10ª Jornada Científica e Tecnológica
7º Simpósio da Pós-Graduação
INSTITUTOS FEDERAIS 10 ANOS 2008|2018
INSTITUTO FEDERAL Sul de Minas Gerais
ISSN 2319-0124

COMPARAÇÃO DE MÉTODOS DE FILTRAGEM COLABORATIVA PARA SISTEMAS DE RECOMENDAÇÃO

João M. S. B. de MORAES[1]; Rodrigo C. EVANGELISTA[2]; Raphael A. P. DIAS[3]

RESUMO

Sistemas de recomendação são importantes nos dias atuais, pois personalizam o conteúdo entregue a cada usuário. Porém, escolher qual algoritmo utilizar pode se tornar difícil devido à quantidade de soluções existentes. Em vista disso, este trabalho tem como objetivo comparar os principais métodos e algoritmos de filtragem colaborativa utilizados em recomendação, a fim de facilitar a escolha da solução mais viável e colaborar para sua implementação.
Palavras-chave: Recomendação; Técnicas; Análise comparativa.

1. INTRODUÇÃO

O volume de dados gerados e requisitados na *Web* diariamente é muito grande, assim como o número de pessoas com acesso à rede. De acordo com Meeker (2015), de 1995 até 2014, a porcentagem da população mundial com acesso à Internet cresceu de 1% a 39%, tendo aproximadamente 3,5 bilhões de usuários em 2014. Dessa forma, a quantidade de opções e de escolhas possíveis em qualquer atividade *online* torna-se igualmente ampla. Nesse contexto, justifica-se o uso de filtros que organizem as informações entregues ao usuário.

Uma tecnologia amplamente utilizada nos dias de hoje para filtrar informações são os sistemas de recomendação. Ricci, Rokach e Shapira (2011) os definem como ferramentas e técnicas com o objetivo de oferecer sugestões que almejam ser úteis ao usuário, auxiliando o processo de escolha em um mundo interconectado.

Há escassez de trabalhos científicos que exponham as técnicas existentes aplicadas a conjuntos de dados de pequenos tamanhos e que estejam em português, evidenciando tal necessidade. Portanto, o objetivo deste trabalho é analisar e comparar os principais métodos de filtragem colaborativa de sistemas de recomendação, a fim de facilitar a escolha da solução mais viável de ser empregada.

[1] IFSULDEMINAS – *Campus* Muzambinho. E-mail: 12151002588@muz.ifsuldeminas.edu.br
[2] IFSULDEMINAS – *Campus* Muzambinho. E-mail: rodrigo.evangelista@muz.ifsuldeminas.edu.br
[3] IFSULDEMINAS – *Campus* Muzambinho. E-mail: raphael.a.p.dias@gmail.com

2. FUNDAMENTAÇÃO TEÓRICA

Os sistemas de recomendação são classificados basicamente em três categorias: filtragem baseada em conteúdo, filtragem colaborativa e filtragem híbrida. Conforme Bobadilla *et al.* (2013), dentro das três categorias, algumas técnicas empregadas incluem o algoritmo dos vizinhos mais próximos, que será o método testado neste trabalho. As propriedades dos sistemas de recomendação são mais detalhadamente discutidas no trabalho de Aggarwal *et al.* (2016).

3. MATERIAL E MÉTODOS

O conjunto de dados (*dataset*) utilizado para a realização dos testes e das avaliações foi o *MovieLens* de 100 mil registros, que possui 6,3% de densidade e conta com mais de 7500 referências no Google Acadêmico.

A linguagem de programação utilizada para comparar os métodos foi o *Python* na versão 3.5.2. Para fazer uso dos algoritmos e compará-los, foi empregada a biblioteca *Surprise*. O pacote *Pandas* foi usado para lidar com estrutura de dados e por ter ferramentas estatísticas. A biblioteca *Matplotlib*, por sua vez, foi utilizada para construir os gráficos comparativos. Para visualizar os resultados de modo simples e prático, fez-se uso do *Jupyter Notebook* na versão 4.4.0.

O paradigma de avaliação utilizado foi o *offline* e a métrica de avaliação foi a métrica de acurácia Raíz do Erro Quadrático Médio, do inglês *Root Mean Square Error* (RMSE). A técnica de validação empregada, por sua vez, foi a validação *k-fold*. Fazendo uso da métrica de acurácia e da validação *k-fold*, foi possível realizar uma análise comparativa de desempenho dos algoritmos, para determinar os mais viáveis ao *dataset* selecionado.

Os algoritmos testados foram o *k*-NN com a função de predição básica e o *k*-NN com a função de predição centrada na média, que leva em consideração as médias das notas de cada usuário. A abordagem utilizada para ambos os métodos foi a baseada em usuário, com a função de similaridade Diferença Quadrática Média, do inglês *Mean Square Difference* (MSD).

