UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE BIBLIOTECONOMIA E COMUNICAÇÃO
DEPARTAMENTO DE CIÊNCIAS DA INFORMAÇÃO
CURSO DE BIBLIOTECONOMIA

JOÃO VITHOR DE SOUZA BAEZ

Proposta de modelo conceitual de recomendação de publicações científicas na Brapci

Porto Alegre
2024

JOÃO VITHOR DE SOUZA BAEZ

Proposta de modelo conceitual de recomendação de publicações científicas na Brapci

Trabalho de conclusão de curso de graduação apresentado como requisito para a obtenção do título de Bacharel em Biblioteconomia da Faculdade de Biblioteconomia e Comunicação da Universidade Federal do Rio Grande do Sul.

Orientador: Prof. Dr. Rene Faustino Gabriel Junior

Porto Alegre

2024

CIP - Catalogação na Publicação

Baez, João Vithor de Souza
Proposta de modelo conceitual de recomendação de publicações científicas na Brapci / João Vithor de Souza Baez. -- 2024.
96 f.
Orientador: Rene Faustino Gabriel Junior.

Trabalho de conclusão de curso (Graduação) -- Universidade Federal do Rio Grande do Sul, Faculdade de Biblioteconomia e Comunicação, Curso de Biblioteconomia, Porto Alegre, BR-RS, 2024.

1. sistemas de recomendação. 2. Brapci. 3. modelo conceitual. 4. publicações científicas. I. Gabriel Junior, Rene Faustino, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

RESUMO

Sistemas de recomendação estão presentes em diferentes espaços na Web, oferecendo sugestões personalizadas de itens ou serviços. Em um contexto de sobrecarga informacional resultante do crescimento exponencial da informação e do predomínio de dados não estruturados em relação à informação organizada, o uso de sistemas de filtragem da informação torna-se cada vez mais necessário, sendo uma tendência imprescindível para atender às necessidades informacionais de usuários na Internet, especialmente na busca por informação especializada. Este estudo propôs um modelo conceitual de recomendação de publicações científicas para a Brapci, precedido por uma pesquisa bibliográfica nas bases de dados nacionais Brapci e OasisBR. A pesquisa analisou dezessete estudos sobre sistemas de recomendação (SR) em áreas como *e-commerce*, *streaming*, catálogos *online* e repositórios digitais, com foco nas técnicas de Filtragem Colaborativa (FC), Filtragem Baseada em Conteúdo (FBC), Filtragem Híbrida e Filtragem Semântica. O modelo conceitual desenvolvido é dividido em três estratégias: itens relacionados, itens citados e itens personalizados. Itens relacionados usam FBC para recomendações baseadas nas consultas, no item visualizado e nos itens selecionados. Itens citados têm duas abordagens: uma não personalizada, com base nos itens mais citados pelo texto lido, e outra que considera a similaridade textual com os itens citados. Itens personalizados utilizam uma abordagem de Filtragem Híbrida que combina FC e FBC, aliadas à técnica de cascata, para alcançar maior precisão e relevância nas recomendações. O modelo também define aspectos relacionados à extração de informações do usuário, ao método de saída das recomendações, ao grau de personalização e à visualização das recomendações e dos itens avaliados no perfil de usuário da Brapci.

Palavras-chave: sistemas de recomendação. Brapci. modelo conceitual. publicações científicas.

ABSTRACT

Recommendation systems are present in various spaces on the Web, offering personalized suggestions for items or services. In a context of informational overload resulting from the exponential growth of information and the predominance of unstructured data over organized information, the use of information filtering systems becomes increasingly necessary, making it an essential trend to meet the informational needs of users on the Internet, especially in the search for specialized information. This study proposed a conceptual model for recommending scientific publications for Brapci, preceded by a literature review of the national databases Brapci and OasisBR. The research analyzed seventeen studies on recommendation systems (RS) in areas such as e-commerce, streaming, online catalogs, and digital repositories, focusing on collaborative filtering (CF), content-based filtering (CBF), hybrid filtering, and semantic filtering techniques. The developed conceptual model is divided into three strategies: related items, cited items, and personalized items. Related items use CBF for recommendations based on queries, the viewed item, and selected items. Cited items have two approaches: a non-personalized one based on the most cited items in the read text, and another considering textual similarity with the cited items. Personalized items use hybrid filtering combining CF and CBF, along with the cascade technique, aiming for precision and relevance in the recommendations. The model also defines aspects related to user information extraction, recommendation output methods, the degree of personalization, and the visualization of recommendations and assessed items in the Brapci user profile.

Keywords: recommender systems. Brapci. conceptual model. scientific publications.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

BOF	*Bag-of-Faces*
BOW	*Bag-of-Words*
Brapci	Base de dados em Ciência da Informação
BRES	Base Brasil/Espanha de Artigos de Periódicos da área em Ciência da Informação
CASER	*ConvolutionaI Sequence Embedding Recommendation*
CI	Ciência da Informação
CNN	Rede Neural Convolucional
CSS	*Cascading Style Sheets*
DC	*Dublin Core*
DM	Mineração de Dados
DOI	Identificador de Objeto Digital
E2PC	Grupo de pesquisa Educação, Pesquisa e Produção Científica
E3PI	Grupo de Pesquisa Educação, Pesquisa e Perfil Profissional em Informação
EM	Algoritmo de maximização de expectativa
FBC	Filtragem Baseada em Conteúdo
FC	Filtragem Colaborativa
FI	Filtragem da Informação
FOAF	*Friend of a Friend*
GDPR	Regulamento Geral sobre a Proteção de Dados
GRU	Unidade Recorrente Controlada
GUI	Interface Gráfica de Usuário
HTML	*Hypertext Markup Language*
IA	Inteligência Artificial
ID	Identificação
IHC	Interação Humano-Computador
IP	Protocolo de rede Internet
ISBN	Número Padrão Internacional de Livro
K-NN	K-Vizinhos mais Próximos
LGPD	Lei Geral de Proteção de Dados
MAE	Erro Médio Absoluto
ML	Aprendizado de Máquina

OAI-PMH	*Open Archives Iniciative – Protocol Metadata Harvesting*
OPAC	Catálogo de Acesso Público Online
P2P	Ponto a Ponto
PLN	Processamento de Linguagem Natural
PPGCIN	Programa de Pós-Graduação em Ciência da Informação
RDF	*Resource Description Framework*
RI	Recuperação da Informação
RMSE	Raiz do Erro Quadrático Médio
SBERT	Sentence-Bert
SR	Sistemas de Recomendação
SRI	Sistemas de Recuperação da Informação
SRS	Sistema de Recomendação Semântico
SRSSC	Sistema de Recomendação Semântico Sensível ao Contexto
SSLIM	*Sparse Linear Method with Side Information*
SVD	*Singular Value Decomposition*
TF-IDF	*Term Frequency - Inverse Document Frequency*
UC3M	Universidade Carlos III de Madrid
UFPR	Universidade Federal do Paraná
UFRGS	Universidade Federal do Rio Grande do Sul
W3C	World Wide Web Consortium
XML	*Extensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

Recuperar a informação necessária tem se mostrado uma tarefa de complexidade crescente em meio à expansão exponencial das informações disponíveis na Internet. Buscas feitas por usuários nem sempre se encontram alinhadas com as expressões de busca mais eficazes ou resultados mais relevantes, tornando este processo longo e trabalhoso, mesmo para profissionais especializados.

Consequentemente, há um interesse em nível internacional em pesquisa e desenvolvimento de tecnologias como *Machine Learning*[1] (aprendizado por máquina) e *Data Mining*[2] (mineração de dados) para lidar com a quantidade massiva de dados na *World Wide Web* a partir de sistemas de recomendação. Nesse sentido, empresas como Amazon, Microsoft e Google apresentam, respectivamente, modelos de recomendação em seus produtos como *Amazon Personalize*, *Intelligent Recommendations* e *Recommendations AI*.

De fato, o processo de busca das informações por meio dos sistemas de recomendação tem sido um avanço para as plataformas de *e-commerce* (comércio eletrônico), mas também se inserem em plataformas de *streaming* (transmissão de conteúdo *online*) ou mesmo em ambientes acadêmicos, como catálogos de bibliotecas e bases de dados. É em relação a estas últimas que o seguinte trabalho discorre, tendo como tema a proposta de um modelo de recomendação de publicações científicas na Base de Dados em Ciência da Informação (Brapci), vista a seguir.

A base de dados Brapci surgiu em 1996, em projeto produzido pelo Grupo de Pesquisa Educação, Pesquisa e Perfil Profissional em Informação (E3PI) da Universidade Federal do Paraná (UFPR). No entanto, sua disponibilização ao público ocorreu apenas em 2008 (Gabriel Junior, 2014a). Desde então, a Brapci estabeleceu-se como a principal base de dados temática da Ciência da Informação (CI) no Brasil, tendo sido constantemente aprimorada pelos membros do E3PI, professora Leilah Santiago Bufrem e professor Rene Faustino Gabriel Junior.

---

[1] *Machine Learning* (ML) é uma área da inteligência artificial (IA) que utiliza algoritmos e dados estatísticos para realizar previsões e auxiliar em decisões. Diferencia-se de outras IAs pela capacidade de aprimoramento não programado a partir do consumo de dados estruturados.

[2] *Data Mining*, ou mineração de dados (DM), refere-se a um amplo espectro de técnicas de modelagem matemática e ferramentas de *software* que são usadas para encontrar padrões em conjuntos de dados. Sistemas de recomendação que incorporam técnicas de mineração de dados fazem suas recomendações com base no conhecimento adquirido a partir das ações e atributos dos usuários (Schafer, 2001, p. 18, tradução nossa).

Assim, por meio desta pesquisa, busca-se avançar no desenvolvimento da Brapci a partir da proposta de um modelo conceitual de recomendação que possa servir de base para aprimorar a eficácia na recuperação da informação. Nesse sentido, as subseções a seguir expõem, na ordem citada, o problema de pesquisa, objetivos geral e específicos e justificativa.

1.1 PROBLEMA DE PESQUISA

A Organização e Recuperação da Informação estão entre as principais áreas de estudo na Ciência da Informação. Contudo, apesar das técnicas empregadas na representação da informação, a recuperação de informação especializada ainda depende do entendimento do usuário sobre as suas necessidades informacionais, cada vez mais exigentes, da terminologia a ser aplicada na construção das estratégias de busca e do entendimento dos sistemas de recuperação da informação utilizados. Tendo em vista as dificuldades apresentadas a respeito do processo de busca e recuperação da informação desejada, bem como de aspectos relacionados à precisão e à revocação dos mecanismos de busca em bases de dados, o uso dos sistemas de recomendação em conjunto dos tradicionais sistemas de recuperação da informação, permite aprimorar a qualidade dos resultados de busca ao incorporar tanto as estratégias de busca quanto às preferências individuais dos usuários. Entretanto, a literatura em CI ainda apresenta poucos estudos acerca dos SR, assim como das técnicas e estratégias de recomendação. Além disso, não foram encontrados estudos de natureza aplicada que abordam a implementação desses sistemas.

Essa lacuna na literatura é particularmente problemática quando se trata de bases de dados especializadas, como a Brapci, que não possui um sistema de recomendação integrado. A ausência desse recurso limita a eficácia da recuperação de informações e a personalização das buscas, tornando mais difícil para os usuários encontrar publicações relevantes de maneira eficiente. A falta de estudos aplicados na literatura destaca a necessidade de desenvolver e implementar tecnologias de recomendação para melhorar a precisão e a relevância dos resultados de busca na Brapci.

Portanto, este trabalho apresenta como problema de pesquisa o seguinte questionamento: quais os requisitos para implementação de um sistema de recomendação poderiam contribuir para aprimorar a busca de informações na Brapci?

1.2 OBJETIVOS

Os objetivos são vistos a seguir, subdivididos em objetivos geral e específicos.

1.2.1 Objetivo geral

Propor um modelo conceitual para recomendação de publicações científicas aplicável na Base de Dados em Ciência da Informação (Brapci).

1.2.2 Objetivos específicos

Para atingir ao objetivo geral desta pesquisa, por meio dos objetivos específicos a seguir, pretende-se:

a) identificar na literatura os modelos de recomendação, as técnicas e as estratégias aplicadas à filtragem de informação;
b) avaliar modelos de recomendação compatíveis com os critérios estabelecidos pela metodologia de pesquisa;
c) sistematizar as decisões sobre os aspectos e requisitos fundamentais de implementação para a proposta do modelo de recomendação com base no referencial teórico, metodologia e resultados da pesquisa;
d) criar estratégias de recomendação dos itens que atendam a diferentes contextos de pesquisa e de buscas por informações;
e) elaborar um modelo de visualização das recomendações de publicações científicas.

1.3 JUSTIFICATIVA

O presente estudo justifica-se, inicialmente, em virtude da contribuição à área da Ciência da Informação a partir do tema e do objeto de estudo abordados neste trabalho. Em relação ao tema, atualmente, os sistemas de recomendação são aplicados em diversos espaços da sociedade, tal como a pesquisa, tornando-se parte da realidade diária dessas diferentes esferas sociais. Nesse sentido, a CI contribui para o desenvolvimento dos sistemas por meio das teorias empregadas, principalmente, na Organização, Representação e Recuperação da

Informação. Assim, considera-se essencial o estudo deste tema na área, uma vez que esteja entre as principais formas de recuperação da informação na Web pelo usuário, tendo em vista o uso da personalização nas recomendações resultantes dos sistemas de recomendação.

No que concerne ao objeto de estudo, esta pesquisa busca contribuir para o desenvolvimento do principal repositório digital nacional na área de CI, a Brapci, sendo de suma importância seu aprimoramento para a comunidade científica, uma vez que esteja diretamente relacionada com os processos de busca e recuperação da informação, além da visibilidade de pesquisas da área, especialmente no Brasil. Dessa forma, a partir dos resultados desta pesquisa, visa-se aprimorar a experiência do usuário e a recuperação da informação na base de dados, com margem para pesquisas futuras fundamentadas na implementação e melhorias do sistema, assim como em seu impacto e avaliações por meio de métricas e estudos de usuários.

Por último, a pesquisa justifica-se também pelo interesse pessoal do autor pelo tema, tendo em vista a sua afinidade e experiências anteriores na área de Tecnologia da Informação, bem como da possibilidade de aprofundar os estudos sobre a Brapci sob orientação de seu coordenador Rene Faustino Gabriel Junior. Dessa forma, houve grande interesse em desenvolver ao longo da pesquisa o conhecimento acerca dos sistemas de recomendação, tendo em vista sua ampla implementação no mercado atual.

2 REFERENCIAL TEÓRICO

Nesta seção serão abordados os capítulos que compõem o referencial teórico desta pesquisa, sendo eles: a produção científica, etapa fundamental na formalização da comunicação científica, onde são abordados os processos da publicação científica; a Brapci, objeto de estudo em que se inserem os requisitos e necessidades a serem examinadas durante a pesquisa, neste tópico aborda-se seu histórico e as principais ferramentas de recuperação da informação na base; e os sistemas de recomendação, na qual são vistas as etapas, classificação e técnicas de filtragem da informação (FI).

2.1 PRODUÇÃO CIENTÍFICA

Entre os alicerces que sustentam a pesquisa no âmbito da ciência, a comunicação científica destaca-se como um processo contínuo de construção do conhecimento científico, constituindo um diálogo entre diferentes pensamentos e permitindo a realização de novas descobertas. Nesse contexto, os avanços obtidos por meio da pesquisa têm origem nas diversas contribuições formadas por cientistas ao longo do tempo, os quais se baseiam em estudos anteriores para fundamentar suas próprias investigações.

Segundo Weitzel (2006), a produção científica consolida os procedimentos formais da comunicação científica por meio dos processos de publicação, tornando o conhecimento público e amplamente disseminado. Dessa maneira, ela se configura como um componente essencial na especialização dos saberes e na prática científica, oferecendo diversos níveis de aprendizagem que remetem às dinâmicas colaborativas da comunidade científica (Weitzel, 2006; Caxias, 2009).

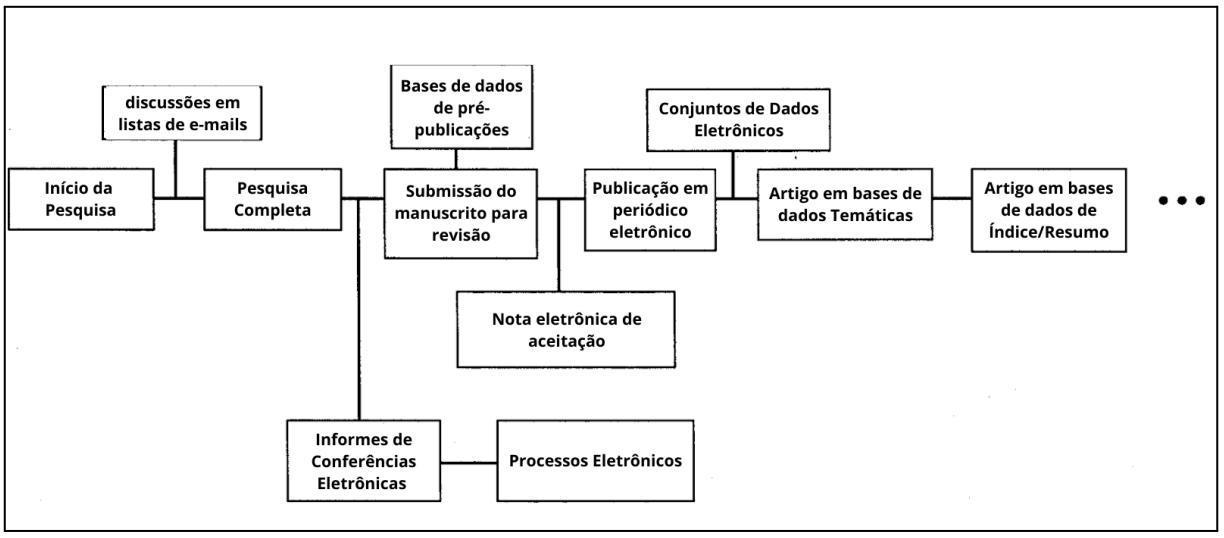
A respeito da formação da publicação científica, também denominada memória científica, Sayão (1996, p. 315) afirma que

> O caráter cumulativo da ciência, que se apropria de uma forma rigorosamente seletiva das contribuições de seus pesquisadores, resulta em um corpo de conhecimento baseado no consenso. Este corpo de conhecimento é representado pela literatura técnico-científica, fruto mais óbvio e mais facilmente sujeito à mensuração da atividade científica. São os livros, os artigos de revistas, os trabalhos de congresso, as patentes, portadoras das inovações tecnológicas, os mais autênticos registros da faina diária dos cientistas.

Nesse sentido, a rigidez metodológica, a revisão por pares e a clareza na apresentação dos dados são alguns dos elementos formais que permeiam a seleção de trabalhos realizados por pesquisadores dos diversos campos científicos, garantindo a validade e a confiabilidade dos resultados das pesquisas. A publicação destes resultados, não apenas documenta os avanços técnico-científicos, mas também promove a interação entre pesquisadores através dos processos de produção, disseminação e uso da informação, contribuindo para a sistematização da comunicação científica.

A figura 1 sintetiza a sistematização do processo cíclico das etapas de comunicação, produção e publicação científica, adaptada por Hurd (1996) a partir do modelo original de Garvey e Griffith (1972).

Figura 1 – Modelo de sistema da comunicação científica



Fonte: Garvey e Griffith (1972, adaptado por Hurd, 1996, tradução nossa).

O sistema da comunicação científica compreende, portanto, os dois processos de comunicação delineados por Le Coadic (2004), com adaptações nos métodos escritos (formais) e orais (informais), inseridos no ambiente eletrônico. No que concerne à comunicação oral, o autor destaca que esta abrange formas de difusão de informações, como conferências, colóquios, seminários, conversas e mensagens. Já a comunicação escrita compreende as publicações primárias (publicações científicas), secundárias e terciárias (publicações de resumos e índices).

Ao abordar os métodos formais, Weitzel (2006) amplia essa perspectiva ao incluir as publicações em meio eletrônico, nas quais as revistas eletrônicas, os repositórios digitais e os provedores de serviços podem ser categorizados, respectivamente, como publicações

primárias, secundárias e terciárias. Assim, essas adaptações refletem a evolução do ambiente científico para a era digital e a diversificação dos meios de comunicação na pesquisa acadêmica impulsionados pelo uso da Internet em todos os níveis de publicações.

Nesse contexto, a seção a seguir aborda a Brapci, repositório digital temático em Ciência da Informação. Classificados como publicações secundárias, os repositórios digitais, em conformidade com as políticas de Acesso Aberto, desempenham um papel significativo ao facilitar a identificação, seleção e utilização da informação (Weitzel, 2006). Essa função essencial contribui para a continuidade do ciclo de produção científica, fortalecendo a disseminação do conhecimento, além de configurar-se como intermédio de movimentos da Ciência Aberta na pesquisa ao proporcionar o Acesso Aberto a publicações científicas.

2.2 BRAPCI

A Brapci é um repositório digital de abrangência temática e nacional que engloba a produção científica de estudos e propostas nas áreas de Ciência da Informação, Biblioteconomia e Arquivologia. Oriunda do projeto de pesquisa "Opções metodológicas em pesquisa: a contribuição da área da informação para a produção de saberes no ensino superior" (Brapci, c2024), caracteriza-se pela sua relevância para a pesquisa acadêmica nacional, com frentes de pesquisa na Organização e Representação do Conhecimento, no uso de Tecnologias da Informação e na Recuperação da Informação (Bufrem; Gabriel Junior, 2022), além de indexadora de materiais como revistas científicas, artigos de periódicos, trabalhos apresentados em eventos, livros e capítulos, em sua maioria disponíveis em Acesso Aberto. Segundo Bufrem e Gabriel Junior (2022, p. 3), a Brapci foi criada originalmente a fim de

> facilitar a localização e obtenção da informação de artigos científicos; oferecer suporte à pesquisa na área em seus diferentes domínios; facilitar a análise de dados; subsidiar estudos na busca pela melhoria da qualidade das publicações periódicas da CI e socializar saberes no ensino superior, criando um observatório da área.

Atualmente, a Brapci é mantida pela Universidade Federal do Rio Grande do Sul (UFRGS), através do Programa de Pós-Graduação em Ciência da Informação (PPGCIN), em parceria com a Universidade Federal Fluminense e a Universidade Federal do Rio de Janeiro (Brapci, c2024). A base de dados brasileira da CI está disponível on-line, desde a sua conversão da base no *software* ProCite para a Web, realizada por um dos bolsistas de graduação da Universidade Federal do Paraná em 2007.

Segundo Gabriel Junior (2014b), as atualizações no número de registros da base de dados são conduzidas por um mecanismo de coleta baseado no protocolo *Open Archives Iniciative – Protocol Metadata Harvesting* (OAI-PMH). Criado por Rene Faustino Gabriel Junior, membro do projeto E3PI, este mecanismo realiza a coleta de metadados dos registros encontrados em diversas fontes de informação disponíveis na Web, tais como autor, título, palavras-chave, Identificador de Objeto Digital (DOI), referências, entre outros, bem como o texto na íntegra em formato PDF. É importante destacar que apenas os textos de registros que tenham políticas de Acesso Aberto, ou aqueles para os quais a Brapci tenha autorização dos editores ou representantes legais da publicação, são coletados automaticamente pela ferramenta, medida adotada para garantir o cumprimento das políticas de direitos autorais e copyright.

Assim, sua importância evidencia-se na fala de Bufrem e Gabriel Junior (2022, p. 3), ao afirmarem que

> Transcendendo seu papel de socializadora do conhecimento, como fonte de informação e referência para a busca e análise da informação, a Brapci tornou-se também, objeto de estudo experimental da própria área, especialmente para os membros do grupo de pesquisa Educação, Pesquisa e Produção Científica que a utilizam como objeto de suas pesquisas.

A Brapci tem contribuído de sobremaneira para o avanço da Ciência da Informação como área do conhecimento, além da pesquisa, não somente no contexto acadêmico e universitário, mas também profissional (Bufrem; Gabriel Junior, 2022), especialmente no cenário brasileiro. Com a disponibilização da base de dados na Internet, aliada às contínuas melhorias integradas desde 1996 com o projeto da Base Brasil/Espanha de Artigos de Periódicos da área em Ciência da Informação (BRES), a Brapci consolida-se como uma fonte de informação internacional para pesquisadores da CI e áreas correlatas, denotado pela crescente adoção do repositório pelos usuários nos últimos anos.

2.2.1 Histórico: da BRES à BRAPCI

Conforme destacado em trabalhos de Gabriel Junior (Gabriel Junior, 2014b; Bufrem; Gabriel Junior, 2022), a origem do atual repositório digital Brapci remonta ao projeto de pós-doutorado de Leilah Santiago Bufrem em 1995. Impulsionado pelos aportes públicos nas universidades, respaldados pelo artigo nº 207 da Constituição Federal, o projeto foi viabilizado por meio de um convênio institucional binacional entre Brasil e Espanha, que

contou com a participação e migração de professores ligados à UFPR e à Universidad Carlos III de Madrid (UC3M). O projeto tinha como objetivo principal a criação de um ambiente tecnológico e operacional que possibilitasse o estudo comparativo da produção brasileira e espanhola em Ciência da Informação (Gabriel Junior, 2014b, p. 59).

No ano seguinte, houve o planejamento para implementação da Base Brasil/Espanha de Artigos de Periódicos da área em Ciência da Informação (BRES). A escolha do *software* para suporte lógico e gerenciamento ocorreu com base na experiência prévia da UC3M, que selecionou o ProCite, um *software* de gerenciamento de referência criado pelo professor Victor Rosenberg, da Faculdade de Biblioteconomia e Estudos de Informação da Universidade de Michigan. Na época, o ProCite havia sido recentemente adquirido pelo Institute for Scientific Information Research Soft, divisão da Thomson Reuters, que desenvolveu a ferramenta no período de 1996 a 2013, sendo posteriormente substituída pelo EndNote. Durante o uso do *software* na base BRES, a ferramenta permitiu a "geração de diversos bancos de dados, a criação de filtros, para busca e recuperação da informação e a emissão de relatórios, possibilitando a exportação de dados" (Gabriel Junior, 2014b, p. 59), além do controle de informações referenciais e da compatibilidade com o protocolo Z39.50, comumente utilizado por bases de dados catalográficas, também chamadas de catálogos *online* ou Catálogo de Acesso Público *Online* (OPAC). As informações coletadas pela base, por sua vez, englobavam metadados do "título, autores, resumo, palavras-chave, localização física ou eletrônica e de identificação da fonte publicadora, com o título do periódico, volume, fascículo e ano" (Gabriel Junior, 2014b, p. 59-60), provenientes de fontes de ambos os países, como "bases de dados *online*, CD-ROM, correio eletrônico, bibliotecas digitais e exemplares disponíveis nos acervos das bibliotecas" (Gabriel Junior, 2014b, p. 59).

Além da implementação da base de dados, o andamento do projeto ainda contemplou uma investigação de dimensões duplas:

> uma delas dirigida à literatura na área, voltada às tendências temáticas e suas raízes teóricas, cujos procedimentos integram estudos métricos com as análises de conteúdo e de domínio, especialmente focadas nos artigos de periódicos e comunicações em eventos; a outra vertente, voltada à comparação entre as tendências verificadas na literatura dos dois países participantes do convênio, incluiu em seu plano de trabalho atividades didáticas e de pesquisa, a partir dos questionamentos encontrados na literatura sobre a situação da CI diante das inovações e das transformações da contemporaneidade (Gabriel Junior, 2014b, p. 60).

Garantido o sucesso do projeto e da BRES ao fornecerem condições para o reconhecimento e análise de expressões de diferentes práticas de investigação por meio da literatura especializada, possibilitando a construção de novas investigações, motivou-se a continuidade deste projeto, agora voltado à construção de uma base de dados temática de caráter nacional, a Brapci.

A implantação da Brapci, inicialmente denominada Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação, nos anos subsequentes (2000-2003), foi decorrente da colaboração existente entre mestrandos do Programa de Pós-Graduação em Ciência, Gestão e Tecnologia da Informação, bolsistas e membros do grupo de pesquisa Educação, Pesquisa e Produção Científica (E2PC) da UFPR, especialmente da coordenadora da E2PC, Leilah Santiago Bufrem, e do bolsista de iniciação científica Francisco Daniel de Oliveira Costa (Gabriel Junior, 2014b). Após o período de implantação, a base já indexava 13 títulos de periódicos, além do acervo físico em CI da Biblioteca do setor de Ciências Sociais Aplicadas da UFPR e dos fascículos solicitados através de editores e de bibliotecas cooperantes do Catálogo Coletivo Nacional de Publicações Seriadas. Os anos seguintes permitiram a indexação de novos títulos, chegando a marca de 27 títulos de periódicos indexados na base até 2008 (Bufrem, 2008).

Além da indexação de novos títulos, a Brapci contou com a aprovação no Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) em dois projetos liderados pela coordenadora do Grupo de Pesquisa E2PC nos anos de 2006 e 2008. Os projetos, intitulados “Opções metodológicas em pesquisa: a contribuição da área da informação para a produção de saberes no Ensino Superior” e “Metodologia para criação de uma base de dados *online* de acesso público: modelizando práticas para a socialização de saberes”, viabilizaram a aquisição de um servidor e recursos de infraestrutura para implantação da Brapci *online*, incluindo a conversão da base do *software* ProCite para a Web, tornando-a uma ferramenta acessível ao público (Gabriel Junior, 2014b).

2.2.2 Ferramentas de recuperação da informação na Brapci

Desde a transição da Brapci para o ambiente Web, ocorreram significativos avanços de natureza científica e tecnológica na base de dados (Bufrem *et al.*, 2010; Freitas; Bufrem; Gabriel Junior, 2010; Bufrem, Gabriel Junior, 2022). A exemplo da implementação do

mecanismo de coleta baseado no protocolo OAI-PMH, concebido por Rene Gabriel Junior em 2009, que viabilizou a coleta automatizada de metadados de diversas publicações periódicas (Gabriel Junior, 2014b). Outras melhorias abrangem a inclusão de artigos de periódicos indexados disponíveis em Acesso Aberto na base, ou daqueles para os quais foi obtida autorização dos detentores dos direitos autorais, tornando a Brapci um repositório da Ciência da Informação. Também houve a incorporação de diferentes modalidades documentais na base, como trabalhos de eventos, livros e capítulos. Além disso, até 2018 foram implementados o modelo conceitual FRBR (modelo de Requisitos Funcionais para Registros Bibliográficos) e o mecanismo de busca ElasticSearch, ampliando a eficiência na pesquisa e recuperação de informações na plataforma (Bufrem; Gabriel Junior, 2022).

