ISSN: 2319-0124
josif 2023
15ª JORNADA CIENTÍFICA E TECNOLÓGICA
12º SIMPÓSIO DE PÓS-GRADUAÇÃO
DO IFSULDEMINAS
REALIZAÇÃO INSTITUTO FEDERAL Sul de Minas Gerais
APOIO FAPEMIG CNPq

RECOMENDAÇÃO DE LIVROS BASEADA EM CLUSTERIZAÇÃO E ALGORITMOS DE FILTRAGEM COLABORATIVA

Heberty A. SILVA[1]Larissa S. PEREIRA[2]; Diego SAQUI[3]

RESUMO

O clustering, uma técnica de agrupamento de dados, tem se destacado como uma abordagem promissora para melhorar a precisão das recomendações. O estudo utiliza o algoritmo K-means para clusterizar livros com base em palavras-chave de suas descrições. O pré-processamento dos dados, como remoção de caracteres não ASCII e stopwords, e a vetorização dos textos dos livros com TF-IDF são realizados. Os resultados revelam uma acurácia de aproximadamente 70% e métricas de desempenho são utilizadas para avaliar o sistema. Embora existam limitações, como a dependência da qualidade das descrições e a falta de consideração de outros fatores, o projeto contribui para a área de recomendação de livros e destaca a importância da avaliação contínua do desempenho.

Palavras-chave:
K-means; Métricas de Avaliação; Inteligência Artificial.

1. INTRODUÇÃO

Com a quantidade de informações disponíveis na internet cada vez maior é perceptível o aumento da demanda por sistemas de recomendação eficientes. Esses sistemas têm como propósito auxiliar os usuários na descoberta de conteúdos relevantes, "Os sistemas de recomendação de preferência são amplamente utilizados em sistemas web para sugerir conteúdo personalizado aos usuários, utilizando algoritmos de filtragem colaborativa e de conteúdo (SILVA et al., 2022)., personalizando as recomendações de acordo com seus interesses e preferências. Analisando as diversas abordagens utilizadas para melhorar a precisão das recomendações, o *clustering*, ou agrupamento, tem se destacado como uma técnica promissora.

O *clustering* é uma técnica de aprendizado não supervisionado que busca agrupar objetos semelhantes em conjuntos distintos,baseando-se em suas características. No contexto de sistemas de recomendação de livros, a implementação de *clustering* permite identificar grupos de usuários com preferências análogas, bem como agrupar os livros de acordo com suas características, como gênero, autor ou tema.

Ao utilizar o *clustering* em um sistema de recomendação de livros, se torna viável melhorar a precisão das recomendações, pois usuários com predileções semelhantes têm maior probabilidade de se interessarem por livros semelhantes. Além disso, a abordagem de *clustering* também permite a

[1] IFSULDEMINAS – *Campus* Muzambinho. E-mail: hebertysilva271@gmail.com

[2] IFSULDEMINAS – *Campus* Muzambinho. E-mail: larissapereirainfo2@gmail.com

[3] Orientador, IFSULDEMINAS – *Campus* Muzambinho. E-mail: diego.saqui@muz.ifsuldeminas.edu.br.

exploração de novos livros, recomendando obras pertencentes a grupos diversos dos quais o usuário já demonstrou interesse.

Diversos estudos e implementações práticas têm comprovado a eficácia do *clustering* em sistemas de recomendação de livros. Segundo uma pesquisa realizada por (RICCI; ROKACH; SHAPIRA, 2015). Os sistemas de recomendação de preferência são utilizados para sugerir livros com base nas preferências de leitura do usuário, desta forma para o desenvolvimento do sistema, teve como base aplicar o método de clusterização, analisando a descrição dos livros e utilizando algoritmos de filtragem colaborativa.

Por meio do método K-means, algumas aplicações podem ser citadas por alguns autores como Zahra et al. (2015) Kant et al. (2018) Wen, Bao, & Ding (2018) e Putriany, Jauhari, & Heroza (2019), que aplicaram a *clusterização* por k-means como forma de agrupar os diferentes itens de um conjunto de dados para o desenvolvimento de sistemas de recomendação..

2. MATERIAL E MÉTODOS

A metodologia utilizada neste estudo teve como propósito realizar a ***clusterização*** dos livros com base nas palavras-chave encontradas em suas descrições. Para isso, foi utilizado o algoritmo de K-Means, que é amplamente utilizado para a *clusterização* de dados em diversas áreas, incluindo a análise de preferências de leitura de livros (SILVA,2021), já foi bastante explorada em trabalhos anteriores para sistemas de recomendação.

O algoritmo K-Means busca identificar k centróides que representam os grupos formados pelos dados. Inicialmente, foram selecionados aleatoriamente k centróides ou gerados randomicamente com base nas coordenadas máximas e mínimas dos objetos analisados. Posteriormente cada livro foi atribuído ao centróide mais próximo, utilizando a distância euclidiana entre os vetores representativos. Essa atribuição foi realizada com base nas palavras-chave encontradas nas descrições dos livros.

Importante ressaltar que, antes da aplicação do algoritmo K-Means, os dados passaram por um pré-processamento para garantir a qualidade e a consistência dos resultados. O pré-processamento de dados é uma etapa fundamental para a aplicação de algoritmos de *clustering* e sistemas de recomendação em sistemas web, envolvendo técnicas de Processamento de Língua Natural (PLN) e mineração de dados (FEITOSA, 2019). Foi realizada a remoção de caracteres não ASCII presentes nas descrições dos livros para evitar problemas de processamento e assegurar que o algoritmo funcione corretamente.Também foi realizada a conversão para letras minúsculas visando não gerar duplicidade e facilitar o processo, assim como a remoção de pontuação e de stopwords para eliminar caracteres que não são relevantes para a análise de similaridade entre os documentos.

