Predição de Surtos de Dengue a Partir de Fatores Climáticos e Socioeconômicos: Uma Abordagem KDD nos Municípios do Brasil

Glenda Maria Oliveira de Oliveira
Universidade Federal do Pará (UFPA)
Belém, Pará, Brasil

Tobias Moraes de Souza
Centro Universitário Estácio de Belém
Belém, Pará, Brasil

Wendia Oliveira de Andrade
Universidade Federal do Pará (UFPA)
Belém, Pará, Brasil

Marcos César da Rocha Seruffo
Universidade Federal do Pará (UFPA)
Belém, Pará, Brasil

Diego Lisboa Cardoso
Universidade Federal do Pará (UFPA)
Belém, Pará, Brasil

Frederico Guilherme Santana da Silva Filho
Centro Universitário Estácio de Belém
Belém, Pará, Brasil

Abstract

The exponential growth of digital libraries poses a challenge to information retrieval, often resulting in information overload and siloed reading. This work proposes a multidisciplinary framework that bridges Library Science and Computer Engineering to enhance scientific paper recommendation. By integrating controlled vocabulary-based semantic analysis with graph-based centrality metrics, the proposed solution addresses traditional limitations such as the cold-start problem and lack of diversity. The framework utilizes a hybrid filtering approach, combining content-based and collaborative methods with advanced bibliographic heuristics. This integration aims to facilitate interdisciplinary knowledge transfer and improve the precision of literature reviews in STEM fields. This framework addresses the limitations of traditional approaches by incorporating semantic analysis and graph-based metrics, showing superior accuracy (89.05%) with hybrid metadata and robust semantic coverage for cold-start scenarios. The inclusion of ontologies and contextual metadata significantly improves collaborative filtering, reinforcing the importance of semantics for performance.

Keywords

Sistemas de Recomendação, Ciência da Informação, Filtragem Híbrida, Bibliometria, Grafos de Citação, Vocabulários Controlados

1 INTRODUÇÃO

A explosão informacional em repositórios digitais afetou diversas áreas do conhecimento, dentre elas a Ciência da Informação (CI). Por ser uma área inerentemente interdisciplinar, a CI necessita de um forte contexto informacional para estruturar o conhecimento; do contrário, os termos recuperados tornam-se polissêmicos, criando barreiras complexas para pesquisadores nas áreas de Ciência, Tecnologia, Engenharia e Matemática (STEM). [5, 6]

Além disso, a revisão da literatura enfrenta o desafio da sobrecarga informacional e da dificuldade em filtrar relevância em meio ao ruído de dados [5]. É neste ecossistema de alta complexidade que os Sistemas de Recomendação (SR) emergem não apenas como ferramentas auxiliares, mas como infraestruturas críticas para a navegação do conhecimento [7]. Sistemas tradicionais de Recuperação da Informação (SRI), focados apenas em palavras-chave ou contagem simples de citações, falham em capturar a semântica profunda dos textos, pois frequentemente trazem dados representados apenas por palavras-chave, o que não reflete o real conteúdo semântico dos documentos, dificultando assim a transferência de conhecimento interdisciplinar [6].

Para enfrentar estas falhas, os Sistemas de Organização do Conhecimento (SOC) contribuem para que a semântica seja complexa, visto que é uma forma de representar e organizar, de forma estruturada e sistemática, um domínio do conhecimento[4]. É esse Sistema de Organização do Conhecimento (SOC) que atua como ponte semântica entre a linguagem natural e polissêmica presente nos documentos e a representação estruturada exigida pelos algoritmos. Sendo assim, a complexidade inerente à representação do conhecimento reside no equilíbrio entre a especificidade necessária para descrever um item único e a generalização necessária para agrupá-lo em categorias recuperáveis.

As questões relativas à análise conceitual, à compreensão da natureza, do significado e do escopo dos conceitos ocupam, portanto, um lugar de destaque nas operações técnicas. Não se trata apenas de atribuir rótulos, mas de mapear domínios de conhecimento, estabelecendo conexões que permitam a recuperação da informação precisa.