O ElasticSearch, segundo a Elastic (c2024a), é um mecanismo de busca e análise de dados distribuídos que permite a realização de diversos tipos de busca em aplicações com grandes volumes de dados, incluindo dados numéricos, textuais, geográficos, estruturados, não estruturados e geoespaciais. A ferramenta foi lançada no ano de 2010, sendo amplamente utilizada por empresas no mundo inteiro, como New York Times, Adobe, Lenovo e Walmart.

Uma das principais vantagens do ElasticSearch é sua estrutura de dados de índice invertido, que permite inferir a similaridade entre textos utilizando técnicas como *Term Frequency - Inverse Document Frequency* (TF-IDF). Essa técnica baseia-se na frequência de um termo em determinado documento em relação à frequência inversa desse termo em um conjunto de documentos, proporcionando uma análise mais eficaz e precisa dos dados armazenados (Beiske, 2013; Elastic, c2024a). Além disso, o ElasticSearch oferece rápida performance para leitura e gravação de dados, escalabilidade para lidar com grandes volumes de dados em um único servidor, suporte a diversas linguagens de programação devido ao uso de *APIs RESTful*[3] e integração com outras ferramentas, como na plataforma Elastic Stack (Elastic, c2024a).

Além desses, Freitas, Bufrem e Gabriel Junior (2010) descrevem outra mudança na recuperação da informação, que busca hierarquizar a distribuição dos resultados nas consultas na Brapci, atribuindo diferentes valores aos campos de busca. A relação de relevância dos campos está ilustrada no quadro 1.

[3] *API RESTful* é uma interface que dois sistemas de computador usam para trocar informações de forma segura pela Internet. A maioria das aplicações de negócios precisa se comunicar com outras aplicações internas e de terceiros para executar várias tarefas. [...] As APIs RESTful suportam essa troca de informações porque seguem padrões de comunicação de *software* seguros, confiáveis e eficientes (Amazon, c2023).

Quadro 1 – Atribuição dos valores aos campos de busca

Bit	Atribuição	Valor decimal
1	Resumo	1
2	Palavras-chave	2
3	Título	4

Fonte: Freitas, Bufrem e Gabriel Junior (2010, p. 49).

Com isso, termos de busca encontrados nos campos de título têm maior impacto do que aqueles encontrados em resumos ou palavras-chave, ou mesmo em ambos. O valor máximo é alcançado quando os termos de consulta aparecem em todos os campos de um artigo, somando sete pontos (1 ponto para o resumo, 2 pontos para as palavras-chave e 4 pontos para o título).

Adicionalmente, a Brapci incorporou o uso de Inteligência Artificial (IA) em seu sistema, como parte do projeto de produtividade conduzido por Rene Gabriel Junior da Universidade Federal do Rio Grande do Sul em 2023. O objetivo principal foi aprimorar a organização das informações, iniciando com a aplicação da IA na tradução automática de trabalhos redigidos em inglês e espanhol (Brapci, c2024).

2.3 SISTEMAS DE RECOMENDAÇÃO

A sobrecarga de informação é um dos principais sintomas causados pelo crescimento informacional. Desencadeada pelos avanços técnico-científicos, a “explosão informacional”, que marcou a metade do século XX após o término da Segunda Guerra Mundial, refere-se a um “crescimento exponencial e ininterrupto de publicações científicas e técnicas, bem como registros de informações de todos os tipos”, resultando em uma revolução técnico-científica (Saracevic, 1999). De acordo com o autor, essa revolução resultou no surgimento de novos campos, tais como a CI e a Ciência da Computação, além do desenvolvimento de áreas como a Recuperação da Informação. Posteriormente, com a criação da World Wide Web por Tim Berners-Lee no início da década de 1990, a disponibilidade de informações na Internet experimentou uma expansão significativa, resultando mais uma vez na sobrecarga informacional.

Durante o período elucidado, diferentes áreas do conhecimento propuseram soluções para lidar com a grande quantidade de dados na Web, a exemplo dos Sistemas de Organização do Conhecimento na CI, a Web Semântica[4] na Ciência da Computação e a Personalização no Marketing. A respeito desta última, compreende-se por Personalização a "técnica utilizada para recomendar produtos aos consumidores com base em seus perfis de consumo" (Torres, 2004a, p. 25). Além disso, Monteiro-Krebs (2013) acrescenta que a técnica de personalização também pode ser aplicada na recomendação de documentos, especialmente no contexto das bibliotecas digitais.

A Internet, por sua vez, tornou-se um ambiente favorável para o amplo uso e desenvolvimento destas soluções, além de possibilitar aplicações diretas resultantes da interdisciplinaridade entre áreas. As demandas no mercado comercial eletrônico, por exemplo, formaram o ambiente ideal para utilização de serviços de personalização e recomendação na Web, provenientes de uma intersecção entre as áreas de Marketing e Ciência da Computação.

Por recomendação, neste contexto de intersecção, assume-se o seguinte significado: "forma particular de filtragem de informação, que explora os comportamentos do passado e semelhanças do usuário para gerar uma lista de itens de informação que é pessoalmente adaptada às preferências de um usuário final" (ACM, c2024, tradução de Monteiro-Krebs, 2013, p. 36-37). Já os serviços de personalização, apresentados por Torres (2004a), são soluções formadas pela integração entre os dados de perfis de consumo de clientes (ou usuários), o sistema (ou site), e o sistema de recomendação, um *software* de personalização baseado em filtragem da informação. Dada a importância dos conceitos de personalização e recomendação para os serviços de personalização, recém elencados, busca-se a seguir conceitualizar os sistemas de recomendação.

Os sistemas de recomendação (SRs) são compostos por *softwares* baseados em ferramentas e técnicas de filtragem da informação, mineração de dados e, em alguns casos, inteligência artificial para processamento e entrega de sugestões personalizadas de produtos ou serviços, visando atender a necessidades e interesses de usuários do sistema (Torres, 2004a; Monteiro-Krebs, 2013; Vieira; Passos; Salm, 2023).

---

[4] A Web Semântica é uma estrutura de representação de dados na Web, desenvolvida pelo World Wide Web Consortium (W3C), que permite a interoperabilidade entre computadores e pessoas por meio de uma Web de Dados Conectados (W3C Capítulo São Paulo, [202?]).

Outras definições incluem aspectos diferentes dos sistemas. De acordo com a *Encyclopedia of Machine Learning*, os sistemas de recomendação podem ser definidos da seguinte maneira:

> O objetivo de um sistema de recomendação é gerar recomendações significativas para um grupo de usuários em relação a itens ou produtos que possam interessá-los. Sugestões de livros na Amazon ou filmes na Netflix são exemplos do funcionamento de sistemas de recomendação de alta performance na indústria [...]. Além disso, o sistema pode ter acesso a atributos de perfil específicos do usuário e do item, como dados demográficos e descrições de produtos, respectivamente. Sistemas de recomendação diferem na forma como analisam essas fontes de dados para desenvolver noções de afinidade entre usuários e itens, que podem ser usadas para identificar pares bem correspondidos. Sistemas de Filtragem Colaborativa analisam apenas interações históricas, enquanto sistemas de Filtragem Baseada em Conteúdo baseiam-se em atributos de perfil; e técnicas híbridas tentam combinar ambos esses designs. A arquitetura de sistemas de recomendação e sua avaliação em problemas do mundo real são áreas ativas de pesquisa (Melville; Sindhwani, 2011, p. 829, tradução nossa).

No âmbito da CI, autores como Monteiro-Krebs (2013), Alvarez *et al.* (2016), Monteiro-Krebs, Rocha e Ribeiro (2017), Monteiro-Krebs *et al.* (2021), Monteiro-Krebs et al. (2022), Souza e Lima (2022), Monteiro-Krebs et al. (2023), Vieira, Passos e Salm (2023), entre outros, abordam os sistemas de recomendação sob diferentes perspectivas. Entre as pesquisas citadas, destacam-se a análise métrica das recomendações em bibliotecas universitárias, o uso desses sistemas em projetos de arquitetura de ambientes informacionais, os SR em plataformas de *streaming* aliados aos Sistemas de Organização do Conhecimento, a proposta de modelos conceituais de recomendação baseadas em filtragem híbrida, bem como a análise da mediação algorítmica, comunicação com os usuários e vieses das recomendações presentes em um SR para redes sociais acadêmicas.

2.3.1 Etapas da recomendação

As predições em forma de recomendações que são feitas pelos sistemas constituem parte essencial da filtragem de informação. No entanto, para que sejam geradas recomendações personalizadas pelo sistema, é imperativo atender a etapas que visam a coleta de dados dos perfis de consumo até que sejam geradas recomendações baseadas nas preferências de um usuário ativo ou corrente. Estas etapas, conforme Torres (2004a), referem-se à identificação do usuário e dos hábitos de consumo e à geração de recomendações aos usuários pelo sistema.

No que concerne à identificação do usuário, podem ser utilizados meios automatizados ou manuais. Entre aqueles que dispensam as ações de um usuário para a sua identificação estão os *cookies,* os identificadores (ID) em sessões e os registros do Protocolo de rede Internet (IP). Segundo Torres (2004, p. 34), um *Cookie* "é um arquivo gravado pelo site no computador do usuário. Dessa forma, pode-se recuperar o perfil do usuário utilizando seu identificador, armazenado nesse arquivo", permitindo uma espécie de *login* automático. Já um endereço IP, ou endereço de rede, "é um endereço configurado por meio do *software* da rede", podendo conter um número de código que identifique a rede, sub-rede ou *host* (Casad; Willsey, 1999, p. 8). A identificação manual, por outro lado, pode ser feita através de *login*, na qual são fornecidas credenciais de acesso na forma de um identificador e senha criados pelo próprio usuário.

Ambos os métodos apresentam vantagens e desvantagens a serem avaliadas pelos gestores destes sistemas. *Cookies* e endereços IPs apresentam abordagens que, normalmente, são despercebidas pelo usuário. No entanto, a gravação de *Cookies* pode ser desabilitada pelo usuário ou apagada no histórico do navegador. Além disso, os *Cookies*, assim como o acesso via *login*, podem ser evitados através da navegação anônima (Monteiro-Krebs, 2013). Por último, outro problema enfrentado pelos métodos automatizados está no uso de diferentes dispositivos para acesso ao sistema por um único usuário, que recebe endereços lógicos e identificadores únicos para cada dispositivo, sendo necessária a sincronização manual dos mesmos pelo usuário.

Quanto à identificação dos hábitos de consumo dos usuários, estes podem ser extraídos por métodos implícitos ou explícitos. A extração de perfil implícita compõe um método de extração automática de perfis de usuários pelo sistema, que monitora hábitos de consumo ou navegação através do rastreamento de ações do usuário, como o histórico de compras em um site, o tempo de leitura de uma página ou sua adição aos *bookmarks* (favoritos), o *download* de um artigo, as expressões de busca solicitadas ao servidor, os cliques e outros movimentos do mouse (*scroll*), entre outros métodos de rastreio (Torres, 2004a).

A extração de perfil explícita, por outro lado, é realizada pelo próprio usuário, que informa ao sistema suas preferências em avaliações (*ratings*) no site (Torres, 2004a). Essas avaliações são frequentes durante o processo de cadastro de usuários e geralmente envolvem escalas numéricas a fim de delinear o perfil de consumo do usuário. Neste tipo de extração de

hábitos de consumo de usuários novos, são comuns as avaliações de itens ou categorias escolhidos de maneira aleatória pelo sistema a fim de criar um perfil de consumo que permita o sistema gerar recomendações precisas.

A respeito das vantagens e desvantagens proporcionadas por estes métodos, Torres (2004a) afirma que as avaliações (extração explícita) compõem um método mais confiável. No entanto, segundo o autor, os métodos para extração explícita de informações podem ser um problema para usuários que desgostam das interações para calibragem de predições do sistema, ao passo que o monitoramento e rastreamento dos hábitos de consumo (extração implícita), muitas vezes imperceptíveis para o usuário, reduzem, ou mesmo anulam, o esforço do usuário para informar suas preferências ao sistema. Em vista disso, alguns sistemas de recomendação proporcionam métodos mistos de extração dos hábitos de consumo, com a presença de métodos de monitoramento e rastreamento e opções para que o usuário altere dados de seu perfil manualmente.

Apesar disso, cabe ressaltar que, ao longo dos últimos anos, a privacidade e a proteção de dados têm sido amplamente discutidas pela sociedade atual, resultando em mudanças significativas tanto no direito, com a implementação da Lei Geral de Proteção de Dados (LGPD) e do Regulamento Geral sobre a Proteção de Dados (GDPR), quanto na percepção e confiança dos usuários em relação às plataformas com as quais interagem. Consequentemente, os debates acerca da mediação algorítmica e seu impacto na sociedade têm atraído o interesse de diversos pesquisadores no meio acadêmico, que destacam alguns dos desafios no desenvolvimento de sistemas algorítmicos, incluindo questões éticas, transparência, viés e *design,* e como esses desafios influenciam diretamente a maneira como os sistemas coletam, armazenam, utilizam e divulgam os dados de seus usuários (Monteiro-Krebs et al., 2019; Storms; Alvarado Rodriguez; Monteiro-Krebs, 2022).

Diante deste cenário, Monteiro-Krebs et al. (2019) apontam como possíveis caminhos para o aprimoramento e desenvolvimento de sistemas fundamentados na Interação Humano-Computador (IHC), incluindo os SR, a utilização de abordagens de *design* centradas no usuário. Enquanto Storms, Alvarado Rodriguez e Monteiro-Krebs (2022) ilustram como as expectativas e impressões dos usuários estão diretamente relacionadas às experiências nas plataformas, além de demonstrar a inter-relação entre a transparência do sistema, o controle dos usuários sobre o uso de seus dados e a satisfação com o sistema.

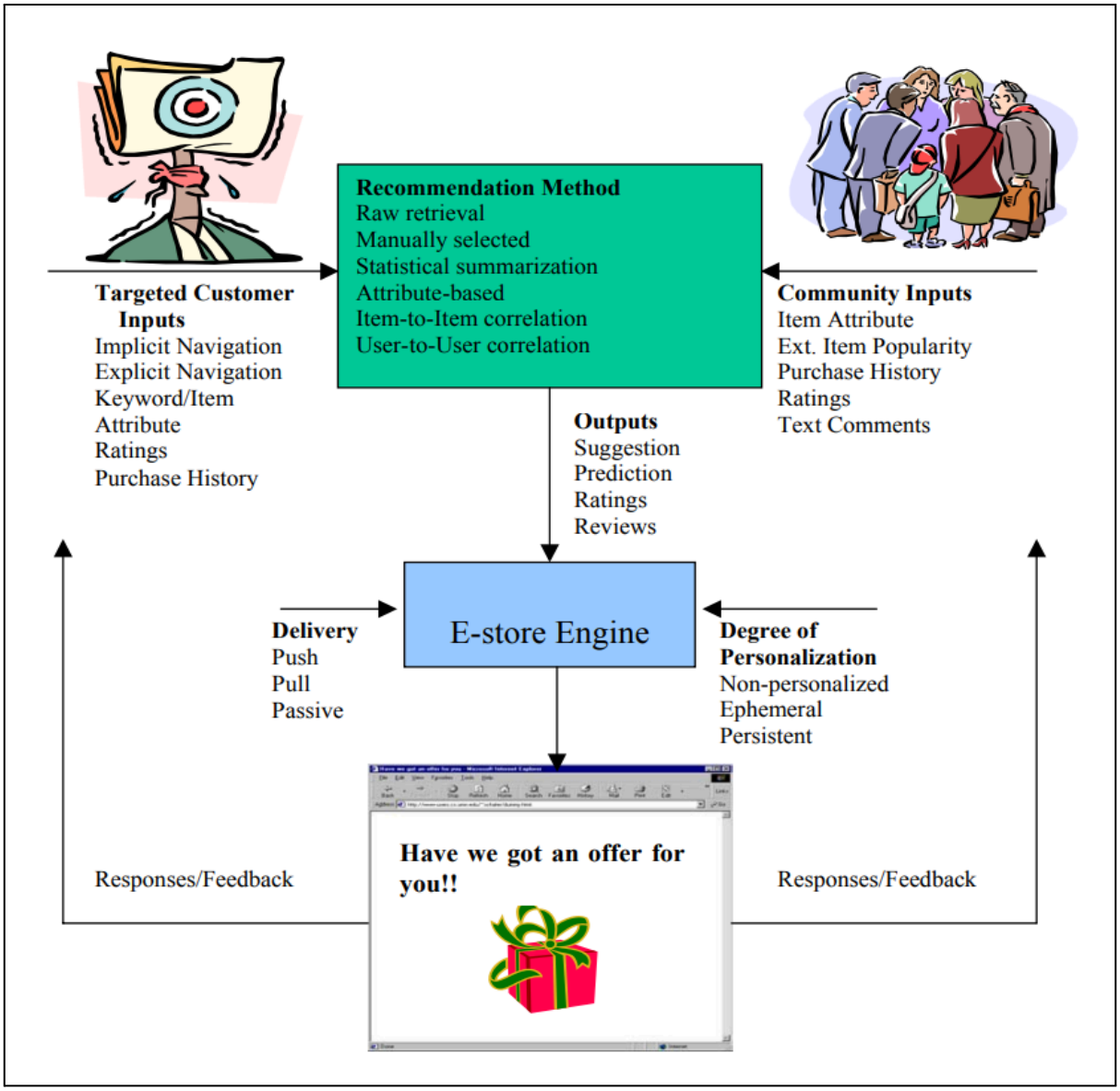
2.3.2 Classificação dos Sistemas de Recomendação

Estudos em universidades acerca dos sistemas de recomendação vêm sendo conduzidos desde a sua disseminação no mercado digital na década de 1990. Neste sentido, Torres (2004a) exemplifica diversos estudos, destacando questões específicas relacionadas ao aprimoramento das recomendações, ao aumento na velocidade das predições do sistema, ao desenvolvimento de técnicas que consigam recomendar produtos e serviços para usuários com poucas informações em seus perfis de consumo, às atualizações das preferências dos usuários do sistema e ao desenvolvimento de novas técnicas de filtragem da informação.

Outro aspecto abordado nas pesquisas desta área diz respeito a suas classificações. Ainda que não haja um consenso na área, pesquisas anteriores abordam aspectos específicos, como a apresentação das recomendações aos usuários e as formas de hibridização de técnicas diferentes em um único sistema (Burke, 2007). No entanto, para os objetivos desta pesquisa, será examinada a taxonomia proposta por Schafer (2001) e revisada por Torres (2004a).

Assim, com base na avaliação das características que compõem os sistemas de recomendação utilizados por empresas do ramo de *e-commerce*, Schafer (2001) classifica os sistemas com base em três grandes categorias: (1) a entrada e saída funcionais, (2) o método de recomendação e (3) outras questões de design. Abaixo, a figura 1 demonstra o processo de recomendação pelo sistema a partir da perspectiva desses três aspectos.

Figura 2 – O processo de recomendação no *e-commerce*



Fonte: Schafer (2001, p. 37).

A análise individual destas categorias será abordada nas alíneas a seguir segundo a taxonomia de Schafer (2001) e apontamentos de Torres (2004a):

a) **entrada e saída funcionais:** relacionada com os fluxos de entrada e saída da informação dentro do sistema. Os tipos de informação utilizados são:

   – **informações do usuário (*targeted customer inputs*):** considera os aspectos para extração de hábitos de consumo e navegação do usuário ativo, vistos na seção 2.3.1 Etapas da recomendação. Práticas comuns para extração dos hábitos de navegação dos usuários incluem a extração implícita e a extração explícita. No

entanto, nem sempre a entrada de informações se limita a um único tipo de extração, havendo a possibilidade de extrair informações de palavras-chave em uma busca ou atributos de um item disponível na página;

- **informações da comunidade de usuários (*community inputs*):** inclui informações sobre uma comunidade inteira de usuários para recomendação de produtos. Apesar deste método de recomendação não incluir personalização, pode ser utilizado para recomendar produtos ou informações para usuários com poucas informações em seus perfis. Exemplos de informações sobre a comunidade utilizadas para gerar recomendações compreendem os atributos de itens, como categorias e rótulos que constituem consenso entre a comunidade; a popularidade externa do item, refletindo a popularidade de itens na comunidade a exemplo dos históricos de acesso, *downloads* ou compras; os comentários de texto, que acrescentam a opinião de outros usuários, apesar de constituírem um grande esforço a usuários e, portanto, são feitos e lidos por uma parcela menor da comunidade; e as avaliações abordadas anteriormente, encorajadas pela maioria dos sites que utilizam comentários;

- **formato de saída ou oferta das recomendações (*outputs*):** ofertas de recomendação variam em tipo, quantidade e apresentação da informação recomendada. O tipo mais comum de recomendação é o formato de sugestão. Quanto à quantidade de recomendações, o menor número de sugestões pode aumentar as chances de consumo de um único item, mas aumentam o risco de o usuário não aproveitar nenhuma recomendação caso esta seja rejeitada. Comumente utilizam-se listas para minimizar os riscos. Já a apresentação da informação recomendada considera diversos aspectos, como a ordem de disposição das recomendações, para que usuários não deixem de conferir outros itens com índices de predição menores; a entrega dos valores de predição, adicionalmente os valores podem ser representados por agrupamentos; e a disponibilização de avaliações (*ratings*) e comentários de texto (*reviews*);

b) **método de recomendação:** referente aos processos específicos usados em sistemas de recomendação. Segundo Schafer, os métodos de recomendação dividem-se em seis

abordagens, além disso apresentam uma segunda classificação proposta por Torres em métodos não personalizados, efêmeros e personalizados. São elas:

- **busca direta (*raw retrieval*):** pode ser classificada como um recomendador nulo, que não apresenta personalização para gerar recomendações. Este método está presente em interfaces de buscas de bases de dados, em que expressões de busca (*queries*) ligadas às linguagens documentárias são usadas para recuperar a informação desejada. Apesar de não apresentar critérios de personalização entre as recomendações, Schafer considera este um método de recomendação por possibilitar que usuários encontrem novos produtos não esperados. Contudo, Torres adverte que este método não compõe um sistema de filtragem da informação, mas sim de recuperação da informação, tornando-o um falso recomendador;

- **seleção manual (*manually selected*):** inclui a seleção manual de produtos ou informações por especialistas (editores, artistas, críticos, etc.) e listas de recomendações por usuários do sistema. Normalmente, encontram-se acompanhadas de comentários de texto. Baseia-se em critérios, muitas vezes, subjetivos como gosto, qualidade, interesses e objetivos. Por apoiar-se apenas em interesses distantes do usuário-alvo, ou das opiniões presentes em uma comunidade, este método pode ser classificado como não personalizado;

- **resumos estatísticos (*statistical summaries*):** são utilizados em recomendações genéricas, para perfis com poucos hábitos de consumo conhecidos pelo sistema. Fundamentam-se em estatísticas de uso, visualizações, compras e avaliações. Assim como o método anterior, os resumos estatísticos constituem um método não personalizado;

- **baseado em atributos (*attribute-based*):** aproxima-se do método de busca direta, porém apresenta características adicionais. Métodos baseados em atributos consideram diversos atributos dos itens recomendados além das presentes na busca direta. Este método pode ser classificado como efêmero, pois pondera apenas a consulta ao item ocorrida durante a navegação pela página;

- **correlação item-item (*item-to-item correlation*):** Ocorre quando há associações entre itens, a exemplo de dados de compras com dois ou mais itens, preferências de usuários comuns e outras métricas associativas. Também considerado um método efêmero, justificado pela consulta a dados recentes sobre a comunidade para embasar recomendações;

- **correlação usuário-usuário (*user-to-user correlation*):** o último método proposto por Schafer fundamenta-se na correlação de similaridade entre perfis de consumo de usuários. Consiste em um método persistente de recomendação que avalia o histórico do perfil de consumo do usuário-alvo para gerar recomendações baseadas em seus vizinhos (usuários com perfis similares);

- outro critério elencado pelo autor é a presença de recomendações computadas inteiramente *online* ou parcialmente *offline*. Recomendações do tipo busca direta, seleção manual, resumos estatísticos e baseadas em atributos requerem menor processamento computacional, ao passo que métodos baseados na correlação item-item ou usuário-usuário exigem uma capacidade de processamento mais intensiva;

c) **outros aspectos do sistema:** incluem diversos aspectos como precisão, utilidade e individualização das recomendações, compreendidos pelo grau de personalização, e entrega de recomendações. A precisão pode ser medida através das recomendações aceitas por usuários, enquanto a utilidade é mensurada mediante a utilização de conceitos como a serendipidade, representando a capacidade de oferecer sugestões inesperadas. A individualização mede a capacidade de gerar recomendações personalizadas. Já o grau de personalização, pode ser aplicado para classificar recomendações geradas por cada método de recomendação em não-personalizadas, efêmeras ou persistentes. O quadro 2 simplifica a visualização da categorização dos métodos de recomendação mencionados. Por fim, a entrega de recomendações é classificada pela forma de apresentação das sugestões em *"push"*, onde são automaticamente “empurradas” ao usuário; *"pull"*, método menos intrusivo no qual as recomendações são inicialmente ocultas do usuário, bastando um clique para que sejam “puxadas” pelo sistema; e orgânica (ou passiva), na qual as recomendações são integradas ao conteúdo do site.

Quadro 2 – Grau de personalização x método de recomendação

Grau de personalização	Métodos de recomendação
Não-personalizado	Busca direta, seleção manual, resumos estatísticos
Efêmero	Correlação item-item e baseado em atributos
Persistente	Correlação usuário-usuário

Fonte: Adaptado de Torres (2004a).

Os métodos de recomendação citados utilizam diversas técnicas aplicadas na recuperação e filtragem da informação. A seção a seguir aprofunda-se nas técnicas mais utilizadas para FI, além de diferenciar os conceitos de recuperação da informação e filtragem da informação.

2.3.3 Técnicas de filtragem da informação

As primeiras técnicas aplicadas em sistemas de recomendação têm origem nos sistemas de recuperação da informação, como os sistemas baseados no modelo de busca direta, que precedem a Web e mesmo os estudos e pesquisas sobre sistemas de recomendação nas universidades. Para entender as técnicas usadas por estes sistemas hodiernamente, é necessário contextualizá-los e realizar uma comparação das principais diferenças existentes entre ambos.

A recuperação da informação (RI) é uma área ligada à Ciência da Computação, Ciência da Informação e Organização do Conhecimento, que estuda a eficiência e eficácia das buscas de informações em sistemas computacionais (Silva; Santos; Ferneda, 2013). Portanto, a RI concentra-se na investigação da representação, armazenamento, organização e acesso da informação. Conforme Cesarino (1985, p. 2), a recuperação da informação abrange o estudo de diversos assuntos como a teoria da informação, os canais de comunicação, o usuário da informação, a seleção e aquisição de documentos, a análise de assunto, as linguagens documentárias, a armazenagem da informação, a formação de base de dados, as estratégias de busca, a disseminação da informação, o planejamento e a avaliação de sistemas de informação. Quanto aos sistemas de recuperação da informação, a autora define-os como:

> [...] um conjunto de operações consecutivas executadas para localizar, dentro da totalidade de informações disponíveis, aquelas realmente relevantes. Para isso, executam as funções de seleção, análise, indexação e busca das informações. Em todas essas etapas a interação usuário x sistema é fundamental, embora tenha se apresentado com muitas falhas (Cesarino, 1985, p. 1).

Nesse contexto, os sistemas de recuperação da informação (SRI) desempenham funções essenciais, incluindo a representação, o armazenamento, a organização e a localização de informações. Contudo, apesar dos avanços nessa área, que incorporam novas técnicas, como os modelos clássicos de recuperação da informação (modelo booleano, modelo vetorial e modelo probabilístico), bem como abordagens mais recentes, incluindo o modelo fuzzy, as Redes Neurais e os Algoritmos Genéticos (Silva; Santos; Ferneda, 2013), a crescente demanda por métodos que aprimorem as buscas de informações na Web tem estimulado o desenvolvimento de novos sistemas e novas técnicas.

Os sistemas baseados em filtragem de informações, por exemplo, foram desenvolvidos com o intuito não apenas de lidar com buscas de informações recentes feitas por meio de *queries* (consultas) em um SRI, mas também integrar os perfis de consumo dos usuários. Dessa maneira, os sistemas de filtragem de informações possibilitam a localização de itens que, embora não se encontrem nas estratégias de busca adotadas, estejam alinhados com os perfis de consumo dos usuários. Tais itens são então apresentados em conjunto com os demais resultados recuperados pela *query* utilizada, ampliando a relevância e, consequentemente, a acurácia e a revocação de itens recuperados. Outro aspecto da filtragem de informações a ser citado é a capacidade dos sistemas em recomendar itens automaticamente, sem a necessidade de criação de consultas para buscas no sistema. Adicionalmente, a personalização proporcionada ao usuário beneficia a sua experiência e permite aprimorar a mediação do conteúdo. Abaixo, o quadro 3 destaca as principais diferenças entre os sistemas baseados em recuperação da informação e FI.

Quadro 3 – Recuperação de Informações x filtragem de Informações

Recuperação de Informações (RI)	Filtragem de Informações (FI)
Uso esporádico por usuários com necessidade de informação momentânea	Uso constante por usuários com objetivos e interesses a longo prazo
Realização de consultas por meio de “queries”, criadas pelos usuários	O sistema constrói o perfil dos usuários e fornece as informações baseadas nele

Fonte: Torres (2004a, p. 76).

As técnicas aplicadas na filtragem de informações variam. Entretanto, as mais utilizadas atualmente são as técnicas de Filtragem Colaborativa, Filtragem Baseada em Conteúdo e Filtragem Híbrida, vistas a seguir. Outras técnicas mencionadas e discutidas na literatura, como filtragem baseada em conhecimento (Souza; Lima, 2022; Vieira; Passos; Salm, 2023), filtragem demográfica (Vieira; Passos; Salm, 2023) e frames de recomendação (Torres, 2004a), não comporão parte do referencial teórico desta pesquisa.

2.3.3.1 Filtragem Colaborativa

A Filtragem Colaborativa (FC) tem sua origem no modelo Tapestry, o primeiro modelo de recomendação comercial, introduzido por Goldberg, Nichols, Oki, e Terry em 1992 (Melville; Sindhwani, 2011). O sistema havia sido criado no intuito de aumentar a "colaboração social", proporcionando maior sofisticação nas recomendações de notícias dadas aos usuários, além de mitigar a sobrecarga de informação trazida pelo grande volume de documentos e assim sobrepujar os modelos mais antigos de filtragem de conteúdo, similares aos sistemas de recuperação da informação.

