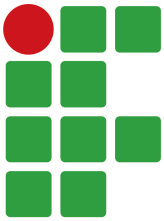


INSTITUTO FEDERAL
Goiás
Câmpus Anápolis

Adriano César de Melo Camargo

Categorização e Avaliação de Algoritmos de Sistemas de Recomendação na base de dados IFG-Produz

Anápolis-GO

2022

Adriano César de Melo Camargo

Categorização e Avaliação de Algoritmos de Sistemas de Recomendação na base de dados IFG-Produz

Trabalho de Curso submetido ao Instituto Federal de Goiás - Câmpus Anápolis como parte dos requisitos necessários para a obtenção do título de Bacharel em Ciência da Computação.

Instituto Federal de Goiás - Câmpus Anápolis

Orientador: Prof. Dr. Daniel Xavier de Sousa

Anápolis-GO
2022

Dados Internacionais de Catalogação na Publicação (CIP)

Camargo, Adriano César de Melo

C173c Categorização e avaliação de algoritmos de sistemas de recomendação na base de dados IFG-Produz. / Adriano César de Melo Camargo. – Anápolis: IFG, 2022.

35 f. il. color.

Orientador: Prof. Dr. Daniel Xavier de Sousa.

Trabalho de Conclusão de Curso – Instituto Federal de Educação, Ciência e Tecnologia de Goiás: Campus Anápolis – Bacharelado em Ciência da Computação, 2022.

1. Sistema de recomendação. 2. recuperação de informação. 3. processamento de linguagem natural. 4. aprendizado de máquinas. 5. publicações científicas.

I. Sousa, Daniel Xavier de (orient.).

III. Título.

CDD 004

Ficha catalográfica elaborada pela bibliotecária Claudineia Pereira de Abreu CRB1/1956
IFG Campus Anápolis

INSTITUTO FEDERAL
Goiás

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

ATA DA SESSÃO PÚBLICA DE APRESENTAÇÃO E DEFESA DO TRABALHO DE CONCLUSÃO DE CURSO

Aos **dezenove** dias do mês de **Dezembro** de **2022**, às **08h00**, por meio de webconferência via Google Meet, foi realizada a sessão pública de apresentação e qualificação do Trabalho de Conclusão de Curso II do graduando **Adriano César de Melo Camargo (matrícula 20191060140142)** do curso de Bacharelado em Ciência da Computação. A banca foi composta pelos seguintes membros: **Dr. Daniel Xavier de Sousa, Dr. Sérgio Daniel Carvalho Canuto, e Dr. Reinaldo Silva Fortes (UFOP)** , sob a presidência do primeiro. O trabalho de conclusão de curso tem como título **Categorização e Avaliação de Algoritmos de Sistemas de Recomendação na base de dados do IFG-Produz**, sob orientação de **Daniel Xavier de Sousa**. Após a apresentação, tendo sido o autor arguido pela Banca Examinadora, a nota obtida foi **9,7** e o Trabalho de Conclusão de Curso foi considerado **APROVADO** pela banca examinadora.

Encerra-se a presente sessão às **09** horas e **20** minutos. Eu, **Daniel Xavier de Sousa**, dato e assino a presente ata que segue assinada por todos os membros da Banca e pelo graduando.

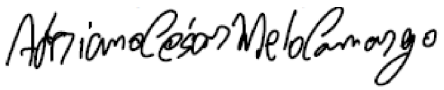
Dr. Daniel Xavier de Sousa
(Assinado Eletronicamente)

Dr. Sergio Daniel Carvalho Canuto
(Assinado Eletronicamente)

Dr. Reinaldo Silva Fortes

REINALDO SILVA FORTES:00521452635
Assinado de forma digital por REINALDO SILVA FORTES:00521452635
Dados: 2022.12.19 10:43:30 -03'00'

Adriano César de Melo Camargo
(Graduando)



Documento assinado eletronicamente por:

- **Sergio Daniel Carvalho Canuto**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 19/12/2022 15:19:50.
- **Daniel Xavier de Sousa**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 19/12/2022 11:02:39.

Este documento foi emitido pelo SUAP em 19/12/2022. 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: 358293
Código de Autenticação: 1267fc22bd

Instituto Federal de Educação, Ciência e Tecnologia de Goiás
Avenida Pedro Ludovico, s/ nº, Reny Cury, ANÁPOLIS / GO, CEP 75131-457
Sem Telefones cadastrados

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
[X] 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. (X) 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 , 01/12/2023.
Local Data

Adriano Cesar Machado Camargo
Assinatura do Autor e/ou Detentor dos Direitos Autorais

Lista de ilustrações

Lista de tabelas

Sumário

Resumo

Com a produção acadêmica científica em crescimento, encontrar publicações relevantes para pesquisadores se tornou uma tarefa desafiadora. Com técnicas de sistemas de recomendações de artigos científicos essa dificuldade é mitigada gerando recomendações específicas para os pesquisadores. Inclusive, como afirma a literatura, algoritmos de recomendação de artigos científicos tendem a ter resultados mais efetivos que a ação dos pesquisadores fazerem buscas em máquinas de busca. Isso ocorre, pois os algoritmos de recomendação conseguem usar dados diversos, como histórico do pesquisador, área de atuação, e artigos publicados, especificando assim os possíveis contextos na associação com outros artigos. Assim, o trabalho apresentado neste documento disponibiliza uma nova base de dados pública e anotada para sistema de recomendação de artigos científicos, com pesquisadores de diversas áreas. Além da nova base de dados, apresentamos uma caracterização da mesma, comparando inclusive com outras bases de dados disponíveis na literatura. Dentre as análises feitas, incluímos a comparação de diversos algoritmos de sistema de recomendação com as bases estudadas. Segundo os resultados experimentais, a base de dados disponibilizada possui melhor efetividade ao aplicar estratégias com filtragem baseada em conteúdo, principalmente devido à riqueza em que os atributos descrevem os artigos e projetos. Os experimentos mostram que bases de dados com menor esparsidade tendem a ter melhor resultado aplicando estratégias com filtragem colaborativa. De forma geral, entendemos que esse trabalho contribui para a área de recomendação de artigos científicos fornecendo uma nova base de dados para futuros trabalhos da comunidade científica, ao mesmo tempo que apresenta um comparativo de técnicas no estado da arte de sistemas de recomendação.

Palavras-chave: Sistema de Recomendação, Recuperação de informação, Processamento de Linguagem Natural, Aprendizado de Máquinas, Publicações Científicas.

Abstract

With the growth of scientific academic productions over the years, finding relevant publications has become a difficult task. Considering scientific article recommendation systems, this difficulty is mitigated by recommending specific research items. Following the claims in the literature, algorithms for recommending scientific articles tend to have more effective results than researchers searching search engines. Due to the recommendation algorithms can use different data, such as the researcher's background, area of expertise, and published articles, thus specifying the possible contexts in association with other articles. Thus, the work presented in this document provides a new public and annotated database for the recommendation system of scientific articles, with research in several areas. Besides the new database, we present a characterization of it, comparing it with other databases available in

the literature. Among the analyses, we included the comparison of several recommendation system algorithms. According to the experimental results, the available database has better effectiveness when applying content-based filtering strategies, mainly due to the good textual description of articles and projects. The experiments show that databases with less sparsity tend to have better results when applying collaborative filtering strategies. In general, this work contributes to the area of recommendation of scientific articles by providing a new database for future works, while presenting a comparison of techniques in the state of the art of recommendation systems.

Keywords: Recommender System, Information Retrieval, Natural Language Processing, Machine Learning, Scientific Productions.

1 Introdução

A produção acadêmica científica tem crescido de forma vigorosa nos últimos anos. Dados descritos em (CENTRO. . . , 2021) mostram que as produções biográficas no mundo cresceram cerca de 27%, em especial no Brasil registra-se um aumento de 32% de 2021 em relação a 2015. Em meio a tantos dados, a tarefa de encontrar publicações relevantes tem se tornado extremamente difícil, impactando os pesquisadores ao realizarem suas pesquisas acadêmicas.

Dentro da área de Recuperação de Informação, uma tarefa que tenta mitigar esta dificuldade é a de Sistemas de Recomendação. De forma específica, existe uma área de pesquisa conhecida como **recomendação de artigos científicos em comunidades científicas** (BAI et al., 2019). Nessa linha, plataformas como ResearchGate (O'BRIEN, 2019) e Google Scholar (JACSÓ, 2005) visam aprender o interesse de seus usuários e recomendar produções bibliográficas mais relevantes possíveis. Por exemplo, pesquisadores em início de carreira com poucas publicações podem exigir um comportamento de recomendação diferente de pesquisadores veteranos (BAI et al., 2019). No caso do primeiro, os sistemas podem recomendar artigos menos alinhados aos artigos do pesquisador e ainda alguns clássicos de uma determinada área para ampliar suas atuações. No caso dos veteranos, os sistemas devem recomendar principalmente artigos mais alinhados às linhas de pesquisa de maior interesse.

Contudo, a aplicação destas ferramentas já existentes apresenta um escopo geral de pesquisadores, com usuários espalhados em todo o mundo e dificilmente aplicadas em um cenário restrito, com o objetivo, por exemplo, de conectar os pesquisadores de uma instituição com várias unidades. De fato, uma tarefa bastante dispendiosa nas instituições acadêmicas de pesquisa é manter seu corpo de pesquisadores conectados, ou sabedores da produção bibliográfica gerada por seus pares. Em instituições multi campus – com várias centenas de pesquisadores e dezenas de unidades – isso fica ainda mais difícil, pois sem algum elemento facilitador um pesquisador pode não ter conhecimento dos trabalhos do colega de mesma área e em campus distante. Em algumas instituições o elemento facilitador permanece como atribuição de um servidor, que após avaliar manualmente as publicações envia e-mails informativos aos colegas sobre as bibliografias recentemente publicadas. Numa tentativa de automatizar esse processo e facilitar a conexão entre os pesquisadores, uma ferramenta existente é o IFG-Produz (FLAUBERTH et al., 2020).