Posteriormente ao pré-processamento dos dados, os textos dos livros foram vetorizados utilizando a técnica *Term Frequency-Inverse Document Frequency* (TF-IDF). Com essa técnica, atribuímos pesos aos termos nos textos dos livros, levando em conta sua frequência e importância relativa, gerando uma representação vetorial numérica para cada livro, onde os vetores refletem a importância dos termos em cada documento. Essa abordagem nos permite comparar os livros de forma mais precisa, considerando suas características individuais e o contexto geral dos textos.

Com os dados pré-processados e vetorizados, o algoritmo K-means analisou o conjunto de dados, fazendo assim, os seres foram agrupados em k clusters, sendo este 567, equivalente a quantidade de categorias dos livros, em que cada grupo representa um conjunto de livros com características semelhantes, com base nas palavras-chave encontradas em suas descrições.

3. RESULTADOS E DISCUSSÕES

Com base em informações, como, a descrição de um livro que fora fornecido pelo usuário, apresentou-se resultados relevantes de recomendações. Ao inserir a descrição de um livro, identificações a respeito de características correspondentes ao texto foram realizadas e a partir de tais dados uma lista contendo um total de 10 livros foram recomendados.

A partir do modelo de clusterização utilizando, uma acurácia significativa de aproximadamente 70%, através da utilização do método *acuraccy_score*, das métricas do *sklearn*, em conjunto dos cluster inseridos em um novo atributo na tabela, pode ser constatada uma possível taxa de veracidade a respeito da classificação indicada pelo sistema. Demonstrando capacidade acima da média de agrupamento e separação dos gêneros literários por meio das características adquiridas mediante os dados obtidos da descrição realizada pelo usuário. Evidenciando uma confiança em que os consumidores do sistema podem ter a respeitos das recomendações fornecidas.

Como base de dados para o sistema, foi utilizado um dataset que apresenta diversas informações a respeito de um conjunto de atributos referentes a características de livros. Dentre os quais conta com os seguintes atributos: os identificadores, título, subtítulo, autores, categorias, URL da miniatura, descrição, ano de publicação, classificação média e número de avaliações, contendo um total de 6810 instâncias.

Apesar de resultados positivos, não se pode deixar de mencionar as limitações que podem ser encontradas em um sistema de recomendação. De início a qualidade das recomendações está fortemente ligada a quão bem descrito foi o texto do usuário. Descrições genéricas com termos semelhantes em diversos textos das descrições do dataset acarretaria em uma apresentação de resultados de forma também genérica. Ademais considerando que o sistema atual utiliza apenas de informações do texto da descrição, sem levar em conta outros fatores, no qual a inclusão de algumas informações adicionais acarretaria em uma aprimoração das recomendações.

4. CONCLUSÕES

Utilizando-se de métricas de avaliação de desempenho do sistema, os resultados apresentados são uma forma de constatar a satisfação a respeito da acurácia da clusterização realizada no que se diz respeito aos livros, afirmando assim um sistema eficaz e com capacidade de recomendar conteúdos seguindo os padrões e a preferência do usuário.

O projeto em questão, quando atribuído a um ambiente relacionado ao de recomendações de livros, apresenta uma contribuição eficaz na área em questão, permitindo um agrupamento de livros com características semelhantes através do processo de clusterização. Destaca-se também a importância de uma avaliação de desempenho para que se qualifique o sistema e ocorra aprimoramentos contínuos.

AGRADECIMENTOS

Agradecemos ao IFSULDEMINAS- Campus Muzambinho pela oportunidade e estrutura concedidas para realização dessa pesquisa.

REFERÊNCIAS

FEITOSA, R. Mineração de dados e sistemas de recomendação. São Luís: UFMA, 2019.

KANT, S.; MAHARA, T.; JAIN, V. K.; JAIN, D. K.; SANGAIAH, A. K. LeaderRank based k-means clustering initialization method for collaborative filtering. Computers & Electrical Engineering, [s.l.], v. 69, p. 598-609, 2018.]

PUTRIANY, V.; JAUHARI, J.; HEROZA, R. I. Item Clustering as An Input for Skin Care Product Recommended System using Content Based Filtering. Journal of Physics: Conference Series, Palembang, v. 1196, n. 1, p. 012004, 2019.

RICCI, F.; ROKACH, L.; SHAPIRA, B. Introduction to recommender systems handbook. Cham: Springer, 2015.

SILVA, A. C. et al. Ciência de dados em políticas públicas: uma experiência de formação. Brasília: Escola Nacional de Administração Pública, 2021.

SILVA, A. C. et al. Sistemas de recomendação de conteúdo personalizado em sistemas web. In: CONGRESSO BRASILEIRO DE INFORMÁTICA, 2022, Florianópolis. Anais... Florianópolis: SBC, 2022. p. 123-130.

WEN, T.; BAO, J.; DING, F. QoS-Aware Web Service Recommendation Model Based on Users and Services Clustering. Proceedings of the International Conference on Information Technology and Electrical Engineering 2018, Xiamen, v. 18, n. 1, p. 1-6, 2018.

ZAHRA, S.; GHAZANFAR, M. A.; KHALID, A.; AZAM, M. A.; NAEEM, U.; PRUGEL-BENNETT, A. Novel centroid selection approaches for KMeans-clustering based recommender systems. Information sciences, [s.l.], v. 320, p. 156-189, 2015.