Comparação de técnicas de similaridade baseada em usuário

Cassiano H. da Silva[1], **Geovani S. Celebrim**[1]

[1]Departamento de Ciência da Computação
Universidade Federal Rural do Rio de Janeiro (UFRRJ)
26.020-740 – Nova Iguaçu – RJ – Brasil

{honoriocassiano, geovanicelebrim}@gmail.com

Resumo. *Sistemas de Recomendação ajudam a personalizar a experiência dos usuários, oferecendo a eles conteúdos específicos que tendem a ser de seu interesse. Existem várias técnicas para construção de um Sistema de Recomendação, cada uma com seus algoritmos, suas vantagens e desvantagens. Este trabalho explora um algoritmo utilizado em Sistemas de Recomendação com Filtragem Colaborativa, o KNN. O principal objetivo do trabalho é realizar uma análise entre técnicas de similaridade existentes na literatura e outras aqui propostas. Ao final, espera-se que seja possível dizer qual técnica é mais eficiente para este conjunto de dados.*

1. Definição

Sistemas de Recomendação exploram características e relações entre os usuários e itens com a finalidade de entender como se dá essa relação, para que seja capaz de realizar uma recomendação de um novo item para um usuário [Tail 2006]. As informações sobre essas características e relações podem ser capturadas do usuário e do item de forma implícita ou explícita, cada uma com suas vantagens e desvantagens.

Existem diferentes tipos de Sistemas de Recomendação e o que caracteriza essa diferença é a forma como é criada a relação da interação entre os usuários e os itens, bem como a forma que a recomendação é processada. Segundo [Burke 1999], existem basicamente quatro tipos de recomendação: Filtragem Colaborativa, Baseada em Conteúdo, Baseados em Conhecimento e Híbridos. Neste trabalho será explorada a técnica de Filtragem Colaborativa com predição baseada em usuário, introduzida por [Resnick et al. 1994], que recomendam itens baseados na utilização desses pela comunidade de usuários.

Os algoritmos de Filtragem Colaborativa podem ser divididos em dois grandes grupos de técnicas: baseados em memória e baseados em modelo. Os algoritmos baseados em memória, também conhecidos como Baseados em Vizinhança, possuem grande destaque pelo pioneirismo [Goldberg et al. 1992]. Para estes sistemas, o interesse do usuário u no item i é calculado usando as avaliações dos seus usuários mais próximos. Para isso, é necessário armazenar as avaliações de cada usuário e obter os K usuários mais próximos de u. Calcular essa proximidade de forma eficiente é fundamental para que se obtenha bons resultados. Na sessão seguinte são apresentadas as técnicas de cálculo de similaridade que foram estudadas neste trabalho.

2. Similaridades

Medidas de similaridade permite selecionar os vizinhos mais pertinentes para o usuário *u*, diminuindo o erro da predição. Além disso, as medidas de similaridade fornecem uma ponderação para o cálculo da predição da recomendação dos itens aos usuários.

A seguir são apresentadas algumas técnicas de similaridade escolhidas para análise neste trabalho. Dentre elas estão técnicas clássicas na literatura e outras propostas a partir da análise do funcionamento de diferentes técnicas existentes.

2.1. Cosseno

A similaridade do cosseno é amplamente conhecida e muito utilizada, devido a sua facilidade de implementação e seus resultados satisfatórios, dependendo do domínio do problema. Ela se fundamenta na ideia de que dados dois vetores não-nulos, o cosseno do ângulo entre eles representa o quão similares eles são. Portanto, quanto mais a solução se aproxima de 1, mais "similares" são os vetores [Shapira et al. 2011]. Essa similaridade pode ser definida como: dado dois vetores de atributos, A e B, a similaridade cosseno, $cos(\theta)$, é obtida através da equação 1.

$$cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

2.2. Cosseno Interseção

Esta medida de similaridade é semelhante à similaridade por cosseno, apresentada na subseção anterior. A diferença entre elas é que na similaridade por cosseno, um valor deve ser atribuído aos valores faltantes, normalmente a média global das avaliações é escolhida para essa substituição. Já no cosseno interseção, é considerado apenas os valores contidos em ambos conjuntos, diminuindo assim a cobertura do método, mas aumentando sua precisão [Shapira et al. 2011]. Para esse método, a equação 1 continua sendo válida, sendo que os valores faltantes devem ser simplesmente desconsiderados.