O IFG-Produz (FLAUBERTH et al., 2020)[1] é uma ferramenta desenvolvida pelo Instituto Federal de Goiás, e conecta seus pesquisadores fazendo o trabalho de processamento

[1] ifgproduz.ifg.edu.br

automatizado e recomendação de produções bibliográficas, segundo o perfil individual de cada pesquisador. Para isso, a ferramenta armazena os dados dos currículos Lattes de todos os pesquisadores do IFG, obtidos via Conselho Nacional de Pesquisa (CNPq). Mediante a atualização de seus currículos, o IFG-Produz identifica pesquisadores e seus perfis, recomendando os trabalhos que possam ser relevantes aos seus pares.

Contudo, embora em funcionamento e amplamente difundida no IFG, a ferramenta carece de diversas ampliações, como exemplo: caracterização da base de dados, avaliação de algoritmos estado da arte na área de Sistemas de Recomendação, aplicação de tarefas que visam ampliar a avaliação dos pesquisadores sobre os itens avaliados, e construção de interfaces que permitam visualizar os dados no processo de recomendação. Ampliações essas que compõem o objetivo deste Trabalho de Conclusão de Curso (TCC) e que tenta avançar com a Plataforma IFG-Produz, desenvolvendo novas funcionalidades e dando mais robustez aos métodos aplicados. Detalhamos um pouco mais as contribuições neste TCC.

1.1 Caracterização e Disponibilização de uma Nova Base de Dados de Recomendação

Considerando a carência de dados públicos rotulados para a tarefa de recomendação de artigos científicos, e ainda, de usuários em diversas áreas do conhecimento que avaliam as recomendações, uma contribuição deste trabalho é a disponibilização pública e caracterização da base de dados IFG-Produz.

Além disso, no intuito de permitir analisar a participação dos pesquisadores sobre as avaliações das recomendações, disponibilizamos a instalação de uma ferramenta de visualização dos dados. Ou seja, aplicamos ferramentas com abordagem de dados não estruturados e sistemas dinâmicos de análise, como é o caso de PowerBI (FERRARI; RUSSO, 2016). O maior objetivo com esta ferramenta é analisar a interação dos usuários com o sistema de forma individualizada por área de atuação do pesquisador e pelo campus em que ele atua.

1.2 Avaliação dos Algoritmos no Estado da Arte em Sistemas de Recomendação

Os algoritmos de recomendação podem ser divididos em três categorias principais: filtragem baseadas em conteúdo, filtragem colaborativa e filtragem mista. Na primeira, os itens são recomendados para os usuários de acordo com o que ele interagiu anteriormente, levando-se em conta o quão parecido os itens são. Na estratégia colaborativa as recomendações são feitas de acordo com o perfil do usuário, então é considerado o comportamento

de cada pessoa analisando os itens avaliados e comparando com outros usuários. No caso da híbrida, a tentativa é misturar as duas primeiras, considerando o padrão entre os usuários e a similaridade dos itens. Atualmente o IFG-Produz utiliza-se de uma estratégia baseada em filtragem baseada em conteúdo, conforme descrito na seção 2.1 e utilizando uma representação de dados com estratégia $bag - of - words$. Como primeira contribuição deste trabalho faremos uma análise do algoritmo de filtragem baseada em conteúdo mas considerando outras representações dos dados, com as estratégias distribucionais ou embeddings. Para além disso, alguns trabalhos descrevem que filtragem colaborativa tem sido mais utilizada, por fornecer recomendações bastante precisas e por possuir algoritmos já bastante maduros (ISINKAYE; FOLAJIMI; OJOKOH, 2015; CHO; KIM, 2004). Desta forma, outra contribuição deste trabalho é explorar também uma análise com uso dos algoritmos do estado da arte em filtragem colaborativa.

Como resultado, a partir dos experimentos realizados e descritos no Capítulo 4, pode-se perceber que apesar das técnicas de filtragem colaborativas serem mais usadas, para os dados do IFG-Produz os melhores resultados foram obtidos utilizando as técnicas baseadas na filtragem por conteúdo, muito provavelmente pela qualidade da descrição dos dados das produções científicas presente na base. Ainda, constatamos que a pouca avaliação das produções pelos pesquisadores tem dificultado a construção de modelos representativos para as estratégias com filtragem colaborativa.

Acreditamos que o trabalho aqui apresentado irá contribuir com uma nova base de dados na área de recomendação de artigos científicos, com avaliação de pesquisadores de diversas áreas. Ainda, como a base de dados está em uso e em constante crescimento, seja de novos artigos e projetos cadastrados, será crescente também as novas avaliações perante as recomendações enviadas.

O texto é organizado dividido em capítulos, onde o Capítulo 2 descreve os métodos e técnicas utilizadas na pesquisa, o Capítulo 3 de resultados parciais descreve os resultados encontrados até agora e o Capítulo 4 de cronogramas descreve os trabalhos que faltam ser realizados.

2 Fundamentação Teórica

Sistemas de Recomendação são estratégias que visam auxiliar o acesso à informações relevantes dentro de um contexto de grande volume de dados(LÜ et al., 2012). Eles ajudam aplicando técnicas de Recuperação de Informação que eliminam tópicos ou dados desnecessários, selecionando conteúdos e itens personalizados e exclusivos. Várias abordagens para construção desses sistemas foram desenvolvidas, que utilizam de diferentes técnicas como(ISINKAYE; FOLAJIMI; OJOKOH, 2015): filtragem colaborativa, filtragem baseada em conteúdo ou filtragem híbrida .

Descreveremos agora em mais detalhes algumas destas técnicas, apontando alguns procedimentos necessários para entendimento deste trabalho.

2.1 Filtragem Baseada em Conteúdo

Nas técnicas de filtragem baseada em conteúdo é dada ênfase na similaridade dos itens para gerar as recomendações, especialmente em cenários de *cold-start*. No contexto do *cold-start*, em que há pouco conhecimento sobre os gostos dos usuários e escassas avaliações dos itens, a estratégia consiste em apresentar novas opções de itens com base em conteúdos similares aos já conhecidos, como mostrado na Figura 1. A estratégia é geralmente utilizada quando se tem pouco conhecimento dos gostos dos usuários e há poucas avaliações dos itens, onde então é utilizado os conteúdos dessas interações anteriores(AGGARWAL, 2016). Uma questão relevante dessa aplicação é que ao usuário geralmente é recomendado mais do mesmo, ou seja, prioritariamente itens bastante associados ao que ele já teve experiência.

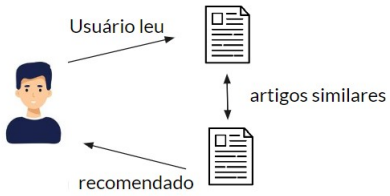


Figura 1 – Diagrama de recomendação Baseada em Conteúdo

Na prática duas partes se destacam na aplicação de algoritmos de Filtragem Baseada em Conteúdo: i) aplicação de algoritmos de agrupamento entre os itens, e como utilizá-los para recomendar itens não vistos, e ii) como representar os itens de forma que o cálculo de similaridade no agrupamento seja o mais assertivo possível.

Para o caso de i), a literatura tem sugerido a utilização de algoritmos que ponderam os itens de interesse e agrupa com outros itens parecidos, como é o caso do algoritmo

Rocchio(JOACHIMS, 1996). Para ii) existem diversas estratégias de representação dos dados, e detalharemos essas representações na próxima seção.

2.1.1 Análise de agrupamento entre itens.

Dentre os algoritmos que agrupam entre os mais similares, o mais utilizado é o Rocchio, descrito no Algoritmo 1. O algoritmo recebe como entrada uma matriz X de similaridade entre os itens, as avaliações dos usuários para cada item e também quais itens são relevantes para os usuários. Seguindo o algoritmo, na linha 1 uma matriz é inicializada, recebendo o nome de *results*, onde as linhas são os usuários e as colunas os itens, sendo preenchida com valores 0 inicialmente. A matriz *results* será responsável por guardar a pontuação calculada para cada combinação. Na linha 3 é feita a iteração sobre todos os itens. Na linha 4 e 5 é inicializado duas variáveis com 0, utilizadas para somatório das pontuações da pontuação (*score*), e para normalização desta pontuação (*sum*). Na linha 6 é feita a iteração sobre todos os itens avaliados para aquele usuário, seguindo na linha 7 a pontuação calculada pela multiplicação da similaridade do item avaliado (S_{ir}) com a nota do usuário para aquele item (X_ur) (detalhado na próxima seção). Na linha 8 é feito o somatório das notas daquele usuário dos itens avaliados, usado para normalização (linha 10). A linha 10 a pontuação do item é obtida, e armazenada na matriz final, linha 11.

