

Willian Alecsander Farias Costa

**Recomendação de Filmes Utilizando
Filtragem Colaborativa**

Niterói - RJ, Brasil

03 de agosto de 2016

Willian Alecsander Farias Costa

**Recomendação de Filmes Utilizando
Filtragem Colaborativa**

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador: Profa. Dra. Jessica Quintanilha Kubrusly

Niterói - RJ, Brasil

03 de agosto de 2016

Willian Alecsander Farias Costa

**Recomendação de Filmes Utilizando
Filtragem Colaborativa**

Monografia de Projeto Final de Graduação sob o título “*Recomendação de Filmes Utilizando Filtragem Colaborativa*”, defendida por Willian Alecsander Farias Costa e aprovada em 03 de agosto de 2016, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Profa. Dra. Jessica Quintanilha Kubrusly
Departamento de Estatística – UFF

Profa. Dra. Ludmilla da Silva Viana Jacobson
Departamento de Estatística – UFF

Prof. Dr. Wilson Calmon Almeida dos Santos
Departamento de Estatística – UFF

Niterói, 03 de agosto de 2016

C837 Costa, Willian Alecsander Farias.
Recomendação de filmes utilizando filtragem colaborativa /
Willian Alecsander Farias. – Niterói, RJ: [s.n.], 2016.

51f.

Orientador: Prof.Drª Jessica Quintanilha Kubrusly.
TCC (Bacharelado em Estatística) – Universidade Federal Fluminense, 2016.

1. Sistemas de recomendação. 2. Filtragem colaborativa. 3. Estatística matemática.
I. Título.

CDD 519.5

Resumo

Com o crescimento da Internet, uma grande quantidade de pessoas buscam rapidez e eficiência na pesquisa, na compra e na aquisição de produtos e serviços que são disponibilizados na Web. Para aumentar a capacidade de indicação de itens para usuários, surgem os Sistemas de Recomendação. Uma das técnicas mais conhecidas para se recomendar itens é a Filtragem Colaborativa, que será usada neste trabalho para comparar 8 diferentes abordagens oriundas desta técnica. O objetivo desta comparação será observar qual ou quais abordagens apresentam os melhores resultados para a recomendação de filmes para telespectadores. As 8 abordagens foram formadas a partir da combinação de escolhas da Filtragem Colaborativa. Entre elas estão: o tipo de semelhança adotado, usuário ou item; a escala de avaliação; e a forma de definição do grupo de semelhança, por número fixo ou constante de similaridade. Verificou-se um resultado ineficiente para as 8 abordagens, devido ao alto índice de recomendação, mesmo para filmes não relevantes, ocasionando uma baixa Especificidade para todas as abordagens. Apesar do resultado ruim, a Abordagem 5, que leva em conta a semelhança entre itens, escala *Likert* e o número fixo de itens para a definição do grupo, apresentou melhor desempenho. A fim de melhorar o desempenho do método, foi formulada uma análise de sensibilidade, que buscou escolher melhor os parâmetros. A análise de sensibilidade indicou valores a serem adotados para os parâmetros do modelo e, quando a Abordagem 5 foi rodada para tais parâmetros, os resultados se mostraram um pouco melhores.

Palavras-chaves: Sistemas de Recomendação, Filtragem Colaborativa, Comparação de Abordagens.

Dedicatória

Dedico este trabalho à minha família, meu pai José Wilson, minha mãe Maria Aparecida e meu irmão Antony, que sempre me apoiaram nos momentos difíceis. A eles, meu muito obrigado.

Agradecimentos

Agradeço primeiramente a Deus por me dar saúde e determinação para continuar estudando e almejando m futuro melhor.

À minha mãe, Maria Aparecida, que sempre fez tudo por mim desde a minha infância até hoje.

À meu pai, José Wilson, por todos os conselhos e incentivos que ele me deu e ainda dá.

À meu irmão, Antony, pela amizade e companheirismo que sempre demonstrou para que eu não desanimasse.

