Explorando Sistemas de Recomendação para Resolução de Desafios em Cauda Longa e Cold Start

Davi Silva da Cruz[1]
[1]Instituto Federal do Maranhão (IFMA), Campus Coelho Neto
Email: davi.cruz@acad.ifma.edu.br
Orientador: Prof. Dr. Diogo Vinícius de Sousa Silva[2]
[2]IFMA, Campus Coelho Neto
Email: diogo.silva@ifma.edu.br

Resumo:

Este trabalho apresenta uma investigação sobre a aplicação de técnicas de clusterização em sistemas de recomendação, com objetivo principal de propor soluções para os desafios de cauda longa e cold start em sistemas de recomendação. Partindo de um sistema de recomendação sequencial, o SimRec, foram integrados métodos de agrupamento de itens — K-Means, Agglomerative Clustering (AGNES) e FasterPAM — utilizando métodos de redução de dimensionalidade nos conjuntos de dados ML-1M e Beauty. Para avaliação utilizaram-se métricas de recomendação (HR@10 e NDCG@10) e de qualidade dos clusters(Silhouette e Dunn). Os resultados indicaram que, em dados densos (ML-1M), a clusterização contribuiu para ganhos de até 4,69% no HR@10 (FasterPAM, k=2000), enquanto em dados esparsos (Beauty), os agrupamentos apresentaram ruído e não superaram a baseline. Conclui-se que a eficácia da clusterização depende da densidade e estrutura dos dados, sendo necessária a adaptação de estratégias para diferentes cenários.

Palavras-chave: Sistemas de recomendação. Cold Start. Cauda longa. Clusterização. Recomendação sequencial.

Financiamento: Instituto Federal do Maranhão (IFMA), campus Coelho Neto.

Introdução:

Os sistemas de recomendação desempenham papel central em plataformas digitais ao reduzir a sobrecarga de informação e personalizar a experiência do usuário (Adomavicius & Tuzhilin, 2005; Ricci et al., 2011). Entretanto, enfrentam desafios

[1] Discente, Instituto Federal do Maranhão (IFMA), Campus Coelho Neto. davi.cruz@acad.ifma.edu.br
[2] Professor Dr, Instituto Federal do Maranhão (IFMA), Campus Coelho Neto. diogo.silva@ifma.edu.br

persistentes como o cold start, em que novos usuários ou itens possuem poucos dados, e a cauda longa, que prejudica a visibilidade de itens pouco populares (Anderson, 2004; Zhou, Zhang e Yang, 2023).

Recentemente, estudos têm demonstrado que técnicas de clusterização podem mitigar esses problemas. Conforme observado por Qin (2021), métodos baseados em agrupamento para a recomendação de itens de cauda longa e novos itens podem funciona, pois, através do agrupamento de itens semelhantes, essas a escassez de dados pode ser superada e a preferência do usuário pode ser melhor capturada. Diante disso, este trabalho investiga a integração de algoritmos de agrupamento ao SimRec, sistema de recomendação sequencial baseado em embeddings de itens desenvolvido por Brody e Lagziel (2024), para avaliar melhorias.

Este trabalho tem como objetivo principal investigar e propor soluções para os desafios de cauda longa e cold start em sistemas de recomendação. Os objetivos específicos foram: 1. Realizar uma revisão abrangente da literatura sobre cauda longa e cold start. 2. Desenvolver algoritmos e técnicas para lidar com esses desafios. 3. Implementar e testar os algoritmos em um ambiente de simulação. 4. Avaliar o desempenho dos sistemas propostos utilizando métricas apropriadas.

Metodologia:

A pesquisa foi desenvolvida seguindo uma abordagem experimental estruturada em algumas etapas principais, iniciando com uma revisão da literatura sobre desafios em sistemas de recomendação, especialmente os relacionados ao cold start e à cauda longa. Esse levantamento permitiu mapear trabalhos recentes que exploram diferentes técnicas de mitigação desses desafios, além de identificar o SimRec como estrutura para extensão. De acordo com Brody e Lagziel (2024) o SimRec se destaca por integrar similaridades, mitigando limitações em contextos sequenciais.

Na sequência, na etapa de desenvolvimento e implementação, realizou-se a incorporação de técnicas de clusterização aplicadas sobre embeddings de itens. Investigaram-se três métodos de clusterização: K-Means, Agglomerative Clustering(AGNES) e FasterPAM. Seguindo para a etapa de experimentação, os testes foram planejados em diferentes cenários de representação dos dados, com e sem redução de dimensionalidade, utilizando as técnicas de análise de componentes principais (Principal Component Analisys - PCA) e aproximação e projeção de

variáveis uniformes (Uniform Manifold Approximation and Projection - UMAP). Nesse contexto, Peng et al. (2024) definem a "maldição da dimensionalidade" como o fenômeno em que o desempenho dos algoritmos de inteligência artificial é afetado devido à alta dimensionalidade dos dados, justificando assim a experimentação de diferentes cenários de dimensionalidade. Os experimentos foram conduzidos com dois conjuntos de dados, o MovieLens 1M (ML-1M) e o Amazon Beauty, suas características são mostradas na Tabela 1, sendo que a densidade mencionada é dada pela razão entre as interações e os itens do conjunto de dados.