Conforme destacado por Torres (2004a) e Ricci, Rokach e Shapira (2015), a FC figura entre as técnicas mais empregadas em sistemas de recomendação, fundamentando-se na similaridade entre os usuários para gerar recomendações. Semelhante à forma como as informações são repassadas entre pessoas no cotidiano da vida real, a técnica apoia-se no pressuposto de que usuários com preferências e comportamentos similares têm maior probabilidade de apreciar e aceitar recomendações uns dos outros. Assim, a Filtragem Colaborativa destaca-se por sua capacidade de explorar padrões coletivos, proporcionando recomendações automatizadas, personalizadas e relevantes aos usuários com base nas interações, afinidades e avaliações compartilhadas na comunidade de usuários do sistema.

Para Melville e Sindhwani (2011), sistemas baseados em Filtragem Colaborativa operam coletando informações dos usuários através de avaliações (*ratings*) dadas aos itens. A partir destas avaliações, os sistemas poderão determinar qual item recomendar com base na similaridade entre o usuário ativo e seus "vizinhos". Nesse contexto, o termo vizinhos refere-se ao método de vizinhança *K-Nearest Neighbors* (K-NN), no qual os usuários que demonstram uma correlação significativamente alta de suas avaliações com as avaliações do usuário ativo são identificados como vizinhos.

A determinação da correlação de similaridade entre dois ou mais usuários envolve cálculos matemáticos, comumente realizados por meio de métricas ou medidas de similaridade como o coeficiente de correlação de Pearson e o cosseno, com ênfase no uso dado ao primeiro (Torres, 2004a). O coeficiente de Pearson é uma métrica usada para avaliar a força do relacionamento entre variáveis em uma escala de [-1;1]. Por outro lado, o cosseno é uma métrica que mede o ângulo entre vetores, fornecendo valores em uma escala de [0;1]. Em ambos os métodos, valores mais baixos à esquerda do intervalo indicam a ausência de correlação entre as variáveis (usuários), enquanto valores mais altos à direita indicam uma forte correlação entre as variáveis. Há ainda outras medidas de similaridade encontradas na literatura da área, como o cosseno ajustado, *Log-likelihood*, a distância euclidiana, correlação de Spearman, etc (Aleixo, 2014).

A correlação de similaridade entre usuários é recalculada a cada consulta ao sistema, possibilitando a formação de novas vizinhanças com base nos valores de similaridade atualizados, o que também aumenta a exigência computacional do método (Torres, 2004a). Essa exigência computacional é acompanhada por uma demanda de recursos em memória e armazenamento de dados intensiva, dificultando a escalabilidade do algoritmo (Elastic, c2024c). Quanto às vizinhanças, Torres (2004a) afirma que elas podem ser geradas de duas maneiras:

a) **similaridade:** o método de similaridade determina a quantidade de vizinhos com base em um grau fixo de correlação entre o perfil do usuário ativo e seus vizinhos. Consequentemente, apenas usuários com perfis sabidamente similares ao usuário ativo são considerados vizinhos. No entanto inclui-se a possibilidade de não haver vizinhos compondo sua vizinhança;

b) **número de vizinhos:** o método de número de vizinhos impõe um intervalo que especifica o número máximo de usuários para formarem vizinhanças. Ao contrário do método anterior, estabelece-se um número de usuários que garanta vizinhos suficientes, ao custo da baixa confiabilidade presente nas correlações de similaridade entre usuários. Dependendo do domínio e dimensões do sistema, o número de vizinhos poderá estar na casa de milhares de usuários.

O autor também ressalta que, em determinadas situações, usuários com uma vizinhança muito reduzida são denominados "ovelhas negras" ("*gray sheep*"), o que pode

complicar o processo de fornecimento de recomendações pelo sistema. Nessas circunstâncias, optar por recomendar itens mais bem avaliados pode representar uma alternativa viável.

Por último, a geração das recomendações pode ser realizada como uma forma de predição das avaliações de um usuário ou como sugestão. Através de outro cálculo matemático, o sistema efetua uma média ponderada entre as avaliações do usuário ativo e todas as avaliações de seus vizinhos para todos os itens, recomendando aqueles com os valores de previsão mais elevados.

Conforme pesquisa realizada por Melville e Sindhwani (2011), os autores destacam que o modelo de vizinhança possui algumas extensões que podem aprimorar a performance do sistema. Algumas dessas extensões incluem:

a) **Filtragem Colaborativa baseada em itens (*item-based Collaborative Filtering*):** o método de FC baseada em itens, também conhecido como Item K-NN, foi proposto por Linden, Smith e York (2003) como alternativa à complexidade computacional de sistemas com milhões de usuários enfrentada pelo uso do método convencional de FC baseado em vizinhanças. Em vez de calcular a similaridade entre usuários, esse método pré-processa a similaridade entre os itens avaliados por cada usuário, resultando em recomendações mais rápidas e eficientes;

b) **ponderação de significância (*significance weighting*):** Esta abordagem é uma ferramenta eficaz para mitigar o problema do "falso" bom vizinho (Torres, 2004a). Os falsos bons vizinhos, frequentemente classificados como maus preditores (Melville; Sindhwani, 2011), são erroneamente adicionados à vizinhança de usuários e sobrepõem vizinhos estáveis devido à alta similaridade gerada por coincidências em um baixo número de avaliações. Nesse sentido, a ponderação de significância relaciona-se a atribuição de um "peso" ao coeficiente de similaridade das avaliações de usuários, reduzindo a inclusão indevida de vizinhos instáveis com perfis altamente correlacionados, baseados em avaliações escassas;

c) **votação padrão (*default voting*):** semelhante a ponderação de significância, a votação padrão é uma técnica utilizada para diminuir o impacto de correlações formadas por poucas avaliações em comum. Dessa forma, atribui-se um valor padrão para itens que não tenham sido coavaliados por ambos os usuários a fim de calcular a correlação dos itens coavaliados;

d) **frequência inversa do usuário (*inverse user frequency*):** A estratégia de frequência inversa do usuário é empregada para evitar sugestões de itens avaliados por uma quantidade excessiva de usuários do sistema, já que essas tendem a ser menos úteis e significativas em comparação com recomendações de itens conhecidos por uma quantidade menor de usuários;

e) **amplificação de caso (*case amplification*):** esta técnica permite aumentar o peso de avaliações geradas por vizinhos que tenham uma correlação de similaridade muito alta com o usuário-ativo;

f) **outras técnicas:** além das técnicas citadas, Melville e Sindhwani (2011, p. 832) citam ainda outras extensões da Filtragem Colaborativa como FC baseada em similaridade e FC impulsionada por imputação.

Adicionalmente, Ricci, Rokach e Shapira (2015, p. 12), ao discorrerem sobre a escolha dos métodos aplicados à técnica de Filtragem Colaborativa baseada em métodos de vizinhança, apontam que:

> Em sistemas de recomendação comerciais típicos, nos quais o número de usuários ultrapassa o número de itens disponíveis, abordagens baseadas em itens devem ser preferidas. Essa preferência decorre de sua capacidade de fornecer recomendações mais precisas, sendo também mais eficientes computacionalmente e exigindo atualizações menos frequentes. Por outro lado, métodos baseados em usuários geralmente proporcionam recomendações mais originais, o que pode conduzir os usuários a uma experiência mais satisfatória.

Além do método de vizinhança, a Filtragem Colaborativa apresenta uma segunda subdivisão, denominada Filtragem Colaborativa baseada em modelos (Melville; Sindhwani, 2011; Vieira; Passos; Salm, 2023). Existem várias técnicas de FC baseadas em modelos, as quais estimam parâmetros de modelos estatísticos para avaliações de usuários. Essas técnicas oferecem características complementares, como o aprendizado contínuo de um modelo preditivo, em comparação ao método de memória clássico baseado em vizinhança de usuários ou itens.

No entanto, técnicas oriundas de métodos baseados em modelos exigem um pré-processamento dos dados que compõem a matriz de avaliações de itens e usuários, conhecido como período de treino, gerando um alto custo computacional. Muitas vezes, esse custo pode ultrapassar o tempo necessário para gerar recomendações nos métodos tradicionais de vizinhança. Por essa razão, métodos de recomendação baseados em modelo geralmente são

executados *offline* para evitar a presença de gargalos no sistema, o que pode ocasionar a recomendação de itens que não se adequem às buscas mais recentes dos usuários no sistema (Aleixo, 2014). Em contrapartida, os métodos baseados em modelo oferecem soluções variadas para problemas de escalabilidade, esparsidade e outros desafios discutidos na literatura.

2.3.3.2 Filtragem Baseada em Conteúdo

A técnica de Filtragem Baseada em Conteúdo (FBC), ao contrário da Filtragem Colaborativa, fundamenta-se no princípio de que usuários que consumiram um item do sistema tendem a consumir itens similares. Dessa forma, esta técnica baseia-se no uso de informações acerca de atributos de itens para medir sua similaridade com outros itens e com o perfil do usuário. Quanto às classificações desta técnica, Vieira, Passos e Salm (2023) indicam duas abordagens de aplicação para a FBC: a abordagem de modelo clássico e a abordagem de modelo aliada a abordagens semânticas.

Essas duas abordagens são destacadas por Ricci, Rokach e Shapira (2015, p. 12, tradução nossa), que fornecem o seguinte resumo a respeito dos modelos aplicados em FBC:

> As técnicas clássicas de recomendação baseadas em conteúdo visam combinar os atributos do perfil do usuário com os atributos dos itens. Na maioria dos casos, os atributos dos itens são simplesmente palavras-chave extraídas das descrições dos itens. As técnicas de indexação semântica representam o item e os perfis do usuário usando conceitos em vez de palavras-chave. [...] Os autores [Gemmis *et al*., 2015] apresentam dois grupos principais de técnicas de indexação semântica: *top-down* [de cima para baixo ou exógena] e *bottom-up* [de baixo para cima ou endógena]. As técnicas do primeiro grupo baseiam-se na integração de fontes externas de conhecimento, tais como: ontologias, conhecimento enciclopédico (como a Wikipedia) e dados da nuvem *Linked Data*, enquanto as técnicas do último grupo dependem de uma representação semântica leve baseada na hipótese que o significado das palavras depende do seu uso em grandes Corpora de documentos textuais.

Torres (2004a) discute o uso do modelo clássico em sistemas de recomendação, afirmando que esses modelos têm suas raízes nos utilizados na Recuperação da Informação décadas atrás. Os sistemas baseados em FBC de modelo clássico fundamentam-se principalmente na técnica TF-IDF, adaptada do Modelo Vetorial e desenvolvida para uso em ferramentas de busca em bibliotecas digitais. Cabe ressaltar que este método é aplicado exclusivamente para medir a similaridade textual entre documentos ou para analisar informações descritivas e avaliativas sobre o conteúdo, não sendo adequado para a análise de

conteúdo de domínios não textuais, como documentos de áudio e imagens (Torres, 2004a; Melville; Sindhwani, 2011).

O modelo aliado a abordagens semânticas, por sua vez, emprega conceitos no lugar de palavras-chave e abrange outras duas técnicas de indexação semântica: a exógena (*top-down*) e a endógena (*bottom-up*). Segundo Gemmis *et al.* (2015), a abordagem de indexação semântica exógena baseia-se na integração com bases externas do conhecimento, como dicionários especializados, taxonomias, tesauros e ontologias. Essa integração possibilita a interpretação da linguagem natural e especializada por máquinas. Por outro lado, os autores explicam que a abordagem endógena tem como alicerce um modelo vetorial ou distributivo, no qual cada palavra e documento são analisados dentro do contexto em que estão inseridos, representando pontos específicos no espaço vetorial, que abrange múltiplos contextos. Para realizar essa análise, esses modelos avaliam grandes Corpora de documentos textuais, buscando interpretar o contexto semântico de uso dos termos.

2.3.3.3 Filtragem Híbrida

Ambas as técnicas de Filtragem Baseada em Conteúdo e Filtragem Colaborativa apresentam vantagens e desvantagens em relação ao seu funcionamento. A respeito delas, Torres (2004a) destaca os seguintes aspectos, conforme o quadro 4 a seguir:

Quadro 4 – Vantagens e desvantagens de FC e FBC

Técnicas de Filtragem	Vantagens	Desvantagens
Filtragem Colaborativa	Independência de conteúdo	Primeiro avaliador
	Uso de qualidade e gosto	Esparsidade
	Serendipity	
Filtragem Baseada em Conteúdo	Não há o problema do primeiro avaliador	Dependência de conteúdo
	Não há esparsidade	Não usa qualidade e gosto
		Superespecialização

Fonte: Torres (2004a, p. 96).

Com base no quadro acima, o autor ressalta que ambas as abordagens apresentam vantagens que se complementam. No contexto da Filtragem Colaborativa, a independência em relação ao conteúdo dos itens é destacada, uma vez que ela se baseia nas avaliações dos usuários para fazer recomendações. Por outro lado, a Filtragem Baseada em Conteúdo depende diretamente do conteúdo textual dos itens para realizar suas recomendações. Uma vantagem adicional da FC é a consideração do aspecto subjetivo das avaliações, o que contribui para determinar a qualidade e o gosto dos produtos avaliados.

Além disso, FC destaca-se pelo uso de *serendipity* (serendipidade), proporcionando recomendações inesperadas. Essa serendipidade pode ser vista como uma solução para a superespecialização de categorias frequentemente associada à FBC, que tende a recomendar itens muito semelhantes ao perfil do usuário, limitando a diversidade de recomendações. A FBC, por outro lado, dissolve as desvantagens associadas à FC ao mitigar problemas relacionados ao primeiro avaliador e à esparsidade, uma vez que qualquer item pode exibir grau de similaridade com algum usuário.

Nesse sentido, em termos de classificação dos sistemas híbridos, Burke (2007) propõe uma classificação para os sete tipos de combinações identificadas entre duas ou mais técnicas de recomendação. O quadro 5 realiza uma breve descrição dos diferentes tipos.

Quadro 5 – Tipos de sistemas híbridos

Tipo	Descrição
Ponderado (*Weighted*)	A pontuação de diferentes componentes de recomendação é combinada numericamente
Alternado (*Switching*)	O sistema escolhe entre os componentes de recomendação e aplica o selecionado.
Misto (*Mixed*)	Recomendações de diferentes recomendadores são apresentadas juntas.
Combinação de Características (*Feature Combination*)	Recursos derivados de diferentes fontes de conhecimento são combinados e fornecidos a um único algoritmo de recomendação.
Aumento de Características (*Feature Augmentation*)	Uma técnica de recomendação é usada para calcular uma característica ou conjunto de características, que é então parte da entrada para a próxima técnica.
Cascata (*Cascade*)	Recomendadores têm prioridade estrita, com os de menor prioridade desempatando nas pontuações de maior prioridade.
Meta-nível (*Meta-level*)	Uma técnica de recomendação é aplicada e produz algum tipo de modelo, que é então a entrada usada pela próxima técnica.

Fonte: Burke (2007, p. 380, tradução nossa).

Diversas combinações entre diferentes técnicas de recomendações foram propostas. Inicialmente para unir as técnicas de FC e FBC em um único sistema de recomendação, mas também podendo unir outras técnicas como Filtragem Baseada em Conhecimento, Filtragem Demográfica, Filtragem Social, etc. Sistemas híbridos tendem a aumentar a cobertura do sistema, gerando um percentual maior de recomendações úteis (Torres, 2004a).

No entanto, sistemas híbridos não resolvem um problema comum em sistemas de recomendação: o *Start-Up* (início). De acordo com Torres (2004a), essa situação ocorre devido ao tempo necessário para que um sistema crie uma base de dados sobre os perfis de consumo e hábitos de seus usuários. Soluções comuns para combater este problema envolvem o uso de recomendações de produtos mais vendidos ou de sugestões com base na extração explícita de dados e perfis de consumo. Além disso, diversas pesquisas em sistemas de recomendação têm buscado soluções para este problema.

3 PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa caracteriza-se pela sua natureza aplicada e abordagem qualitativa. Segundo Silveira e Córdova (2009, p. 35), a pesquisa de natureza aplicada "objetiva gerar conhecimentos para aplicação prática, dirigidos à solução de problemas específicos", indo ao encontro dos objetivos tencionados por esta pesquisa em propor um modelo conceitual de recomendação de publicações científicas.

No que concerne à abordagem utilizada, a abordagem qualitativa demonstrou-se a mais propícia, uma vez que "enquanto exercício de pesquisa, não se apresenta como uma proposta rigidamente estruturada, ela permite que a imaginação e a criatividade levem os investigadores a propor trabalhos que explorem novos enfoques" (Godoy, 1995, p. 21), favorecendo o uso de métodos qualitativos na análise dos modelos de recomendação abordados na literatura.

Quanto ao objetivo da pesquisa, destaca-se como pesquisa metodológica. A pesquisa metodológica destaca-se como um processo sistemático de investigação, empregando métodos específicos e rigorosos para a coleta, análise e interpretação dos dados, conforme descrito por Melo *et al*. (2017). Nesse contexto, buscou-se por meio de um estudo metodológico, a elaboração e validação de um modelo conceitual de recomendação.

Em relação aos procedimentos adotados para a pesquisa, esta foi delineada a partir da pesquisa bibliográfica. A pesquisa bibliográfica, segundo Gil (2002), ocorre exclusivamente a partir de fontes bibliográficas, como livros e publicações periódicas, que oferecem um tratamento analítico prévio. Sua utilização na pesquisa apresenta como principais vantagens a possibilidade de atingir uma variedade de dados e informações não disponíveis ao investigador por meio de outros procedimentos.

Com respeito ao objeto de estudo delimitado por esta pesquisa, foram englobados tanto os requisitos necessários para propor um modelo conceitual de recomendação na Brapci quanto os modelos de recomendação da literatura analisada. Dessa forma, a investigação concentrou-se, não apenas na análise das características essenciais para a concepção de um modelo conceitual em sistemas de recomendação de publicações científicas, mas também na revisão aprofundada dos diferentes modelos disponíveis na literatura especializada.

No tocante às etapas metodológicas deste trabalho, serão discutidos a seguir os procedimentos adotados nesta investigação. A fase inicial, associada ao primeiro objetivo

específico da pesquisa, foi realizada por meio do levantamento bibliográfico nas bases de dados Brapci e OasisBR, com o propósito de identificar os sistemas, técnicas e algoritmos de recomendação utilizados na filtragem da informação.

As consultas nas bases de dados incluíram os termos: sistemas de recomendação e modelos de recomendação, considerando variações dos termos no plural e suas respectivas traduções para o idioma inglês e espanhol. Dessa forma, a estratégia de busca foi montada utilizando os seguintes termos: “sistemas de recomendação”, "sistema de recomendação", "recommender system", "recommender systems", "recommendation system”, "recommendation systems", "sistema de recomendación", "sistemas de recomendación", “modelos de recomendação”, "modelo de recomendação", "recommender model", "recommender models", "recommendation model", "recommendation models", "modelo de recomendación" e "modelos de recomendación".

Além disso, para cada uma das bases, foram adotados procedimentos específicos. Na Brapci todos os documentos disponíveis foram recuperados. Na OasisBR, devido à margem de documentos recuperados pela estratégia de busca, os resultados foram filtrados utilizando o campo de assunto na busca avançada, com o objetivo de obter uma análise mais precisa.

Referente à coleta dos dados desta pesquisa, após a recuperação dos documentos nas bases, foram exportados os metadados e textos em PDF através do *software* Zotero Web para a biblioteca *online* pessoal do investigador. O Zotero é um gerenciador de referências de *software* livre que possibilita, entre outras funções, a coleta, armazenagem, citação e compartilhamento de documentos físicos ou digitais.

Na análise dos dados, em consonância com a abordagem de pesquisa metodológica, optou-se pelo uso de técnicas qualitativas, sendo, portanto, utilizada a análise de conteúdo. Acerca desta técnica, consoante Moraes (1999, p. 1),

> a análise de conteúdo constitui uma metodologia de pesquisa usada para descrever e interpretar o conteúdo de toda classe de documentos e textos. Essa análise, conduzindo a descrições sistemáticas, qualitativas ou quantitativas, ajuda a reinterpretar as mensagens e a atingir uma compreensão de seus significados num nível que vai além de uma leitura comum.

Nesse sentido, por meio da análise de conteúdo, a etapa de avaliação dos modelos de recomendação teve como objetivo descrever e interpretar os modelos selecionados com base no quadro teórico desta pesquisa, bem como identificar os elementos necessários para a construção de um modelo conceitual de recomendação, adequado ao contexto da Brapci. Para isso, foram considerados critérios de recomendação baseados em acoplamento,

palavras-chave e autores, de modo a garantir que o modelo proposto atendesse às necessidades específicas de recuperação da informação e recomendação de publicações científicas na base de dados.

A terceira etapa metodológica envolveu a discussão e sistematização dos aspectos relacionados à construção do sistema, fundamentada na taxonomia para SR de Schafer (2001) apresentada no referencial teórico da pesquisa. Nessa etapa, foram abordados aspectos como a extração de informações do usuário, o método de saída das recomendações, o grau de personalização e as técnicas de filtragem da informação. Também foram considerados os aspectos relacionados ao modo de computação e ao método de recomendação empregado.

Na quarta etapa, foi proposto o modelo conceitual para recomendação. Assim, as decisões acerca dos aspectos que embasaram a construção do modelo de recomendação resultaram na criação de três estratégias de recomendação, que incluem a recomendação de itens relacionados, itens citados e itens personalizados. Essas estratégias foram detalhadas nos resultados da pesquisa.

Por fim, a proposta de visualização das recomendações considerou o modelo conceitual de recomendação como base para a criação da interface visual, empregando ferramentas do modo desenvolvedor do navegador e a inspeção dos elementos em *Hypertext Markup Language* (HTML) e *Cascading Style Sheets* (CSS) da Brapci.

4 RESULTADOS

Neste capítulo, são apresentados os resultados obtidos a partir da coleta e análise de dados deste estudo. Os resultados estão organizados de acordo com a sequência dos objetivos específicos e amparados pela metodologia estabelecida nas seções anteriores, visando fornecer uma compreensão clara e manter a uniformidade dos mesmos no contexto do processo de investigação.

4.1 LEVANTAMENTO BIBLIOGRÁFICO

No intuito de alcançar o proposto pelo primeiro objetivo específico e identificar modelos de recomendação, técnicas e estratégias aplicadas à filtragem de informação na literatura, realizou-se o levantamento bibliográfico por meio das bases de dados Brapci e OasisBR. No que tange aos procedimentos de busca, as consultas iniciais às bases de dados foram realizadas entre o período de 27 de março e 03 de abril de 2024.

A análise dos documentos considerados pertinentes ao estudo foi conduzida através da avaliação dos títulos, resumos e palavras-chave dos documentos recuperados. Em alguns casos, procede-se com a análise do sumário e listas de figuras e tabelas. Os resultados das consultas em cada uma das bases e o total de documentos relevantes encontrados são ilustrados na Tabela 1. Durante a pesquisa, também se buscou a dissertação de Torres (2004b), que não havia sido recuperada pela estratégia de busca utilizada em ambas as bases, sendo posteriormente incluída na análise deste trabalho.

Tabela 1 – Resultados das buscas em bases de dados

Bases de dados	Resultados	Documentos Relevantes
Brapci	39	9
OasisBR	612	37

Fonte: Elaborado pelo autor a partir dos dados da pesquisa (2024).

Entre o total de documentos relevantes, foram selecionados dezessete trabalhos, incluindo duas teses, oito dissertações, dois trabalhos de conclusão de curso e seis artigos. A data de publicação desses trabalhos varia entre os anos de 2004 e 2023, abrangendo textos nos três idiomas pesquisados: português, inglês e espanhol.

A seleção parcial dos documentos de interesse foi inicialmente baseada nas similaridades entre certas pesquisas, que incluem o uso ou estudo dos mesmos modelos, técnicas e algoritmos de recomendação em diferentes cenários. Também foram evitados os trabalhos acadêmicos que pouco exploram a arquitetura dos sistemas de recomendação utilizados ou que se distanciam do esperado pelo modelo de recomendação. Assim, o objetivo não foi encontrar o maior número possível de pesquisas relevantes, mas sim identificar aquelas consideradas mais pertinentes à proposição do modelo.

Sob essa perspectiva, o quadro 6 resume características relevantes nos sistemas de recomendação analisados, proporcionando uma compreensão aprofundada desses modelos. O quadro inclui informações sobre os autores, enfoques e cenários de recomendação, abordagens ou técnicas de FI, classificações dos métodos utilizados, o grau de personalização de acordo com a taxonomia de Schafer (2001), e as técnicas de recomendação empregadas. Ressalta-se, contudo, que um dos textos (Peis; Morales-Del-Castillo; Delgado-Lopez, 2008), uma revisão bibliográfica, foi excluído do quadro por não apresentar um modelo em específico, mas uma classificação para os modelos semânticos e diversos modelos propostos por autores com trabalhos relacionados.

Os parágrafos a seguir dedicam-se ao resumo e à descrição dos sistemas de recomendação selecionados com base na leitura dos dezessete textos de interesse.

Torres (2004b) desenvolve uma série de algoritmos em um sistema de recomendação para artigos científicos nomeado TechLens+. Os testes no sistema utilizam múltiplas abordagens, como Filtragem Colaborativa, Filtragem Baseada em Conteúdo e Filtragem Híbrida a fim de avaliar a acurácia de cada algoritmo.

Dentre os algoritmos baseados em Filtragem Colaborativa, dois se destacam: o algoritmo FC-puro, um método tradicional de K-NN que utiliza uma matriz formada por artigos e suas citações para gerar recomendações de trabalhos semelhantes com base nas citações em comum; e o FC-denso, que aborda o problema da esparsidade no método anterior ao incorporar citações indiretas, considerando não apenas as citações iniciais, mas também as citações feitas pelos artigos citados. No contexto da Filtragem Baseada em Conteúdo FBC, são descritos os algoritmos FBC-puro, junto com suas extensões FBC-separado e FBC-combinado. Estes dois métodos não apenas exploram o texto do artigo ativo, mas também analisam o texto dos artigos citados.

Quadro 6 – Resumo dos sistemas de recomendação analisados

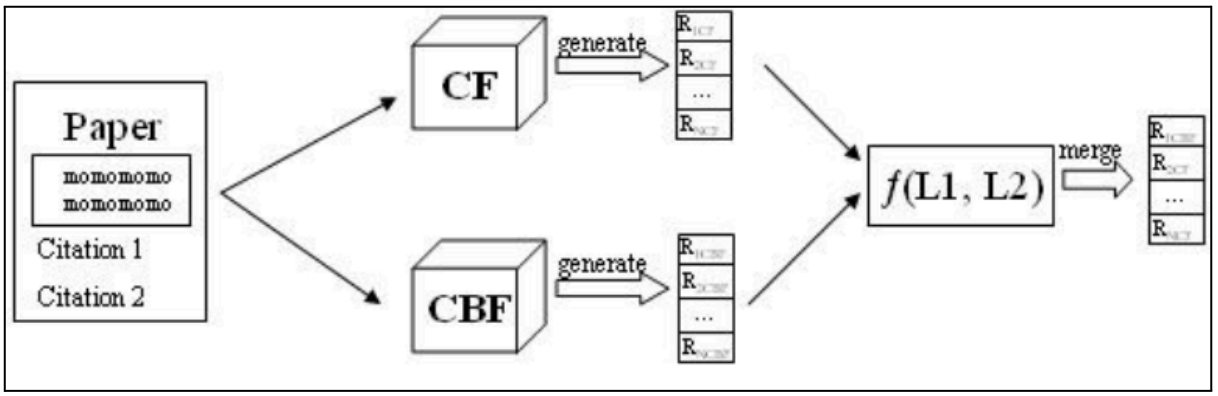
Trabalhos analisados	Enfoque de recomendação*	Abordagem de recomendação	Classificação do método	Grau de Personalização	Técnicas utilizadas
Torres (2004b)	Publicações científicas	Híbrido	Aumento de características/ Misto	Efêmero	K-NN e TF-IDF
Lopes (2007)	Publicações científicas	FBC	Vetorial	Efêmero	TF-IDF
Aleixo (2014)	Filmes	FC	Memória	Personalizado	Item K-NN
Nóbrega (2014)	Filmes, produtos e relacionamentos	FC	Modelo	Personalizado	Fatoração de Matrizes
Oliveira (2014)	Filmes	Híbrido	Aumento de características	Personalizado	Clusterização e Mineração de dados
Franco, Sanchez e Serna (2015)	Ambientes organizacionais	Semântico	SRSSC	Personalizado	–
Silva, Schreiber e Nara (2015)	Artigos de notícias	FC	Modelo	Personalizado	Clusterização e Redes Bayesianas
Conceição e*t al.* (2016)	Vídeos	FBC	Vetorial	Efêmero	*Sparse Linear Method with Side Information* (SSLIM), TF-IDF e BOW
Furtado (2016)	Biblioteca Universitária	Híbrido	Vetorial	Efêmero	BOW, TF-IDF
Monteiro-Krebs, Rocha e Ribeiro (2017)	Biblioteca Universitária	FC e FBC	Memória / Vetorial	Efêmero / Personalizado	K-NN, Extração de metadados no sistema

Trabalhos analisados	Enfoque de recomendação*	Abordagem de recomendação	Classificação do método	Grau de Personalização	Técnicas utilizadas
Mourão (2018)	Filmes / música	Híbrido	Misto	Personalizado	Fatoração de matrizes, distribuição normal multivariada
Costa (2020)	Música	FC	Redes Neurais	Personalizado	Redes Neurais Recorrentes (GRU) e Convolucionais (CNN)
Cunha (2021)	*E-commerce*	Híbrido	Modelo	Não personalizado/ efêmero / personalizado	Estatísticas de compras, fatoração de matrizes, regras de associação, Clusterização, TF-IDF
Neves (2022)	Artigos de notícias e teses	FBC	Redes Neurais	Efêmero	*Word embeddings*, *Sentence embedding*, K-NN
Souza e Feitosa (2022)	Publicações científicas	FBC	Modelo	Personalizado	TF-IDF, Clusterização
Vieira, Passos e Salm (2023)	Conceitual	Híbrido	Cascata	Personalizado	–

Nota: o enfoque refere-se à aplicação utilizada durante os resultados de cada trabalho. SR baseados em FBC normalmente podem ser aplicados em qualquer domínio rico textualmente, enquanto SR baseados em FC podem ser usados em contextos em que haja dados suficientes sobre o histórico de usuários e itens.
Fonte: Elaborado pelo autor a partir dos dados da pesquisa (2024).