Algorithm 1 Rocchio

Require: U como todos usuários
Require: I como todos os itens
Require: R são itens avaliados para cada usuário
Require: X como a matriz da avaliação de cada usuário pelos itens
Require: S como a matriz de similaridade entre cada item

1: $results \leftarrow zeros(U, I)$
2: **for** u $\leftarrow$ 1 **to** $|U|$ **do**
3: **for** i $\in I$ **do**
4: $sum \leftarrow 0$
5: $score \leftarrow 0$
6: **for** r $\in R_u$ **do**
7: $score \leftarrow score + S_{ir} * X_{ur}$
8: $sum \leftarrow sum + X_{ur}$
9: **end for**
10: $score \leftarrow score/sum$
11: $results_{ui} \leftarrow score$
12: **end for**
13: **end for**

2.1.2 Representação dos Itens.

Um ponto importante nos algoritmos baseado em conteúdo é o cálculo da similaridade entre os itens, que é dependente de como os itens são representados. Considerando

seus atributos, espera-se representá-los de forma que possamos computar a distância de similaridade com efetividade, conseguindo assim recomendar itens de interesse do usuário mas ainda não avaliados. Dentre os métodos de representação, neste trabalho avaliamos duas categorias, bag-of-words e embeddings dinâmicos ou *autoencoders*, respectivamente TF-IDF e padrões que seguem Bidirectional Encoder Representation from Transformers (BERT). Daremos mais detalhes deles na sequência.

O TF-IDF é uma técnica de bag-of-words que atribui peso para as palavras (ou termos) de um texto, considerando a interdependência de cada termo. A ideia é que cada termo seja considerado como uma dimensão independente, aplicando um peso de frequência e inverso da frequência do corpus. Basicamente, o TF-IDF funciona determinando a frequência relativa dos termos em um documento específico, comparado com a proporção inversa desse termo em todo corpus de documentos, e o cálculo determina o quão importante cada termo é para o documento (RAMOS et al., 2003). Considerando uma coleção de documentos D, um termo w e um documento individual d ∈ D (Equação 2.2), o cálculo se dá pela multiplicação da frequência inversa do termo w nos documentos em D (Equação 2.1) pela frequência do termo w e d. A multiplicação é resumida em 2.2.

$$idf(w, D) = log(\frac{|D|}{tf(w, D)}) \tag{2.1}$$

$$tf - idf(w, d, D) = tf(w, d) * idf(w, D) \tag{2.2}$$

Considerando que tf(w, d) é o número de vezes que a palavra w aparece no documento d, e |D| é o tamanho do corpus. A representação criada então a partir desse cálculo é um vetor onde se tem um valor para cada termo do corpus. Esses valores conseguem dar menos peso para palavras que são frequentes em muitos documentos e logo não tem muita relevância ou possui menor poder de discriminação. Ainda, o método atribui maior peso às palavras que podem aparecer muito em certo documento, mas que não é muito frequente em todo o corpus, o que significa que é uma palavra que tem maior potencial de discriminação.

A representação bag-of-words criada é utilizada no cálculo de similaridade. Por exemplo, a Equação 2.3 abaixo aplica a distância dos cossenos, considerando como vetores a representação vetorial de dois artigos A e B. A partir deste cálculo algoritmo Rocchio (descrito acima), consegue calcular os itens mais relevantes e que serão recomendados.

$$cos(A, B) = \frac{A \cdot B}{||A||||B||} \tag{2.3}$$

Uma outra forma de representar os itens, a partir dos seus atributos, é utilizando um conceito mais recente denominado de *Bidirectional Encoder Representations from*

Transformers(DEVLIN et al., 2018) ou BERT. Bastante utilizado na área de Processamento de Linguagem Natural, BERT utiliza-se do contexto das palavras e gera uma representação bastante enriquecida de termos e sequências. Na prática o BERT recebe vetores que são representações distribucionais de termos, construído a partir de estratégias de embeddings estáticos como Word2Vec(RONG, 2014). Com os embeddings estáticos o BERT aplica o análise de contexto, por meio de Transformers (HOANG; BIHORAC; ROUCES, 2019), para analisar os termos de uma sentença independente da direção do processamento, se da esquerda para a direita ou vice-versa. O BERT gera a nova representação aplicando duas estratégias de treinamento semi-supervisionada: Masked Language Models (MLM) e Next Sentence Prediction (NSP).

A MLM seleciona aleatoriamente termos que serão consideradas máscaras, e o treinamento será a tarefa de, dada uma sentença, predizer qual seria o termo correto para máscara (HOANG; BIHORAC; ROUCES, 2019). Segundo o trabalho original do BERT(DEVLIN et al., 2018), a base de dados para este treinamento terá 80% das sentenças com espaços vazios para o termo ditos com máscaras, que deverão receber a nova predição, 10% o termo será uma palavra aleatória, e outros 10% as máscaras terão seus termos mantidos. Assim, como o BERT gera seu próprio conjunto rotulado a partir de uma base sem rótulos, denomina-se uma estratégia semi-supervisionada.

Outra fase importante do treinamento do BERT é denominada Next Sentence Prediction. Neste caso, sentenças são comparadas umas com as outras, e o modelo tenta predizer se uma é sequência da outra. Para o treinamento nessa fase o modelo recebe duas sentenças, sendo que em alguns casos a segunda é uma sequência da primeira, casos positivos, e em outras entradas a segunda não é uma sequência da primeira, caso negativos. Novamente aplicando uma estratégia semi-supervisionada.

Tanto as fases de MLM e NSP são processadas com diversas base de dados e tamanhos simplesmente gigantescos. No artigo original (DEVLIN et al., 2018), o treinamento foi executado com 800 milhões de palavras do BookCorpus(ZHU et al., 2015) e a Wikipédia em inglês, com 2,5 milhões de palavras, onde foram extraídas apenas passagens de texto.

Uma das estratégias bastante utilizadas atualmente para se construir uma nova representação dos dados, é uma tradução do dados iniciais para uma representação distribucional com análises de contextos dinâmicos. Uma dessas técnicas é chamada de Zero-Shot, ou feature-based. O Zero-Shot é feito em duas fases. Primeira se aproveita de um modelo que foi treinado uma grande base de dados, depois é feita a geração da nova representação usando o modelo pré-treinado com um novo conjunto de dados. O resultado desse modelo para o novo conjunto, é uma representação nova, que considerando o grande volume de dados utilizado, possui forte potencial de encapsular na nova representação uma maior representação dos dados. Logo, o Zero-Shot depende da existência de uma base de dados de treinamento, e a expectativa é que o conhecimento adquirido pelo modelo

com o grande volume de dados seja utilizado ao gerar a nova representação. Em alguns casos, o conjunto de treino utiliza textos de diversas línguas e gramáticas, gerando também novas representações enriquecidas em línguas diferentes. Como descrito em (HOANG; BIHORAC; ROUCES, 2019), a representação com dados utilizando BERT ou mesmo outras configurações de transformers como RoBERT, aumentam substancialmente a acurácia em diversas tarefas de aprendizado de máquina.

Além da nova representação obtida com os algoritmos baseados em BERT, é possível ajustar ainda mais o modelo, aplicando algo conhecido como $fine - tunning$, ou ajuste fino. A ideia é fornecer mais alguns dados para o sistema complementar o ajuste dos modelos, mas agora considerando o novo conjunto de dados.

Assim como TF-IDF, a representação BERT usando zero-shot ou fine-tunning, também pode ser usada em um cálculo de similaridade como entrada da Equação 2.3. Consequentemente o algoritmo Rocchio poderá igualmente ser aplicado pela representação gerada pelo BERT, mas neste caso usando a semântica entre os termos.

2.2 Filtragem Colaborativa

Filtragem Colaborativa é uma técnica de recomendação onde não é feita principalmente pelo conteúdo dos itens, mas principalmente na relação das avaliações dos itens entre os usuários. Logo, de forma geral essa filtragem cria uma matriz de itens-usuários que indica as preferências dos usuários para cada item. Então para a recomendação são utilizadas estratégias para achar usuários de interesses comuns, e construir uma representação que seja capaz de associar esses usuários. Por exemplo, usuários podem ser associados por gostarem dos mesmos atores ou gênero de filme. Essa representação é então utilizada para recomendar itens que um usuário ainda não interagiu, mas que os outros usuários similares já deram uma nota positiva.

Diferente da estratégia baseada em conteúdo em que a representação dos itens tem forte peso, agora o objetivo maior é em como construir o mapeamento entre usuários de interesses comuns. Na prática as estratégias baseadas em filtragem colaborativa apresentam duas estratégias diferentes (ISINKAYE; FOLAJIMI; OJOKOH, 2015): i) avaliação de usuários parecidos com aplicações de clusterings, e ii) aplicação de modelos supervisionados que criam uma representação a partir dos rótulos, ou grau de relevância dos itens para os usuários.

Neste trabalho avaliamos as técnicas consideradas no estado-da-artes de filtragem colaborativa, caso ii). Em suma, elas recebem como entrada uma matriz de interação entre usuários e itens, e fornecem as notas ou grau de relevância dos itens que não foram avaliados pelos usuários ainda. Descrevemos na sequência a técnica de NMF, ou NonNegative Matrix Factorization(BERRY et al., 2007) e a estratégia Denominada

GraphRec(RASHED; GRABOCKA; SCHMIDT-THIEME, 2019).

2.2.1 Fatoração de Matriz

De forma básica a fatoração de matriz é uma técnica que recebe uma matriz de relações de itens e usuários, e o objetivo é conseguir outras duas matrizes que, quando combinadas, conseguem se aproximar da matriz original. Modelos de fatoração de matriz mapeiam os usuários U e itens I para um espaço de atributos latente de dimensionalidade F, onde cada iteração de item-usuário pode ser modelada como produto interno nesse espaço (KOREN; BELL; VOLINSKY, 2009). Cada item i pode ser associado a um vetor $q_i \in Q_{FxI}$ e todo usuário u associado ao vetor $p_u \in Q_{UxF}$, onde são as matrizes que se quer inferir a partir da matriz original R, de dimensionalidade |U| x |I|. Os elementos de q_i representam o valor que cada item tem em cada fator f, e p_u o interesse do usuário em itens que têm valores altos em cada fator. O produto escalar, $q_i^T p_u$, é a interação do usuário u e o item i, conseguindo assim aproximar a nota que o usuário daria para o item, e uma matriz $\hat{R}$ que consegue aproximar a original.