Aos colegas de UFF, que sempre me ajudaram quando foi preciso e também ofereceram momentos de descontração.

A minha orientadora Jessica Quintanilha Kubrusly, que me propôs o tema deste trabalho e me ajudou muito a buscar os resultados desejáveis deste.

A todos os meus professores, por todos os ensinamentos passados.

A todos estes, Minha gratidão!

Sumário

Lista de Figuras

Lista de Tabelas

1	Introdução	p. 12
2	Objetivos	p. 15
3	Materiais e Métodos	p. 16
3.1	Materiais	p. 16
3.2	Filtragem Colaborativa	p. 17
3.2.1	Matriz de Utilidade	p. 17
3.2.2	Medida de Similaridade	p. 19
3.2.3	Matriz de Similaridade	p. 21
3.2.4	Como Prever a Avaliação de um Item por um Usuário	p. 23
3.3	Diferentes abordagens para a Recomendação	p. 25
3.4	Diferentes formas de Realizar uma Recomendação	p. 26
3.5	Medidas de Comparação para uma Recomendação	p. 27
3.5.1	<i>Recall</i> , <i>Precision</i> e Medida F	p. 29
3.5.2	Sensibilidade e Especificidade	p. 30
4	Análise dos Resultados	p. 31
4.1	Descrição da Amostra	p. 31
4.2	Detalhes da Implementação	p. 32

4.3	Resultados	p. 34
4.4	Resultados com critérios ajustados	p. 38
4.4.1	Escolha das 100 avaliações que serão retiradas da amostra	p. 39
4.4.2	Apresentação dos novos resultados	p. 40
5	Conclusões	p. 45
	Referências	p. 48
	Anexo A – Código do trabalho no software R	p. 49

Lista de Figuras

1	Abordagem para a recomendação de itens	p. 26
---	--	-------

Lista de Tabelas

1	Matriz de Utilidade	p. 17
2	Exemplo de uma Matriz de Utilidade	p. 18
3	Matriz de Similaridade entre usuários	p. 21
4	Matriz de Similaridade entre usuários do exemplo	p. 21
5	Matriz de Similaridade entre itens	p. 22
6	Matriz de Similaridade entre itens do exemplo	p. 23
7	Tabela de classificação	p. 28
8	Medidas Descritivas para o Número de Filmes assistidos por Usuário . . .	p. 31
9	Medidas Descritivas para o Número de Usuários que assistiram cada filme	p. 31
10	Comparação entre número de entradas avaliadas e não avaliadas	p. 32
11	Opinião dos usuários em relação aos filmes disponíveis no questionário . .	p. 32
12	Tabela de classificação para a Abordagem 1	p. 34
13	Tabela de classificação para a Abordagem 2	p. 34
14	Tabela de classificação para a Abordagem 3	p. 35
15	Tabela de classificação para a Abordagem 4	p. 35
16	Tabela de classificação para a Abordagem 5	p. 35
17	Tabela de classificação para a Abordagem 6	p. 36
18	Tabela de classificação para a Abordagem 7	p. 36
19	Tabela de classificação para a Abordagem 8	p. 36
20	Métricas de classificação	p. 37
21	Resultados ajustados para a Abordagem 1	p. 41
22	Resultados ajustados para a Abordagem 2	p. 41

23	Resultados ajustados para a Abordagem 3	p.41
24	Resultados ajustados para a Abordagem 4	p.41
25	Resultados ajustados para a Abordagem 5	p.42
26	Resultados ajustados para a Abordagem 6	p.42
27	Resultados ajustados para a Abordagem 7	p.42
28	Resultados ajustados para a Abordagem 8	p.42
29	Análise de sensibilidade detalhada - Abordagem 5	p.44

1 Introdução

A ascensão da Internet nos últimos anos proporcionou vários benefícios à sociedade. Dentre eles pode-se citar: a agilidade e velocidade nas trocas de informação; a superação de distâncias; a facilidade de pesquisa; a liberdade de escolha de acesso a todos; e aplicações na Web. Como exemplos de aplicações na Web podem ser citados as compras on-line e as pesquisas de artigos, músicas, livros, entre outros tipos de itens. Neste contexto, cada usuário busca por informação que possa auxiliá-lo na realização de suas tarefas de maneira rápida e eficiente, porém mesmo com ferramentas de busca o usuário tende a despende muito tempo nesta busca.