Todas as três técnicas utilizam a abordagem TF-IDF para calcular a relevância e similaridade entre os artigos. Em seguida, são descritos os cinco algoritmos híbridos criados, que usam as estratégias de hibridização de aumento de características e fusão (ou misto segundo a taxonomia de Burke), combinando as técnicas de FC e FBC recém mencionadas. A figura 3 ilustra a arquitetura do algoritmo Híbrido de Fusão.

Figura 3 – Algoritmo Híbrido de Fusão TechLens+



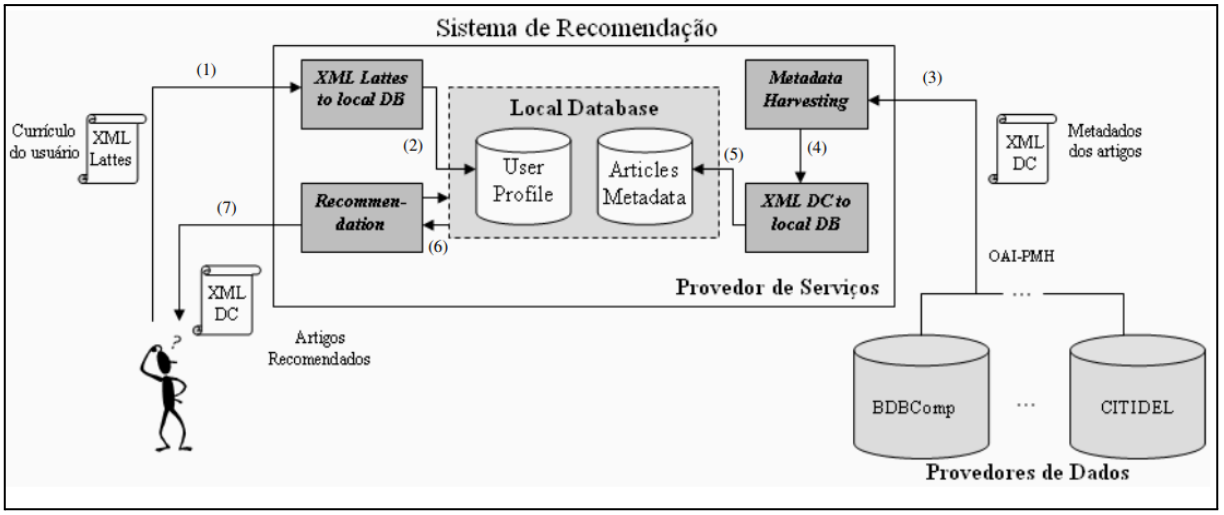
Fonte: Torres (2004b, p. 33).

Os algoritmos são avaliados em duas etapas. Na etapa *offline*, Torres avalia o percentual de acertos e a posição das citações faltantes entre as recomendações geradas. Nesta etapa os algoritmos melhor avaliados foram os algoritmos não híbridos FBC-Separado, FC-puro e os híbridos FC -FBC separado, FBC combinado -FC e Fusão, especialmente os algoritmos de Fusão e FC-puro. Já na etapa *online*, foram testados apenas estes cinco algoritmos utilizando o sistema TechLens+ com a participação de 110 usuários, entre os quais encontram-se alunos de pós-graduação, professores e pesquisadores do Brasil, Estados Unidos e outras regiões. Os resultados deste último teste mostram que os algoritmos foram aceitos pela maioria dos usuários, especialmente entre os alunos de mestrado e doutorado. Além disso, o desempenho dos algoritmos também varia de acordo com o tipo de artigo (novo, autoritativo, introdutório ou *Survey*).

Seguindo o enfoque de recomendação para artigos científicos, Lopes (2007) apresenta um sistema de recomendação com Filtragem Baseada em Conteúdo para bibliotecas digitais sob a perspectiva da Web Semântica (figura 4). Neste estudo, a autora combina a coleta de metadados de artigos de bibliotecas digitais com a extração de dados dos usuários provenientes de seus currículos na plataforma Currículo Lattes, visando fornecer recomendações alinhadas ao perfil de cada usuário. Para alcançar esse objetivo, o sistema

utiliza tecnologias da Web Semântica, como a *Extensible Markup Language*[5] (XML) e o esquema de metadados *Dublin Core*[6] (DC).

Figura 4 – Arquitetura do SR baseado em perfis do Currículo Lattes



Fonte: Lopes (2007, p. 37).

A formação do perfil destes usuário considera variáveis diversas, entre as quais se destacam o nome, ano de graduação, título da monografia, palavras-chave, área e orientador (referentes à monografia), além de informações sobre proficiência em línguas estrangeiras e publicações científicas, incluindo títulos, palavras-chave, língua e ano (referentes às publicações científicas). Para isso, o sistema extrai os metadados das variáveis mencionadas em formato XML, seguindo os padrões recomendados pela Comunidade para Ontologias em Ciência, Tecnologia e Informações de Aperfeiçoamento de Nível Superior, que são mapeados em uma base do sistema intitulada User Profile. Após a coleta das informações dos usuários, os metadados dos artigos são coletados nas bibliotecas digitais por meio do protocolo OAI-PMH. Em seguida, utiliza-se a técnica TF-IDF para criar um índice invertido com a frequência dos termos em cada artigo a ser armazenado na base de dados Articles Metadata.

Para ambos os conjuntos de dados, é atribuído um peso final que será utilizado para calcular a similaridade entre o perfil do usuário e os documentos recomendados. As

[5] XML é uma linguagem de representação de dados na Web criada pelo W3C, que possibilita a criação de marcas hierárquicas para representar e qualificar a informação, tornando os conteúdos e significados de documentos e outras informações compreensíveis tanto para pessoas quanto para computadores (Bax, 2001).

[6] *Dublin Core* é um esquema de metadados para descrição padronizada de recursos na Web criado em 1995. Atualmente o DC conta com um vocabulário expansível formado por quinze elementos principais (Dublin Core Metadata Initiative, c2024).

recomendações são então organizadas em ordem decrescente de acordo com o grau de similaridade entre o usuário e suas recomendações.

Em Peis, Morales-Del-Castillo e Delgado-Lopez (2008), os autores realizam uma revisão da literatura acerca dos sistemas de recomendação semânticos (SRS). No artigo, são elencados alguns critérios para a classificação dos SR, englobando a escolha entre sistemas centralizados ou baseados em redes "Ponto a Ponto" (P2P)[7], técnicas de filtragem ativa e passiva, coletas de informações explícita e implícita, entre outros aspectos relevantes ao optar por um sistema de recomendação.

Os autores definem os SRS como sistemas que operam a partir de bases de conhecimento, como ontologias e esquemas conceituais (taxonomias e tesauros, por exemplo), e que são auxiliadas pelo uso de tecnologias da Web Semântica (Peis; Morales-Del-Castillo; Delgado-Lopez, 2008). Assim, os SRS podem ser classificados em duas principais vertentes, que incluem: os SRS baseados em ontologias ou esquemas conceituais, e aqueles que adicionam filtros de informação, sendo estes os sistemas adaptáveis ao contexto ou sistemas baseados em redes de confiança, também conhecidos como Context-Aware Recommender Systems, ou Sistemas de Recomendação Sensíveis ao Contexto (SRSSC).

Segundo o artigo em questão, diversos trabalhos na área utilizam ontologias ou esquemas conceituais para suas aplicações. Middleton *et al.* (2002) apresentam um sistema de recomendação de artigos científicos integrado a uma ontologia contendo informações extraídas automaticamente de diversos bancos de dados disponíveis na Web. Jung *et al.* (2005) propõem um SRS a partir do uso de triplas *Resource Description Framework*[8] (RDF) para representar os serviços web e os perfis de usuários. Nesse sistema, os serviços web oferecidos por cada empresa são armazenados em documentos de triplas RDF e utilizados na recuperação da informação com base nos perfis de cada usuário. Díaz-Avilés (2005) descreve um sistema de recomendação (SR) baseado no modelo de vizinhança e em tecnologias da Web Semântica, fundamentado em uma rede P2P descentralizada. Dessa forma, o sistema mantém uma arquitetura distribuída, em que cada componente da rede é responsável por

[7] P2P, ou Peer-to-Peer, é uma arquitetura de rede onde todos os computadores possuem o mesmo nível hierárquico e podem compartilhar e transmitir dados diretamente entre si, sem a necessidade de administradores ou servidores definidos (Casad; Willsey, 1999).

[8] RDF é um modelo de representação semântica recomendado pelo W3C para descrição de recursos na Web, que pode ser aplicado em diversas áreas para garantir a interoperabilidade entre aplicações Web. O RDF baseia-se no modelo de triplas: sujeito, predicado e objeto, que permite tanto a representação dos objetos como de seus relacionamentos em formato de grafos (Lassila; Swick, 1999).

gerenciar parte dos dados e executar localmente o método de recomendação (Peis; Morales-Del-Castillo; Delgado-Lopez, 2008).

Além destes, entre os sistemas adaptativos ao contexto ou baseados em redes de confiança, Szomszor *et al.* (2007) desenvolveram um algoritmo de recomendação para filmes, integrado a uma base de conhecimento semântica estruturada em torno de uma folksonomia de avaliações de filmes, com o objetivo de armazenar informações sobre a descrição e categorização dos itens, bem como representar os interesses e opiniões dos usuários.

Massa e Avesani (2004), por sua vez, sugerem um modelo híbrido apoiado no conceito da "*Web of Trust*" e FC. Nesse sistema, é utilizada uma matriz de similaridade entre usuários e uma "matriz de confiança", composta por declarações autênticas dos usuários, para aprimorar a confiabilidade das recomendações por meio de previsões mais precisas (Peis; Morales-Del-Castillo; Delgado-Lopez, 2008). Golbek (2005), Kruk e Decker (2005) exploram o uso de ontologias baseadas no vocabulário *Friend of a Friend*[9] (FOAF) em sistemas de recomendação para redes sociais, permitindo que os usuários gerenciem seus perfis por meio da interação com perfis de amigos. Segundo os autores, técnicas de Filtragem Colaborativa Social Semântica melhoram a acurácia das recomendações em comparação com a abordagem de Filtragem Colaborativa tradicional.

Há ainda os sistemas adaptativos ao contexto, que visam analisar diversos fatores ligados ao contexto de utilização e adaptar as recomendações para atender às necessidades imediatas de usuários (Peis; Morales-Del-Castillo; Delgado-Lopez, 2008). Um exemplo é o SRS introduzido por Loizou e Dasmahapatra (2006), que utiliza informações contextuais armazenadas em uma ontologia para analisar tanto os itens recomendados quanto o processo de recomendação.

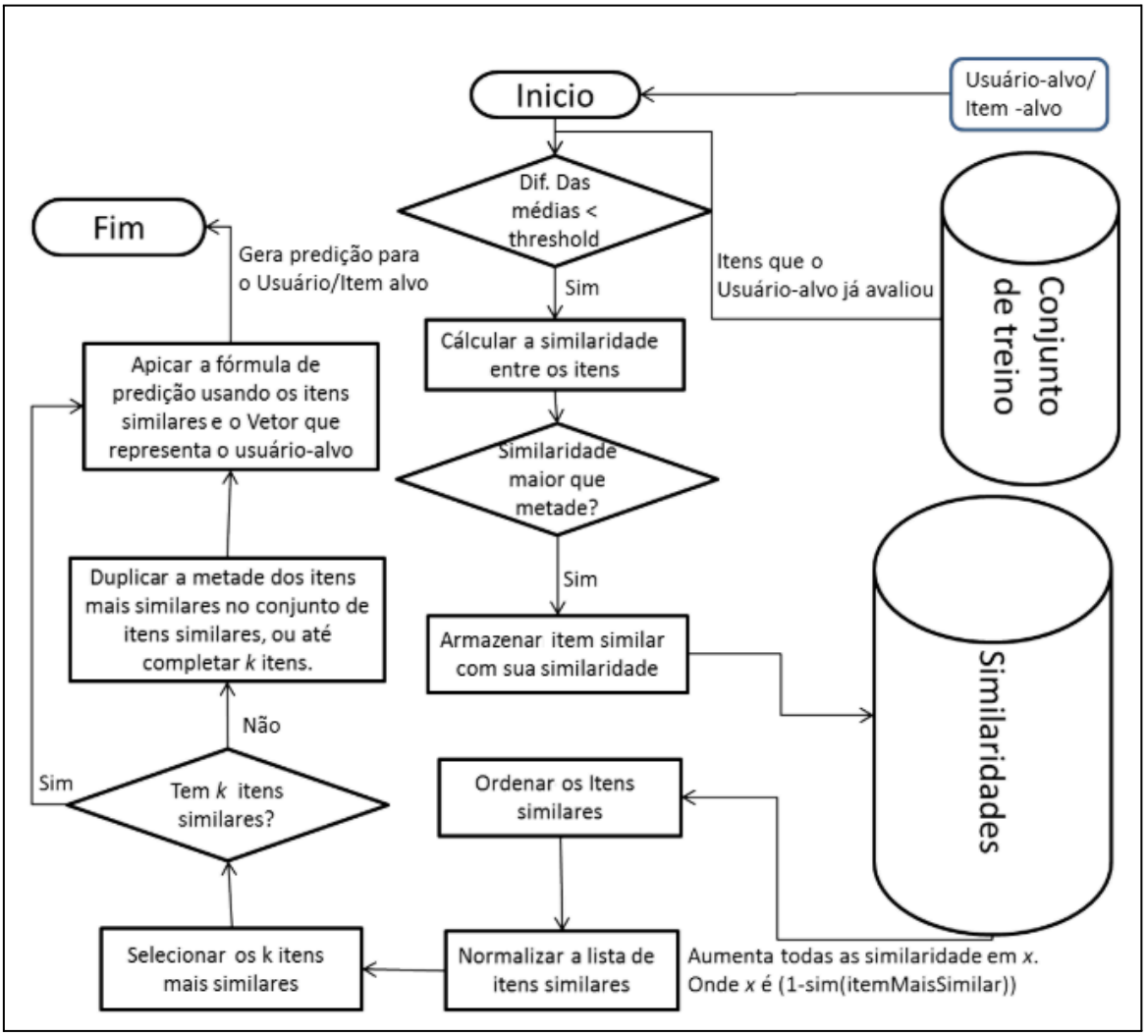
Aleixo (2014), por sua vez, propõe quatro modificações ao método de recomendação de Filtragem Colaborativa baseado em memória de itens em um algoritmo denominado Item-Based-ADP. Nesta dissertação, o autor discorre a respeito das quatro adaptações realizadas no algoritmo, com o objetivo de melhorar a acurácia e o tempo de resposta das recomendações.

A primeira adaptação consiste na introdução de uma nova etapa que reduz o número de itens no conjunto de vizinhos potenciais. Essa etapa seleciona apenas os pares de itens cuja diferença entre suas médias seja inferior ou igual a um limite δ (delta), diminuindo assim o

[9] FOAF é um vocabulário de ontologias na Web descrito em RDF focado na descrição de pessoas, objetos e suas relações no contexto de redes sociais (Wikipedia contributors, 2023).

tempo de processamento do algoritmo. A segunda etapa, por outro lado, está relacionada com a acurácia do sistema de recomendação. Dessa forma, apenas os vizinhos que estejam acima de um limite mínimo de cinquenta por cento do intervalo permitido pela medida de similaridade (como Cosseno, Coeficiente de Correlação de Pearson e Log-likelihood[10]) são aceitos pelo conjunto de vizinhos do item-alvo. As etapas seguintes, portanto, visam normalizar os valores de similaridade encontrados, forçando o sistema a considerar vizinhos que contenham baixo grau de similaridade, e, por fim, possibilitar a repetição dos valores de similaridade quando o número de K-vizinhos não for atingido. A figura 5 demonstra em detalhes o fluxograma de funcionamento do algoritmo elaborado pelo autor.

Figura 5 – Fluxograma do algoritmo Item-Based-ADP



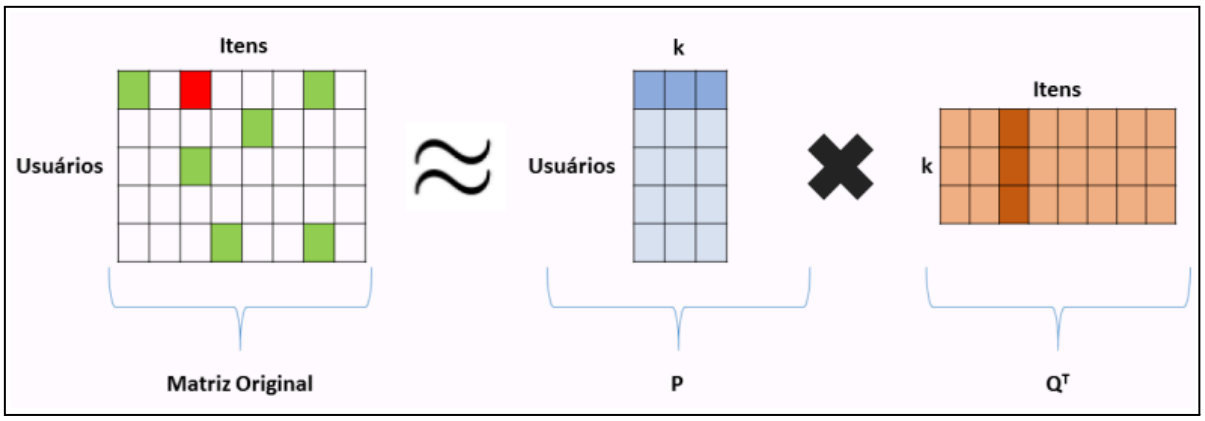
Fonte: Aleixo (2014, p. 55).

[10] *Log-likelihood* é uma medida de similaridade que considera as ações de usuários ao invés das avaliações. Dessa forma, são calculadas, respectivamente, as relações entre os itens avaliados por ambos usuários, por um único usuário e dos itens não avaliados por ambos (Aleixo, 2014).

Além de descrever o algoritmo modificado, Aleixo ainda conduz uma série de avaliações deste algoritmo com as versões tradicionais de FC baseada em memória de itens e usuários, medindo o tempo de execução dos algoritmos e a acurácia a partir de métricas como o Erro Médio Absoluto (MAE) e a Raiz do Erro Quadrático Médio (RMSE). Estas avaliações foram conduzidas utilizando os conjuntos de dados da MovieLens 100K[11], contendo uma matriz com mais de 100 mil avaliações, e um subconjunto de dados da Netflix Prize[12] com um total de 1000 filmes em uma matriz com mais de 50 mil avaliações.

Em outro dos trabalhos revisados nesta pesquisa, Nóbrega (2014) realizou um estudo sobre sistemas de recomendação que utilizam a técnica de fatoração de matrizes, com o objetivo de melhorar esses algoritmos ao reduzir o número de iterações iniciais necessárias para o funcionamento do sistema. A fatoração de matrizes é amplamente empregada em sistemas de Filtragem Colaborativa baseados em modelos. Semelhante a outras abordagens classificadas dessa maneira, a fatoração de matrizes requer o treinamento do sistema para desenvolver novos modelos preditivos. Em essência, esse método reduz a dimensionalidade de uma matriz original, como avaliações em uma matriz de usuários e itens, em matrizes de fatores latentes (figura 6). Estas matrizes buscam representar as dimensões da matriz original com matrizes de usuários k-dimensionais e matrizes de itens k-dimensionais.

Figura 6 – Ilustração da Fatoração de Matrizes como solução para completar valores ausentes na matriz original



Fonte: Nóbrega (2014, p. 16).

O uso dessa técnica resulta em uma significativa redução no custo computacional, diminuindo o tempo necessário para treinar os modelos preditivos. Além disso,

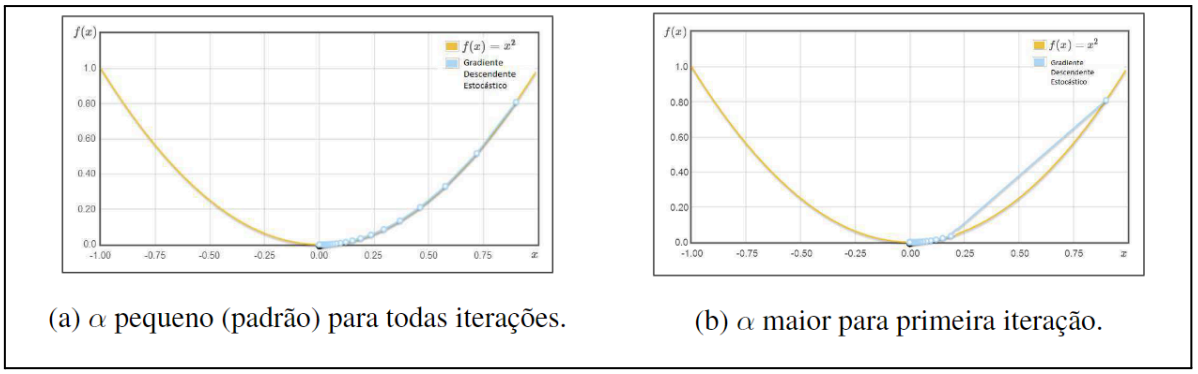
[11] Disponível em: https://grouplens.org/datasets/movielens/100k/.

[12] O conjunto de dados utilizado na competição Netflix Prize, iniciada em 2006, é composto por mais de 100 milhões de avaliações feitas pelos usuários do serviço de *streaming*. (Wikipedia Contributors, 2024).

frequentemente possibilita recomendações baseadas na similaridade entre os novos vetores gerados. Contudo, para obter as matrizes de fatores latentes, é crucial determinar o valor ideal para a taxa de aprendizagem do algoritmo por meio de uma função de perda. Portanto, são empregadas técnicas de otimização como o gradiente descendente estocástico para convergir para um mínimo local (ponto de convergência). Nesse método, é necessário ajustar hiperparâmetros, incluindo a taxa de aprendizagem.

Diante deste contexto, Nóbrega apresenta uma estratégia para minimizar a taxa de aprendizagem do gradiente descendente estocástico, reduzindo o número de iterações do algoritmo em até 40% em uma das oito bases de dados utilizadas na pesquisa, em comparação com o algoritmo padrão de fatoração de matrizes. Essa estratégia valida a hipótese de que há uma relação entre o valor inicial da taxa de aprendizagem do gradiente descendente estocástico e o número mínimo de iterações necessárias para que o sistema atinja a convergência. A figura 7 ilustra esse processo, comparando o uso de um valor fixo padrão para a taxa de aprendizagem α (alfa) no gráfico A com a estratégia utilizada pelo autor no gráfico B.

Figura 7 – Ilustração do gradiente descendente estocástico usando valores de α igual a (a) 0,1 e (b) 0,01



(a) α pequeno (padrão) para todas iterações.

(b) α maior para primeira iteração.

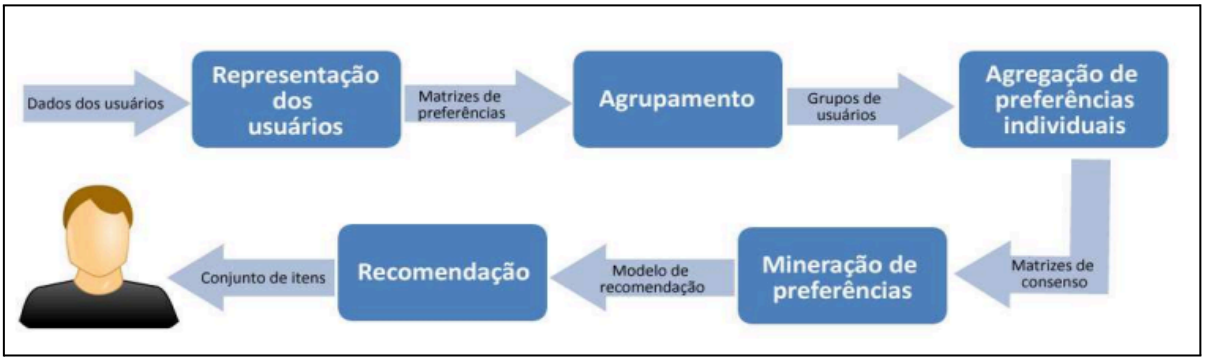
Fonte: Nóbrega (2014, p. 6).

Em outro estudo, Oliveira (2014) apresenta uma metodologia para o sistema de recomendação híbrido PrefRec, que emprega as técnicas de FC e FBC. Segundo a autora, essa metodologia é dividida em cinco etapas, que vão desde a representação dos usuários até o momento de recomendação dos itens ao usuário-alvo, conforme ilustrado na figura 8.

Na primeira etapa, os dados dos usuários presentes nas bases de dados são representados por meio de uma matriz de preferências. Ao contrário da abordagem convencional de representar avaliações em vetores, frequentemente usada na Filtragem

Colaborativa, as matrizes de preferências organizam as avaliações de um usuário-alvo sobre itens com base em suas preferências de um item em relação a outro. Na segunda etapa, estes usuários são organizados em grupos por meio da técnica de clusterização, ou agrupamento. De acordo com Han, Kamber e Pei (2011), a clusterização é uma técnica pré-processamento e redução de dados que tem como objetivo classificar objetos em grupos (*clusters*) com alta similaridade entre os itens do mesmo grupo e baixa similaridade com itens de outros grupos. Dessa forma, a técnica contribui para a otimização do processamento computacional necessário ao lidar com grandes volumes de dados.

Figura 8 – Arquitetura geral da Metodologia PrefRec



Fonte: Oliveira (2014, p. 54).

Diversos algoritmos são empregados no uso da técnica de agrupamento, com diferentes especificações. Entre os algoritmos citados pela autora estão o K-Means, que está entre os mais utilizados pela técnica de clusterização, o DBScan e o Cure. Devido a metodologia utilizada, o algoritmo escolhido a partir dos testes realizados pela autora foi o DBScan. Diferente dos outros dois algoritmos elencados, o DBScan não requer o número de *clusters* como parâmetro, mas o valor do raio e a quantidade mínima de pontos no raio de cada grupo.

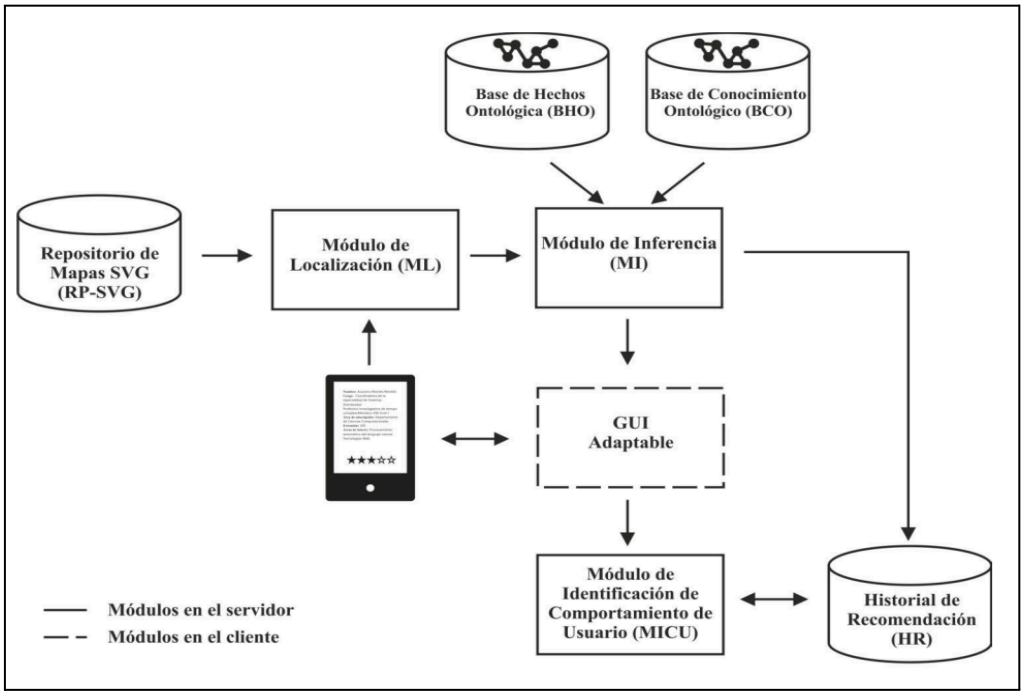
Na terceira etapa são construídos os perfis de preferência de cada grupo por meio da média aritmética, agregando as características comuns e individuais de cada usuário em um conjunto de dados que represente as preferências de cada *cluster*. Na etapa seguinte, os perfis de preferência recém-criados são aplicados no algoritmo de mineração de preferências CprefMiner, gerando um modelo de recomendações de itens a ser utilizado de acordo com o *cluster* de cada perfil formado na etapa anterior.

Todas as quatro etapas são efetuadas em modo *offline*, de maneira assíncrona, durante intervalos de menor tráfego no sistema. Dessa forma, apenas a última etapa é

executada *online*, de maneira síncrona. Nesta fase, o sistema é responsável por calcular a matriz de preferências do usuário-alvo com base em seu histórico de avaliações, associá-lo ao perfil de preferências de maior similaridade e recomendar itens não avaliados que fazem parte do modelo de recomendações do perfil alvo.

Em artigo da revista cubana *Ciencias de la Información*, Franco, Sanchez e Serna (2015) apresentam o Sistema de Recomendação Semântico Sensível ao Contexto FIND-IT! para ambientes organizacionais. O artigo descreve uma ampla metodologia que abrange todos os módulos utilizados pelo modelo (figura 9). Especificamente, ele apresenta conceitos para a formação de um modelo a partir do uso de duas ontologias relacionadas. A primeira é uma base de fontes para criar as redes internas e estruturas da ontologia, formada por ontologias menores que podem ou não utilizar a importação de fontes externas, como repositórios. A segunda ontologia é específica para o armazenamento de regras a serem aplicadas por um módulo de inferência, que gera recomendações aplicando as regras da segunda ontologia sobre a primeira.

Figura 9 – Arquitetura do SRSSC FIND-IT!



Fonte: Franco, Sanchez e Serna (2015, p. 12).

Além disso, o sistema conta com uma interface gráfica de usuário (GUI) adaptável para os grupos predefinidos de estudantes, empresários e professores, variando em conteúdo,

funções e desenho. Em relação ao conteúdo, o sistema visa recomendar itens de acordo com a categorização de nível do usuário.