Para o algoritmo da fatoração de matriz não negativa, que é avaliado no trabalho, temos o seguinte algoritmo para a optimização de P e Q, onde se quer minimizar a função **min** $||R - PQ||^2_{loss}$.:

Algorithm 2 Optimização NMF

Require: **R** como a matriz de usuários e itens,
Require: **F** dimensão dos fatores latentes.
Require: **max_iter** número máximo de iterações.
Require: α taxa de aprendizado.

1: **P** $\leftarrow$ random(|U|, F) > 0
2: **Q** $\leftarrow$ random(F, |I|) > 0
3: **for** e $\leftarrow$ 1 **to** max_iter **do**
4: $\quad \hat{\mathbf{R}} \leftarrow PQ$
5: $\quad$ **for** u $\leftarrow$ 1 **to** $|U|$ **do**
6: $\quad\quad$ **for** i $\leftarrow$ 1 **to** $|I|$ **do**
7: $\quad\quad\quad \mathbf{P}_u \leftarrow P_u + \alpha * 2(R_{ui} - \hat{R_{ui}})Q_i$
8: $\quad\quad\quad \mathbf{Q}_i \leftarrow Q_i + \alpha * 2(R_{ui} - \hat{R_{ui}})P_u$
9: $\quad\quad$ **end for**
10: $\quad$ **end for**
11: **end for**

O algoritmo funciona por inicializar as matrizes P e Q com valores aleatórios maiores do que zero, como visto na linha 1 e 2. Para cada usuário u e item i é calculado os novos pesos de P e Q utilizando o gradiente da função de custo ($e_{ui} = (r_{ui} - \hat{r_{ui}})^2$), mostrado nas linhas 7 e 8. Tendo P e Q calculado a predição se dá por $\hat{R} = PQ$. Ao final uma nova matriz é gerada, mas agora com valores nos campos em que o usuário não

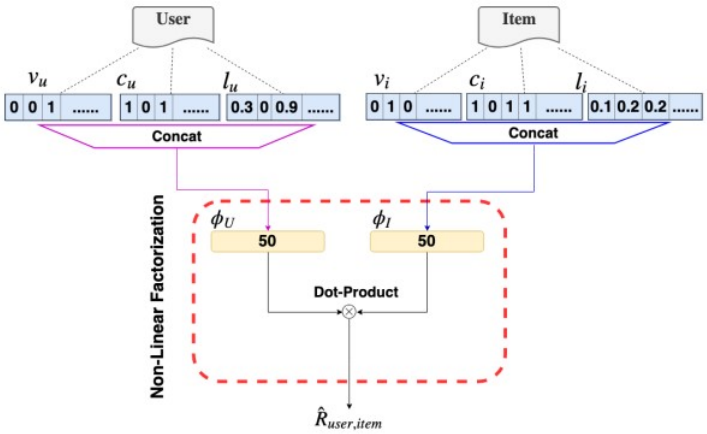
havia avaliado alguns itens. A recomendação será fornecida ordenando os itens com maior pontuação para cada usuário.

2.2.2 GraphRec

Abordagens baseadas em fatoração de matriz têm uma boa performance, mas elas não podem receber atributos externos diretamente como informações adicionais, e para isso é preciso fazer modificações, e essas modificações geram modelos híbridos que podem não generalizar bem em casos de atributos externos escassos (RASHED; GRABOCKA; SCHMIDT-THIEME, 2019). O GraphRec para resolver esse problema constrói atributos latentes de item e de usuário como a fatoração de matriz, mas utiliza uma função de embedding não linear, como as redes neurais.

O funcionamento do GraphRec pode ser visto na Figura 2, onde o modelo é dividido em dois componentes principais. O primeiro é o modelo de Fatorização não linear que recebe as avaliações de item usuário e extrai vetores latentes de itens e usuários. O segundo componente é um conjunto de atributos baseados em grafos que permitem ao modelo aprender atributos latentes que são mais expressivos, especialmente quando os atributos externos de item e usuário não estão disponíveis. Os parágrafos a seguir descrevem os componentes.

Figura 2 – Diagrama da arquitetura do GraphRec



A definição do problema para o componente de fatorização não linear pode ser dada por: tendo um conjunto de usuários $U = \{u_1, u_2, ..., u_N\}$ e itens $I = \{i_1, i_2, ..., i_M\}$, o vetor latente dos usuários pode ser definido como $P = \phi_U : \mathbb{R}^{NxK}$, onde a função ϕ_U recebe os atributos específicos de usuários concatenada com os atributos externos, do perfil do usuário $C_u \in \mathbb{R}^{NxL}$. O vetor latente dos itens, similarmente pode ser definido por $Q = \phi_I : \mathbb{R}^{MxK}$, onde ϕ_I que recebe a variável indicadora do item e os atributos adicionais

do conteúdo $C_i \in \mathbb{R}^{MxJ}$. Onde a predição é data por $\hat{r}_{ui} = P_u^T Q_i$. Os parâmetros do modelo são os pesos θ_U e θ_I das redes neurais ϕ_U e ϕ_I.

2.2.3 Métrica de avaliação

NDCG A tarefa de recomendação espera construir uma lista em que os itens mais relevantes estão no topo desta lista. Dessa forma, utilizamos a métrica NDCG (Normalized Discounted Comulative Gain) para avaliar os modelos em nosso trabalho. O NDCG funciona por calcular um ganho acumulativo dos itens mais relevantes, acumulando a posição em que os itens relevantes ocorrem. Inicialmente se calcula o DCG, que é a soma do ganho acumulativo levando em consideração a posição dos itens. Depois esse valor é normalizado levando em conta o cenário ideal, ou seja, em que todos os documentos relevantes estejam no topo da lista.

Considerando f como uma função de ranqueamento e S_k uma consulta, o ganho descontado cumulativo é descrito na 2.4, onde o ganho cumulativo calculado pelos resultados de f em S_k é descontado com o log(r + 1), onde r seria a posição daquele item na lista.

$$DCG(f, S_k) = \sum_{r=1}^{k} \frac{2^{f(y_{(r)})}}{log(r+1)} \tag{2.4}$$

Considerando que o DCG ideal (IDCG) da melhor função de ranqueamento em S_k seja calculado com o DCG utilizando os valores reais, o NDCG é calculado como descrito em 2.5

$$NDCG(f, S_k) = \frac{DCG(f, S_k)}{IDCG(S_k)} \tag{2.5}$$

MRR Também levando em conta o ranqueamento de uma lista, utilizamos outra métrica de avaliação chamada Mean Reciprocal Rank (MRR). O ranque recíproco de uma consulta é o multiplicativo inverso do primeiro item relevante. O ranque recíproco médio é a média do resultado de todas as consultas. Considerando um conjunto de consultas Q, e rank a posição do primeiro item relevante na consulta i, a equação para o cálculo do MRR é descrito em 2.6.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{2.6}$$

RMSE Além das métricas utilizadas para avaliação de métodos de ranqueamento, foi aplicado o uso também do Root Mean Square Error (RMSE). A métrica calcula o erro médio entre dois conjuntos, onde um são os valores reais e outro a predição. O cálculo é dado pela equação 2.7

$$RMSE(Y, \hat{Y}) = \sqrt{\sum_{i=1}^{|Y|} \frac{(\hat{y}_i - y_i)^2}{|Y|}} \tag{2.7}$$

3 Categorização da Base de Dados IFG-Produz

Uma importante contribuição deste trabalho é disponibilizar uma base de dados anotada de sistemas de recomendação no contexto de artigos científicos com pesquisadores de diversas áreas. Desta forma, apresentamos neste capítulo uma caracterização da base de dados IFG-Produz, considerando suas características sobre os usuários, itens de interesse e avaliações dos usuários sobre os itens.

Mesmo mantendo o foco sobre a base de dados IFG-Produz, incluímos também nesta categorização outras duas bases de dados bastante conhecidas na comunidade científica de Sistemas de Recomendação. São elas a Movielens100k[1](HARPER; KONSTAN, 2015) e BookCrossing[2](ZIEGLER et al., 2005). O objetivo ao incluir essas bases é permitir uma comparação do nosso trabalho com outros da literatura, considerando tanto as características da categorização das bases, como as análise de efetividades para os diversos algoritmos, este último descrito no próximo capítulo.

A escolha da base de dados MovieLens100k se justifica por apresentar uma grande quantidade de avaliações entre usuários e os itens, situação diferente da base IFG-Produz. Ainda, a base de dados MoviLens100k apresenta pouca descrição dos itens, e como veremos, também é uma característica bem distinta do IFG-Produz. No caso da base BookCrossing, nosso objetivo é tentar avaliar uma base de dados com características mais próximas à base IFG-Produz. Desta forma, para aumentar as semelhanças, fizemos algumas alterações nas características dessa base para que a relação de quantidades de usuários, itens e avaliações ficassem mais próximas à base IFG-Produz.

Em suma, nosso trabalho fará comparações com outras duas bases disponíveis na literatura, a MoviLens100k que não foi alterada e nos permitirá comparar nossos resultados com os da literatura, e uma adaptação da base BookCrossing (daqui em diante chamaremos de ABC ou Adaptação BookCrossing) que se aproxime da quantidade de avaliações da base IFG-Produz. Como veremos no próximo capítulo, a base de dados MovieLens100k, por ter maior avaliação dos itens pelos usuários, possui melhor efetividade que as bases IFG-Produz e ABC em algoritmos que usam filtragem colaborativa. Por outro lado, ABC e IFG-Produz, por mostrarem maior descrição dos itens, tiveram melhores resultados em métodos de filtragem baseada em conteúdo. Percebemos que a avaliação com as três bases selecionadas oferecem maior sustentação às hipóteses levantadas com os experimentos.