2.3. Correlação de Pearson

A Correlação de Pearson permite remover os efeitos da média e da variância das avaliações dos usuários e mede a linearidade entre duas variáveis [Resnick et al. 1994]. Para se obter as correlações w_{uv} entre os usuários u e v, são feitas as correlações das avaliações que ambos os usuários fizeram aos mesmos itens, de acordo com a equação 2.

$$w_{uv} = \frac{\sum_{i \in T} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in T} (r_{ui} - \bar{r}_u)^2}\sqrt{\sum_{i \in T} (r_{vi} - \bar{r}_v)^2}} \tag{2}$$

2.4. Silva-Celebrim Dice

A similaridade Silva-Celebrim Dice é baseada no coeficiente de Sørensen-Dice [Sørensen 1948, Dice 1945] com uma alteração da função de interseção. Na função original, os valores considerados são aqueles que estão na interseção entre os itens avaliados por dois usuários. Mas esse método não considera as notas das avaliações. Para solucionar esse problema, os valores considerados passaram a ser avaliados de acordo com um certo limiar t de diferença entre notas, não somente pelo fato dos itens terem sido avaliados ou não. Dados dois conjuntos de itens avaliados A e B, a fórmula é dada por:

$$P(u, v) = \frac{2R}{|A| + |B|}, \text{ onde } R = |(A \cap B)| \text{ e } |a_i - b_i| \leq t \tag{3}$$

3. Metodologia

Para a realização deste trabalho foi utilizado um conjunto de dados do *MovieLens*, disponível em `http://grouplens.org/datasets/movielens`. Os dados foram divididos em 5 partes, cada uma com exatamente 20% dos dados. Foram realizados 5 testes para cada parâmetro e em cada teste uma das partes foi separada para teste e o restante para treino. Os resultados finais para cada parâmetro são uma média aritmética dos resultados de todas as partes.

Uma vez que os dados foram devidamente separados, o conjunto de teste foi escondido enquanto o algoritmo utilizava o conjunto de treino para se adequar às características e relações entre usuário e item. Posteriormente, o conjunto de treino foi utilizado para que fosse analisada a acurácia da técnica aplicada.

Buscando obter o melhor de cada medida de similaridade, o algoritmo passou por diversos testes, variando seus parâmetros de forma que pudesse ser observado como o KNN se comporta com determinada medida de similaridade, tendo seus parâmetros variados. O principal parâmetro testado foi o número de vizinhos que foi considerado para a predição. Outros parâmetros como a tolerância de similaridade, da técnica apresentada na seção 2.5, também foram submetidos a testes.

A seção seguinte apresenta de forma objetiva algumas características de cada um dos experimentos realizados para cada medida de similaridade, bem como os resultados alcançados com o experimento.

4. Experimentos e Resultados

Esta seção apresenta os resultados obtidos através dos experimentos realizados para cada uma das técnicas de similaridade apresentadas anteriormente. Para analisar a acurácia das técnicas de similaridade, duas das medidas mais comuns para avaliar o desempenho das recomendações foram utilizadas: *Mean Absolute Error* (MAE), dada pela equação 4 e *Root Mean Squared Error* (RMSE), dada pela equação 5.

$$MAE(f) = \frac{\sum_{r_{ui} \in R_{test}} |f(u,i) - r_{ui}|}{|R_{test}|} \tag{4}$$

$$RMSE(f) = \sqrt{\frac{\sum_{r_{ui} \in R_{test}} (|f(u,i) - r_{ui}|)^2}{|R_{test}|}} \tag{5}$$

Os gráficos da figura 1 apresentam uma comparação entre as métricas de similaridade utilizando o MAE, no gráfico da esquerda e o RMSE, no gráfico da direita. No caso da métrica SC-Dice, o valor de limiar utilizado foi 1. Para o cálculo da similaridade, foram considerados somente vizinhos que possuíam mais de 10 itens em comum.