No que concerne ao módulo de inferência, responsável por gerar as recomendações aos usuários, há poucas informações presentes no artigo a respeito dos algoritmos ou técnicas empregados no sistema. Além da divisão do módulo em etapas de pré-filtragem de itens e geração das recomendações devido a demanda de processamento ao utilizar ontologias, não há muitos detalhes. Dessa forma, foram descritas apenas as características referentes aos módulos mais relevantes para esta pesquisa, como o uso de bases ontológicas e a interface adaptável aos grupos de usuários, tendo em vista que os SRSSC dispõem de mais variáveis (ou vetores) relacionadas ao histórico de itens e usuários para gerar recomendações.

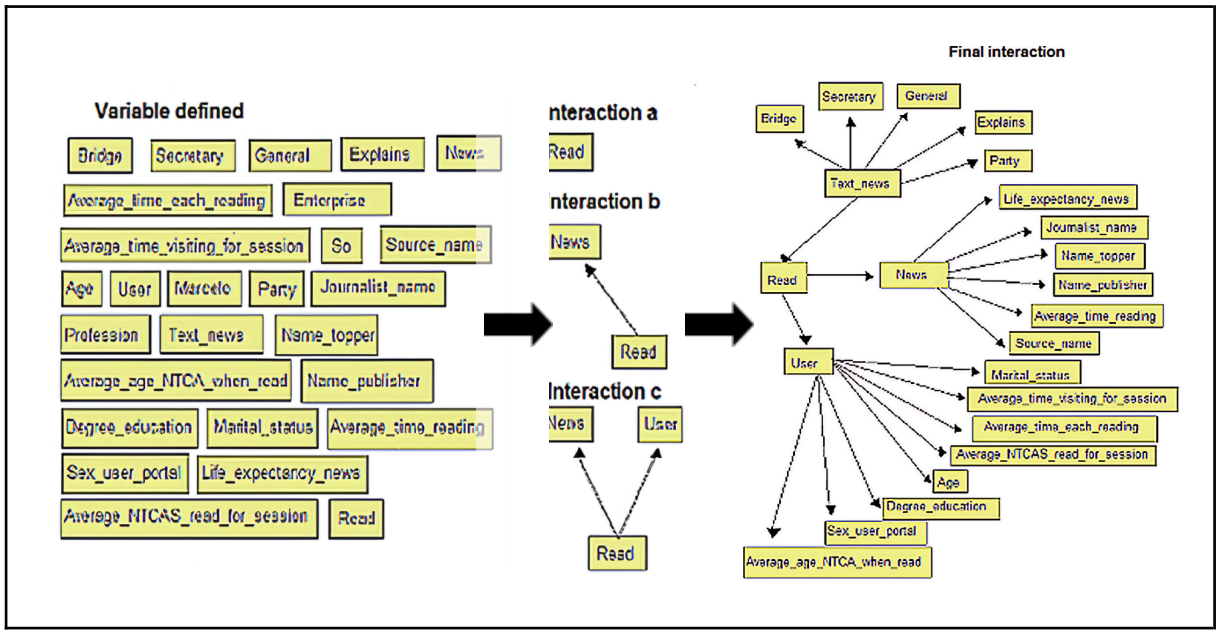
Outro método de recomendação foi revisado em Silva, Schreiber e Nara (2015). Neste artigo, os autores desenvolvem um sistema de recomendação probabilístico baseado em técnicas de redes bayesianas e *clustering* para um jornal *online*. No âmbito dos sistemas de recomendação, redes bayesianas são usadas com o intuito de capturar relações entre elementos relevantes, como o histórico de leitura do usuário e preferências de conteúdo, e calcular a probabilidade de aceitabilidade das recomendações do sistema. Com essa abordagem, o sistema analisa padrões de comportamento de múltiplos usuários para inferir preferências individuais, fornecendo recomendações personalizadas que visam aumentar a relevância e a satisfação do leitor.

Segundo os autores, a construção do modelo foi dividida em quatro fases, começando pela seleção das variáveis usadas pela rede bayesiana. Nesta etapa, foram analisadas as variáveis que possivelmente influenciam em decisões de leitores do sistema, sendo criados um total de quatro conjuntos de variáveis. Algumas das variáveis consideradas pelo sistema incluem características do usuário, como nível de educação, tempo médio de visita por sessão, tempo médio dedicado à leitura de cada notícia, quantidade de notícias lidas por sessão e idade da notícia no momento da leitura. As variáveis relacionadas às notícias abrangem a fonte, editora, tempo de leitura e expectativa de vida (pico de leituras). Em seguida, os termos presentes em cada artigo foram transformados em vetores booleanos, constituindo um terceiro grupo de variáveis na rede bayesiana. Por fim, o quarto grupo incluiu as variáveis calculadas pelo *software* WEKA através do algoritmo K-Means, como *clusters* de usuários, *clusters* de termos em artigos e a probabilidade de leitura de itens.

Após a definição das variáveis, estas são organizadas e incorporadas à rede, conforme proposto na modelagem feita pelo especialista do domínio (autor do texto). Durante a fase de estruturação, cada variável é então transformada em um nó que se conecta com outros nós da mesma rede. Além disso, os estados de cada nó são determinados com base nas informações armazenadas na base de dados do jornal *online*. A figura 10 ilustra o procedimento de estruturação da rede bayesiana.

Em contrapartida às fases anteriores, a etapa subsequente foi executada automaticamente pelo algoritmo de maximização de expectativa (EM) do *software* Netica, responsável por calcular e definir a probabilidade condicional de cada estado de nó com base nas entradas de *log* (acesso) do sistema.

Figura 10 – Processo de estruturação de rede Bayesiana para artigos de notícias



Fonte: Adaptado de Silva, Schreiber e Nara (2015).

Assim como outros métodos de recomendação baseados em modelos, redes bayesianas também apresentam um período de treino ou pré-processamento de dados. O sistema é, portanto, subdividido pelos autores em dois estágios: a modelagem da rede e aprendizado de probabilidades e a formação das recomendações. Na primeira etapa ocorrem os processos descritos anteriormente para modelagem da rede bayesiana, seguidos pelas etapas de aprendizado. Em detalhes, ocorrem no sistema a busca por dados em todos os artigos da base de dados, a vetorização binária das palavras encontradas em artigos, o processamento de dados para ajustes de otimização dos possíveis valores para variáveis, a

divisão em subtabelas e agrupamento dos dados usados pelo quarto grupo de variáveis da rede até a sua redefinição em tabela única. Já na estruturação de rede, há as fases de definição da estrutura, estados de nós e o aprendizado de probabilidades condicionais por meio do algoritmo EM.

Por fim, as sugestões de recomendações são geradas a partir da probabilidade de interesse do usuário em ler algum artigo. Para isso, o sistema vasculha por todos artigos em dado período de tempo com o propósito de organizar uma lista de sugestões baseada na probabilidade de leitura do usuário.

No estudo de Conceição *et al.* (2016), é explorada a aplicação de um método de recomendação multimodal em um sistema que combina recomendações baseadas em conteúdo textual e visual de vídeos. Segundo os autores, o estudo foi motivado pelo problema de *Cold-Start*, frequentemente presente nos métodos baseados em Filtragem Colaborativa. Este problema é comum em sistemas com informações limitadas ou esparsas sobre seus usuários ou itens, ocorrendo principalmente no caso de novos itens ou usuários.

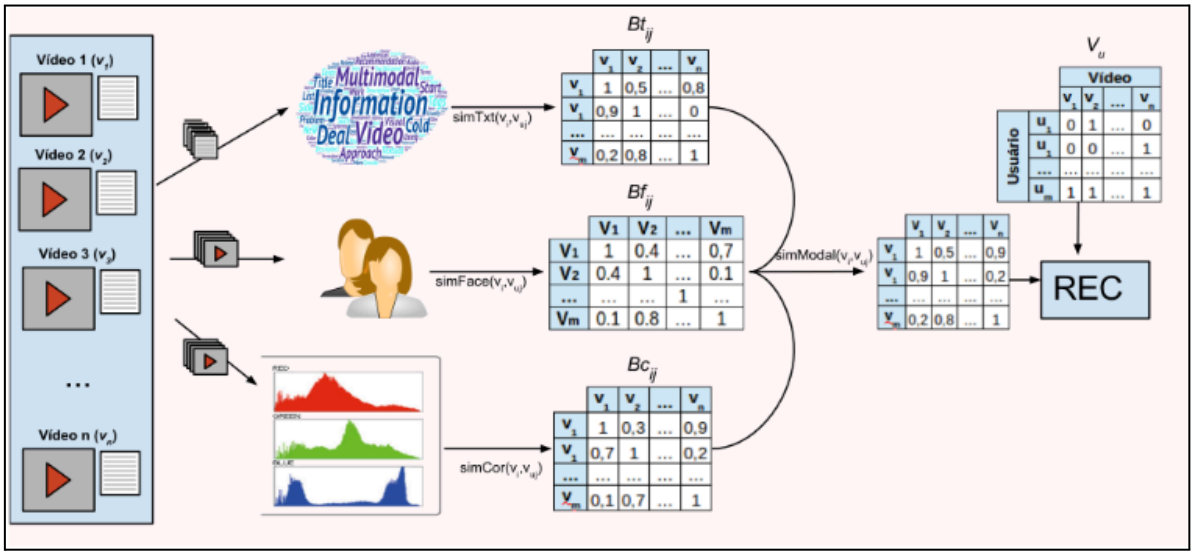
Nesse sentido, o sistema de recomendação multimodal proposto busca confrontar o problema de *Cold-Start* ao gerar sugestões baseadas no conteúdo dos itens de maneira análoga ao modelo baseado em vizinhança. Para isso, são criadas as matrizes de similaridade textual e visual do conjunto de vídeos como dados de entrada na matriz multimodal.

No quesito textual, utiliza-se uma técnica vetorial simples, chamada *Bag-of-Words* (BOW), para representar palavras em vetores binários. Estas palavras são então adicionadas ao conjunto de palavras de cada vídeo em conjunto dos pesos de similaridade destas, calculados por meio do TF-IDF e da distância do cosseno. A similaridade visual, no entanto, é representada pelos conjuntos de similaridade de cores e faces. Diferente de palavras em textos, cada vídeo é composto por um conjunto de quadros de cores, representados pela média dos descritores dos histogramas e pelo conjunto de faces, que utiliza a técnica *Bag of Faces* (BOF) para representação de faces em vídeos. Ambas matrizes utilizam a distância do cosseno para o cálculo de similaridade entre vetores.

Por outro lado, o método multimodal consiste em uma técnica que combina os valores de similaridade de cada matriz criada anteriormente. Dessa forma, é aplicada uma fórmula que adiciona o valor máximo entre as matrizes textual, de cor e de faces à matriz multimodal. O diagrama do modelo de recomendação pode ser visto na figura 11.

Por fim, a matriz multimodal é usada como entrada pelo método *Sparse Linear Method with Side Information* (SSLIM), responsável por gerar recomendações do sistema com base na matriz multimodal e nas informações laterais adicionais presentes na matriz de avaliações binárias composta pelos vídeos e usuários.

Figura 11 – Método de recomendação multimodal

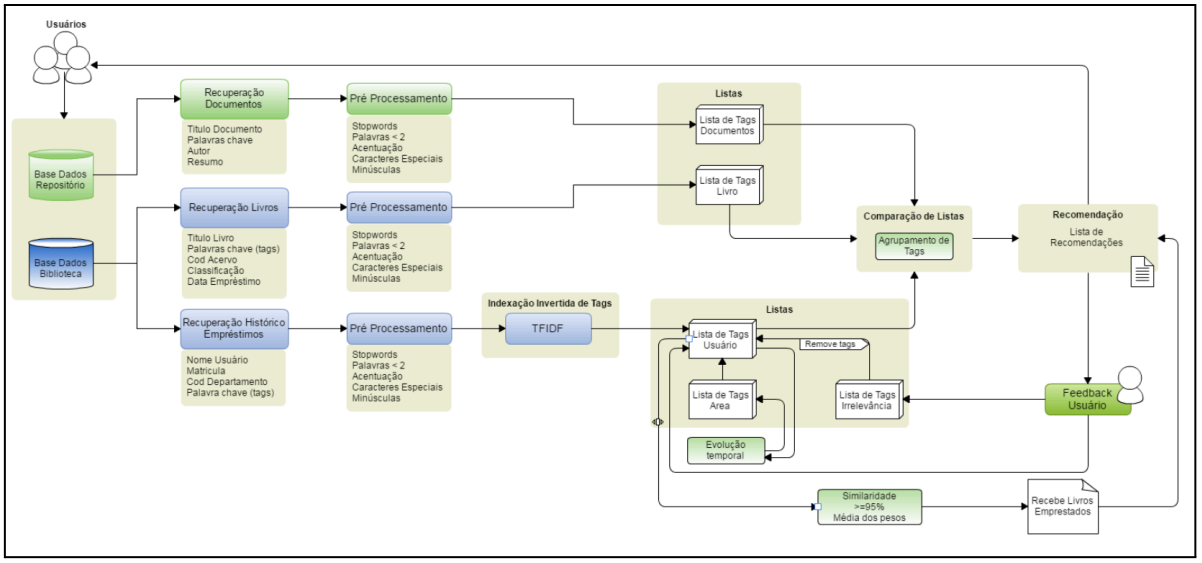


Fonte: Conceição *et al.* (2016, p. 216).

Em Furtado (2016), o autor apresenta um sistema de recomendação híbrido destinado ao ambiente de uma biblioteca digital universitária (figura 12). Além disso, este sistema apresenta uma arquitetura formada com base em cinco listas de *tags* (termos), combinando técnicas de Filtragem Baseada em Conteúdo, que analisa o histórico de empréstimos para criar listas de recomendações para serem avaliadas, e Filtragem Colaborativa, que sugere itens a usuários com perfis semelhantes, considerados vizinhos.

Das cinco listas em questão, duas são "fixas" e representam os itens disponíveis: uma lista de tags para livros, extraída da base de dados da Biblioteca, e outra lista de tags para documentos, provenientes do repositório institucional. Ambas as listas são representadas por matrizes de vetores de itens por tags. Por outro lado, as listas de tags de usuários, áreas e irrelevantes, que são dinâmicas, são derivadas dos empréstimos de itens pelos usuários. Essas listas são calculadas usando a técnica TF-IDF ou feedback dos usuários.

Figura 12 – Arquitetura do SR para biblioteca universitária



Fonte: Furtado (2016, p. 62).

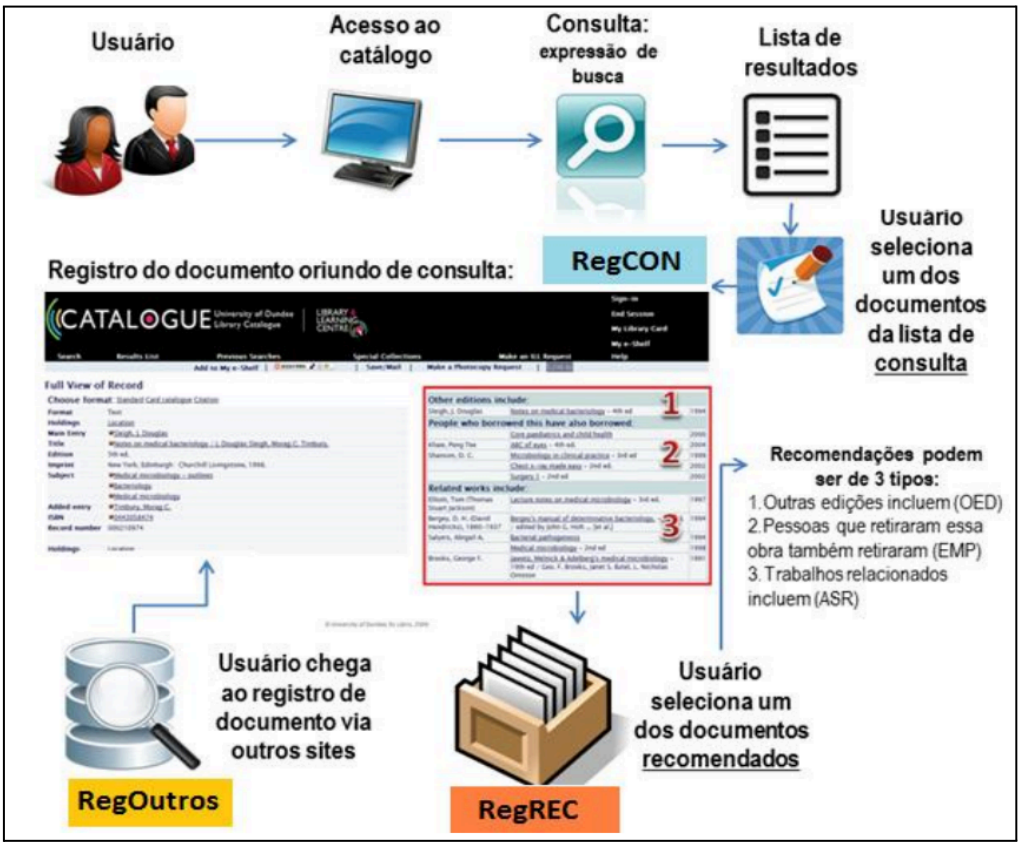
No caso das tags de usuários e áreas, a medida TF-IDF é aplicada considerando apenas os itens emprestados pelo usuário ativo (para tags de usuários) ou pelos usuários associados a um curso específico em um determinado momento (para tags de áreas). As tags irrelevantes são identificadas com base no feedback dos usuários em relação às recomendações do sistema. Para isso, um e-mail é enviado aos usuários para que avaliem as recomendações em uma escala de 0 a 3. Com base nessas avaliações, o sistema ajusta as tags no perfil do usuário, incluindo as tags de artigos considerados pouco relevantes na lista de tags irrelevantes e removendo-as da lista de tags do usuário. Esses termos são então utilizados para refinar as recomendações geradas pelo sistema.

Outro sistema de recomendação inserido no contexto das bibliotecas universitárias é analisado por Monteiro-Krebs, Rocha e Ribeiro (2017). Este artigo apresenta um sistema de recomendação para catálogos *online* chamado *"Related Books in Aleph OPAC"*. Esta é uma extensão do sistema Aleph para bibliotecas, que oferece três tipos de recomendações: “Pessoas que retiraram esta obra também retiraram”, para empréstimos, utilizando a técnica de Filtragem Colaborativa; “Trabalhos relacionados incluem”, para assuntos; e “Outras edições incluem”, para edições, ambas utilizando a técnica de Filtragem Baseada em Conteúdo. A figura 13 ilustra o esquema de utilização do OPAC proposto pelos autores, assim como as recomendações de itens no catálogo.

Em relação ao primeiro tipo baseado em empréstimos, os autores explicam que as recomendações são formuladas com base no item visualizado pelo usuário-alvo, sugerindo

apenas itens que foram emprestados por outros usuários que também retiraram o mesmo item anteriormente. No entanto, esse tipo de recomendação é aplicado somente a itens que foram retirados por pelo menos cinco usuários da biblioteca. Quanto às recomendações de assunto e de outras edições, ambas utilizam FBC para extrair metadados dos registros de itens, como termos descritores ou o Número Padrão Internacional de Livro (ISBN). No tipo de assunto são geradas recomendação para itens que tenham pelo menos três descritores de assunto ou números de classificação em comum com o item visualizado. Enquanto a recomendação por edições considera apenas documentos que possuem o mesmo ISBN para serem sugeridos.

Figura 13 – Esquema de navegação do usuário no catálogo, consulta e recomendação da extensão *Related Books in Aleph OPAC*

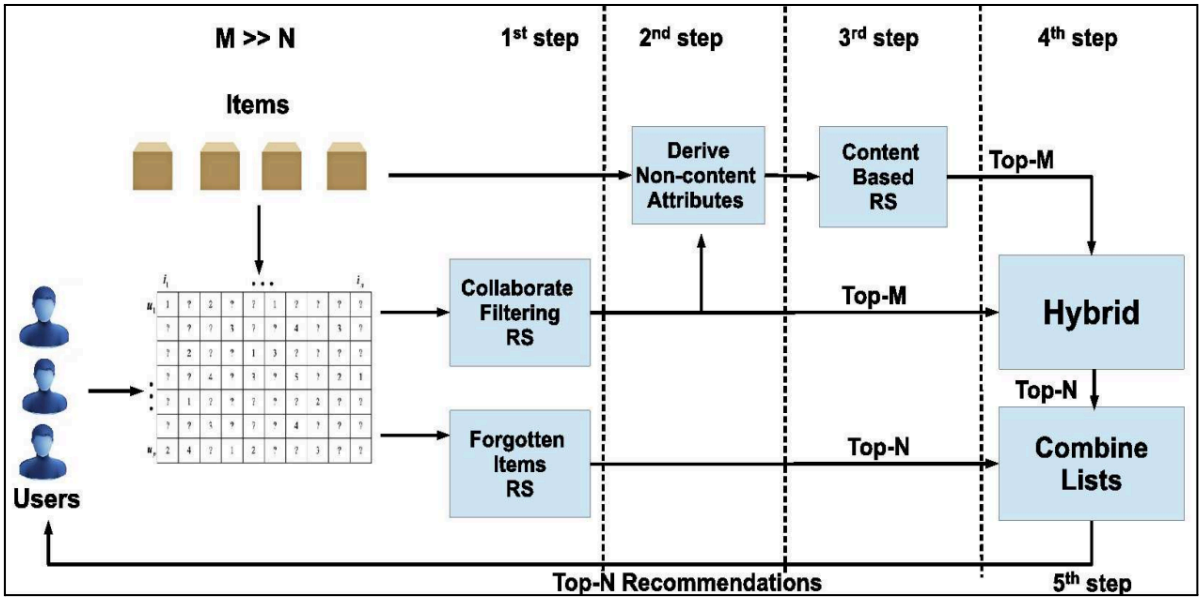


Fonte: Monteiro-Krebs, Rocha e Ribeiro (2017, p. 156).

Cabe ressaltar que o sistema não realiza recomendações em tempo real, optando por atualizar a lista de recomendação em períodos de uma semana para a lista de empréstimos, uma quinzena para as recomendações de assunto e um mês para outras edições do mesmo documento.

Para resolver problemas algorítmicos, como a ausência de recomendações de itens consumidos há muito tempo pelos usuários e as limitações na captura de preferências implícitas de consumo de itens anteriores que se relacionam ao consumo de itens atuais, Mourão (2018) propõe o sistema de recomendação ForNonContent. ForNonContent é um SR híbrido, baseado em FC e FBC, dividido em cinco etapas (figura 14).

Figura 14 – ForNonContent: método híbrido para recomendação de itens reconsumíveis esquecidos e atributos não relacionados ao conteúdo



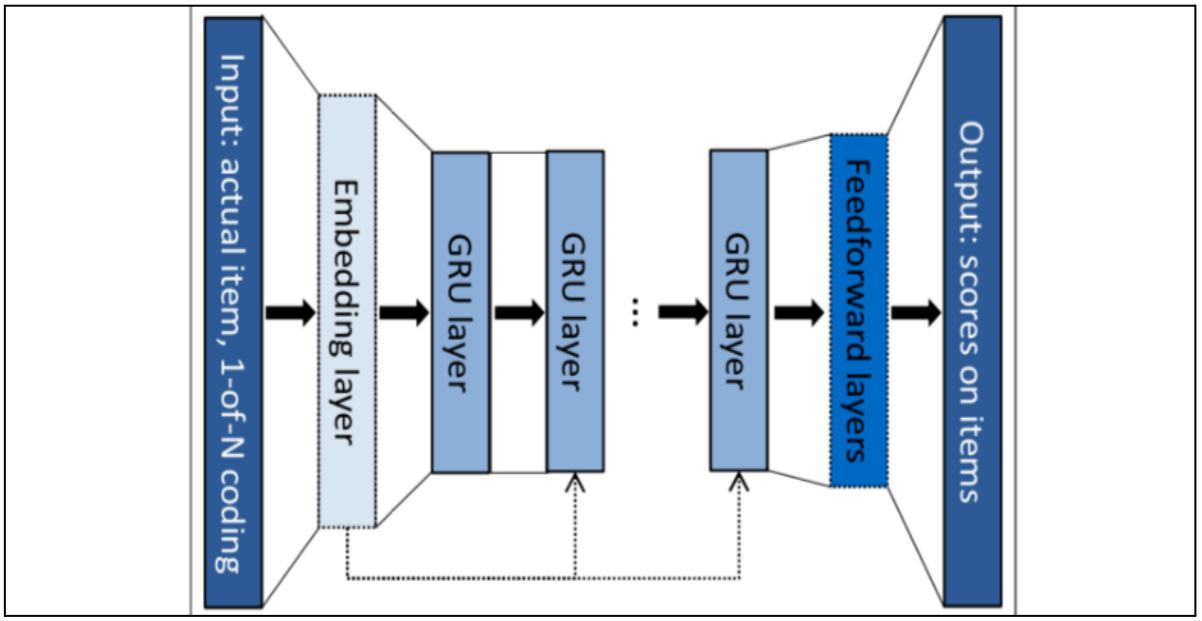
Fonte: Mourão (2018, p. 5).

Este sistema realiza um cálculo para os resultados Top-M e Top-N mais relevantes, iniciando com a execução de dois algoritmos: um para recomendação de itens já vistos, mas que podem ser recomendados novamente, e outro de FC, cuja matriz possui um tamanho muito maior. Na segunda etapa, os atributos não relacionados ao conteúdo, derivados da técnica de FC, são vetorizados em atributos baseados na popularidade, atualidade e similaridade dos itens. Na etapa seguinte de FBC, utiliza-se a distribuição Gaussiana multivariada para calcular tanto a preferência quanto a variabilidade não baseada no conteúdo dentro do espaço vetorial. Na quarta etapa, as avaliações da matriz formada pela técnica de FC são combinadas com as probabilidades da técnica de FBC, gerando as recomendações de atributos não relacionados ao conteúdo. Por fim, essas recomendações são combinadas com a lista de recomendações do algoritmo de itens já vistos e entregues ao usuário de maneira intercalada.

No intuito de analisar o cenário de recomendações baseadas em sessões, em que usuários não são identificados pelo sistema, Costa (2020) avalia a performance de cinco modelos de recomendação, além de apresentar uma implementação alternativa de um algoritmo de redes neurais recorrentes.

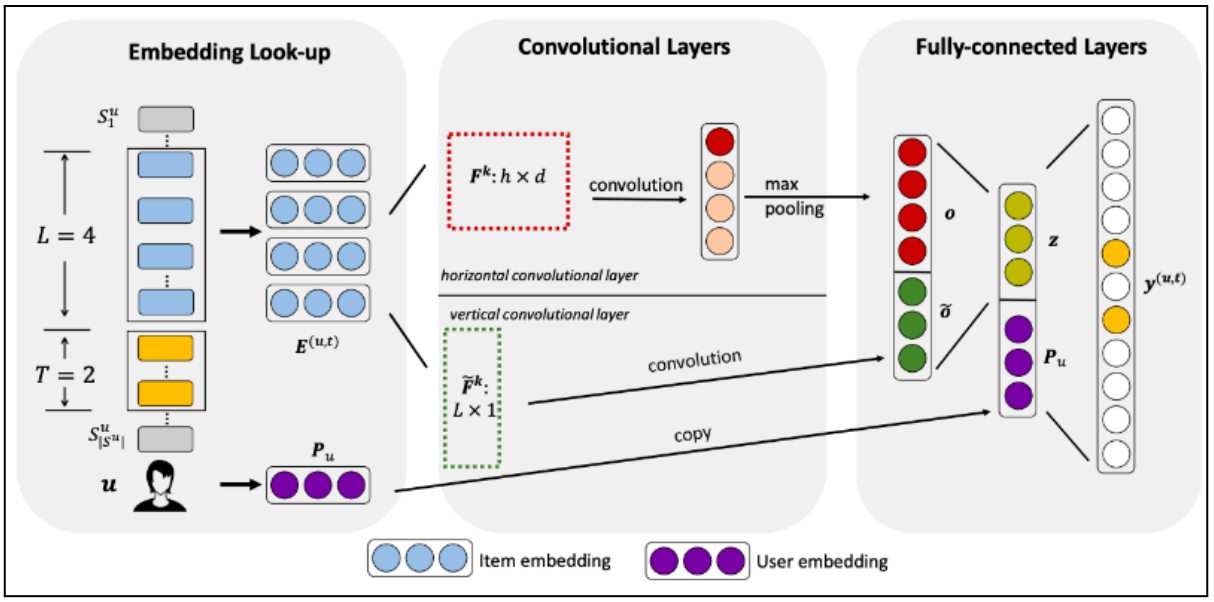
Entre os sistemas incluídos na avaliação estão os modelos: TopPolular e SPop, métodos ingênuos que recomendam itens predefinidos independente das ações dos usuários, estes dois métodos diferenciam-se por gerar recomendações baseadas nos itens mais frequentes e populares dentro de um conjunto inteiro de treinamento (TopPopular) ou de uma sessão atual (SPop); Item K-NN, um modelo de FC baseado em memória, cujo desenvolvimento deriva do algoritmo K-NN (ver seção 2.3.3.1 Filtragem Colaborativa); GRU4Rec (figura 15), um algoritmo de recomendação baseado em sessões e no modelo de Redes Neurais Recorrentes com Unidade Recorrente Controlada (GRU); *ConvolutionaI Sequence Embedding Recommendation* (Caser), ou Recomendação de Incorporação de Sequência Convolucional (figura 16), um modelo recomendação sequencial de Rede Neural Convolucional (CNN); e a implementação alternativa do autor para o modelo GRU4Rec, utilizando a biblioteca *Tensorflow Recommender*. Os resultados da análise realizada pelo autor, utilizando o conjunto de dados da Last.FM, uma plataforma de músicas, mostraram que os algoritmos com melhor desempenho em recomendações baseadas em sessões são, respectivamente, os modelos Caser, GRU4Rec e Spop.

Figura 15 – Modelo de recomendação baseado em sessões GRU4Rec



Fonte: Costa (2020, p. 7).

Figura 16 – Modelo de recomendação sequencial Caser



Fonte: Costa (2020, p. 8).