[1] <https://grouplens.org/datasets/movielens/100k/>
[2] <http://www2.informatik.uni-freiburg.de/~cziegler/BX/>

3.1 Descrição dos itens e usuários

A base de dados IFG-Produz é formada por usuários que são servidores/pesquisadores do Instituto Federal de Goiás, a partir de processamento automatizado de currículos da Plataforma Lattes de todos eles. Na base os itens recomendados são de dois tipos: artigos científicos e projetos de pesquisa. Ambos são representados por um campo textual que descreve o título de cada um.

Os usuários ou pesquisadores são descritos pela sua área de atuação (Ciências Exatas, Ciências Humanas, etc.), artigos publicados e projetos cadastrados (executados ou em execução). As informações de área de atuação são descritas pelo próprio pesquisador em seu currículo Lattes, no campo "Grande Área de Concentração"na seção "Área de Atuação". Os artigos e projetos considerados são todos os registrados também no currículo Lattes. A tabela 1 resume os atributos dos itens recomendados na base de dados IFG-Produz e nas outras bases de comparação.

No caso de MovieLens100k, ocupação e gênero descrevem a profissão do usuário e o gênero masculino ou feminino, respectivamente. No caso de itens, existe a descrição do título do filme e a categoria do filme, podendo ser romance, drama, ação ou outros. No caso da base ABC existe somente a informação de Idade do usuário e título do livro.

Dataset	Atributos dos usuários	Atributos dos itens
IFG-Produz	Grande Área do Conhecimento do Docente, artigos publicados e projetos cadastrados	Título dos artigos e Títulos dos projetos
MovieLens100k	Idade, Ocupação, Gênero	Título do filme, Categoria
ABC	Idade	Título do livro

Tabela 1 – Atributos das bases

Detalhando um pouco mais, a tabela 2 descreve a composição dos itens nas bases de dados. No caso do IFG-Produz o dicionário de termos para publicações e projetos são de respectivamente 11.130 e 6.661 termos. Tratamos esses itens de forma independente, pois as recomendações são apresentados aos usuários igualmente de forma independente. Para a descrição dos itens, percebemos que a base IFG-Produz apresenta uma descrição mais rica que as demais bases, com média de 13 palavras por título, enquanto MoviLens100k e ABC possuem respectivamente 3,92 e 5,82 palavras por título. No caso do campo categoria para a base de MovieLens100k um único termo é adicionado.

A tabela 3 descreve os quantitativos de termos para os campos que detalham os usuários. Conforme mostrado na tabela 2, IFG-Produz possui a descrição das áreas de atuação, sendo a grande área de conhecimento, logo o número 9 na tabela indica a quantidade total de áreas na base, sendo que alguns usuários têm mais formações em

diferentes áreas, logo tendo mais de uma, como mostra a média e o máximo de termos. Na MovieLens100k, temos a profissão, o gênero da pessoa e a idade, logo cada usuário possui 3 termos, onde com todos os termos únicos temos 84 no total, e a base ABC possui apenas o campo de idade, contando no total com 60 valores únicos.

Considerando a quantidade de termos que descrevem os itens de uma base de dados, percebemos que o IFG-Produz possui uma quantidade de ao menos cinco vezes maior que as demais bases. Enquanto o IFG-Produz possui cerca de 11.130 termos que descrevem os títulos de artigos, a base ABC possui 2.525 termos e MoviLens100k 2.304 termos.

Dataset	Média de palavras por Título	Mínimo de palavras	Máximo de palavras	Tamanho do dicionário
IFG-Produz (título artigo)	13.2	1	121	11.130
IFG-Produz (título projeto)	13.9	1	41	6.661
MovieLens100k	3.92	1	16	2.304
ABC	5.82	1	51	2.525

Tabela 2 – Atributos dos itens

Uma observação relevante sobre a base IFG-Produz, é que a descrição dos títulos possui grande preocupação em descrever os itens, no caso os artigos científicos ou projetos. Desta forma, é mais fácil encontrar termos chaves na composição do título, como por exemplo "Machine Learning"ou "Recommendation". Algo diferente nas outras bases de dados, que estão mais preocupados em títulos sucintos e chamativos, logo toleram maior redução na tentativa de explicar o próprio item (filme ou livro).

Dataset	Média de termos por usuário	Mínimo de termos	Máximo de termos	Tamanho do dicionário
IFG-Produz	1.41	1	5	9
MovieLens100k	3	3	3	84
ABC	1	1	1	60

Tabela 3 – Descrição dos usuários

Além dos usuários e itens, faz-se necessário o conhecimento das avaliações dos itens. Na tabela 4 resumimos então as quantidades de usuários, itens e avaliações realizadas até o momento. No caso específico do IFG-Produz, inicialmente foi feita uma recomendação de artigos e projetos considerando uma relação de similaridade textual entre artigos e projetos de autoria do próprio pesquisador com artigos/projetos de outros autores. Para essa similaridade usamos o mesmo algoritmo descrito na Seção 2.1, considerando uma representação TF-IDF no título dos trabalhos. Cada recomendação é então avaliada pelo pesquisador, dando uma nota que varia entre 0 e 5 de grau de relevância, sendo 5 o maior grau.

A coluna esparsidade na tabela mostra o quão preenchida está a matriz de avaliações, considerando o número de itens e usuários e os campos em que o usuário avalia um item. Desta forma, maior esparsidade significa que temos menos informação sobre os interesses dos usuários. Segundo a tabela, a base MovieLens100k apresenta informação mais rica entre usuários e itens, com 100.000 avaliações, mesmo considerando que tem quase quatro vezes mais usuários que a base IFG-Produz, e um pouco menos conjunto de itens. Em função da adaptação feita, temos maior semelhança entre as bases IFG-Produz e ABC, tanto do ponto de vista de quantidade de usuários, itens e esparsidades.

Dataset	# Usuários	# Itens	# Avaliações	Esparsidade
IFG-Produz	194	2220	2948	99.31%
MovieLens100k	943	1682	100000	93.69%
ABC	120	2527	3000	99.01%

Tabela 4 – Características das bases

A tabela 5 visa descrever as bases em relação ao número de avaliações realizadas por usuário. A base IFG-Produz, por exemplo, tem 25,7 avaliações em média por usuário, máximo de 77 e mínimo de 5. Os primeiros 25% dos usuários possuem 10 avaliações cada, chegando a 38 avaliações cada usuário nos últimos 75% dos usuários. A tabela mostra que comparada às bases MovieLens100k e ABC, IFG-Produz ainda possui poucas avaliações. Considerando o primeiro quartil (25%) o IFG-Produz tem 10 avaliações, enquanto a MovieLens100k tem 98. Ou seja, já no primeiro quartil a base MovieLens100k ultrapassa o máximo de avaliações por usuário em relação ao IFG-Produz. Novamente, devida a adaptação que fizemos, a base ABC apresentou a relação de avaliações entre usuário e avaliações um pouco mais próxima da base IFG-Produz, mas mesmo assim com maior média e valor máximo.

Dataset	Média	Máximo	Mínimo	25%	50%	75%
IFG-Produz	25.7	77	5	10	19	38
MovieLens100k	202	737	20	98	181	278
ABC	217	890	5	5	27	95

Tabela 5 – Quartis das avaliações por usuários das bases

Uma característica importante das bases de dados de artigos científicos é que podemos partir de itens que foram criados pelos próprios usuários. Assim, evitamos os problemas de *cold-start*, pois podemos assumir que os usuários tem interesse pelos assuntos que publicam. Logo, a tabela 6 mostra as informações sobre a quantidade de trabalhos que cada pesquisador publicou, tendo uma média de 51 publicações ou projetos por usuário. Importante destacar que fizemos uma filtro, por somente pesquisadores que realizaram mais de uma publicação cadastrada no currículo Lattes.

Dataset	Média	Máximo	Mínimo	25%	50%	75%
IFG-Produz	51.6	148	1	19	39	69

Tabela 6 – Quartis das publicações por usuários

3.2 Caracterização dos usuários do IFG-Produz

Como já descrito, os usuários da base IFG-Produz são servidores pesquisadores do Instituto Federal de Goiás. Desta forma, cabe descrever em nosso trabalho quais câmpus (unidades) são esses pesquisadores e quais suas áreas de atuação. Desta forma, fizemos uso da ferramenta de visualização de dados conhecida por PowerBI (FERRARI; RUSSO, 2016) para agilizar a construção dos gráficos. Desenvolvida pela Microsoft, a ferramenta é destinada à visualização de dados e inteligência de negócios, fornecendo funcionalidades para facilitar a análise, agregação e compartilhamento dos dados. Uma vantagem com o uso desta ferramenta, é que será mais fácil manter as os gráficos atualizados.

Iniciamos com a figura 3 que mostra uma segmentação das áreas de atuação dos usuários e o quantitativo de avaliações na ferramenta IFG-Produz. A figura do lado direito mostra as porcentagens reais de servidores em cada área de atuação. Para essa análise, a atribuição de área de atuação de um pesquisador é definida pela área do conhecimento informado no currículo Lattes, no campo "Área de Atuação". Como mostra a figura, os pesquisadores da área da Educação são os que possuem maior participação nas avaliações das recomendações feitas no IFG-Produz. Essa participação é justificada por ser essa área a mais frequente dentro do IFG. Um padrão parecido para os pesquisadores da área de Ciência da Computação, em segundo lugar. Contudo, algumas áreas como Letras, possui uma porcentagem de servidores menor que Química, Engenharia Elétrica e Artes, e mesmo assim participam mais das avaliações das recomendações feitas.