Historicamente, bibliotecas digitais tentaram mitigar esse problema com algoritmos de Filtragem Colaborativa (CF) e Baseada em Conteúdo (CBF). No entanto, abordagens isoladas sofrem limitações técnicas severas: a CF enfrenta a esparsidade de dados e o problema de *cold-start* (novos itens sem interação), enquanto a CBF muitas vezes ignora a qualidade intrínseca e o impacto bibliométrico da obra [10, 15].

Para superar tais barreiras, a literatura sugere a hibridização de métodos. Pesquisas seminais, como o sistema *Scienstein*, demonstraram que a união de análise de citações, análise de autoria e classificações implícitas supera motores de busca convencionais [8]. Este trabalho propõe o *Framework Semântico-Híbrido (FSH-Rec)*, que operacionaliza esses conceitos, unindo a organização do conhecimento com a capacidade de processamento algorítmico.

2 ARQUITETURA DO SISTEMA PROPOSTO

O framework, denominado *Hybrid-Semantic Scientific Recommender* (HSSR), opera através de um pipeline sequencial. Para estruturar o processamento, o sistema foi dividido em quatro estágios funcionais, cujas entradas e processos de transformação estão detalhados na Tabela 1. A arquitetura utiliza instrumentos de controle de vocabulário para superar a ambiguidade linguística e a esparsidade de dados.

Tabela 1: Fluxo de Processamento: Módulos, Entradas e Saídas do HSSR

Módulo	Dados de Entrada	Técnica/Processamento	Saída (Output)
1. Ingestão	Metadados Brutos (Web/APIs)	Normalização e Limpeza de Dados	Dataset Estruturado
2. Semântico (MPS)	Título, Resumo, Keywords	Expansão via Vocabulário Controlado (CSO) e Classificação ID3	Vetor de Perfil (P_a)
3. Grafo (AGC)	Lista de Referências e Citações	Ponderação Contextual de Seções e PageRank	Score de Centralidade
4. Fusão (Híbrido)	Vetores P_a, Scores de Grafo e Histórico	Combinação Linear com Fator de Diversidade (γ)	Lista Rankeada S_{final}

2.1 Módulo de Processamento Semântico (MPS)

Este módulo trata o conteúdo textual para gerar o vetor de perfil do artigo (P_a). Diferente de abordagens puramente estatísticas (TF-IDF), o MPS implementa uma estratégia de **normalização terminológica baseada em vocabulários controlados**, adaptando as abordagens de [3] e [12]:

(1) **Pré-processamento e Stemming:** Aplica-se a remoção de *stopwords* e o algoritmo de Porter para redução de palavras aos seus radicais. Esta etapa prepara os termos livres para serem confrontados com a linguagem controlada.

(2) **Expansão via Vocabulário Controlado:** Cada termo extraído é mapeado contra uma estrutura hierárquica padronizada, utilizando a *Computer Science Ontology* como tesauro de referência. Se um termo t é validado no vocabulário, seus Termos Genéricos (TG) imediatos são incorporados ao vetor de características com um peso de decaimento $\lambda = 0.5$. Isso garante que um artigo indexado com o termo específico "LSTM" seja recuperado em uma busca pelo termo autorizado mais amplo "Redes Neurais Recorrentes".

(3) **Classificação Supervisionada (ID3):** Para categorizar a área de conhecimento macro do artigo, utilizamos uma implementação do algoritmo ID3 (Iterative Dichotomiser 3), conforme validado por [11]. O ID3 calcula o ganho de informação dos termos normalizados para construir uma árvore de decisão que classifica o documento em categorias temáticas pré-estabelecidas, reduzindo o ruído informacional.

2.2 Motor de Grafos e Ponderação de Citações

Enquanto o MPS garante a consistência terminológica, este motor analisa a estrutura de citação. O sistema modela a base de dados como um grafo direcionado $G = (V, E)$. A inovação proposta reside na **Ponderação Contextual de Arestas**. Baseando-se na heurística de [9], o peso w_{ij} de uma citação do artigo i para o artigo j é calculado conforme a seção estrutural onde a citação ocorre:

$$w_{ij} = \sum_{k=1}^{N} \delta(s_k) \cdot \text{Impacto}(s_k) \quad (1)$$

Onde $\delta(s_k)$ indica a presença da citação na seção s_k e $\text{Impacto}(s_k)$ segue a distribuição: Metodologia (0.5), Resultados (0.3), Introdução (0.1) e Outros (0.1). Isso privilegia obras que fornecem embasamento metodológico real, filtrando citações meramente protocolares. Além disso, aplicam-se métricas de centralidade (PageRank) para identificar autores de referência na rede [16].