Grandes portais na Web recebem uma grande quantidade de usuários e guardam informações sobre seu histórico. Isso gera um conjunto de dados que pode ser útil para análises de consumo, que podem servir para recomendar itens aos usuários. Esta recomendação facilita na compra e aquisição dos produtos e, por consequência, o aumento do consumo e a satisfação dos clientes. Visando melhorar a recomendação para produtos de interesse surgem os Sistemas de Recomendação, que aumentam a capacidade e a eficácia do processo de indicação de itens. Veja o artigo de Resnick et al. (RESNICK, 1997)[1]

Mas por que realizar estas recomendações? Em particular, para as lojas que oferecem a opção de compras on-line é muito interessante que o cliente seja provocado a gastar mais comprando uma maior quantidade de produtos, visto que via Internet não é possível que haja um profissional de vendas trabalhando para isso. Com boas recomendações, o site será acessado com mais frequência, chamando a atenção de anunciantes interessados em fazer propaganda no site e gerando renda. Como exemplos de portais que usam estes sistemas de Recomendações podemos citar a *Amazon*, que recomenda produtos para seus clientes, a *Netflix*, que recomenda filmes e séries de televisão em seu site, e sites de notícias, que recomendam manchetes a seus leitores.

Existem dois tipos de sistemas de recomendação. O primeiro é baseado no perfil de cada usuário, avaliando variáveis como estado civil, número de filhos, idade, etc. O outro

é baseado no padrão de consumo, que é o que será abordado neste trabalho. Dentre algumas técnicas usadas para recomendar itens a usuários, pode-se citar a Filtragem de Informação, a Filtragem Baseada em Conteúdo, a Filtragem Híbrida, a Mineração de textos entre outros (CAZELLA, 2010) [2].

Outra técnica é a Filtragem Colaborativa, que constitui-se em uma das técnicas de maior sucesso na web no que tange a recomendação. Esta técnica baseia-se na chamada filtragem social, (SHARDANAND, 1995) [3], onde as opiniões de usuários “vizinhos” ao usuário-alvo da recomendação são fundamentais no cálculo da recomendação a ser feita. Esses ditos “vizinhos” são geralmente identificados através do uso de técnicas estatísticas que são aplicadas a um grupo de usuários que possuem um histórico de preferências similares ao usuário-alvo, (BALABANOVIC, 1997) [4], por exemplo, usuários que consumiram itens semelhantes ao usuário alvo e apresentaram avaliações semelhantes para estes itens. Algumas vantagens desta técnica é a velocidade na recomendação, já que esta utiliza apenas avaliações e não exige nenhuma outra informação adicional sobre usuários ou itens.

Uma empresa que utiliza esta técnica para recomendar produtos a seus clientes é a *Amazon*, que usa algoritmos de recomendação item-por-item juntamente com a técnica de Filtragem Colaborativa para personalizar a loja online para cada consumidor. Diferentemente do uso tradicional da técnica da Filtragem Colaborativa, o algoritmo produz recomendações baseadas na semelhança entre itens e não na semelhança entre usuários, indicando itens em tempo real e tendo resultados de alta qualidade.

A técnica utilizada pelo site da *Amazon* consiste na identificação de itens vizinhos, utilizando-os para recomendar ou não um item-alvo a algum cliente, ou seja, produtos que foram avaliados de maneira parecida pelos consumidores. Para determinar se os itens são “vizinhos” entre si ou não, o algoritmo calcula alguma métrica para verificar a semelhança entre dois itens. Na sequência, se constrói uma matriz quadrada com dimensão igual ao número de produtos colocados à venda. Em cada uma das entradas desta matriz está um valor que representa a similaridade entre dois itens separadamente. Após a formação da matriz, o algoritmo acha itens vizinhos, agrega-os, e então faz as recomendações destes itens mais correlacionados para cada um dos usuários. Esta computação é muito rápida, dependendo somente do número de itens que o usuário comprou ou avaliou, e não o total de itens a venda (LINDEN, 2003) [5].