Importante destacar que a Ferramenta IFG-Produz tem enviado as recomendações mensalmente, e espera-se que com o passar do tempo o número de avaliações cresça, à medida que os pesquisadores avaliam as recomendações. Até o momento, a ferramenta já fez um total de 42 mil recomendações, entre projetos e artigos, e 3057 foram avaliadas, cerca de 7,2%.

Avaliamos também a participação das avaliações considerando os servidores e seus câmpus de lotação. Como apresentado na figura 4, o campus Goiânia possui maior participação nas recomendações, devido justamente possuir maior número de servidores. Na sequência o campus Anápolis é quem mais participa das avaliações. Neste caso, o padrão de quantidade de pesquisadores por campus está bastante alinhado com as porcentagem de avaliações das recomendações. Diferenças pequenas ocorrem em campus que têm porcentagens parecidas, como o caso dos câmpus Formosa e Inhumas, que se invertem nos rankings.

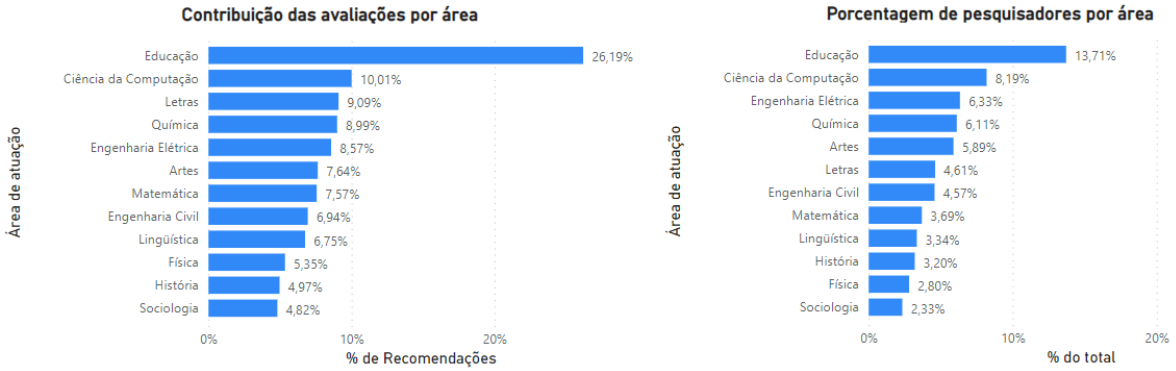


Figura 3 – Gráfico de barras mostrando participação nas recomendações de acordo com área

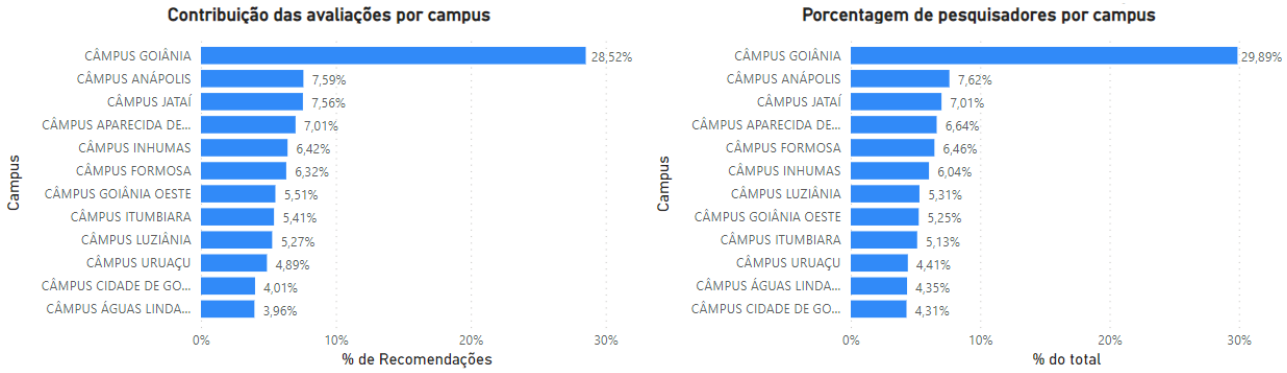


Figura 4 – Gráfico mostrando a participação de cada campus nas avaliações.

Visando analisar a percepção da qualidade das recomendações por parte dos pesquisadores, mostramos na Figura 5 uma distribuição das avaliações considerando as notas de 0 à 5, como intervalos de 0.5. A figura mostra a distribuição das avaliações para a recomendações dos artigos (figura à esquerda) e dos projetos (figura à direita). Como apresentado pela figura, é possível afirmar que os usuários estão tendo boa aceitação da ferramenta, já que a avaliação com nota 5 aparece mais frequentemente tanto para a recomendação de projetos como a recomendação de artigos.

No próximo capítulo daremos mais ênfase a esta análise, considerando diversos algoritmos e métricas tradicionais em Sistemas de Recomendação.

Em suma, sobre a caracterização da base de dados IFG-Produz podemos concluir:

- Em relação às bases ABC e MoviLens100k, a base IFG-Produz é quem possui descrição mais rica dos itens, com até quatro vezes mais termos por item em média, e tamanho de dicionário até cinco vezes maior;
- Em comparação à MoviLens100k, a base IFG-Produz possui bem menos informação

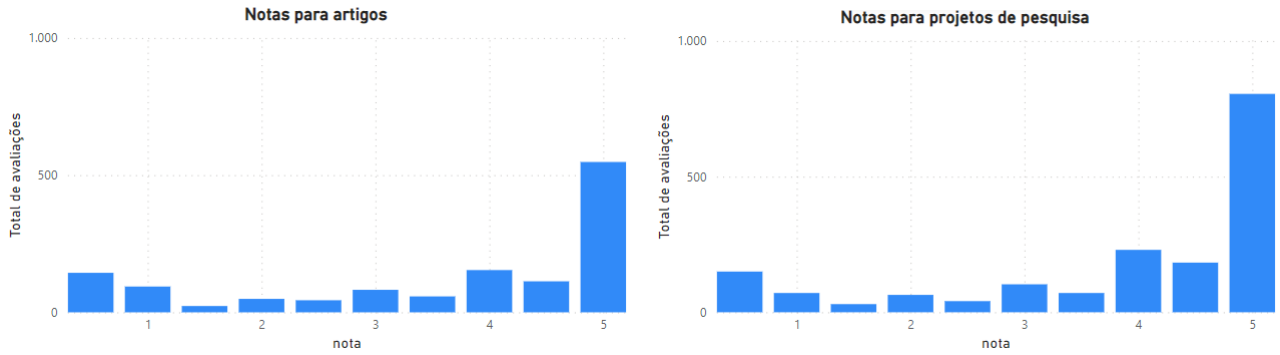


Figura 5 – Gráfico mostrando distribuição de notas

dos gostos do usuário, com esparsidade 6% maior, e uma média de avaliação por usuário 8 vezes menor.

- Sobre a utilização atual da ferramenta, os dados mostram que todos os câmpus do IFG têm feito as avaliações das recomendações seguindo a mesma porcentagem de pessoas por câmpus, e que as área da Educação e Ciência da Computação são as áreas que mais avaliam, seguindo a predominância das áreas no IFG;
- Ainda, que a julgar pelas notas de relevância, o sistema tem sido assertivo com o uso de recomendação baseada em conteúdo.

4 Resultados

Após categorizarmos a base de dados IFG-Produz, no capítulo anterior, nos dedicamos neste capítulo a comparar a performance das bases de dados com algoritmos tradicionais e estado da arte em Sistemas de Recomendação. Nós consideramos duas principais categorias de algoritmos, filtragem baseada em conteúdo e filtragem colaborativa. No que tange a filtragem baseada em conteúdo, como a representação dos itens tem forte impacto na análise de similaridade, consideramos duas estratégias: i) TF-IDF e ii) BERT. O uso do TF-IDF é devido a ser uma técnica simples e assumindo independência entre os termos. Com o BERT, avaliamos o estado-da-arte em Processamento de Linguagem Natural e a verificação se a representação semântica é capaz de enriquecer a acurácia dos modelos. No caso da aplicação com BERT consideramos o aproveitamento de dados pré-treinados, com a técnica Zero-Shot, usando o pré-processamento das bases BERT multilingual (DEVLIN et al., 2018). Por fim, abordamos também a técnica de filtragem colaborativa, considerando o trabalho GraphRec[1] (RASHED; GRABOCKA; SCHMIDT-THIEME, 2019) e NMF (FéVOTTE; IDIER, 2011)[2]. Todos esses algoritmos já foram apresentados em nosso Capítulo 2.

4.1 Configurações de Experimentação

Nesta seção detalhamos como preparamos as bases de dados, métricas de avaliação e como configuramos os hiperparâmetros.

Configuração das bases de dados Em nossos experimentos avaliamos as bases de dados considerando duas configurações. A primeira considera em uma recomendação somente os itens avaliados pelos usuários, entre relevantes e não relevantes. Contudo, quando uma base de dados possui poucas avaliações dos itens por usuário, e sendo boa parte destas avaliações relevantes, qualquer que seja a permutação de itens irá alterar pouco o resultado de desempenho do algoritmo. Desta forma, uma segunda estratégia que adiciona itens não vistos pelo usuário como itens não relevantes. Esta técnica é bastante utiliza, como mostra (HE et al., 2017) e (KOREN, 2008), e se baseia na estratégia que se forem pegos diversos documento não vistos pelo usuário a chance é alta de que muitos destes sejam não relevantes, e a tarefa passa a ser colocar no topo das recomendações os itens positivo avaliados pelo usuário. Assim, em nossos experimentos, foi utilizado a avaliação com 100 itens irrelevantes para cada relevante. No caso do IFG-Produz, para os

[1] No link <https://github.com/znehAC/GraphRec-example> temos uma implementação e exemplo de uso do GraphRec.