2.3 Mecanismo de Fusão Híbrida e Diversidade

O núcleo do recomendador combina os scores dos termos controlados com a análise bibliométrica. Para mitigar o problema da "bolha de filtro" e garantir a transferência de conhecimento interdisciplinar [6], o score final S_{final} para um par usuário-artigo (u,a) é dado por:

$$S_{final}(u,a) = \alpha \cdot \text{Sim}(u, P_{controlado}) + \beta \cdot \text{Rank}(G_a) + \gamma \cdot \text{Div}(a, R_u) \quad (2)$$

Onde:

- $\text{Sim}(u, P_{controlado})$: Similaridade cosseno entre os termos de interesse do usuário e o vetor normalizado do artigo.
- $\text{Rank}(G_a)$: Score de centralidade normalizado do artigo no grafo de citações.
- $\text{Div}(a, R_u)$: Fator de diversidade que penaliza itens muito similares aos últimos N artigos lidos pelo usuário (R_u). Esta abordagem de utilizar o histórico de leitura para refinar a precisão semântica segue o modelo proposto por [2].
- α, β, γ: Hiperparâmetros ajustáveis (0.4, 0.4, 0.2).

O uso de vocabulários controlados neste estágio são o ponto chave para resolver o problema de sinonímia e polissemia que frequentemente degrada a performance de sistemas baseados apenas em palavras-chave livres [13].

2.4 Arquitetura Terminológica

Nesta etapa a Organização do Conhecimento (OC) se estabelece não apenas como uma técnica auxiliar, mas como a disciplina fundamental para a arquitetura de sistemas inteligentes [14].

Como disciplina basilar na arquitetura de sistemas inteligentes, a OC provê os métodos para o desenvolvimento de SOCs. Tais instrumentos atuam na zona de intersecção entre a comunicação humana (inerentemente subjetiva e polissêmica) e as estruturas de dados rígidas, garantindo a exatidão terminológica exigida pelo processamento computacional. [14]

Para melhorar a distribuição dentro dos SOCs e sua adequação à variedade de complexidades existentes, as principais categorias discutidas na literatura e sua relevância para o processo de conversão descrito por Barbosa et al. [4] estão sintetizadas na Tabela 2:

Tabela 2: [4]Tipologia e Complexidade dos Sistemas de Organização do Conhecimento (SOCs)

Tipo de SOC	Estrutura	Características Principais	Exemplos Típicos	Relevância (Linked Data/SKOS)
Listas de Termos	Plana	Listas alfabéticas, glossários. Baixo controle.	Glossários técnicos, Dicionários.	Conversão trivial para `skos:ConceptScheme`.
Arquivos de Autoridade	Lista Controlada	Controle de nomes e desambiguação.	*LC Name Authority*, VIAF.	Identificação de entidades (MADS/RDF ou SKOS).
Taxonomias	Hierárquica	Divisão ordenada. Relação estrita *is-a*.	Taxonomias bio, E-commerce.	Mapeáveis para `skos:broader/narrower`.
Cabeçalhos de Assunto	Pré-coordenada	Frases com subdivisões. Hierarquia implícita.	LCSH (*Library of Congress*).	Desafiam SKOS (pré-coordenação).
Tesauros	Relacional	Relações de Equivalência, Hierarquia e Associação.	Tesauro da UNESCO, AGROVOC.	Candidato ideal para SKOS e Web Semântica.
Ontologias	Lógica Formal	Classes, axiomas e inferência.	Gene Ontology, CIDOC-CRM.	Mais complexas que SKOS (Raciocínio).

Na presente Revisão, o SOC é considerado parte fundamental na contribuição das habilidades de normalização terminológicas, de Vocabulários Controlados, e os Tesauros a qual é um dispositivo de controle terminológico usado para traduzir a linguagem natural dos documentos, dos indexadores ou dos utilizadores para uma linguagem de sistema mais restrita [14], abrangendo definições claras

de conceitos fornecidos por Termos Gerais padronizados hierarquicamente [4] que foram adaptados para a interação com os Sistemas de Recomendação.