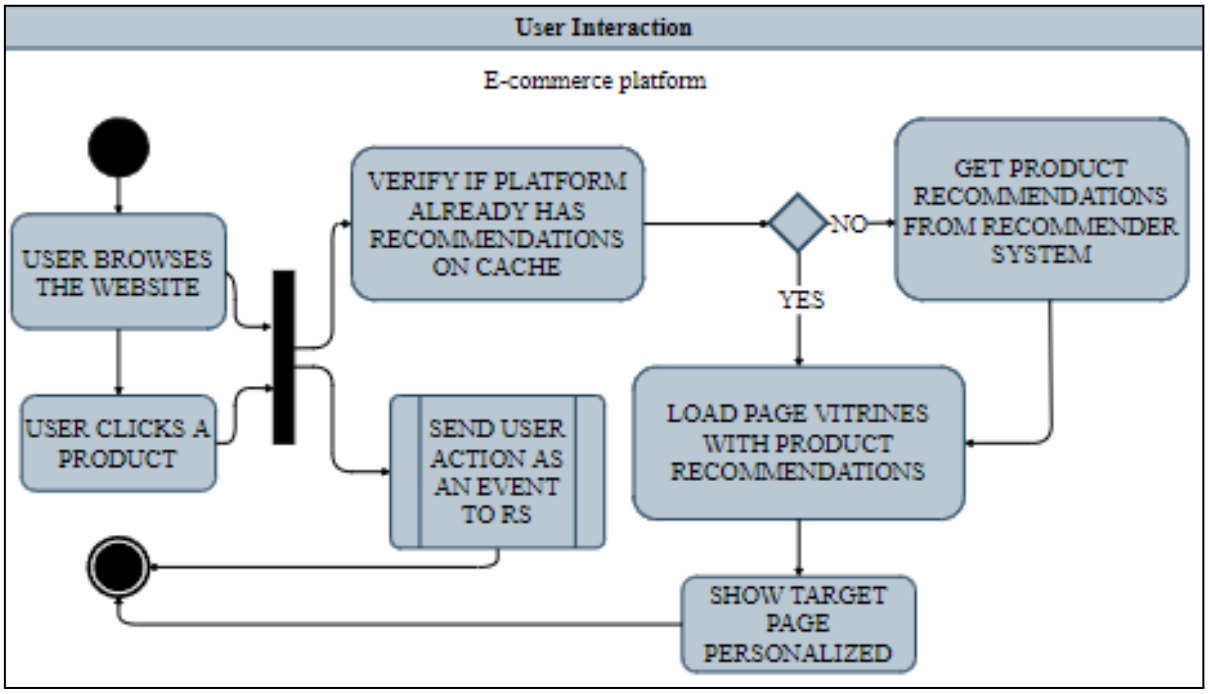
No entanto, Costa considera alguns aspectos durante a análise dos algoritmos. O algoritmo Item K-NN, em um conjunto de dados extremamente esparso, com muitos itens e poucos usuários, apresentaria um desempenho superior no cenário oposto. Da mesma forma, o algoritmo K-NN tradicional teria um desempenho melhor neste cenário. Por outro lado, o algoritmo GRU4Rec foi o modelo mais rápido avaliado pelo autor, superando a acurácia dos algoritmos tradicionais de vizinhança e popularidade. O Caser, por sua vez, obteve os melhores resultados, devido à sofisticação do modelo ao utilizar representações vetoriais de palavras (*word embeddings*) em redes neurais convolucionais.

Cunha (2021) apresenta a arquitetura de um sistema de recomendação destinado a diversas plataformas de *e-commerce*. O sistema inclui quatro tipos distintos de recomendação: popularidade, híbrido, produtos similares e produtos complementares. No tipo de popularidade, são sugeridos os itens mais populares com base em interações dos clientes na plataforma. O tipo híbrido combina Filtragem Colaborativa, baseada em conteúdo e segmentação de clientes para melhorar a acurácia do sistema, empregando, respectivamente, o algoritmo *Singular Value Decomposition* (SVD) para classificação, a técnica TF-IDF para calcular a frequência de termos e o algoritmo K-means para agrupamento de clientes em *clusters*. Produtos similares são identificados com base nas semelhanças entre itens ou textos de descrição dos produtos, também utilizando TF-IDF. Por fim, a categoria de produtos

complementares faz uso de Regras de Associação e do algoritmo Apriori para gerar recomendações de itens frequentemente comprados em conjunto.

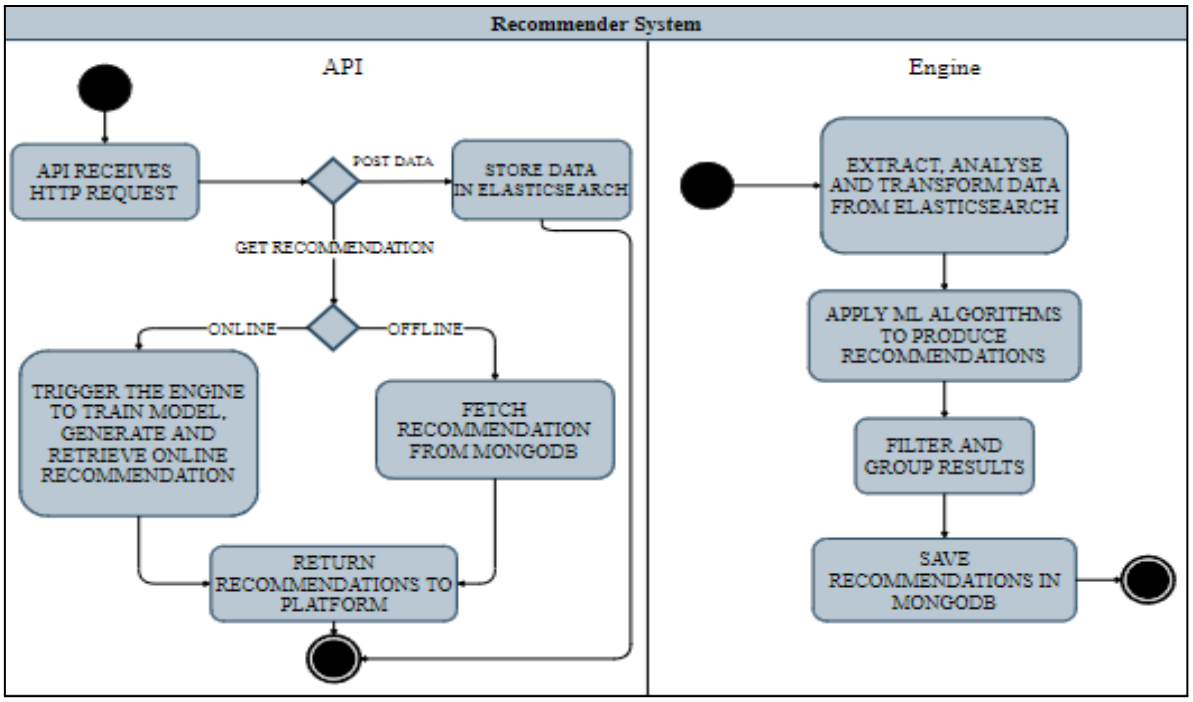
Além disso, esta dissertação aborda ainda outras questões, incluindo a proposta de um modelo com recomendações em tempo quase real (*nearline recommendations*), que combina os modos de recomendação *online* e *offline*. Segundo o autor, essa abordagem proporciona maior flexibilidade ao sistema, atualizando as recomendações em novos treinos do modelo em ciclos de vinte minutos. Isso evita problemas com as poucas atualizações em recomendações *offline*, que devem ser realizadas em períodos de menor atividade do sistema, e também com a escalabilidade do sistema em recomendações *online*, que podem resultar em gargalos com grandes volumes de requisições de usuários. Outro aspecto importante encontrado foi a proposta de um sistema de recomendação que estivesse diretamente conectado às bases de dados ElasticSearch e MongoDB, proporcionando uma melhoria significativa em termos de desempenho e eficácia, ao mesmo tempo em que possibilita reduzir possíveis sobrecargas na API do sistema decorrentes da manipulação excessiva de recursos. As figuras 17 e 18 possibilitam a visualização dos diagramas de atividades do sistema realizadas durante uma interação com o usuário e o funcionamento do SR.

Figura 17 – Diagrama de atividades de interação do usuário em sistema de e-*commerce*



Fonte: Cunha (2021, p. 47).

Figura 18 – Diagrama de atividades em SR de e-*commerce*

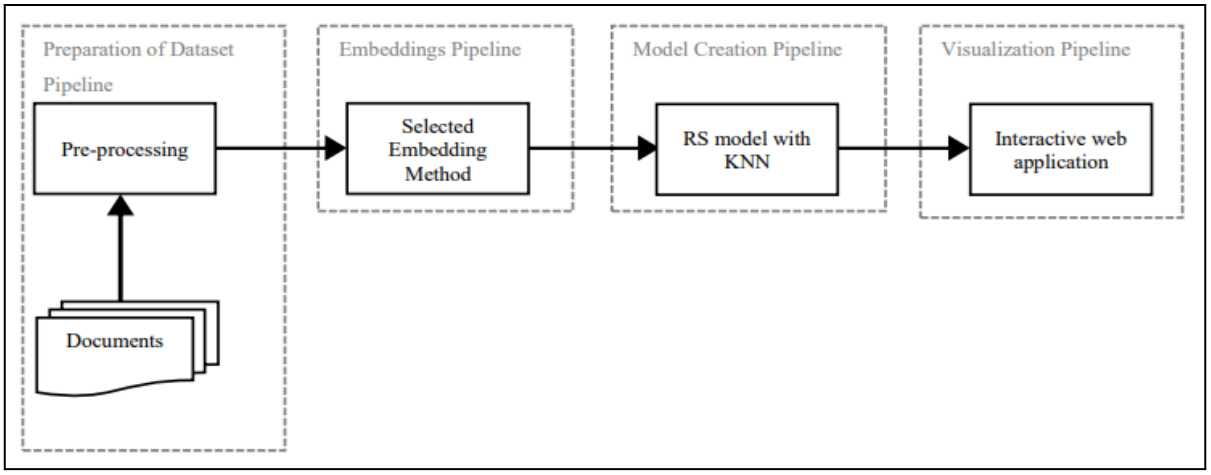


Fonte: Cunha (2021, p. 47).

Tendo em vista a utilização dos SR para domínios ricos em conteúdo textual, Neves (2022) propõe um SR amparado nas técnicas de FBC e *embeddings*. Nesta dissertação, a autora avalia a performance de cinco técnicas de *word embeddings* e *sentence embedding* (técnica de vetorização de sentenças), incluindo modelos vetoriais como o TF-IDF e modelos de redes neurais como Word2Vec, InferSent, Universal Center Encoder e Sentence-Bert (SBERT). Com base nos resultados, Neves seleciona o algoritmo SBERT para desenvolver o sistema de recomendação.

O SR proposto compreende um método dividido em quatro etapas distintas (figura 19). Inicialmente, ocorre o pré-processamento dos dados, que envolve a seleção dos conjuntos de dados e a aplicação de técnicas específicas de pré-processamento. Na segunda fase, o método de *embedding* SBERT é treinado utilizando dois conjuntos de dados formados por notícias e teses, que serão reutilizados na etapa seguinte. Em seguida, um SR que utiliza o modelo K-NN é aplicado para mensurar a similaridade entre os itens pré-processados pelo SBERT e a *query* atual, gerando uma lista de sugestões. A última etapa consiste na aplicação deste SR na web, utilizando o *software* gráfico Gephi, a aplicação web de teses e dissertações da Nova IMS, ThesesViz, e o repositório GitHub para a hospedagem do sistema (Neves, 2022).

Figura 19 – Diagrama de fluxo de SR baseado em *embedding*



Fonte: Neves (2022, p. 10).

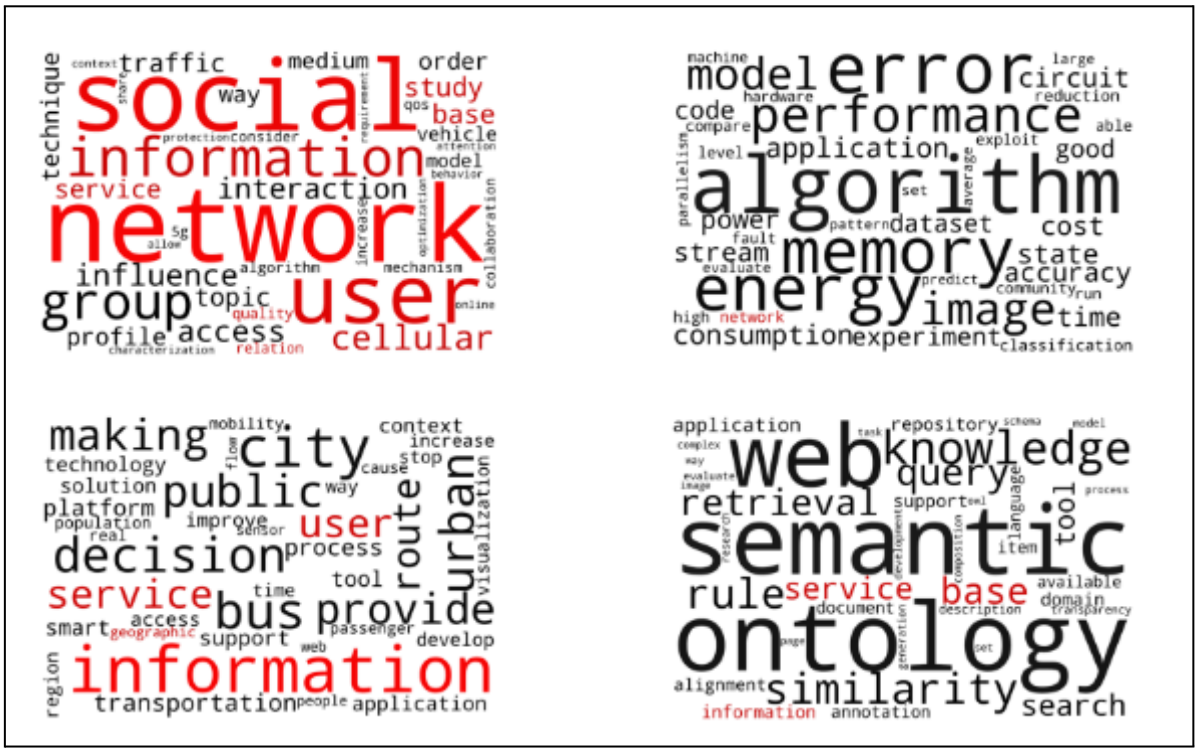
Em Souza e Feitosa (2022), os autores desenvolveram um algoritmo de recomendação para artigos científicos, baseado em clusterização, utilizando K-Means e o conceito de *Word Clouds*. Inicialmente, os autores executaram o algoritmo K-Means para gerar os *clusters* a partir do conjunto de dados de treino. Em seguida, aplicam técnicas de pré-processamento textual para tratar os artigos e usam a técnica TF-IDF para vetorização das palavras nos resumos.

O objetivo desse sistema é identificar o *cluster* mais adequado para cada artigo de teste e calcular a distância (dissimilaridade) em relação ao centroide e a outros documentos do grupo. Para isso, é necessário vetorizar o resumo e calcular a distância euclidiana de cada novo item, a fim de gerar recomendações baseadas nessa nova publicação. Além disso, o modelo incorpora a criação de *Word Clouds* (nuvens de palavras) formadas pelas principais palavras de cada *cluster*. As *Word Clouds* são usadas com o intuito de facilitar a visualização de palavras em determinado grupo, aumentando em tamanho conforme sua importância no contexto da nuvem. Dessa forma, palavras do item selecionado que aparecem na nuvem de palavras são destacadas em vermelho, permitindo visualizar a similaridade do artigo com o *cluster* selecionado.

Adicionalmente, este algoritmo é avaliado com base na sua performance (tempo de execução) e precisão (número de acertos), sendo posteriormente comparado com o algoritmo K-NN clássico. Embora o algoritmo demonstre desempenho inferior ao algoritmo K-NN, um segundo aspecto deste modelo a ser analisado é a criação das *Word Clouds*, que permitem

visualizar a similaridade do artigo com a nuvem mais próxima criada e compará-la com nuvens formadas por palavras de outros *clusters*, conforme mostrado na figura 20.

Figura 20 – Visualização de *Word Clouds* em Modelo de recomendação baseado em *Clustering*



Fonte: Souza e Feitosa (2022, p. 22).

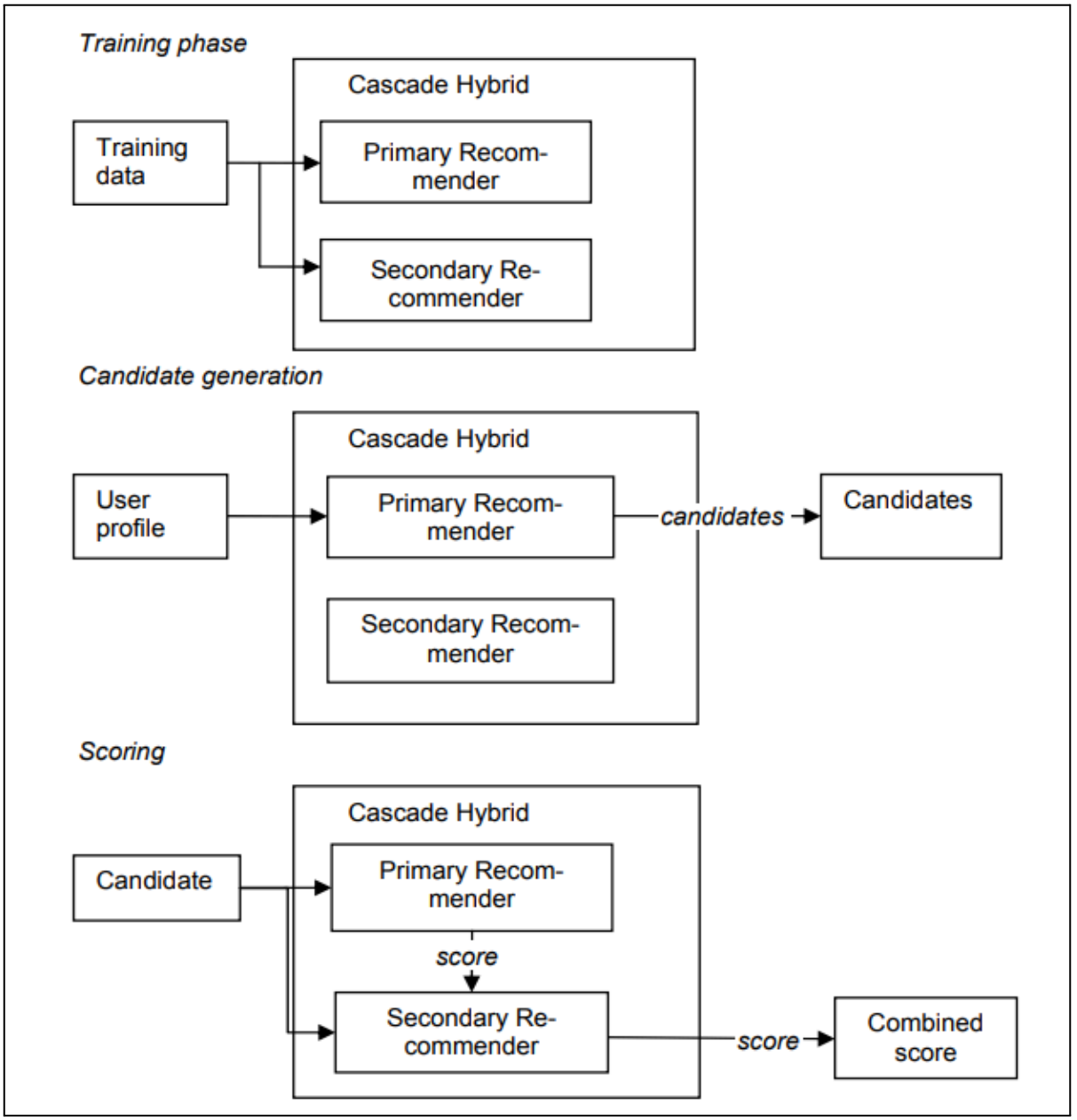
Vieira, Passos e Salm (2023) conduzem uma revisão bibliográfica sobre sistemas de recomendação (SR) aplicados em bibliotecas universitárias, analisando sistemas baseados em Filtragem Colaborativa, Filtragem Baseada em Conteúdo e sistemas híbridos. Ao final, propõem um modelo híbrido fundamentado nas estratégias de hibridização em SR descritas por Burke (2007), conforme ilustrado no Quadro 5.

As autoras afirmam, com base no artigo de Burke, que as três estratégias de recomendação que apresentam os melhores resultados são: a estratégia de Aumento de Características, combinada com FBC e Filtragem Baseada em Conhecimento; a estratégia de cascata, em conjunto com FC e FBC; e a estratégia de cascata associada à Filtragem Baseada em Conhecimento e FC. Nos resultados finais do artigo, a estratégia de recomendação híbrida destacada foi a estratégia de cascata, utilizando as técnicas de FC e FBC (figura 21).

Neste modelo, há dois recomendadores: o recomendador primário, baseado em Filtragem Colaborativa (FC), e o recomendador secundário, que utiliza Filtragem Baseada em

Conteúdo (FBC). Ambos passam por um período de pré-processamento do conjunto de dados de treino. Na etapa seguinte, usa-se apenas a técnica de FC para gerar itens candidatos à próxima fase. Em seguida, o recomendador primário pontua a similaridade entre os itens e o perfil do usuário, organizando-os em ordem decrescente. Em caso de empate, as pontuações dos itens são revisadas pela técnica de FBC, que aplica um critério de desempate.

Figura 21 – Técnica de Cascata para SR Híbrido



Fonte: Burke (2007, p. 390).

Os capítulos a seguir dessa seção compreendem a análise dos trabalhos descritos e a proposta de um modelo conceitual de recomendação que tenha como base o referencial teórico desta pesquisa.

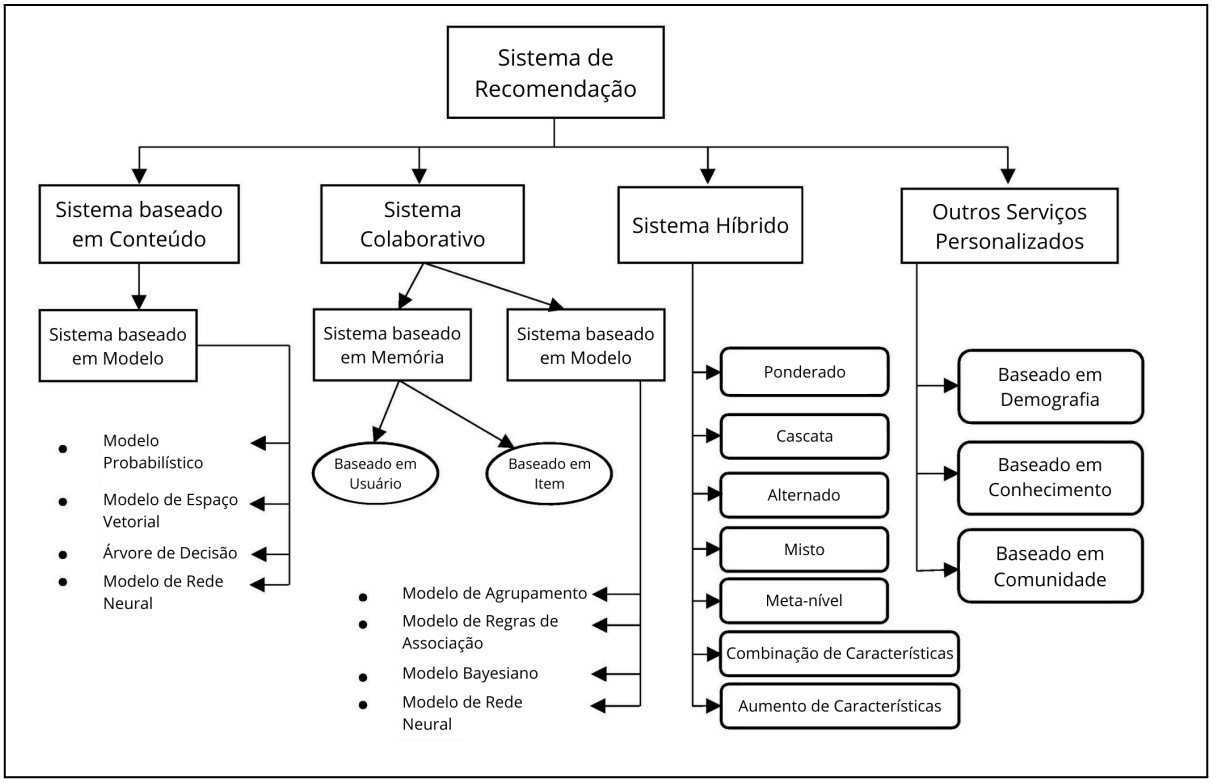
4.2 ANÁLISE DOS MODELOS

A partir da descrição dos documentos recuperados, percebe-se que o estado da arte continua alinhado com as técnicas de filtragem da informação apresentadas ao longo do referencial teórico desta pesquisa. Embora novas técnicas de FI sejam utilizadas por diferentes sistemas de recomendação na literatura, variando principalmente quanto aos objetivos e cenários de aplicação, há um grande enfoque em três técnicas principais: FC, FBC e Filtragem Híbrida, com destaque para a Filtragem Colaborativa.

No entanto, o desenvolvimento de novos modelos de recomendação não apenas amplia a gama de opções, mas também aumenta a complexidade desses sistemas, tornando a escolha do método de recomendação e algoritmos uma tarefa tão crucial quanto selecionar a técnica de FI. Nesse sentido, a figura 22 exemplifica alguns dos principais métodos de recomendação encontrados na literatura. Além dos métodos de recomendação apresentados na figura, há ainda diversos outros SR, entre os quais podem ser citados os sistemas baseados em modelos de regressão, fatoração de matrizes, algoritmos genéticos, entre outros.

Por outro lado, diversos estudos buscam não apenas a criação de novos sistemas de recomendação, mas também aprimorar algoritmos e sistemas existentes (Aleixo, 2014; Nóbrega, 2014; Costa, 2020). Isso inclui desde testes com diferentes medidas de similaridade e a aplicação dos métodos em diferentes domínios até ajustes nos algoritmos utilizados, com o intuito de melhorar o desempenho em métricas específicas como precisão e performance, bem como reduzir o custo computacional. Também são exploradas soluções para problemas como *Cold-Start*, falsos vizinhos, escalabilidade e esparsidade.

Figura 22 – Classificação para sistemas de recomendação



Fonte: Sinha; Dhanalakshmi (2019, p. 1049, tradução nossa).

É importante ressaltar que, ao avaliar o desempenho dos métodos, é necessário interpretar com cautela a maioria das execuções dos sistemas analisados no capítulo anterior, devido às diferentes performances e métricas utilizadas. Isso se deve ao fato de que foram utilizados conjuntos de dados de fontes externas como GitHub, Netflix, MovieLens, Youtube ou Last.FM (Aleixo, 2014; Nóbrega, 2014; Oliveira, 2014; Conceição *et al.*, 2016; Mourão, 2018; Costa, 2020; Souza; Feitosa, 2022). Além disso, em alguns casos, foram empregadas avaliações *online* (Torres, 2004b; Lopes, 2007; Franco; Sanchez; Serna, 2015; Silva; Schreiber; Nara, 2015; Furtado, 2016; Mourão, 2018; Cunha 2021) e testes *offline* com conjuntos de dados de outras plataformas, como bibliotecas digitais e sites (Torres, 2004b; Furtado, 2016; Monteiro-Krebs; Rocha; Ribeiro, 2017; Neves, 2022).

No contexto da Brapci, a indexação automática de publicações provenientes de diferentes fontes, como trabalhos de eventos e artigos de periódicos variados, pode resultar na introdução de ruídos durante a extração dos metadados e impactar negativamente na acurácia das recomendações. Cabe ressaltar, entretanto, que este problema tem sido alvo de revisões automatizadas e supervisionadas pelo projeto da Brapci IA.

Neste sentido, métodos focados em metadados das publicações científicas, como referências, palavras-chave, resumos e autores, assim como no acoplamento, em sua maioria propostos nos procedimentos metodológicos desta pesquisa, demonstram uma afinidade natural com as técnicas de FBC e SRS. Isso ocorre pela facilidade com que esses métodos apresentam ao lidar com dados ricos em texto (FBC), semântica e contexto (SRS). Entre os modelos analisados, observa-se que algoritmos mais avançados, como redes neurais, superam a técnica TF-IDF em termos de desempenho, apesar de exigirem mais recursos computacionais, e aproximam-se dos modelos classificados como endógenos (Gemmis *et al.,* 2015). Em contrapartida, nos SRS, o uso de filtros de informação ou esquemas conceituais caracteriza-os como modelos exógenos, proporcionando informações confiáveis e contextualizadas sobre as necessidades informacionais dos usuários, o que eleva a qualidade das recomendações (Peis; Morales-Del-Castillo; Delgado-Lopez, 2008; Gemmis *et al.,* 2015).

Em contrapartida, métodos fundamentados em FC oferecem vantagens na personalização e relevância das recomendações, alinhadas com o perfil e histórico de consultas dos usuários. Como apontado anteriormente, os métodos baseados em FC são divididos em duas abordagens principais (Melville; Sindhwani, 2011; Vieira; Passos; Salm, 2023). A abordagem orientada por vizinhança de usuários ou itens proporciona recomendações em tempo real, mas pode causar gargalos no sistema devido a problemas de escalabilidade. Alternativamente, os métodos baseados em modelos podem solucionar vários problemas associados aos métodos de vizinhança, utilizando técnicas como clusterização e fatoração de matrizes, que reduzem o custo computacional síncrono dos SR, além de abordarem questões de esparsidade e *Cold-Start*. As redes neurais, entretanto, constituem uma técnica avançada na área de aprendizagem profunda, frequentemente apresentando resultados superiores a outros sistemas de recomendação em várias métricas avaliadas (Costa, 2020; Neves, 2022). Além disso, o modelo de FC-Puro proposto por Torres (2004b) também pode ser aplicado no contexto das recomendações por acoplamento uma vez que utiliza as citações dos próprios artigos como base para a recomendação de novos itens ao leitor.

A Filtragem Híbrida, por sua vez, possibilita a união destas técnicas para que as vantagens de uma sobreponham os aspectos negativos de outra, aprimorando a precisão e a performance dos sistemas. Em Burke (2007), os métodos híbridos aplicados a um conjunto de dados com aproximadamente 50 mil sessões apresentaram, em sua maioria, uma classificação média para recomendações corretas superiores aos métodos isolados, especialmente quando

utilizados com estratégias de aumento de características ou cascata. Ambas as estratégias, assim como a estratégia de hibridização mista, são aplicadas aos modelos híbridos descritos nesta seção, conforme mostrado no quadro 6, evidenciando que a aplicação e a relevância dessas estratégias se estende a múltiplas áreas de interesse, incluindo *e-commerce*, *streaming* de mídia e serviços de informação.