Tabela 1 – Descrição dos conjuntos de dados utilizados.

Dataset	Usuários	Itens	Interações	Densidade
Beauty	52133	57226	0,4 M	6,9
ML-1M (MovieLens)	6040	3416	1 M	292,6

Fonte: Adaptada de Brody e Lagziel (2024)

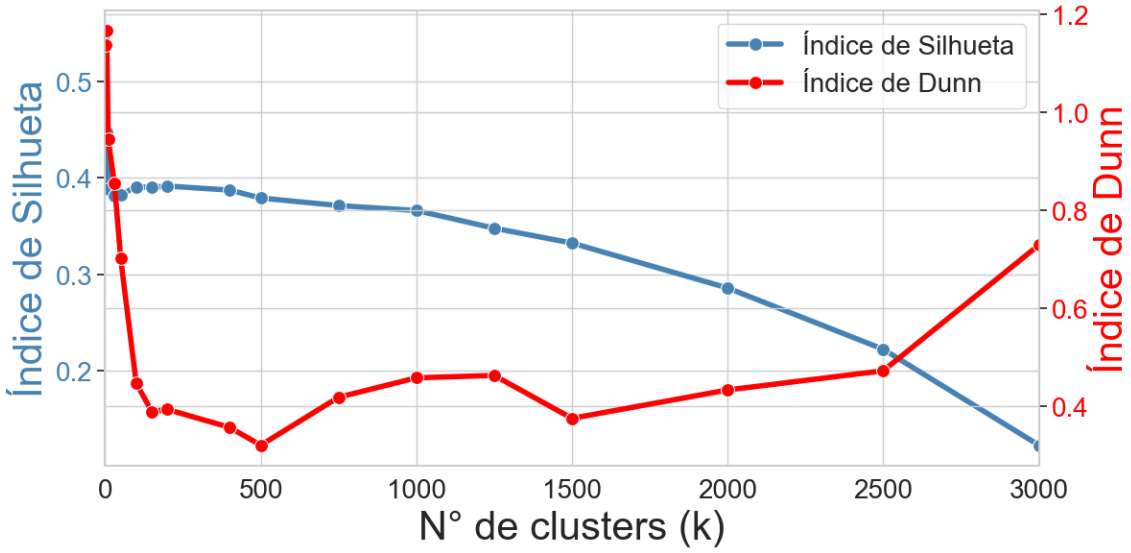
A etapa de análise de resultados contemplou dois tipos de métricas. No campo de recomendações, foram utilizadas Taxa de acerto (Hit Rate – HR) e Ganho cumulativo descontado normalizado(Normalized Discounted Cumulative Gain – NDCG), ambas nos dez primeiros itens recomendados (HR@10 e NDCG@10). Para análise da qualidade dos agrupamentos, aplicaram-se o Índice de silhueta (Silhouette Score) e Índice de Dunn (Dunn Index), a fim de verificar coesão e separação dos clusters formados. Por fim, testou-se o impacto de um fator de impulso, que multiplica os escores de itens pertencentes ao mesmo grupo do último item consumido pelo usuário.

Resultados e Discussão:

Na análise de cenários de dimensionalidade dos dados, foi definido como cenário ótimo a representação com UMAP utilizando 32 componentes, já que esse cenário alcançou os melhores valores de índice de silhueta (0,3414) e de índice de dunn (0,5479) em comparação com as demais representações de dimensionalidade. A partir desse cenário ótimo, as técnicas de clusterização foram executadas, escolhendo diferentes valores do número de clusters (k) a partir de seus resultados nas métricas de clusterização. As Figuras 1 e 2 mostram dois dos gráficos de desempenho dos algoritmos de agrupamento testados. As linhas mostram a variação da pontuação de silhueta e (azul, eixo y