O objetivo deste trabalho é estudar e aplicar a Filtragem Colaborativa para fazer recomendações de filmes para usuários a partir de um banco de dados onde foram ava-

liados 50 filmes por 525 entrevistados. Estas recomendações serão realizadas a partir de 8 abordagens, sendo estas formadas pela combinação de escolhas dentro da Filtragem Colaborativa. Dentro da Filtragem Colaborativa existem algumas escolhas sobre a metodologia a se fazer. Combinando tais escolhas é possível criar diferentes abordagens. Após isto, estas abordagens serão comparadas entre si para verificar quais, ou qual, tem melhor desempenho quando usadas, ou usada, para recomendar filmes para telespectadores. As comparações serão feitas usando um banco de dados de 50 filmes e 525 entrevistados, criado para a monografia de Chavez (CHAVEZ, 2014) [6].

Este trabalho está dividido em cinco capítulos, sendo o primeiro a introdução. No Capítulo 2, são apresentados os objetivos geral e específico. Em seguida, no Capítulo 3 apresentamos a metodologia da Filtragem Colaborativa, quais as ferramentas necessárias para a sua aplicação e como avaliar seus resultados, incluindo seus diversos desdobramentos. No Capítulo 4, são apresentados os resultados da recomendação dos itens e por fim, no Capítulo 5 são apresentadas as conclusões finais.

2 Objetivos

Objetivo Geral

Discutir a metodologia da Filtragem Colaborativa sobre a recomendação de filmes.

Objetivos Específicos

Comparar diferentes Abordagens da Filtragem Colaborativa, a fim de mostrar qual delas apresenta os melhores resultados para a recomendação.

3 Materiais e Métodos

3.1 Materiais

O material que será objeto de análise pelo método descrito a seguir é composto pela avaliação de um grupo de usuários referente a um grupo de itens. É importante ressaltar que neste método nem todos os itens devem ser avaliados. O banco de dados foi coletado pela aluna Juliana Chaves Giglio para a sua monografia (CHAVEZ, 2014) [6], na ocasião a aluna disponibilizou um questionário na Internet com 50 filmes para a avaliação. Ao fim da coleta de dados, 530 usuários haviam respondido as questões referentes às informações pessoais e às avaliações dos filmes. Entre eles, 4 não assistiram filme algum e por isso serão desconsiderados nesta análise.

Cada usuário foi orientado para avaliar os filmes de acordo com a escala Likert (CAZELLA, 2009) [7] de 5 pontos, onde cada item é classificado como: “Gostei muito”, “Gostei”, “Indiferente”, “Não gostei” ou “Detestei”. Desta forma, foi possível criar uma tabela com 526 linhas e 50 colunas, onde cada linha representa um usuário e cada coluna representa um filme. As entradas desta tabela indicam a avaliação que cada usuário deu para os filmes que assistiu. Toda a programação que gerou os resultados foi implementada no software estatístico *R*.

Vale ressaltar que existem diferentes formas de se avaliar um item dependendo do tipo de cada um. Por exemplo, um item pode ser avaliado através da opinião de cada usuário sobre ele, se o usuário adquiriu-o ou não, se ele funciona ou não, entre outras formas de avaliação.

Como já foi dito, existem diversos tipos de itens que podem ser avaliados. Dentre eles se destacam filmes, livros, eletroeletrônicos, artigos de vestuário, eletrodomésticos, etc.

3.2 Filtragem Colaborativa

Esta técnica se baseia na análise de preferências comuns em um grupo de pessoas, assim os conteúdos que podem ser recomendados são filtrados com base nas avaliações (feedback) feitas pelos usuários sobre os mesmos itens. Para cada usuário procura-se identificar um conjunto de “vizinhos próximos”, que são assim classificados por possuírem um comportamento semelhante (ADOMAVICIUS, 2005) [8].