[2] <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html>

itens irrelevantes, além de selecionar aqueles que o usuário não interagiu, foram escolhidos os que também não tivessem a mesma área do usuário, reduzindo assim o risco de pegar um documento que seja relevante.

Métricas de performance Para as avaliações de performance foram utilizadas as métricas NDCG@1, NDCG@5, MRR e MSE. Em todos os casos consideramos a média entre os 5-folds e utilizamos uma análise de significância estatística das nossas medidas pela confiança de 95% do teste pareado T-test.

Hiperparâmetros Referente ao uso dos hiperparâmetros dos nossos modelos, fizemos avaliações *grid-search* e seleção de parâmetros dos artigos originais. Para o modelo Graph-Rec os hiperparâmetros para a base MovieLens100k foram obtidos a partir de (RASHED; GRABOCKA; SCHMIDT-THIEME, 2019), utilizando o modelo com atributos externos dos itens e dos usuários. Ainda, aplicamos a técnica de atributos baseados em grafos, onde eles são: $\lambda_u = 0.05$, $\lambda_i = 0.02$, tamanho de fatores latentes = 50, taxa de aprendizagem = 0.00003 e número de iterações = 195 para o Movielens100k. Para as demais bases os hiperparâmetros foram encontrados a partir de *grid-search*, em que λ_u e λ_i foram analisados valores entre 0.02 e 0.08, o tamanho dos fatores latentes entre 10 a 100, para a taxa de aprendizagem foram testados os valores 0.00003, 0.0002 e 0.0005, o número de iterações foi encontrado com o melhor ajuste. No IFG-Produz foram encontrados $\lambda_u = 0.08$, $\lambda_i = 0.02$, fatores latentes = 50, taxa de aprendizagem = 0.0002, número de iterações = 550, utilizando dos atributos dos itens e usuários, mas sem atributos baseados em grafo. Na ABC foi $\lambda_u = 0.06$, $\lambda_i = 0.02$, fatores latentes = 30, taxa de aprendizagem = 0.0002, número de iterações = 200, sem atributos baseados em grafos e sem atributos externos de itens e usuários. No caso da aplicação do algoritmo NMF, os hiperparâmetros também foram obtidos a partir de grid search. O intervalo procurado foi número de componentes entre 10 e 100, alpha_w e alpha_h com valores 0.1, 0.01, 0.001 e 0.0001, e a taxa L1 como 0, 0.5 e 1. No IFG-Produz o número de componentes foi 10, alpha_w e alpha_h com 0.001 e taxa L1 em 0. No MovieLens100k o número de componentes é 30, alpha_w e alpha_h 0.001 e taxa L1 em 0. Na ABC os valores seguem o mesmo do MoviLens100k, com exceção no número de componentes em 20.

4.2 Resultados Experimentais

A apresentação dos resultados faremos em duas seções. Na primeira considerando a base de dados que não considera itens relevantes e usando exclusivamente as avaliações dos itens dos usuários. Na segunda seção mostramos os resultados com o conjunto de dados que além dos itens avaliados tem também itens adicionados como não relevantes.

4.2.1 Base contendo exclusivamente itens avaliados

A tabela 7 apresenta a análise dos algoritmos com filtragem baseado em conteúdo, usando representação TF-IDF e BERT-MultiLinguagem, e filtragem colaborativa, com NMF e GraphRec. O valor em negrito apresenta o maior valor absoluto em cada base de dados, e o "*"mostra quando existe uma significância estatística na diferença entre o valor apontado e o resultado com maior valor absoluto.

Olhando inicialmente na Tabela, em especial para o conjunto de dados no IFG-Produz, percebemos que a estratégia filtragem baseada em conteúdo, usando a representação com BERT apresenta melhor resultado para as métricas NDCG1, NDCG5 e MRR. Ainda existe um empate estatístico ao usar a mesma estratégia mas com representação TF-IDF. Para o IFG-Produz, os modelos com filtragem colaborativa apresentaram resultados semelhantes, na grande maioria inferiores em relação à filtragem baseada em conteúdo. Para a métrica RMSE, os resultados apontam que o GraphRec apresenta performance melhor, isso se da ao fato de não ser uma técnica baseada em conteúdo. A justificativa disso ainda merece maior estudo.

No caso da análise usando MRR, verificamos que não há diferença estatística entre os modelos para a IFG-Produz. De fato, como a métrica MRR avalia a posição do primeiro item mais relevante da lista de recomendação, e existem muitos itens relevantes nas recomendações do IFG-Produz, variações na permutação não geram forte impacto na performance. Ou seja, mesmo por sorte, é fácil acertar um item relevante nas primeiras posições da lista. Diferente da MoviLens100k, que por possuir muitos itens, gera uma análise mais confiável.

Uma justificativa para que a filtragem baseada em conteúdo tenha melhor resultado na IFG-Produz, se deve principalmente pela rica descrição textual dos itens, artigos e projetos contidas na base. Por outro lado, por ter poucas avaliações dos itens, fica mais fácil explicar os resultados ruins com os modelos NMF e GraphRec. Inclusive, o número de avaliações dos itens recomendados na base de dados MovieLens100k justifica os bons resultados com os algoritmos de filtragem colaborativa, que inclusive segue o trabalho publicado em (RASHED; GRABOCKA; SCHMIDT-THIEME, 2019), que aponta os melhores resultados para o algoritmo GraphRec[3].

Por fim, observamos que a base de dados ABC possui resultados próximos com a IFG-Produz. Ou seja, como retiramos algumas informações sobre a avaliação dos itens, e mantivemos as informações que descrevem os itens, os algoritmos com filtragem baseado em conteúdo tiverem resultados absolutos melhores, de forma mais consistente, que os algoritmos com filtragem colaborativa, apesar dos empates estatísticos.

[3] Destacamos que possíveis variações dos nossos resultados em comparação aos descritos em (RASHED; GRABOCKA; SCHMIDT-THIEME, 2019) se justifica, pois fizemos a aplicação da estratégia cross-validations 5-fold, mas ao executar a base da mesma forma, conseguirmos resultados semelhantes.

Métrica	TFIDF	BERT	NMF	GraphRec
IFG-Produz				
NDCG@1	0.82	**0.86**	0.66	0.70*
NDCG@5	0.86	**0.87**	0.71*	0.73*
MRR	0.77	**0.81**	0.71	0.76
RMSE	3.92	3.78	3.57	**1.45**
MovieLens100k				
NDCG@1	0.46*	.49*	0.51*	**0.76**
NDCG@5	0.55*	.59*	0.63*	**0.79**
MRR	0.43*	.47*	0.5*	**0.76**
RMSE	3.68	3.24	2.74	**0.88**
ABC				
NDCG@1	0.33	**0.37**	0.19*	0.31
NDCG@5	0.49	**0.53**	0.38	0.45
MRR	0.41	**0.58**	0.37	0.44
RMSE	4.45	4.36	4.45	**4.05**

Tabela 7 – Resultados usando base de dados contendo exclusivamente itens avaliados

4.2.2 Base contendo itens avaliados e itens não relevantes

Nesta seção avaliamos as bases de dados considerando em cada recomendação os itens avaliados pelos usuários e também itens aleatórios, impondo uma dificuldade maior para ordenar os documentos relevantes no topo da lista.

A Tabela 8 está organizada da mesma forma que a Tabela 7. Neste caso, a estratégia com TF-IDF tem um resultado melhor que usando a representação do BERT, para filtragem baseada em conteúdo. No caso da métrica NDCG@1 a diferença foi ainda com significância estatística. Uma possível justificativa para esse resultado, posta uma configuração da base mais difícil, é que a estrategia de informação de contexto imposta pelo BERT possa ter reduzido o poder de discriminação. Já com o uso do TF-IDF, como os termos tendem a explicar bem os itens, apresentou um resultado mais relevante para ranquear os itens recomendados. Novamente seguindo o padrão da base IFG-Produz, a ABC apresenta melhores resultados com a técnica de similaridade, usando a representação TF-IDF.

Mesmo com resultados menores, em relação a subseção anterior, a estratégia MovieLens100k continua tendo melhores resultados com o algoritmo GraphRec, de filtragem colaborativa. Algo que segue o padrão descrito em (RASHED; GRABOCKA; SCHMIDT-THIEME, 2019), em que bases com uma matriz de avaliação menos esparsa, tendem a tirar proveito dos algoritmos com filtragem colaborativa.

Sumarizando os resultados, na base IFG-Produz, a performance das técnicas de filtragem colaborativa são significativamente piores do que as baseadas em conteúdo. Resultado que pode ser devido ao baixo número de avaliações de itens por usuário,

Métrica	TFIDF	BERT	NMF	GraphRec
IFG-Produz				
NDCG@1	**0.61**	0.53*	0.11*	0.20*
NDCG@5	**0.74**	0.68*	0.13*	0.26*
MRR	**0.72**	0.47*	0.09*	0.14*
RMSE	**0.39**	0.40	2.61	3.1
MovieLens100k				
NDCG@1	0.12*	0.10*	0.10*	**0.30**
NDCG@5	0.23*	.22*	0.24*	**0.35**
MRR	0.06*	.07*	0.10*	**0.12**
RMSE	**0.36**	0.56	2.04	2.9
ABC				
NDCG@1	**0.30**	0.21*	0.11*	0.20*
NDCG@5	**0.36**	0.24*	0.17*	0.25*
MRR	**0.73**	0.71*	0.69*	0.66*
RMSE	**0.47**	0.49	0.63	0.91

Tabela 8 – Resultados das métricas para as bases na configuração com 100 itens irrelevantes

dificultando algoritmos de filtragem colaborativa. Por outro lado, a riqueza dos dados presentes nos títulos dos itens, onde temos textos capazes de gerar boas representações para comparação, favorece as técnicas baseadas em conteúdos avaliadas.