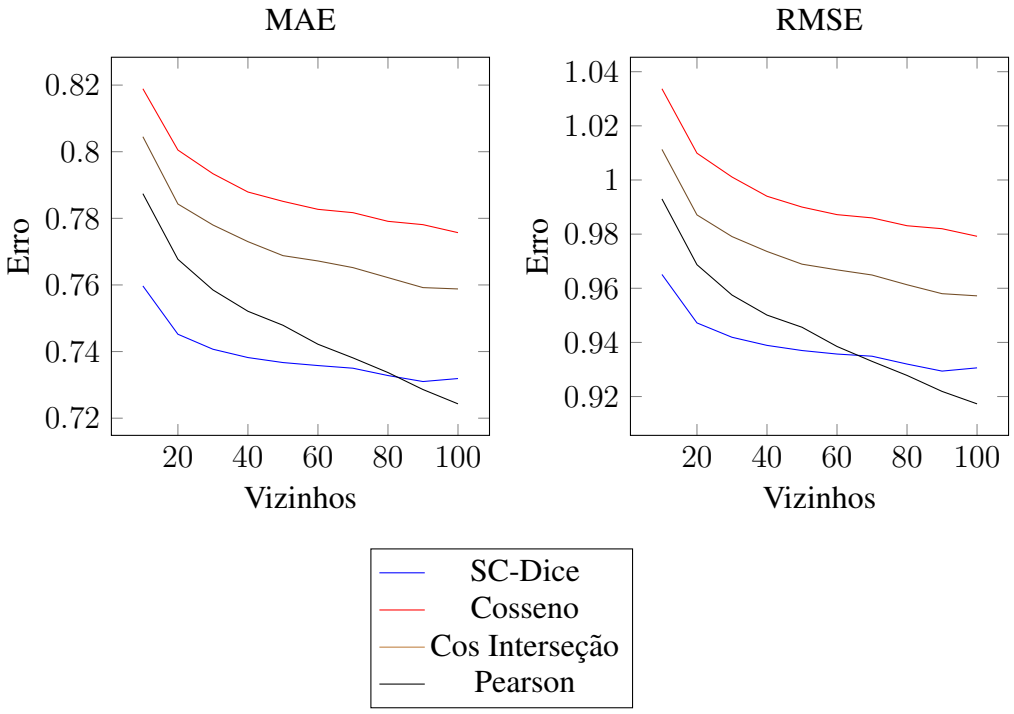

Figura 1. Comparação das métricas de similaridade.

Os gráficos apresentados na figura 1 indicam que a melhor métrica de similaridade até 60 vizinhos é a SC-Dice. Para mais vizinhos a similaridade de Pearson se sai melhor. Vale lembrar que quando o número de vizinhos aumenta muito, a similaridade de Pearson pode enfrentar o problema de *overfitting*, quando os dados se adequam demais ao conjunto de treino.

O gráfico da figura 2 mostra testes realizados com diferentes limiares para a métrica SC-Dice. O valor 0 indica que somente vizinhos com notas iguais para um mesmo item serão consideradas, e o valor 4 considera apenas que os usuários avaliaram ou não um item. O melhor valor encontrado foi para o limiar 1.

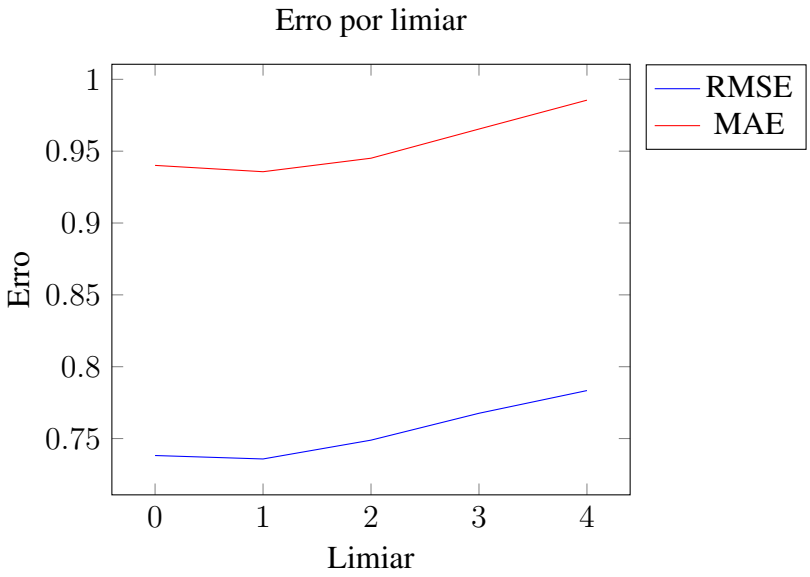

Figura 2. Comparação de valores de limiar para o SC-Dice.