Para isto, usa-se a similaridade entre itens ou a similaridade entre usuários para fazer as recomendações. Ainda neste capítulo serão definidos formalmente os conceitos de similaridade entre usuários ou itens, mas de forma geral entende-se por usuários semelhantes aqueles que avaliam diferentes itens de forma semelhante. Já os itens semelhantes são aqueles avaliados de forma semelhante por diferentes usuários (REATEGUI, 2005) [9].

Primeiramente, com os dados brutos se monta uma matriz, chamada de Matriz de Utilidade, e a partir desta se calcula a similaridade entre usuários ou itens, como apresentado nas Seções 3.2.1 e 3.2.2. Mais a frente, veremos como esta similaridade é medida e interpretada pela Filtragem Colaborativa.

3.2.1 Matriz de Utilidade

A matriz de utilidade representa o conjunto de todas as entradas referentes às avaliações dadas pelos usuários. Nesta, os usuários são representados pelas linhas e os itens são representados pelas colunas como mostra a Tabela 1 abaixo.

Tabela 1: Matriz de Utilidade

	Item 1	Item 2	...	Item m
Usuário 1	$x_{1,1}$	$x_{1,2}$...	$x_{1,m}$
Usuário 2	$x_{2,1}$	$x_{2,2}$...	$x_{2,m}$
...
Usuário n	$x_{n,1}$	$x_{n,2}$...	$x_{n,m}$

Nesta matriz, cada entrada $x_{i,j}$, $i = 1, \dots, n$ e $j = 1, \dots, m$, representa a avaliação do usuário i ao item j . Se o usuário i não avaliou o item j temos $x_{i,j} = 0$. Veja que cada usuário pode ser considerado um vetor no espaço \mathbb{R}^m representado pela sua linha de matriz. Da mesma forma, pode-se considerar cada item como um vetor no espaço \mathbb{R}^n representado pela sua coluna da matriz.

Dentre as diferentes formas de preenchimento da matriz se destaca o uso da escala

Likert (CAZELLA, 2009) [7] de 5 pontos, onde cada item é classificado como: “Gostei muito”, “Gostei”, “Indiferente”, “Não gostei” ou “Detestei”. As notas atribuídas para cada classificação são 1, 2, 3, 4 ou 5 respectivamente. Assim, na Matriz de Utilidade $x_{i,j} \in \{1, 2, 3, 4, 5\}$.

Outra forma de preenchimento é usar 0 ou 1 em vez dos números da Escala Likert (CAZELLA, 2009) [7]. Esta forma se aplica, por exemplo para sites de compra, que sabem apenas os itens comprados por cada usuário e não a opinião sobre estes itens. Assim, os itens já comprados pelos usuários são indicados por 1 e os demais por 0.

Vale ressaltar que essa matriz é muito grande e que a maioria das entradas são nulas. Por exemplo, o DEC Systems Research Center, um laboratório de pesquisas americano, testou o serviço de recomendação por 18 meses junto com o algoritmo de Filtragem Colaborativa. Durante esse tempo, 1.628 filmes e vídeos foram colocados em avaliação com 72.916 usuários fornecendo avaliações para estes itens. Porém, de um total de $1.628 \times 72.916 = 118.707.248$ entradas, apenas 2.811.983 foram preenchidas, ou seja, aproximadamente 97.6% das entradas eram vazias e a matriz de utilidade era bastante esparsa, ou seja, com muitas entradas vazias (MCJONES, 1997) [10].

Exemplo 3.2.1 Temos na Tabela 2 um exemplo de uma Matriz de Utilidade que representa a avaliação de 6 usuários referentes a 5 filmes. Neste exemplo, os itens a serem avaliados são os filmes. Para esta avaliação será utilizada a escala Likert de 5 pontos. Entradas em branco significam que o usuário não avaliou aquele filme.

