendation**. 2018. Disponível em: <https://arxiv.org/abs/1808.09781>. Acesso em: 16 out 2024. 13, 55, 56, 57, 58, 60, 64, 69

KIPF, T. N.; WELING, M. **Semi-Supervised Classification with Graph Convolutional Networks**. 2017. Disponível em: <https://arxiv.org/abs/1609.02907>. Acesso em: 16 out 2024. 46

KREUTZ, Christian K.; SCHENKEL, Ralf. **Scientific paper recommendation systems: a literature review of recent publications**. 2022. 335–369 p. Disponível em: <https://doi.org/10.1007/s00799-022-00339-w>. 12, 18, 19, 20, 22

L, Hao; LIU, Shijun; PAN, Li. Paper recommendation based on author-paper interest and graph structure. *In: 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. [S.l.: s.n.], 2021. p. 256–261. 17, 25, 26

RASHED, A.; GRABOCKA, J.; SCHMIDT-THIEME, L. **Attribute-Aware Non-Linear Co-Embeddings of Graph Features**. RecSys '19: Proceedings of the 13th ACM Conference on Recommender Systems, 2019. 314–321 p. Disponível em: <https://dl.acm.org/doi/10.1145/3298689.3346999>. Acesso em: 10 out 2024. 13, 44, 45, 46, 47, 64, 71

SANCHEZ-LENGELING, Benjamin; REIF, Emily; PEARCE, Adam; WILTSCHEK, Alexander B. A gentle introduction to graph neural networks. **Distill**, 2021. <https://distill.pub/2021/gnn-intro>. 12, 15

SUGIYAMA, Kazunari; KAN, Min-Yen. Scholarly paper recommendation via user's recent research interests. *In: Proceedings of the 10th Annual Joint Conference on Digital Libraries*. New York, NY, USA: Association for Computing Machinery, 2010. (JCDL '10), p. 29–38. ISBN 9781450300858. Disponível em: <https://doi.org/10.1145/1816123.1816129>. 12, 21, 23, 24, 38, 63

SUN, Fei; LIU, Jun; WU, Jian; PEI, Changhua; LIN, Xiao; OU, Wenwu; JIANG, Peng. **BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer**. 2019. Disponível em: <https://arxiv.org/abs/1904.06690>. Acesso em: 16 nov 2024. 13, 58, 59, 61, 64

TANG, Hao; LIU, Baisong; QIAN, Jiangbo. Content-based and knowledge graph-based paper recommendation: Exploring user preferences with the knowledge graphs for scientific paper recommendation. **Concurrency and Computation: Practice and Experience**, v. 33, n. 13, p. e6227, 2021. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.6227>. 48

W., Taylor. **"Cloze Procedure": A New Tool For Measuring Readability**. 1953. Disponível em: <https://gwern.net/doc/psychology/writing/1953-taylor.pdf>. Acesso em: 16 nov 2024. 61

WANG, Chong; BLEI, David M. Collaborative topic modeling for recommending scientific articles. *In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2011. (KDD '11), p. 448–456. ISBN 9781450308137. Disponível em: <https://doi.org/10.1145/2020408.2020480>. 23, 31, 38, 40

WANG, Hao; CHEN, Binyi; LI, Wu-Jun. Collaborative topic regression with social regularization for tag recommendation. *In: IJCAI*. [S.l.: s.n.], 2013. 12, 18, 21, 22, 23, 31, 40, 63

WANG, X.; HE, X.; CAO, Y.; LIU, M.; CHUA, T. **KGAT: Knowledge Graph Attention Network for Recommendation**. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019. 950-958 p. Disponível em: <https://arxiv.org/pdf/1905.07854>. Acesso em: 14 out 2024. 13, 47, 48, 49, 50, 51, 64, 72

WU, S.; TANG, Y.; ZHU, Y.; WANG, X. Xie L.; TAN, T. **Session-based Recommendation with Graph Neural Networks**. 2019. Disponível em: <https://arxiv.org/pdf/1811.00855>. Acesso em: 16 out 2024. 13, 17, 52, 53, 54, 55, 57, 64, 69

YAN, Jihong; XU, Chen; LI, Na; GAO, Ming; ZHOU, Aoying. Optimizing model parameter for entity summarization across knowledge graphs. **Journal of Combinatorial Optimization**, v. 37, 01 2019. 49