A eficácia do tesauro reside na sua estrutura relacional, que permite ao sistema de recomendação "navegar" pelo conhecimento e inferir relevância mesmo quando os termos exatos não coincidem.

3 LITERATURA CORRELATA

Para fundamentar a arquitetura proposta, realizou-se uma revisão sistemática da literatura, sumarizada na Tabela 3. Esta análise destaca modelos híbridos, de Machine Learning (ML) e semânticos, considerando os resultados reportados por esses trabalhos em seus respectivos benchmarks para identificar tendências e lacunas.

Tabela 3: Comparativo Métricas de Desempenho de Modelos Relatos em Sistemas de Recomendação Híbridos e Semânticos
Métricas extraídas dos estudos citados, refletindo o desempenho reportado por seus autores em seus respectivos contextos e benchmarks. Selecionadas por sua relevância na avaliação de sistemas de recomendação, incluindo cold-start e diversidade.

Referência	Modelo/Técnica	Métrica Principal	Resultado
Rúbio & Gulo [11]	Algoritmo ID3 (ML)	Acurácia	89.05%
Wayesa et al. [17]	Híbrido (Padrão + Semântica)	F1-Score	0.521 (Híbrido)
	Word Embedding	F1-Score	0.569 (Embedding)
Sakib et al. [12]	Híbrido (CBF + CF)	Precision/Recall	> Baselines*
Al-hassan et al. [1]	Ontologia + CF	Significância (p-value)	< 0.001
Cunningham et al. [6]	GraphSAGE / ComBSAGE	Recall	Alto (Interdiscip.)

*Superou baselines em F1-measure, MAP e MRR (valores exatos variam por dataset).

A análise dos resultados reportados na Tabela 3 indica que abordagens baseadas em classificação supervisionada (ex: ID3 [11]) alcançam alta acurácia (89.05%). Embora os estudos utilizem bases de dados distintas, a síntese destes resultados permite inferir que a hibridização de metadados oferece uma cobertura semântica mais robusta para cenários de cold-start do que métodos puramente colaborativos [17]. Para o desenvolvimento do HSSR, planeja-se a replicação destes métodos sob um benchmark comum (DBLP) para garantir a comparabilidade direta do desempenho.

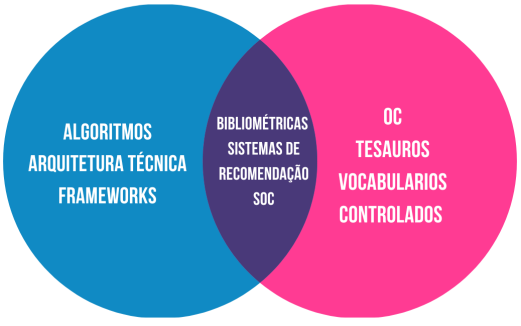

Figura 1: Interseção multidisciplinar

A necessidade dessa abordagem híbrida é ilustrada na Figura 1, que demonstra a interseção fundamental entre os algoritmos de engenharia e os métodos de Organização do Conhecimento (OC) para a eficácia dos sistemas de recomendação.

4 CONSIDERAÇÕES FINAIS

A convergência entre a curadoria de dados da Ciência da Informação e o processamento algorítmico da Engenharia permite a criação de sistemas de recomendação mais resilientes. A proposta apresentada utiliza as melhores práticas identificadas na revisão sistemática, especificamente a hibridização [12] e a classificação bibliométrica de [11] para justificar um modelo que atua como facilitador epistêmico. Contribuindo para a validação dos vocabulários controlados nas diversas áreas do conhecimento, estruturando a informação de forma clara e significativa, que culminem o ponto de interseção entre a Biblioteconomia e a Engenharia da Computação.