O gráfico da figura 3 mostra a cobertura do algoritmo de acordo com o número de vizinhos. Os dados foram obtidos pela métrica do Cosseno Interseção. As outras métricas seguem um comportamento semelhante, decaindo de modo quase linear à medida que o número de vizinhos aumenta.

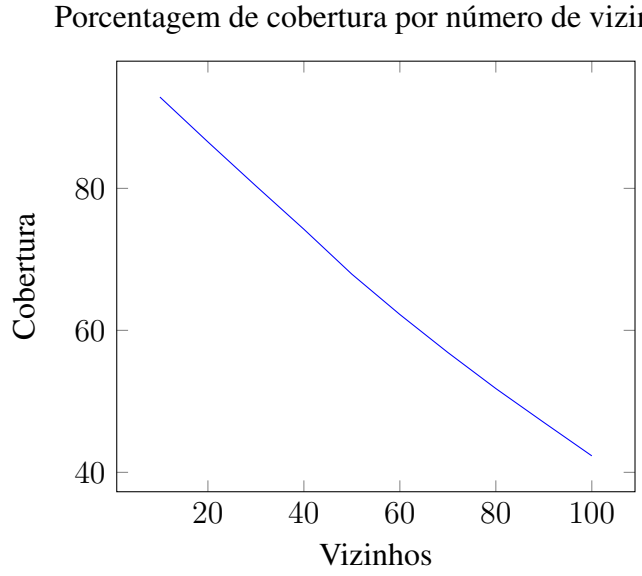

Figura 3. Porcentagem de cobertura por número de vizinhos (Cosseno Interseção).

5. Conclusão

O objetivo principal deste trabalho foi propor uma análise da qualidade da predição de recomendações utilizando o algoritmo KNN, que é uma amplamente utilizado nos sistemas de recomendação com filtragem colaborativa. Foram apresentadas quatro técnicas de similaridade já conhecidas da literatura e foi proposta uma nova medida de similaridade, que se mostrou bastante eficiente.

Pode-se observar que o ajuste dos parâmetros do algoritmo é fundamental para que ele apresente bons resultados. O principal parâmetro deste algoritmo é o número de vizinhos k a ser considerado. Encontrar o k ótimo não é uma tarefa trivial e requer muitos testes empíricos. Testes estes que demoram para ser computados, dada a grande quantidade de dados e complexidade do algoritmo. Nos testes aqui realizados podemos concluir que os resultados melhoram a medida que k aumenta. No entanto, isso implica na diminuição de cobertura. A medida que apresentou o melhor resultado foi a de Pearson, com MAE = 0.7185 e RMSE = 0.9145, no entanto ela cobriu apenas 32.00% da base de dados. Observou-se também, que para um número de vizinhos menor que 60, em média, a similaridade proposta, SC-Dice, apresentou os melhores resultados e com uma cobertura semelhante à de Pearson.

Por fim, conclui-se que encontrar uma boa medida de similaridade e realizar os ajustes corretos dos parâmetros é fundamental para que o sistema alcance bons resultados. Saber ponderar entre maior cobertura e melhor predição é fundamental, afinal, aumentar um implica em diminuir o outro.

Referências

Burke, R. (1999). Integrating knowledge-based and collaborative-filtering recommender systems. In *Proceedings of the Workshop on AI and Electronic Commerce*, pages 69–72.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Goldberg, D., Nichols, D., Oki, B. M., and Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. (1994). Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 175–186. ACM.

Shapira, B., Ricci, F., Kantor, P. B., and Rokach, L. (2011). Recommender systems handbook.

Sørensen, T. (1948). {A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons}. *Biol. Skr.*, 5:1–34.

Tail, L. (2006). Why the future of business is selling less of more.