Portanto, o capítulo a seguir busca apresentar o desenvolvimento de um modelo conceitual de recomendação de publicações científicas para o repositório digital Brapci, idealizado nesta pesquisa a partir da discussão dos modelos revisados nesta seção.

4.3 PROPOSTA DO MODELO CONCEITUAL DE RECOMENDAÇÃO

A partir da análise dos modelos de recomendação discutidos na seção anterior, ressalta-se novamente a viabilidade das abordagens de FBC e Filtragem Semântica para o desenvolvimento de um modelo voltado às publicações científicas, alinhado aos objetivos desta pesquisa. A FC, por sua vez, permite a criação de recomendações personalizadas, adaptadas ao perfil dos usuários do sistema, promovendo também a serendipidade. Já a filtragem híbrida possibilita a combinação de várias técnicas, aproveitando as vantagens de cada método em um único sistema de recomendação.

Com base nisso, esta seção propõe um modelo conceitual dividido em três estratégias para diferentes contextos de pesquisa na Brapci. Portanto, o capítulo foi estruturado para, inicialmente, abordar as características do desenvolvimento do sistema de acordo com a taxonomia de Schafer (2001). Em seguida, são apresentadas as estratégias de recomendação do modelo, aprofundando-se na análise dos aspectos do sistema considerados durante a revisão da taxonomia. Após essa análise, cada estratégia será descrita em detalhes, incluindo considerações sobre a estratégia abordada. Por fim, discute-se o protótipo de visualização das recomendações na Brapci, utilizando as estratégias mencionadas.

As decisões que orientaram o desenvolvimento do modelo conceitual de recomendação foram fundamentadas na análise dos aspectos relacionados ao objeto de estudo - a Brapci, no referencial teórico e na avaliação dos modelos, assegurando que cada escolha contribuísse para atender às necessidades e objetivos da pesquisa. Dessa forma, os aspectos considerados definitivos para a construção do modelo foram previamente estabelecidos segundo a metodologia, com o objetivo de orientar tanto a elaboração do modelo conceitual

quanto das estratégias de recomendação. Nesse sentido, baseado na classificação de Schafer (2001), o quadro 7 resume algumas das características propostas para o modelo conceitual, que serão discutidas a seguir.

Quadro 7 – Decisões sobre o modelo conceitual de recomendação

Aspectos considerados	Decisões do modelo
Extração de informações do usuário	Mista (implícita e explícita)
Método de saída das informações	Sugestões ranqueadas
Grau de personalização	Misto (não personalizado a personalizado)
Modo de computação	Misto (*online* e *offline*)
Técnica de filtragem da informação	A depender da estratégia de recomendação
Método de recomendação	A depender da estratégia de recomendação

Fonte: Elaborado pelo autor (2024).

A respeito da extração de informações do usuário, esta será realizada de maneira mista. A coleta implícita ocorrerá por meio de sessões no navegador, permitindo ao sistema armazenar credenciais e dados de navegação sem exigir *login* direto na Brapci. Entre os dados armazenados, estão incluídos o ID da sessão, *timestamp* ou *log access* (registro de acessos), consultas realizadas, tempo de visualização da página e *downloads*. Já a coleta explícita incluirá o *login* do usuário e as avaliações dos itens, com a vantagem de armazenar dados de diferentes sessões acessadas. Os itens avaliados ficarão listados no perfil do usuário, permitindo a visualização e acesso aos documentos considerados pelo SR ao gerar as sugestões relacionadas à estratégia de recomendação de itens personalizados. A coleta de informações deverá ser precedida por uma política de privacidade do repositório.

Quanto ao método de saída, optou-se pelas sugestões ranqueadas. Estas serão formadas por ranques compostos por três sugestões de itens para cada uma das três estratégias de recomendação, organizadas e dispostas conforme a similaridade dos itens com o proposto pelo tipo de recomendação. O grau de personalização, segundo a classificação de Schafer (2001), pode variar entre não personalizado, efêmero ou personalizado, de acordo com a estratégia utilizada, visto que são apresentados diferentes técnicas e métodos.

O modo de computação dependerá do tipo de recomendação, podendo ser classificado como *online* ou *nearline* no caso das recomendações personalizadas, ou

parcialmente *offline* nas sugestões de itens relacionados e itens citados. As recomendações *online* apresentam como vantagem as atualizações em tempo real, enquanto as recomendações *offline* minimizam a possibilidade de gargalos no sistema. Vale destacar que o sistema poderá realizar o pré-processamento *offline* dos itens em períodos de menor tráfego ou em outras máquinas, garantindo escalabilidade e evitando problemas de desempenho.

A classificação das técnicas e métodos de recomendação, mencionada no quadro 7 varia conforme a estratégia de recomendação analisada. Portanto, as alíneas a seguir introduzem a proposta de segmentação do modelo conceitual de recomendação:

a) **itens relacionados**: sugere itens com conteúdo semelhante à consulta do usuário, ao item-ativo e aos itens previamente selecionados. Essa estratégia de recomendação, baseada na abordagem de FBC, utiliza atributos textuais dos itens do repositório e da *query* de consulta do usuário para recomendar itens afins;

b) **itens citados**: recomenda itens citados pelo item-ativo, utilizando também a abordagem de FBC. A estratégia considera as citações síncronas realizadas no texto do item-ativo, podendo ser aplicada de acordo com duas vertentes distintas, seja pela contagem de citações ou pela similaridade textual;

c) **itens personalizados**: estratégia fundamentada na expressão de busca e no histórico de leituras e avaliações dos usuários, que utiliza a abordagem de Filtragem Híbrida que combina a técnica de cascata com as técnicas de FC e FBC. Essa estratégia visa aumentar a relevância das recomendações, alinhando-se tanto às preferências individuais quanto ao comportamento coletivo dos usuários, resultando em uma experiência de busca mais personalizada e eficaz.

Em seguida, o quadro 8 apresenta as estratégias de recomendação recém descritas, assim como as classificações presentes no quadro anterior, permitindo a revisão dos aspectos que ainda não haviam sido aprofundados.

Quadro 8 – Estratégias de recomendação utilizadas na proposta do modelo

	Itens relacionados	**Itens citados**	**Itens personalizados**
Técnica de filtragem da informação	FBC	FBC	Híbrido
Método de filtragem	Vetorial, rede neural ou probabilístico	Resumo estatístico de citações ou vetorial	Cascata, aliado à FC e à FBC
Grau de personalização	Efêmero	Não personalizado	Personalizado
Modo de computação	*Offline*	*Offline*	*Online*

Fonte: Elaborado pelo autor (2024).

Como visto no quadro acima, foram escolhidas as abordagens de FBC nos modelos de itens relacionados e itens citados, enquanto a abordagem Híbrida (FC e FBC) foi selecionada para o modelo de itens personalizados. Entre os métodos de filtragem contemplados pelas abordagens mencionadas, há diversas possibilidades, embora apenas os métodos destacados na revisão do referencial teórico e na análise dos modelos tenham sido considerados.

Quanto ao grau de personalização, cada estratégia proposta revela seu próprio nível de personalização. As recomendações baseadas em itens relacionados sugerem publicações científicas alinhadas com a consulta atual e o histórico recente de consultas do usuário, sendo, portanto, consideradas efêmeras. Por outro lado, a estratégia de itens citados visa fornecer acesso aos trabalhos citados pelo item-ativo, e é considerada não personalizada, uma vez que suas recomendações independem do perfil do usuário. Já as recomendações personalizadas consideram critérios subjetivos, como gosto ou qualidade, e baseiam-se nas avaliações e no histórico de consultas do usuário, gerando sugestões coerentes com as consultas de outros pesquisadores. Essa estratégia é considerada personalizada por utilizar o perfil histórico de consultas dos usuários para oferecer novas sugestões de itens.

Como mencionado na revisão após o quadro 7, o modo computacional pode ser classificado de acordo com as abordagens de recomendação de cada estratégia. Recomendações *online* são adequadas para a estratégia de itens personalizados, pois permitem atualizações em tempo real, acumulando dados contínuos para oferecer sugestões

mais precisas. Em contraste, recomendações *offline,* aplicáveis aos itens relacionados e itens citados, são atualizadas em intervalos regulares, baseando-se em dados acumulados até o momento da última atualização, o que pode ser mais eficiente em termos de processamento, embora normalmente não ofereça a mesma agilidade em tempo real. Cabe, no entanto, ressaltar a importância da etapa de pré-processamento, que pode ser incluída no Processamento de Linguagem Natural (PLN) para vetorização de textos de documentos na base Brapci.

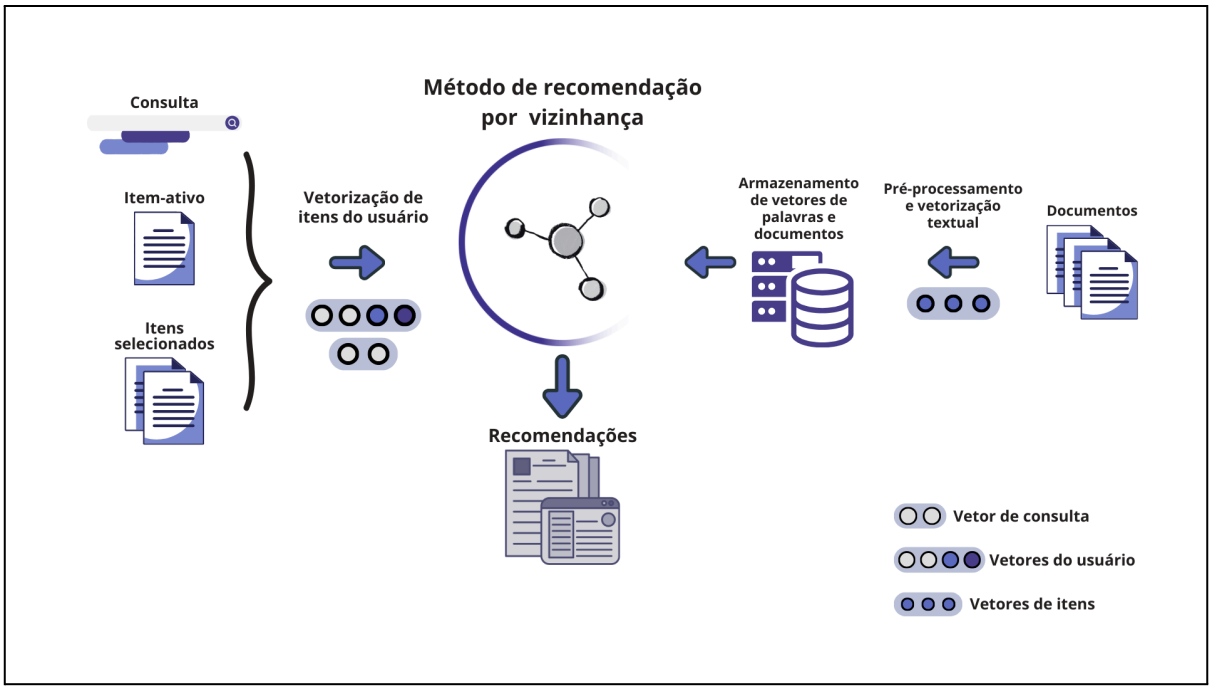
Dessa forma, as subseções a seguir descrevem cada uma das estratégias de recomendação do modelo proposto nesta pesquisa.

4.3.1 Estratégia de recomendação de itens relacionados

Na primeira estratégia de recomendação, o modelo de recomendação proposto para itens relacionados apoia-se na abordagem de FBC e métodos baseados em *embeddings*, utilizando dados de navegação da sessão atual do usuário, tais como a *query*, ou consulta empregada, o item-ativo e os itens selecionados durante a busca. O modelo pressupõe a vetorização textual destes atributos por uma técnica de PLN, que pode ser realizada por meio de abordagens vetoriais, probabilísticas ou de redes neurais. Esses vetores são então incorporados em conjunto em uma *embedding* única que corresponde aos atributos de itens do usuário considerados pelo método de recomendação.

O método de recomendação, embasado no modelo de vizinhança K-NN para atributos de itens, opera em duas etapas. Na primeira etapa, um algoritmo K-NN de aproximação designa os K-vizinhos (itens) mais próximos da consulta realizada pelo usuário. Na etapa seguinte, a consulta é novamente considerada, juntamente com o item-ativo e os itens selecionados pelo usuário. Essa divisão em etapas tem como objetivo limitar o número de artigos candidatos a vizinhos na última etapa, reduzindo o processamento computacional do algoritmo ao considerar uma maior quantidade de atributos em um número reduzido de documentos. Para isso, o algoritmo poderá calcular a similaridade entre os itens utilizando medidas tradicionais como o coeficiente de correlação de Pearson, cosseno ou a distância euclidiana. A figura 23 ilustra o processo descrito para o modelo de itens relacionados.

Figura 23 – Modelo conceitual de recomendação para itens relacionados



Fonte: Elaborado pelo autor (2024).

A estratégia de recomendação recém descrita pode ser vista como uma extensão do mecanismo de busca da Brapci, gerando recomendações de itens com conteúdo similar à consulta, ao item-ativo e itens selecionados durante a sessão do usuário. Dessarte, considera-se que o modelo conceitual proposto nesta estratégia atende aos critérios de recomendação por acoplamento, palavras-chave e autores, uma vez que esses atributos textuais podem ser vetorizados juntamente com os resumos e títulos de cada item. Quanto às técnicas de FBC, justificam-se pela eficiência proporcionada pelos algoritmos de PLN. Por último, o método de vizinhança K-NN é empregado para calcular a similaridade entre os vetores de itens do usuário e os documentos pré-processados.

Embora o algoritmo K-NN seja tradicionalmente conhecido como uma técnica de FC, ele pode ser eficazmente utilizado para a classificação de relevância e recomendação de textos quando combinado com algoritmos de PLN para a vetorização de itens em *embeddings*. Ademais, a facilidade e adaptabilidade de implantação do K-NN contribui para sua aplicação (Elastic, c2024b; Elastic, c2024c). No entanto, conforme mencionado na seção de referencial teórico (Elastic, c2024c), K-NN é computacionalmente intensivo, especialmente quando aplicado a grandes conjuntos de dados. Consequentemente, avaliar a performance do algoritmo será uma etapa crucial na implementação do modelo de recomendação. Além disso,

a exploração de algoritmos classificados como métodos baseados em modelos pode solucionar as dificuldades técnicas decorrentes da proposta deste modelo.

4.3.2 Estratégia de recomendação de itens citados

A estratégia proposta para gerar as recomendações de itens citados possui duas vertentes. A primeira abordagem fundamenta suas recomendações no número de citações presentes no item-ativo, partindo do pressuposto de que publicações com maior número de citações são mais relevantes para o item lido. Dessa forma, novos itens são sugeridos ao usuário com base na contagem pré-processada de citações síncronas incluídas no item atual, recomendando-se os itens que possuem maior número de citações.

Alternativamente, a segunda abordagem de recomendações deste tipo de recomendação adota FBC para incorporar a similaridade textual entre o texto lido e os textos citados, com a finalidade de selecionar os itens mais próximos do item lido. Nesse caso, o sistema recomendará trabalhos com níveis de conteúdo semântico mais próximos ao do item-ativo, por meio das técnicas e algoritmos de PLN, como TF-IDF, Word2Vec ou Doc2Vec. A seguir, o quadro 9 examina as vantagens e limitações de cada vertente apresentada.

Quadro 9 – Comparativo entre os métodos de recomendação para itens citados

Abordagem proposta	Vantagens	Desvantagens
Contagem de citações	Considera a relevância de outros estudos	Não considera contexto
	Fácil implantação	Dependência da qualidade de extração dos metadados
Similaridade textual	Considera a proximidade semântica	Não considera a relevância
	Precisão de conteúdo	Superespecialização

Fonte: Elaborado pelo autor (2024).

Apesar de não utilizar técnicas avançadas de análise do histórico de usuários ou itens, a estratégia proposta visa facilitar o acesso dos usuários às principais citações dos itens lidos, alinhando-se aos objetivos da pesquisa. Além disso, ambas as abordagens são

consideradas não personalizadas e, portanto, não precisam de atualizações recorrentes devido à simplicidade de manutenção do método de recomendação. O quadro 9 (acima) examina as vantagens e limitações de cada vertente apresentada.

Convém destacar que este modelo pode ainda aderir a outros critérios de similaridade, como a participação de um ou mais autores entre os trabalhos citados, a incorporação de citações diacrônicas (citações recebidas pelo texto observado) ao modelo, ou ser completamente substituído por uma lista das referências do item-ativo, permitindo a criação de outra estratégia de recomendação.

4.3.3 Estratégia de recomendação de itens personalizados

A estratégia de recomendação de itens personalizados foi desenvolvida a partir da abordagem híbrida de cascata que combina FC e FBC. Nessa modalidade, as recomendações são feitas com base em usuários que possuem interesses de pesquisa semelhantes aos do usuário-alvo. Para isso, o sistema reconhece como similares os usuários que, além de utilizarem consultas semelhantes, leem e avaliam itens de forma parecida.

No caso de usuários logados, o sistema considera todas as sessões de busca e outras informações do perfil coletadas ao longo do tempo, enquanto para usuários não logados, as recomendações são baseadas apenas nas informações da sessão atual.

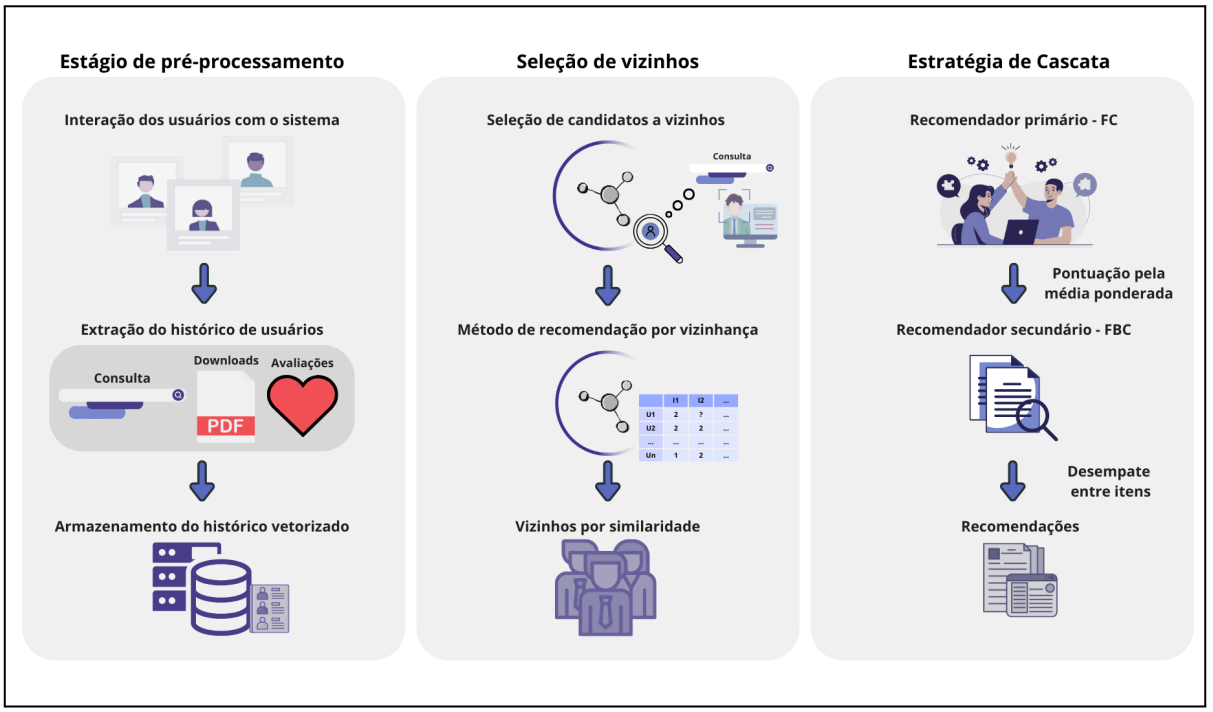
Para gerar as recomendações em tempo real, o sistema define as consultas realizadas ao longo das sessões como dados de entrada para o algoritmo de recomendação, selecionando apenas as sessões de usuários que tenham utilizado estratégias de busca semelhantes. Para tal, serão utilizadas técnicas de vetorização textual das consultas como quesito de similaridade. Além de vetores idênticos, a indexação exógena combinada com a abordagem semântica pode ser usada para identificar relações hierárquicas, associativas ou de sinonímia, permitindo novas maneiras de definir a similaridade entre os conceitos. Uma possibilidade prevista seria a integração com o índice de assuntos da Brapci e o tesauro semântico “Ciência da Informação” incluído no *Software* Thesa, para elaboração de tesauros.

Dessa forma, todas as *queries* são vetorizadas e armazenadas na base de dados para serem reutilizadas em novas recomendações. Essa abordagem reduz significativamente o número de usuários candidatos a vizinhos e o número de itens a serem considerados em uma matriz de entrada ao algoritmo K-NN.

A matriz é então organizada de maneira que os itens baixados pelos usuários recebem o valor um, indicando a leitura dos textos, enquanto os itens avaliados recebem o valor dois, sinalizando um nível mais alto de associação do usuário com o documento. Em seguida, o modelo calcula a similaridade entre o usuário-alvo e os candidatos a vizinhos usando o coeficiente de correlação de Pearson, tendo como alternativa o cosseno, a distância euclidiana ou *Log-likelihood*, visando a redução do custo computacional. Após isso, calcula-se a média ponderada das leituras e avaliações do usuário-alvo e de seus vizinhos para atribuir uma pontuação embasada no valor de predição dos documentos.

O tipo de hibridização definido neste algoritmo fundamenta-se na estratégia de cascata descrita por Burke (2007) e revisada por Vieira; Passo e Salm (2023). Assim, o algoritmo K-NN recém descrito foi escolhido como o recomendador primário, enquanto o recomendador secundário, responsável por desempatar a pontuação dos itens, recorre à técnica TF-IDF para calcular a similaridade textual entre a consulta e os documentos empatados. Essa estratégia permite ao sistema aperfeiçoar as pontuações dos itens com base na consulta atual, especialmente quando há poucos vizinhos disponíveis, e prioriza a sugestão de documentos alinhados com o histórico de consulta de outros usuários. A figura 24 ilustra o modelo proposto para esta estratégia.

Figura 24 – Proposta do modelo conceitual de recomendação personalizada



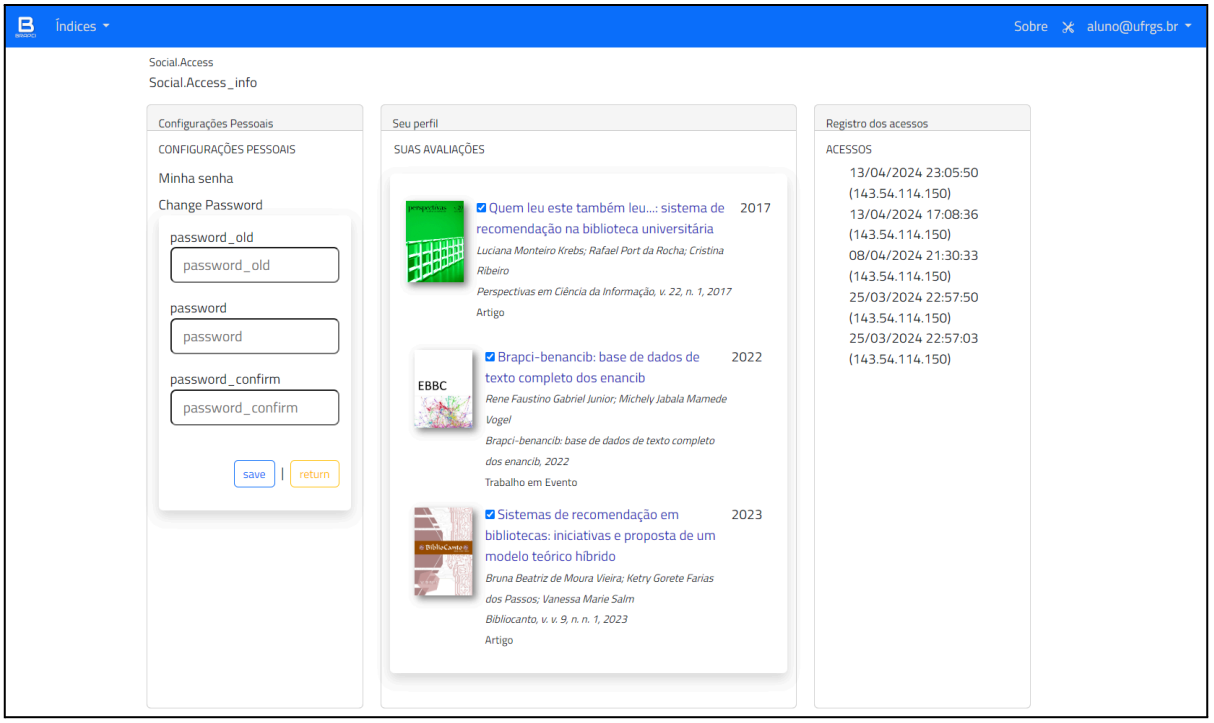
Fonte: Elaborado pelo autor (2024).

Diferente das outras duas estratégias, o tipo de recomendação de itens personalizados busca aprimorar a relevância e a precisão das sugestões por meio de um sistema híbrido, adaptado ao histórico de consultas e às informações do usuário. Ademais, esse tipo de recomendação busca incorporar o conceito de serendipidade nas recomendações de maneira mais intensa do que nas outras duas estratégias, ao considerar aspectos como o gosto e a relevância dada por usuários vizinhos, possibilitando o encontro de itens não esperados. Além disso, usuários logados recebem recomendações refinadas, uma vez que o sistema é capaz de reunir informações de múltiplas sessões em conjunto por meio do *login*.

Entretanto, na ausência de informações sobre o usuário-alvo ou de vizinhos, o sistema pode optar por uma estratégia alternativa, gerando recomendações com base apenas no item atual e seus metadados. Nesse caso, utiliza-se a similaridade textual em conjunto com a abordagem de FBC para elaborar as sugestões. Essa alternativa teria por finalidade solucionar o problema de *Cold-Start* em situações onde o acesso às páginas da Brapci ocorre por meio de redirecionamentos externos, como motores de busca, ou quando a consulta utilizada difere de todas as *queries* armazenadas no sistema.

Quanto às avaliações e ao perfil de usuários, a Brapci apresenta uma opção de *login* com uma página de acesso às informações pessoais, na qual os usuários podem gerenciar suas configurações de senha e consultar um registro detalhado dos acessos com as últimas sessões no sistema, porém não há um espaço definido para a interação com os documentos avaliados na base. Nesse sentido, o novo modelo visual proposto na figura 25 enriquece significativamente essa funcionalidade ao integrar novas informações sobre o perfil do usuário, informando quais itens foram avaliados em interações anteriores com o sistema. Destaca-se que o modelo ainda é um *wireframe* (rascunho), com a proposta de implementação.

Figura 25 – Modelo de visualização do perfil de usuário com avaliações (*wireframe*)



Fonte: Elaborado pelo autor (2024).

Desse modo, o usuário poderá analisar quais publicações científicas estão sendo consideradas nas recomendações e ajustar parcialmente seu perfil no sistema ao desmarcar avaliações realizadas anteriormente. Por fim, o próximo capítulo explora a proposta do modelo de visualização das recomendações no sistema, detalhando a exibição das estratégias discutidas nesta seção.

4.4 PROPOSTA DE VISUALIZAÇÃO DAS RECOMENDAÇÕES

Consoante ao modelo de recomendação apresentado, propõe-se que a visualização das recomendações seja organizada em listas com três sugestões para cada um dos três tipos de recomendação. Ao visualizar um item, essas sugestões serão exibidas ao usuário no lado direito da tela, sendo consideradas recomendações orgânicas por estarem integradas ao conteúdo da página (Schafer, 2001; Torres, 2004a). Essa abordagem permite que os itens sugeridos sejam visualizados sem interferir na leitura do conteúdo atual do documento e possibilita a interação com os itens recomendados conforme as necessidades de cada usuário. A figura 26 ilustra o *wireframe* do modelo visual de recomendações, revelando a disposição dos títulos de cada estratégia e as recomendações associadas à estratégia selecionada.

Figura 26 – Modelo proposto de visualização das recomendações (*wireframe*)

Fonte: Elaborado pelo autor (2024).

Assim como a anterior, a figura acima é considerada um rascunho que ilustra a organização das recomendações do modelo proposto para implementação na plataforma Brapci, oferecendo uma visão preliminar de como os itens sugeridos serão vistos pelos usuários. O *layout* (leiaute) pode ainda ser ajustado e refinado de acordo com as expectativas de uso futuras dos usuários e coordenação da Brapci. Por padrão, a página web exibe aos usuários as sugestões de itens relacionados, mas essa configuração pode ser reavaliada para priorizar a recomendação dos itens citados, com o objetivo de ocultar o processo computacional em tempo real das outras estratégias, que demandam maior capacidade de processamento.

Logo, o modelo de visualização das recomendações foi projetado para garantir uma interface clara e intuitiva, organizando as sugestões de itens conforme as estratégias estabelecidas. Além disso, o modelo visa proporcionar uma navegação fluida e integrada ao conteúdo da página, permitindo que os usuários interajam com as recomendações fornecidas pelo sistema, ao mesmo tempo em que podem descobrir itens inesperados. Por fim, o *wireframe* apresentado oferece uma visão preliminar do modelo, servindo como base para a implementação do sistema de recomendação.

5 CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo propor um modelo conceitual de recomendação de publicações científicas aplicável na Brapci. Para isso, foi realizado o levantamento bibliográfico em duas bases de dados nacionais, Brapci e OasisBR, sendo encontrados um total de 651 pesquisas em sistemas de recomendação.

Da pesquisa nas bases de dados, foram selecionados ao todo quarenta e seis estudos, dos quais dezessete fizeram parte do escopo de análise, incluindo SR com diferentes enfoques, como *e-commerce*, *streaming* de mídia, catálogos *online* e bibliotecas digitais. Entre os trabalhos analisados, foram descritos SRs baseados em quatro técnicas de filtragem da informação: FC, FBC, Filtragem Híbrida e Filtragem Semântica.

Na avaliação desses dezessete modelos, foram revisados, à luz do referencial teórico, o estado da arte e as principais técnicas, métodos e modelos de recomendação presentes na literatura. A análise considerou aspectos fundamentais das quatro técnicas de filtragem da informação supracitadas, contemplando suas vantagens e limitações e a compatibilidade destes com as recomendações por acoplamento, palavras-chave e autores, fornecendo a base para o desenvolvimento de um modelo conceitual de recomendação.