4. RESULTADOS E DISCUSSÕES

Os testes foram realizados no *dataset MovieLens* de 100 mil registros. A comparação entre o *k*-NN com predição básica e o *k*-NN com predição na média é indicada pela Figura 1, que mostra as taxas de erro RMSE (eixo Y) para vários valores *k* de vizinhos (eixo X).

Para o *k*-NN básico, nota-se que a taxa de erro é alta para valores pequenos de X, se estabilizando quando X assume valores acima de 10. Para o *k*-NN na média, no entanto, o resultado é diferente com relação ao valor *k* de vizinhos. Nele, a RMSE (Y) diminuiu à medida em que o

número de vizinhos (X) aumentou, evidenciando o seguinte fato: considerar as avaliações dos usuários aos itens afeta positivamente a precisão do método.

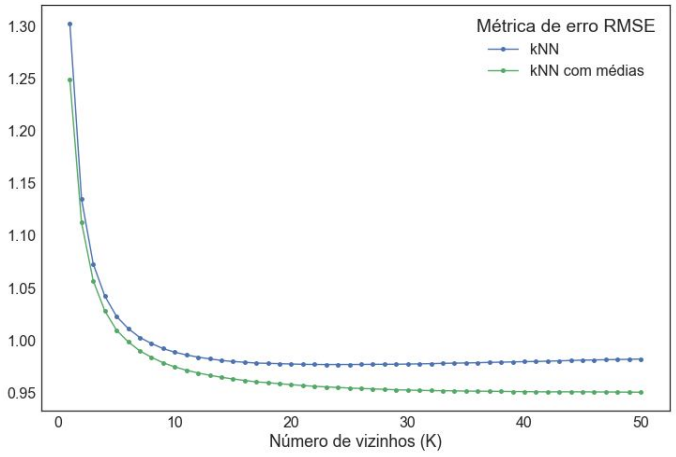

Figura 1: Relação entre o nº de vizinhos (X) e as taxas de erro (Y) para os dois métodos

Embora o *k*-NN com predição na média tenha obtido menores taxas de erro em geral, é necessário analisar também os tempos de execução. A Tabela 1 mostra, para ambos os algoritmos, os valores mínimos da taxa de erro RMSE e os valores mínimos dos tempos de treino e de teste.

Algoritmos	Taxa de erro RMSE	Tempo de treino	Tempo de teste
k-NN básico	0.976958	1.000216	5.748463
k-NN na média	0.950553	1.078361	7.351482

Tabela 1: Valores mínimos de erro e de tempo para os dois algoritmos

Com relação ao *k*-NN com predição na média, o *k*-NN básico apresentou menores tempos de execução, mas teve uma taxa de erro maior. Isso se deve ao fato de a predição básica não considerar a forma como os usuários avaliam os itens, afetando sua precisão. A predição centrada na média, por outro lado, uma vez que faz uso das médias das avaliações dos usuários aos itens, possui uma taxa de erro menor e, consequentemente, torna-se viável para conjuntos de dados de pequeno tamanho e densidade.

5. CONCLUSÕES

Em conformidade com os testes e a análise realizada, o *k*-NN com predição centrada na média se apresenta como uma melhor escolha, pois, embora tenha obtido um tempo maior de execução em geral, forneceu resultados mais precisos.

Desse modo, para *datasets* pequenos e com baixa densidade, como é o caso do *MovieLens* de 100 mil registros, conclui-se que é vantajoso considerar o modo como os usuários avaliam os

itens, isto é, centrar a função na média, pois isso aumenta a acurácia da predição e, consequentemente, a qualidade da recomendação.

Portanto, espera-se que o presente trabalho contribua, de modo geral, para a seleção dos métodos mais viáveis de serem utilizados em sistemas de recomendação com pequenos *datasets* de baixa densidade.

AGRADECIMENTOS

Ao Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais – Campus Muzambinho pela oportunidade de realizar esse projeto.

Aos professores Rodrigo César Evangelista e Raphael Antônio Prado Dias, pela orientação na elaboração desse projeto.

REFERÊNCIAS

AGGARWAL, Charu C *et al*. **Recommender systems**. [S.l.]: Springer, 2016.

BOBADILLA, Jesús *et al*. Recommender systems survey. **Knowledge-based systems**, Elsevier, v. 46, p. 109–132, 2013.

MEEKER, Mary. Internet trends 2015-code conference. **Glokalde**, v. 1, n. 3, 2015.

RICCI, Francesco; ROKACH, Lior; SHAPIRA, Bracha. Introduction to recommender systems handbook. In: **Recommender systems handbook**. [S.l.]: Springer, 2011. p. 1–35.