Tabela 2: Exemplo de uma Matriz de Utilidade

	Filme 1	Filme 2	Filme 3	Filme 4	Filme 5
Usuário 1	5		2	4	1
Usuário 2		1		2	
Usuário 3	2	1	4		5
Usuário 4	4		2	3	3
Usuário 5			3	1	
Usuário 6		3			3

Pode-se perceber na Tabela 2 há uma boa quantidade de espaços em branco, ou seja, filmes que não foram avaliados. Na prática a matriz é ainda mais esparsa. O objetivo da Filtragem Colaborativa é prever estes espaços em branco ou entradas nulas.

3.2.2 Medida de Similaridade

Para definir a similaridade entre usuário ou itens, se faz necessária a definição de uma medida de similaridade para os cálculos que seguem. De todas as medidas de similaridade que podem ser calculadas, a que será usada neste texto é a **Similaridade Cosseno**. Esta calcula a proximidade entre dois vetores a partir do cosseno do ângulo formado entre eles. Quanto maior o cosseno, menor o ângulo entre eles e maior a proximidade entre os vetores.

Considere \mathbf{x} e \mathbf{y} vetores no espaço \mathbb{R}^n , e α é o ângulo entre eles. A similaridade entre \mathbf{x} e \mathbf{y} a partir do cosseno é dada por:

$$S(\mathbf{x}, \mathbf{y}) = \cos(\alpha) = \frac{\mathbf{x}'\mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}. \quad (3.1)$$

Veja que o cosseno varia entre -1 e 1. Porém, como estas avaliações sempre serão positivas, o cosseno não irá assumir valores negativos, variando de 0 a 1. Quanto mais próximo estiver de 1, menor o ângulo entre \mathbf{x} e \mathbf{y} , sendo maior a proximidade entre os vetores.

Medida de similaridade entre usuários

Para o cálculo de similaridade entre usuários, vai se considerar cada usuário como um vetor no \mathbb{R}^m , sendo, estes, vetores-linha da Matriz de Utilidade. Por exemplo, o usuário i será representado pela i -ésima linha da Matriz de Utilidade (\mathbf{x}_i); Já o usuário k será representado pela k -ésima linha da Matriz de Utilidade (\mathbf{x}_k). A partir daí, calculam-se as similaridades cossenos entre os usuários i e k pela expressão abaixo:

$$s_{i,k}^U = S(\mathbf{x}_i, \mathbf{x}_k) \quad (3.2)$$

Exemplo 3.2.2 *A partir desta medida apresentada na Equação 3.2, qual seria a similaridade cosseno entre os usuários 1 e 3 da tabela 2?*

$$S_{1,3}^U = \frac{(5 \times 2) + (2 \times 4) + (1 \times 5)}{\sqrt{5^2 + 2^2 + 4^2 + 1^2} \sqrt{2^2 + 1^2 + 4^2 + 5^2}} = 0,5,$$

e qual seria a similaridade entre os usuários 1 e 4?

$$S_{1,4}^U = \frac{(5 \times 4) + (2 \times 2) + (4 \times 3) + (1 \times 3)}{\sqrt{5^2 + 2^2 + 4^2 + 1^2} \sqrt{4^2 + 2^2 + 3^2 + 3^2}} = 0,9328$$

Os usuários 1 e 4 deram classificações semelhantes aos filmes que assistiram em comum, já quando comparamos os usuários 1 e 3, vemos que as notas para os filmes que ambos classificaram são opostas. Logo, é razoável que a similaridade cosseno entre os usuários 1 e 4 possua um valor mais alto que a entre os usuários 1 e 3, indicando maior similaridade entre os usuários 1 e 4 quando comparados com a similaridade entre os usuários 1 e 3.