O modelo de recomendação proposto foi segmentado em três estratégias de recomendação distintas, com a intenção de abordar diferentes aspectos na geração de sugestões aos usuários. As estratégias foram subdivididas em itens relacionados, itens citados e itens personalizados.

No primeiro tipo de recomendação, utiliza-se a técnica FBC para gerar recomendações alinhadas com as consultas formuladas por usuários, considerando o item-ativo e os itens selecionados durante a sessão de navegação.

Na segunda proposta, propôs-se duas abordagens: a primeira, não personalizada, pressupõe que os itens mais citados pelo item pesquisado são os mais relevantes; enquanto a segunda considera a similaridade textual entre o artigo lido e os itens citados como critério de recomendação dos textos.

A terceira e última estratégia, de itens personalizados, apoia-se no uso de Filtragem Híbrida para gerar recomendações mais precisas e relevantes. O método proposto inclui o uso das técnicas de FC e FBC, aliados à técnica de cascata. O recomendador primário, baseado em FC, é responsável por gerar vizinhanças a partir das consultas, leituras e avaliações dos

usuários, atribuindo as pontuações iniciais dos documentos. Já o recomendador secundário, sustentado pela técnica de FBC, realiza o desempate das pontuações dos itens recomendados.

Além disso, o modelo conceitual proposto inclui a categorização dos principais aspectos considerados na sua construção. Foram propostas maneiras para realizar a extração de informações do usuário, o método de saída das informações, o grau de personalização e o modo de computação das recomendações. Também foram criados modelos visuais para a visualização das recomendações e dos itens avaliados no perfil do usuário na Brapci.

Entre as limitações do estudo, destaca-se a ausência de literatura internacional na aplicação da metodologia e estratégias de buscas simples. Embora a pesquisa tenha considerado o contexto de outros países, não foi realizada uma busca em bases de dados internacionais. A inclusão de fontes internacionais, assim como de outros termos de busca, poderia ter ampliado a compreensão e enriquecido a análise de conteúdo, oferecendo uma perspectiva mais abrangente e diversificada.

Portanto, para estudos futuros, vislumbra-se a possibilidade de realizar investigações tanto de natureza básica quanto aplicada. As pesquisas básicas podem ser realizadas por meio de revisões bibliográficas além da literatura nacional e utilizando estratégias de busca mais sofisticadas, incluindo termos relacionados às técnicas de filtragem. Nos estudos de natureza aplicada, espera-se que o modelo proposto nesta pesquisa possa ser implementado e avaliado quantitativamente, por intermédio de métricas de desempenho como acurácia, precisão, revocação, *F1-score* e custo computacional das recomendações, e qualitativamente, com o auxílio das avaliações de opinião *online* de usuários. Além disso, considerando as discussões acerca da privacidade e proteção de dados dos usuários, visualiza-se a possibilidade de estudo dos critérios necessários para a elaboração de uma política de privacidade aplicável ao contexto da Brapci, bem como explorar novas abordagens de *design* centradas no usuário para aprimoramento do modelo de recomendação.

Por fim, acredita-se que os resultados encontrados neste trabalho, desde o levantamento bibliográfico até a análise dos sistemas de recomendação sob a perspectiva da pesquisa em CI, bem como a concepção do modelo conceitual de recomendação, atendam aos objetivos estabelecidos nas seções anteriores. Espera-se que o esforço despendido nesta pesquisa sirva como base para futuras investigações sobre os SR na CI, podendo contribuir tanto para a construção quanto para a avaliação destes em repositórios digitais e outras bases de dados da área, além dos catálogos de bibliotecas, e principalmente para a Brapci.

REFERÊNCIAS

ALEIXO, Everton Lima. **Item-based-ADP**: análise e melhoramento do algoritmo de filtragem colaborativa item-based. Orientador: Thierson Couto Rosa. 2014. 96 f. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Goiás, Goiânia, 2014. Disponível em: https://repositorio.bc.ufg.br/tede/items/6ee71a88-1ea6-4df7-a76f-f50a6919e766. Acesso em: 21 jul. 2024.

ALVAREZ, Edgar Bisset; SIRIANI, Allan Lincoln Rodrigues; VIDOTTI, Silvana Aparecida Borsetti Gregorio; CARVALHO, Angela Maria Grossi de. Os sistemas de recomendação, arquitetura da informação e a encontrabilidade da informação. **Informação & Tecnologia**, Campinas, v. 28, n. 3, p. 275-286, 2016. Disponível em: https://doi.org/10.1590/2318-08892016000300003. Acesso em: 24 jan. 2024.

AMAZON. Amazon Web Services. **O que é uma API RESTful?**. Seattle, c2023. Disponível em: https://aws.amazon.com/pt/what-is/restful-api/. Acesso em: 27 jan. 2024.

ASSOCIATION FOR COMPUTER MACHINERY. RecSys. **The ACM Conference Series on Recommender Systems**. Nova Iorque, c2024. Disponível em: https://recsys.acm.org/. Acesso em: 16 jan. 2024.

BAX, Marcello Peixoto. Introdução às linguagens de marcas. **Ciência da Informação**, v. 30, n. 1, p. 32-38, jan./abr. 2001. Disponível em: https://brapci.inf.br/index.php/res/v/17684. Acesso em: 24 ago. 2024.

BEISKE, Konrad. Similaridade no Elasticsearch. *In*: ELASTIC. **Blogue**. Mountain View, CA, 26 nov. 2013. Disponível em: https://www.elastic.co/blog/found-similarity-in-elasticsearch. Acesso em: 27. jan. 2024.

BRAPCI. **Sobre a Brapci**. Porto Alegre, c2024. Repositório digital. Disponível em: https://cip.brapci.inf.br/about. Acesso em: 5 jan. 2024.

BUFREM, Leilah Santiago; COSTA, Francisco Daniel de Oliveira; , GABRIEL JUNIOR, Rene Faustino; PINTO, José Simão de Paula. Modelizando práticas para a socialização de informações: a construção de saberes no ensino superior. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 15, n. 2, p. 22–41, maio/ago. 2010. Disponível em: https://periodicos.ufmg.br/index.php/pci/article/view/23631. Acesso em: 27 jan. 2024.

BUFREM, Leilah Santiago; GABRIEL JUNIOR, Rene Faustino. Da BRES à BRAPCI: memória e construção social da Base de Artigos de Periódicos em Ciência da Informação (Brapci). *In:* ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 22., 2022, Porto Alegre. **Anais** [...]. Porto Alegre: UFRGS, 2022. ISSN: 2177-3688. Disponível em: https://enancib.ancib.org/index.php/enancib/xxiienancib/paper/view/1180. Acesso em: 10 jan. 2024.

BUFREM, Leilah Santiago. Práticas de organização e divulgação da produção intelectual em Ciência da Informação no Brasil. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 13, n. 1, p. 36–53, 2008. Disponível em: https://doi.org/10.5007/1518-2924.2008v13nesp1p36. Acesso em: 23 jul. 2024.

BURKE, Robin. Hybrid Web Recommender Systems. *In*: BRUSILOVSKY, Peter; KOBSA, Alfred; NEJDL, Wolfgang. **The Adaptive Web**: methods and strategies of web personalization. Berlim: Springer, 2007. (Lecture Notes in Computer Science, v. 4321). Disponível em: https://doi.org/10.1007/978-3-540-72079-9_12. Acesso em: 26 jan. 2024.

CASAD, Joe; WILLSEY, Bob. **Aprenda em 24 horas TCP/IP**. Rio de Janeiro: Campus, 1999.

CAXIAS, Rodrigo Silva. Das tecnologias da informação à comunicação científica: críticas à nova cultura da pesquisa em Educação. **Em Questão**, Porto Alegre, v. 14, n. 2, p. 301–315, 2009. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/6470. Acesso em: 28 jan. 2024.

CESARINO, Maria Augusta da Nóbrega. Sistemas de recuperação da informação. **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v. 14, n. 2, p. 157-168, 1985. Disponível em: https://periodicos.ufmg.br/index.php/reb/article/view/36507. Acesso em: 23 jan. 2024.

CONCEIÇÃO, Felipe Leandro Andrade da; PÁDUA, Flávio Luis Cardeal; MACHADO, Adriano César; LACERDA, Anísio Mendes; DALIP, Daniel Hasan. Metodologia para recomendação de vídeos baseada em descritores de conteúdo visuais e textuais. **Tendências da Pesquisa Brasileira em Ciência da Informação**, [*S. l.*], v. 9, n.1, p. 208-225, jan./ago. 2016. Disponível em: https://revistas.ancib.org/index.php/tpbci/article/view/382. Acesso em: 20 jul. 2024.

COSTA, Júlio Barreto Guedes da. **Natural language processing techniques for session based recommendation**. Orientador: Leandro Balby Marinho. 2020. 13 f. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) - Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande, Campina Grande, 2020. Disponível em: http://dspace.sti.ufcg.edu.br:8080/jspui/handle/riufcg/20337. Acesso em: 20 jul. 2024.

CUNHA, Gil Fernando Ferreira da. **Data analysis and recommender system architecture for e-commerce platforms**. Orientadores: Hugo Daniel Abreu Peixoto e José Manuel Ferreira Machado. 2021. 123 f. Dissertação (Mestrado em Engenharia Informática) - Escola de Engenharia, Universidade do Minho, Braga, 2021.
Disponível em: https://repositorium.sdum.uminho.pt/handle/1822/81087. Acesso em: 21 jul. 2024.

DÍAZ-AVILÉS, Vladimir Ernesto. **Semantic peer-to-peer recommender systems**. Orientador: Lars Schmidt-Thieme. 2005. 112 f. Dissertação (Mestrado em Ciências da Computação) - Institute of Computer Science, Albert Ludwigs University of freiburg, Freiburg, FR, 2005. Disponível em: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=71c49ffdbb1228a9727955e2f1812498fc0e5873. Acesso em: 9 ago. 2024.

DUBLIN CORE METADATA INITIATIVE. **DCMI metadata terms**. [*S. l.*], c2024. Disponível em: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#section-1. Acesso em 24 jul. 2024.

ELASTIC. **Elasticsearch**: o coração do Elastic Stack gratuito e aberto. Mountain View, CA, c2024a. Disponível em: https://www.elastic.co/pt/elasticsearch. Acesso em: 27 jan. 2024.

ELASTIC. **K-nearest neighbor (kNN) search**. Mountain View, CA, c2024b. Disponível em: https://www.elastic.co/guide/en/elasticsearch/reference/current/knn-search.html. Acesso em: 8 ago. 2024.

ELASTIC. **O que é kNN?**. Mountain View, CA, c2024c. Disponível em: https://www.elastic.co/pt/what-is/knn. Acesso em: 8 ago. 2024.

FRANCO, Nimrod González; SANCHEZ, Hugo Omar Alejandres; SERNA, Juan Gabriel González. Arquitectura de un sistema de recomendación semántico sensible al contexto para entornos tipo campus. **Ciencias de la Información** (Cuba), Havana, v. 46, n. 1, 2015. Disponível em: https://brapci.inf.br/#/v/58854, Acesso em: 20 jul. 2024.

FREITAS, Lazzarotto Juliana; BUFREM, Santiago Leilah; GABRIEL JUNIOR, Rene Faustino. Proposta de metodologia para a recuperação da produção científica em Ciência da Informação na base Brapci. **PontodeAcesso**, Salvador, v. 4, n. 3, p. 45–67, dez. 2010. Disponível em: https://periodicos.ufba.br/index.php/revistaici/article/view/4629. Acesso em: 17 jul. 2024.

FURTADO, Thiago Bellotti. **Abordagem híbrida de recomendação de conteúdo baseado em tags adaptativas aplicada em bibliotecas digitais**. Orientador: Ahmed Ali Abdalla Esmin. 2016. 93 p. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Lavras, Lavras, 2016. Disponível em: http://repositorio.ufla.br/jspui/handle/1/31821. Acesso em: 21 jul. 2024.

GABRIEL JUNIOR, Rene Faustino. Aproximação da bibliometria e recuperação de informação na Brapci. *In*: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 4., 2014, Recife. **Anais** [...]. Recife: UFPE, 2014a. ISSN: 2675-5939. Disponível em: http://hdl.handle.net/20.500.11959/brapci/45770. Acesso em: 27 nov. 2023.

GABRIEL JUNIOR, Rene Faustino. **Geração de indicadores de produção e citação científica em revistas de Ciência da Informação**: estudo aplicado à base de dados BRAPCI. Orientadora: Ely Francina Tannuri de Oliveira. 2014. 145 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2014b. Disponível em: http://hdl.handle.net/11449/123338. Acesso em: 21 jul. 2024.

GARVEY, William. D.; GRIFFITH, Belver. C. Communication and information processing within scientific disciplines: empirical findings for psychology. **Information Storage and Retrieval**, [*S. l.*], v. 8, p. 123-126, jun. 1972.

GEMMIS, Marco de; LOPS, Pasquale; MUSTO, Cataldo; NARDUCCI, Fedelucio; SEMERARO, Giovanni. Semantics-Aware Content-Based Recommender Systems. *In*: Ricci, Francesco; ROKACH, Lior; SHAPIRA, Bracha. **Recommender Systems Handbook**. Boston: Springer, 2015. Capítulo 4. p 119-159. Disponível em: https://doi.org/10.1007/978-1-4899-7637-6_4. Acesso em: 26 jan. 2023.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

GODOY, Arilda Schmidt. Pesquisa qualitativa: tipos fundamentais. **RAE - Revista de Administração de Empresas**, São Paulo, v. 35, n. 3, p. 20-29, maio/jun. 1995.

GOLBECK, Jennifer. **Semantic web interaction through trust network recommender systems**. 2005. Disponível em: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=ac72ab4d4f0ed9d8942336100d5402189ff6a315. Acesso em: 8 ago. 2024.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining**: Concepts and Techniques. 3. ed. San Francisco, CA: Morgan Kaufmann, 2011.

HURD, Julie M. Models of Scientific Communications Systems. *In*: CROWFORD, Susan Y.; HURD, Julie M.; WELLER, Ann C. (Orgs.). **From Print to Electronic**: the transformation of scientific communication. Medford: ASIS, 1996. p. 9-33.

JUNG, Kwanho; HWANG, Myunggwun; KONG, Hyunjang; KIM, Pankoo. RDF triple processing methodology for the recommendation system using personal information *In*: INTERNATIONAL CONFERENCE ON NEXT GENERATION WEB SERVICES PRACTICES, 5., 2005, Seul. **Anais** [...]. Seul: Institute of Electrical and Electronic Engineers, 2005. p. 1-6. ISBN: 0-7695-2452-4. Disponível em: https://doi.org/10.1109/NWESP.2005.66. Acesso em: 8 ago. 2024.

KRUK, Sebastian Ryszard; DECKER, Stefan. Semantic Social Collaborative Filtering with FOAFRealm. *In*: INTERNATIONAL SEMANTIC WEB CONFERENCE, 4., 2005, Galway. **Anais** [...]. Galway: Springer, 2005. Disponível em: https://researchrepository.universityofgalway.ie/server/api/core/bitstreams/0cee8f34-77c9-4da4-a8cf-dbc7d0ba27a7/content. Acesso em: 09 ago. 2024.

LASSILA, Ora; SWICK, Ralph R. **Resource Description Framework (RDF) model and syntax specification**. Cambridge, US-MA: W3C, 1999. Disponível em: http://www.w3.org/TR/1999/REC-rdf-syntax-19990222. Acesso em: 24 jul. 2024.

LE COADIC, Yves-François. **A Ciência da Informação.** Tradução Maria Yêda Falcão Soares de Filgueiras Gomes. 2. ed. Brasília, DF: Briquet de Lemos, 2004.

LINDEN, Greg; SMITH, Brent; York, J. Amazon.com recommendations: item-to-item collaborative filtering. **IEEE Internet Computing**, Washington, DC, v. 7, n. 1, p. 76–80, 2003. Disponível em: https://doi.org/10.1109/MIC.2003.1167344. Acesso em: 8 jul. 2024.

LOIZOU, Antonis; DASMAHAPATRA, Srinandan. **Recommender systems for the semantic web**. 2006. Disponível em: http://eprints.soton.ac.uk/id/eprint/262584. Acesso em: 9 ago. 2024.

LOPES, Giseli Rabello. **Sistema de recomendação para bibliotecas digitais sob a perspectiva da web semântica**. Orientadora: Maria Aparecida Martins Souto. 2007. 69 f. Dissertação (Mestrado em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2007. Disponível em: http://hdl.handle.net/10183/10747. Acesso em: 20 jul. 2024.

MASSA, Paolo; AVESANI, Paolo. Trust-Aware Collaborative Filtering for Recommender Systems. *In*: MEERSMAN, Robert; TARI, Zahir. **On the Move to Meaningful Internet Systems 2004**: CoopIS, DOA, and ODBASE. Berlim: Springer, 2004. (Lecture Notes in Computer Science, v. 3290). Disponível em: https://doi.org/10.1007/978-3-540-30468-5_31. Acesso em: 9 ago. 2024.

MELO, Wesley Soares de; OLIVEIRA, Paulo Jorge Ferreira de; MONTEIRO, Flávia Paula Magalhães; SANTOS, Francisca Carla dos Angelos; SILVA, Maria Janaína Nogueira da; CALDERON, Carolina Jimenez; FONSECA, Lilian Nara Amaral da; SIMÃO, Ana Adélia Chaves. Guia de atributos da competência política do enfermeiro: estudo metodológico. **Revista Brasileira de Enfermagem**, v. 70, n. 3, p. 526–534, maio/jun. 2017. Disponível em: https://doi.org/10.1590/0034-7167-2016-0483. Acesso em: 29 jan. 2024.

MELVILLE, Prem; SINDHWANI, Vikas. Recommender Systems. *In*: SAMMUT, Claude; WEBB, Geoffrey. (ed.) **Encyclopedia of Machine Learning**. Springer: Boston, MA. 2011. p. 829-838. Disponível em: https://doi.org/10.1007/978-0-387-30164-8_705. Acesso em: 16 jan. 2024.

MIDDLETON, Stuart Edward; ALANI, Harith; SHADBOLT, Nigel R.; ROURE, David Charles de. Exploiting synergy between ontologies and recommender systems. *In*: SEMANTIC WEB WORKSHOP, 3.; INTERNATIONAL WORLD WIDE WEB CONFERENCE, 11., 7-11 Maio 2002, Havaí. **Anais** [...]. Havaí: Association for Computing Machinery, 2002. ISBN: 978-1-58113-449-0. Disponível em: https://oro.open.ac.uk/20058/1/www-paper.pdf. Acesso em: 8 ago. 2024.

MONTEIRO-KREBS, Luciana; ALVARADO RODRIGUEZ, Oscar Luis; DEWITTE, Pierre; AUSLOOS, Jef; GEERTS, David; NAUDTS, Laurens; VERBERT, Katrien. Tell me what you know: GDPR implications on designing transparency and accountability for news recommender systems. *In*: CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 19., 2019, Glasgow. **Proceedings** [...]. Nova Iorque: Association for Computing Machinery, 2019. p. 1-6. Disponível em: https://doi.org/10.1145/3290607.3312808. Acesso em: 30 ago. 2024.

MONTEIRO-KREBS, Luciana; ROCHA, Rafael Port da; RIBEIRO, Cristina. Quem leu este também leu…: sistema de recomendação na biblioteca universitária. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 22, n. 1, p. 151-169, jan./mar. 2017. Disponível em: https://doi.org/10.1590/1981-5344/2496. Acesso em: 21 ago. 2024.

MONTEIRO-KREBS, Luciana. **Sistemas de recomendação para bibliotecas universitárias**. Orientador: Rafael Port da Rocha. Coorientadora: Maria Cristina de Carvalho Alves Ribeiro. 2013. 95 f. Trabalho de Conclusão de Curso (Bacharelado em Biblioteconomia) - Faculdade de Biblioteconomia e Comunicação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013. Disponível em: http://hdl.handle.net/10183/78367. Acesso em: 21 jul. 2024.

MONTEIRO-KREBS, Luciana; ZAMAN, Bieke; CAREGNATO, Sônia Elisa; GEERTS, David; GRASSI-FILHO, Vicente; HTUN, Nyi-Nyi. Trespassing the gates of research: identifying algorithmic mechanisms that can cause distortions and biases in academic social media. **Online Information Review**, Dublin, v. 46, n. 5, p. 993-1013, dez. 2022. Disponível em: https://doi.org/10.1108/OIR-01-2021-0042. Acesso em: 28 ago. 2024.

MONTEIRO-KREBS, Luciana; ZAMAN, Bieke; GEERTS, David; CAREGNATO, Sônia Elisa. Every word you say: algorithmic mediation and implications of data-driven scholarly communication. **AI & Soc**, Londres, v. 38, p. 1003–1012, maio 2023. Disponível em: https://doi.org/10.1007/s00146-022-01468-1. Acesso em: 28 ago. 2024.

MONTEIRO-KREBS, Luciana; ZAMAN, Bieke; HTUN, Nyi-Nyi; CAREGNATO, Sônia Elisa; GEERTS, David. Depicting recommendations in academia: how research gate communicates with Its users (via design or upon request) about recommender algorithms. *In*: EAI INTERNATIONAL CONFERENCE, 2., 2021, Switzerland. **Proceedings** [...]. Gewerbestrasse: Springer, 2021. p. 3-25. Disponível em: https://doi.org/10.1007/978-3-030-77417-2_1. Acesso em: 29 ago. 2024.

MORAES, Roque. Análise de conteúdo. **Revista Educação**, Porto Alegre, v. 22, n. 37, p. 1-12, 1999.

MOURÃO, Fernando Henrique de Jesus. **A hybrid recommendation method that combines forgotten items and non-content attributes**. Orientador: Wagner Meira Júnior. 2014. 102 f. Tese (Doutorado em Ciência da Computação) - Universidade Federal de Minas Gerais, Belo Horizonte, 2014. Disponível em: https://repositorio.ufmg.br/handle/1843/ESBF-9TELK3. Acesso em: 21 jul. 2024.

NEVES, Carolina Domingos. **Um sistema de recomendação baseado em conteúdo para enfrentar o desafio do Cold Start na área de informação científica e técnica**. Orientador: Fernando José Ferreira Lucas Bação. 2022. 41 f. Dissertação (Mestrado em Ciência de Dados e Métodos Analíticos Avançados) - Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa, 2022. Disponível em: http://hdl.handle.net/10362/142639. Acesso em: 21 jul. 2024.

NÓBREGA, Caio Santos Bezerra. **Uma estratégia para predição da taxa de aprendizagem do gradiente descendente para aceleração da fatoração de matrizes**. Orientador: Leandro Balby Marinho. 2014. 76 f. Dissertação (Mestrado em Ciência da Computação) - Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande, Campina Grande, 2014. Disponível em: http://dspace.sti.ufcg.edu.br:8080/jspui/handle/riufcg/362. Acesso em: 21 jul. 2024.

OLIVEIRA, Cleiane Gonçalves. **PrefREC**: uma metodologia para desenvolvimento de sistemas de recomendação utilizando algoritmos de mineração de preferências. Orientadora: Sandra Aparecida de Amo. 2014. 100 f. Dissertação (Mestrado em Ciência da Computação) - Faculdade de Ciência da Computação, Universidade Federal de Uberlândia, Uberlândia, 2014. Disponível em: https://repositorio.ufu.br/handle/123456789/12550. Acesso em: 21 jul. 2024.

PEIS, Eduardo; MORALES-DEL-CASTILLO, Jose. M.; DELGADO-LOPEZ, Juan A. Semantic recommender systems. analysis of the state of the topic. **Hipertext.net** (Espanha), Barcelona, n. 6, 2008. Disponível em: https://arxiu-web.upf.edu/hipertextnet/numero-6/recomendacion.html. Acesso em: 21 jul. 2024.

RICCI, Francesco; ROKACH, Lior; SHAPIRA, Bracha. Recommender systems: introduction and challenges. *In*: Ricci, Francesco; ROKACH, Lior; SHAPIRA, Bracha (eds.). **Recommender Systems Handbook**. Boston: Springer, 2015. Capítulo 1. p 1-34. Disponível em: https://doi.org/10.1007/978-1-4899-7637-6_1. Acesso em: 25 jan. 2023.

SARACEVIC, Tefko. Information Science. **Journal of the American Society for Information Science**, v. 50, n. 12, p. 1051-1063, 1999. Disponível em: https://comminfo.rutgers.edu/~tefko/JASIS1999.pdf. Acesso em: 16 jan. 2024.

SAYÃO, Luís Fernando. Bases de dados: a metáfora da memória científica. **Ciência da Informação**, Brasília, DF, v. 25, n. 3, p. 314-318, 1996. Disponível em: http://revista.ibict.br/ciinf/article/view/629/633. Acesso em: 14 jan. 2024.

SCHAFER, John Benjamin. **MetaLens**: a framework for multi-source recommendations. Coorientadores: Joseph A. Konstan e John T. Riedl. 2001. 192 f. Tese (Doutorado em Filosofia) - Faculty of the Graduate School, University of Minnesota, Minnesota, 2001. Disponível em: https://www.cs.uni.edu/~schafer/research.html. Acesso em: 20 jan. 2024.

SILVA, Jossandro Balardin; SCHREIBER, Jacques Nelson Corleta; NARA, Elpídio Oscar Benitez. Bayesian approach to news recommendation systems. **Ciência da Informação,** Brasília, DF, v. 44, n. 3, p. 416-429, set./dez. 2015. Disponível em: https://brapci.inf.br/#/v/21894. Acesso em: 21 jul. 2024.

SILVA, Renata Eleutério da; SANTOS, Plácida Leopoldina Ventura Amorim da Costa; FERNEDA, Edberto. Modelos de recuperação de informação e web semântica: a questão da relevância. **Inf. Inf.**, Londrina, v. 18, n. 3, p. 27-44, set./dez. 2013. Disponível em: https://doi.org/10.5433/1981-8920.2013v18n3p27. Acesso em: 23 jan. 2024.

SILVEIRA, Denise Tolfo; CÓRDOVA, Fernanda Peixoto. Unidade 2 - A pesquisa científica. *In*: GERHARDT, Tatiana Engel; SILVEIRA, Denise Tolfo (org.). **Métodos de pesquisa**. Porto Alegre: Editora da UFRGS, 2009.

SINHA, Bam Bahadur; DHANALAKSHMI, R. Evolution of recommender paradigm optimization over time. **Journal of King Saud University - Computer and Information Sciences**, Riade, v. 34, n. 4, p. 1047-1059, 2022. Disponível em: https://doi.org/10.1016/j.jksuci.2019.06.008. Acesso em: 2 jul. 2024.

SOUZA, Gabriel Justino de; LIMA, Vânia Mara Alves. As plataformas de streaming e seus sistemas de recomendação. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 22., 2022, Porto Alegre. **Anais** [...]. Porto Alegre: UFRGS, 2022. ISSN: 2177-3688. Disponível em: http://hdl.handle.net/20.500.11959/brapci/207204. Acesso em: 25 out. 2023.

SOUZA, Gabriel Rafael Guedine de Jesus; FEITOSA, Lucas Leão. **Recomendação e visualização de conjuntos similares de artigos com base na análise e classificação de seus resumos**. 2022. 29 f. Trabalho de Conclusão de Curso (Graduação em Sistemas de Informação) - Universidade Federal Fluminense, Niterói, 2022. Disponível em: http://app.uff.br/riuff/handle/1/25534. Acesso em: 21 jul. 2024.

STORMS, Elias; ALVARADO RODRIGUES, Oscar Luis; MONTEIRO-KREBS, Luciana. 'Transparency is meant for control' and vice versa: learning from co-designing and evaluating algorithmic news recommenders. **Proceedings of the ACM on Human-Computer Interaction**, v. 6, n. CSCW2, t. 405, p. 1-24 , nov. 2022. Disponível em: https://doi.org/10.1145/3555130. Acesso em: 30 ago. 2024.

SZOMSZOR, Martin; CATTUTO, Ciro; ALANI, Harith; O'HARA, Kieron; BALDASSARRI, Andrea; LORETO, Vittorio; SERVEDIO, Vito. Folksonomies, the semantic web, and movie recommendation. *In*: EUROPEAN SEMANTIC WEB CONFERENCE, 4., 2007, Innsbruck. **Anais** [...]. Innsbruck: Springer, 2007. Disponível em: http://www.kde.cs.uni-kassel.de/ws/eswc2007/proc/Folksonomies.pdf. Acesso em: 9 ago. 2024.

TORRES, Roberto Dias. **Personalização na Internet**: como descobrir os hábitos de consumo dos seus clientes, fidelizá-los e aumentar o lucro de seu negócio. São Paulo: Novatec, 2004a.

TORRES, Roberto Dias. **Sistema de recomendação para bibliotecas digitais sob a perspectiva da web semântica**. Orientadora: Mara Abel. Co-orientador: John Riedl. 2004. 66 f. Dissertação (Mestrado em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2004b. Disponível em: http://hdl.handle.net/10183/5887. Acesso em: 21 jul. 2024.

VIEIRA, Bruna Beatriz de Moura; PASSOS, Ketry Gorete Farias dos; SALM, Vanessa Marie. Sistemas de recomendação em bibliotecas: iniciativas e proposta de um modelo teórico híbrido. **BiblioCanto**, Natal, v. 9, n. 1, p. 43 – 66, 2023. Disponível em: https://doi.org/10.21680/2447-7842.2023v9n1ID32504. Acesso em: 19 jan. 2024.

WEITZEL, Simone da Rocha. O papel dos repositórios institucionais e temáticos na estrutura da produção científica. **Em Questão**, Porto Alegre, v. 12, n. 1, p. 51–71, 2006. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/19. Acesso em: 28 jan. 2024.

WIKIPEDIA CONTRIBUTORS. FOAF. *In*: WIKIPEDIA: the free encyclopedia [San Francisco, CA: Wikimedia Foundation], 18 jul. 2023. Online. Disponível em: https://en.wikipedia.org/w/index.php?title=FOAF&oldid=1165941964. Acesso em: 26 Jul. 2024.

WIKIPEDIA CONTRIBUTORS. Netflix Prize. *In*: WIKIPEDIA: the free encyclopedia [San Francisco, CA: Wikimedia Foundation], 28 Feb. 2024. Online. Disponível em: https://en.wikipedia.org/w/index.php?title=Netflix_Prize&oldid=1210733068. Acesso em: 24 Jul. 2024.

W3C CAPÍTULO SÃO PAULO. **Web semântica**. [202?]. Online. Disponível em: https://www.w3c.br/padroes/web-semantica. Acesso em: 24 jul. 2024.