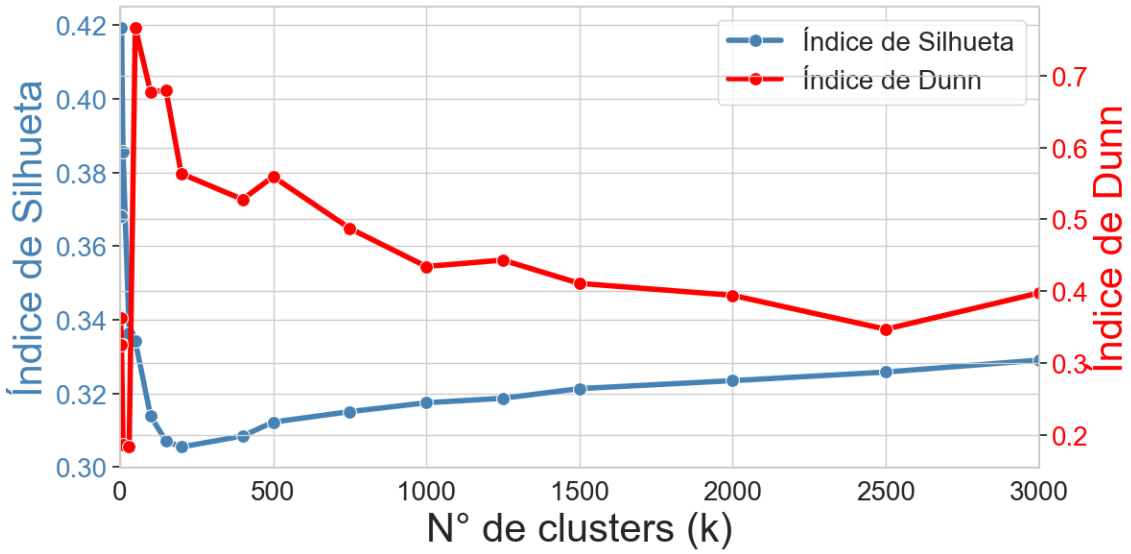
esquerdo) e o índice de dunn (eixo vermelho, direito) em função do número de clusters (k).

Figura 1 - Desempenho K-Means em ML-1M por número de clusters (K)



Fonte: Elaborado pelo autor (2025)

Figura 2 - Desempenho Fasterpam em Beleza por Número de Clusters (K)



Fonte: Elaborado pelo autor (2025)

Com base nessas análises, foram definidos três valores de k para cada algoritmo e conjunto de dados, considerando o ponto de pico, o método de cotovelo e a interseção das curvas de Silhueta e Dunn, a fim de capturar diferentes cenários de agrupamentos. Por fim, os clusters selecionados foram integrados ao SimRec, aplicando um fator de

impulso sobre os itens pertencentes ao mesmo grupo do último item consumido pelo usuário. Os experimentos mostram que no dataset ML-1M, a integração de clusters gerou ganhos em relação ao baseline, alcnçando até 4,69% de melhoria em HR@10 com o K-Means com k = 400 (HR@10 = 0,8200). Já no dataset Beauty, os resultados ficaram abaixo do SimRec original, sugerindo que os clusters gerados foram ruidosos e não contribuiram para a recomendação.

Conclusões:

A integração de técnicas de clusterização ao SimRec demonstrou-se eficaz em cenários de dados densos, como MovieLens 1M, promovendo ganhos em métricas de recomendação. Porém, em contextos esparsos, como o do conjunto de dados Amazon Beauty, a clusterização não resultou em melhorias consistentes, sendo superada em diversos cenários pelo modelo original.

Assim, conclui-se que a eficácia da clusterização em sistemas de recomendação depende das características do conjunto de dados. Como perspectivas futuras, sugere-se o aprofundamento no estudo de métodos de clusterização sensíveis ao contexto, capazes de ajustar-se dinamicamente às peculiaridades de cada domínio de aplicação.

Agradecimentos:

Os autores agradecem ao Instituto Federal do Maranhão (IFMA), Campus Coelho Neto, pelo fomento à pesquisa e pelo apoio financeiro fornecido. Esse apoio foi essencial para o desenvolvimento e a conclusão desta pesquisa.

Referências:

ADOMAVICIUS, G.; TUZHILIN, A. **Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions**. IEEE Transactions on Knowledge and Data Engineering, v. 17, n. 6, p. 734-749, 2005.

ANDERSON, C. **The Long Tail**. Wired Magazine, v. 12, n. 10, p. 1-10, 2004.

BRODY, S.; LAGZIEL, S. **SimRec: Mitigating the Cold Start Problem in Sequential Recommendation by Integrating Item Similarity**. arXiv preprint, arXiv:2410.22136 [cs.IR], 2024. Disponível em: https://arxiv.org/abs/2410.22136.

PENG, D.; GUI, Z.; WU, H. **Interpreting the Curse of Dimensionality from Distance Concentration and Manifold Effect**. arXiv preprint, 2024. Disponível em: https://doi.org/10.48550/arXiv.2401.00422.

QIN, J. **A survey of long-tail item recommendation methods. Wireless Communications and Mobile Computing**, v. 2021, p. 7536316, 2021. Disponível em: https://doi.org/10.1155/2021/7536316.

RICCI, F.; ROKACH, L.; SHAPIRA, B. **Introduction to recommender systems handbook**. In: RICCI, F.; ROKACH, L.; SHAPIRA, B.; KANTOR, P. (org.). Recommender systems handbook. Boston: Springer, 2011. p. 1-35.

ZHOU, Z.; ZHANG, L.; YANG, N. **Contrastive Collaborative Filtering for Cold-Start Item Recommendation**. In: Proceedings of the ACM Web Conference, 2023.