O BERT que se mostra nos resultados é o modelo zero-shot do pré-treinamento BERT multilingual. Foi feita uma comparação entre esse pré-treinamento de múltiplas línguas contra um pré treinamento focado apenas na língua portuguesa, o BERTimbau Base (SOUZA; NOGUEIRA; LOTUFO, 2020). Contudo o resultado foi pior, principalmente pois a base possui tanto exemplos em português como em outras línguas, como o inglês.

Nas bases IFG-Produz e ABC se observa melhores resultados para as técnicas baseadas em conteúdo, que pode ser devido aos textos presentes nas bases, onde na IFG-Produz o resultado é ainda mais superior aos demais, sendo também a com o maior número de palavras nos títulos. Em contrapartida na MovieLens100k os resultados favorecem o GraphRec, técnica de filtragem colaborativa, onde apesar de não ter muito texto para os métodos de similaridade, ela possui bastante informações de relevância, que permitem a filtragem colaborativa se sair melhor.

5 Conclusão

A recomendação de artigos científicos em comunidades acadêmicas é um domínio de aplicação de Sistemas de Recomendação. Considerando que o número de novas publicações acadêmicas têm crescido consideravelmente nos últimos anos, nosso trabalho visa mitigar a dificuldade dos pesquisadores em acessar artigos relevantes. De forma geral, fazemos isso disponibilizando uma nova base de dados de recomendação em artigos científicos, junto com uma categorização e avaliações de algoritmos de sistema de recomendação.

Mais detalhadamente, neste trabalho entregamos um estudo na base de dados de recomendações do IFG-Produz, fazendo assim uma caracterização da mesma. No intuito de ampliar nossas análises e comparações, incluímos outras duas bases, a MovieLens100k e uma adaptação da BookingCrossing (ABC). Também foi feita a utilização da ferramenta PowerBI para gerar visualizações dos dados, ampliando o entendimento sobre os mesmos. Outra contribuição foi uma análise de métodos de recomendação para a plataforma, onde foram testadas técnicas tanto de filtragem baseada em conteúdo como filtragem colaborativa.

Na caracterização dos dados observamos que diferente das outras, a base IFG-Produz apresenta uma rica e representativa descrição textual dos seus itens. Contudo, a base de dados ainda apresenta pouca avaliação dos itens por parte dos usuários, com maior esparsidade na relação tabela-item, avaliação e usuário. Cenário que pode ser alterado ao longo do tempo, mediante os usuários receberem novas recomendações e consequentemente possam avaliá-las. Seguindo os estudos feitos, percebemos que os câmpus do IFG têm participado das avaliações de forma igualitária, ou seja, de forma geral, seguindo a mesma porcentagem de servidores por campus. Por fim, nossos estudos mostram que a ferramenta tem tido bons resultados em suas recomendações, ou seja, boa efetividade. Isso fica claro quando avaliamos a quantidade de avaliações e percebemos que notas ponderando os itens como relevantes aparecem de forma majoritária.

Em relação às experimentações com os métodos de recomendações nas bases o objetivo era responder algumas perguntas, como: i) Na base IFG-Produz, como é a performance das técnicas de filtragem colaborativas comparada à técnicas baseadas em conteúdos?; ii) Como a técnica de representação utilizando o BERT se sai comparado ao TF-IDF nas bases estudadas?; iii) Considerando as técnicas de filtragem colaborativa e filtragem por conteúdo analisadas, quais apresentaram melhores resultados?

Para i) as técnicas de filtragem baseado em conteúdo se saíram consistentemente melhores que as baseadas em filtragem colaborativa, onde é explicado pelos dados dos itens do IFG-Produz serem ricos em conteúdo. Para ii), a técnica utilizando a representação

BERT, no geral, se saiu pior que a utilização do TF-IDF, possivelmente pois o BERT ao utilizar dos pré-processados de outros contexto, pode diminuir o poder de discriminar a relevância de forma mais específica por pesquisador. Para a pergunta iii), nos casos em que a base de dados apresentava menor esparsidade, ou seja, mais informação entre usuários e seus interesses pelos itens, os algoritmos com filtragem colaborativa apresentaram resultados mais efetivos.

Para trabalhos futuros pode ser considerado um estudo de outras técnicas de recomendações que utilizam abordagens diferentes das analisadas nesse trabalho. Inclusive, entendemos que após algum tempo e a base de dados IFG-Produz tiver diminuído sua esparsidade, sugerimos que os algoritmos com filtragem colaborativa sejam re-executados.

Referências

AGGARWAL, C. C. Content-based recommender systems. In: *Recommender systems*. [S.l.]: Springer, 2016. p. 139–166. Citado na página 10.

BAI, X. et al. Scientific paper recommendation: A survey. *Ieee Access*, IEEE, v. 7, p. 9324–9339, 2019. Citado na página 7.

BERRY, M. W. et al. Algorithms and applications for approximate nonnegative matrix factorization. *Computational statistics & data analysis*, Elsevier, v. 52, n. 1, p. 155–173, 2007. Citado na página 14.

CENTRO DE GESTÃO E ESTUDOS ESTRATÉGICOS- CGEE. Panorama da ciência brasileira: 2015-2020. Boletim Anual OCTI, Brasília, v. 1, p. 196, 2021. Citado na página 7.

CHO, Y. H.; KIM, J. K. Application of web usage mining and product taxonomy to collaborative recommendations in e-commerce. *Expert systems with Applications*, Elsevier, v. 26, n. 2, p. 233–246, 2004. Citado na página 9.

DEVLIN, J. et al. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. Disponível em: <http://arxiv.org/abs/1810.04805>. Citado 2 vezes nas páginas 13 e 26.

FERRARI, A.; RUSSO, M. *Introducing Microsoft Power BI*. [S.l.]: Microsoft Press, 2016. Citado 2 vezes nas páginas 8 e 23.

FLAUBERTH, M. S. et al. Plataforma inteligente para reduzir distâncias entre pesquisadores de instituições multicampus. 2020. Citado na página 7.

FéVOTTE, C.; IDIER, J. Algorithms for Nonnegative Matrix Factorization with the β-Divergence. *Neural Computation*, v. 23, n. 9, p. 2421–2456, 09 2011. ISSN 0899-7667. Disponível em: <https://doi.org/10.1162/NECO_a_00168>. Citado na página 26.

HARPER, F. M.; KONSTAN, J. A. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, ACM New York, NY, USA, v. 5, n. 4, p. 1–19, 2015. Citado na página 19.

HE, X. et al. Neural collaborative filtering. In: *Proceedings of the 26th international conference on world wide web*. [S.l.: s.n.], 2017. p. 173–182. Citado na página 26.

HOANG, M.; BIHORAC, O. A.; ROUCES, J. Aspect-based sentiment analysis using bert. In: *Proceedings of the 22nd nordic conference on computational linguistics*. [S.l.: s.n.], 2019. p. 187–196. Citado 2 vezes nas páginas 13 e 14.

ISINKAYE, F. O.; FOLAJIMI, Y. O.; OJOKOH, B. A. Recommendation systems: Principles, methods and evaluation. *Egyptian informatics journal*, Elsevier, v. 16, n. 3, p. 261–273, 2015. Citado 3 vezes nas páginas 9, 10 e 14.

JACSÓ, P. Google scholar: the pros and the cons. *Online information review*, Emerald Group Publishing Limited, 2005. Citado na página 7.

JOACHIMS, T. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization.* [S.l.], 1996. Citado na página 11.

KOREN, Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.* [S.l.: s.n.], 2008. p. 426–434. Citado na página 26.

KOREN, Y.; BELL, R.; VOLINSKY, C. Matrix factorization techniques for recommender systems. *Computer*, IEEE, v. 42, n. 8, p. 30–37, 2009. Citado na página 15.

LÜ, L. et al. Recommender systems. *Physics reports*, Elsevier, v. 519, n. 1, p. 1–49, 2012. Citado na página 10.

O'BRIEN, K. Researchgate. *Journal of the Medical Library Association: JMLA*, Medical Library Association, v. 107, n. 2, p. 284, 2019. Citado na página 7.

RAMOS, J. et al. Using tf-idf to determine word relevance in document queries. In: CITESEER. *Proceedings of the first instructional conference on machine learning.* [S.l.], 2003. v. 242, n. 1, p. 29–48. Citado na página 12.

RASHED, A.; GRABOCKA, J.; SCHMIDT-THIEME, L. Attribute-aware non-linear co-embeddings of graph features. In: *Proceedings of the 13th ACM conference on recommender systems.* [S.l.: s.n.], 2019. p. 314–321. Citado 6 vezes nas páginas 15, 16, 26, 27, 28 e 29.

RONG, X. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014. Citado na página 13.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear).* [S.l.: s.n.], 2020. Citado na página 30.

ZHU, Y. et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: *Proceedings of the IEEE international conference on computer vision.* [S.l.: s.n.], 2015. p. 19–27. Citado na página 13.

ZIEGLER, C.-N. et al. Improving recommendation lists through topic diversification. In: *Proceedings of the 14th international conference on World Wide Web.* [S.l.: s.n.], 2005. p. 22–32. Citado na página 19.