O próximo passo envolve a implementação do protótipo e validação com bases abertas, como a DBLP *computer science bibliography*.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Malak Al-Hassan, Haiyan Lu, and Jie Lu. 2015. A semantic enhanced hybrid recommendation approach: A case study of e-Government tourism service recommendation system. *Decision Support Systems* 72 (4 2015), 97–109. doi:10.1016/j.dss.2015.02.001

[2] Bushra Alhijawi, Nadim Obeid, Arafat Awajan, and Sara Tedmori. 2018. Improving collaborative filtering recommender systems using semantic information. In *2018 9th International Conference on Information and Communication Systems (ICICS)*. IEEE, 127–132. doi:10.1109/iacs.2018.8355454

[3] Mohamed Uvaze Ahmed Ayoobkhan and Liayakath Ali Khan Subair Ali. 2022. WEB PAGE RECOMMENDATION SYSTEM BY INTEGRATING ONTOLOGY AND STEMMING ALGORITHM. *INTERNATIONAL JOURNAL OF ADVANCES IN SIGNAL AND IMAGE SCIENCES* 8, 1 (jan 1 2022), 9–16. doi:10.29284/ijasis.8.1.2022.9-16

[4] Everton Rodrigues Barbosa, Moisés Lima Dutra, Angel Freddy Godoy Viera, and Douglas Dyllon Jeronimo de Macedo. 2021. Thesaurus and subject heading lists as Linked Data. *Transinformação* 33 (2021), e200077.

[5] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2015. Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries* 17, 4 (jul 26 2015), 305–338. doi:10.1007/s00799-015-0156-0

[6] Eoghan Cunningham, Barry Smyth, and Derek Greene. 2025. Facilitating interdisciplinary knowledge transfer with research paper recommender systems. *Quantitative Science Studies* 6 (2025), 854–875. doi:10.1162/qss.a.9

[7] Zeshan Fayyaz, Mahsa Ebrahimian, Dina Nawara, Ahmed Ibrahim, and Rasha Kashef. 2020. Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities. *Applied Sciences* 10, 21 (2020). doi:10.3390/app10217748

[8] Bela Gipp, Jöran Beel, and C. Hentschel. 2009. Scienstein : A Research Paper Recommender System. (2009).

[9] Ying KANG, Aiqin HOU, Zimin ZHAO, and Daguang GAN. 2021. A Hybrid Approach for Paper Recommendation. *IEICE Transactions on Information and Systems* E104.D, 8 (aug 1 2021), 1222–1231. doi:10.1587/transinf.2020bdp0008

[10] Bamshad Mobasher, Xin Jin, and Yanzan Zhou. 2004. *Semantically Enhanced Collaborative Filtering on the Web*. Springer Berlin Heidelberg, 57–76. doi:10.1007/978-3-540-30123-3_4

[11] Thiago R. P. M. Rubio and Carlos A. S. J. Gulo. 2016. Enhancing academic literature review through relevance recommendation: Using bibliometric and text-based features for classification. In *2016 11th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 1–6. doi:10.1109/cisti.2016.7521620

[12] Nazmus Sakib, Rodina Binti Ahmad, Mominul Ahsan, Md. Abdul Based, Khalid Haruna, Julfikar Haider, and Saravanakumar Gurusamy. 2021. A Hybrid Personalized Scientific Paper Recommendation Approach Integrating Public Contextual Metadata. *IEEE Access* 9 (2021), 83080–83091. doi:10.1109/access.2021.3086964

[13] Qusai Shambour and Jie Lu. 2011. A Hybrid Multi-criteria Semantic-Enhanced Collaborative Filtering Approach for Personalized Recommendations. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. IEEE, 71–78. doi:10.1109/wi-iat.2011.109

[14] Leonardo Fernandes Souto. 2003. Recuperação de informações em bases de dados: usos de tesauro. *Transinformação* 15, 1 (jan./abr. 2003), 73–81. 1.

[15] Roberto Torres, Sean M. McNee, Mara Abel, Joseph A. Konstan, and John Riedl. 2004. Enhancing digital libraries with TechLens+. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 228–236. doi:10.1145/996350.996402

[16] Waleed Waheed, Muhammad Imran, Basit Raza, Ahmad Kamran Malik, and Hasan Ali Khattak. 2019. A Hybrid Approach Toward Research Paper Recommendation Using Centrality Measures and Author Ranking. *IEEE Access* 7 (2019), 33145–33158. doi:10.1109/access.2019.2900520

[17] Fikadu Wayesa, Mesfin Leranso, Girma Asefa, and Abduljebar Kedir. 2023. Pattern-based hybrid book recommendation system using semantic relationships. *Scientific Reports* 13, 1 (mar 6 2023). doi:10.1038/s41598-023-30987-0