Medida de similaridade entre itens

Já para os itens, vai se considerar cada item como um vetor no \mathbb{R}^n , sendo, estes vetores-coluna da Matriz de Utilidade. Por exemplo, o item j será representado pela j -ésima coluna da Matriz de Utilidade (\mathbf{x}_j); Já o item l será representado pela l -ésima coluna da Matriz de Utilidade (\mathbf{x}_l). A partir daí, calculam-se as similaridades cosseno entre os itens j e l pela expressão abaixo:

$$s_{j,l}^I = S(\mathbf{x}_j, \mathbf{x}_l) \quad (3.3)$$

Exemplo 3.2.3 A partir da medida apresentada na Equação 3.3, qual seria a similaridade cosseno entre os itens 1 e 4?

$$s_{1,4}^I = \frac{(5 \times 4) + (4 \times 3)}{\sqrt{5^2 + 2^2 + 4^2} \sqrt{4^2 + 2^2 + 3^2 + 1^2}} = 0,8709,$$

e qual seria a similaridade entre os itens 1 e 5?

$$s_{1,5}^I = \frac{(5 \times 1) + (2 \times 5) + (4 \times 3)}{\sqrt{5^2 + 2^2 + 4^2} \sqrt{1^2 + 5^2 + 3^2 + 3^2}} = 0,6068$$

Veja que os filmes 1 e 4 possuem duas avaliações em comum e ambas são parecidas, já os filmes 1 e 5 também possuem duas avaliações em comum, porém estas são completamente opostas. Logo, a similaridade cosseno do Filme 1 para o Filme 4 é menor que a do Filme 1 para o Filme 5.

A similaridade cosseno é bastante eficiente para calcular a similaridade entre usuários ou itens, pois os vetores linhas e colunas da Matriz de Utilidade têm muitos zeros e, com isso, o produto interno entre eles é mais fácil de calcular, diminuindo bastante os custos computacionais. Porém, para vetores múltiplos (com ângulo 0°) a similaridade cosseno é máxima, apesar de não serem iguais.

Existem outras medidas de similaridade, entre elas pode-se citar, por exemplo a distância de Jaccard. (RAJARAMAN, 2012) [11].

3.2.3 Matriz de Similaridade

No caso da Filtragem Colaborativa estamos interessados em calcular a similaridade entre usuários ou itens. Assim, também será para a Matriz de Similaridade.

Matriz de similaridade entre usuários

A Matriz de Similaridade entre usuários terá dimensão $n \times n$, sendo n o número de usuários (linhas) da matriz de utilidade, onde a posição (i, k) é a similaridade entre os usuários i e k . Veja a Tabela 3.

Tabela 3: Matriz de Similaridade entre usuários

	Usuário 1	Usuário 2	Usuário 3	...	Usuário n
Usuário 1	1	$s_{1,2}^U$	$s_{1,3}^U$...	$s_{1,n}^U$
Usuário 2		1	$s_{2,3}^U$...	$s_{2,n}^U$
Usuário 3			1	...	$s_{3,n}^U$
⋮					⋮
Usuário n					1

Além disso, esta é uma matriz simétrica e com todos os elementos da diagonal principal iguais a 1, ou seja, $s_{i,i} = 1$, $i = 1, \dots, n$. Isto acontece porque a similaridade entre os usuários (i, k) é a mesma entre usuários (k, i) . Em relação à diagonal principal, percebe-se que a similaridade entre um usuário com ele mesmo é máxima, tendo valor igual a 1.

Exemplo 3.2.4 A Matriz de Similaridade entre usuários para a matriz da Tabela 2 é apresentada abaixo:

Tabela 4: Matriz de Similaridade entre usuários do exemplo

	Usuário 1	Usuário 2	Usuário 3	Usuário 4	Usuário 5	Usuário 6
Usuário 1	1	0,5275	0,5	0,9328	0,4663	0,1042
Usuário 2		1	0,0659	0,4353	0,2828	0,3162
Usuário 3			1	0,7415	0,5595	0,6255
Usuário 4				1	0,4617	0,3441
Usuário 5					1	0
Usuário 6						1

Veja que a similaridade entre os usuários 1 e 3, calculada no Exemplo 3.2.2 está apresentada na posição $(1, 3)$. O mesmo vale para a similaridade entre os usuários 1 e 4, apresentada na posição $(1, 4)$.

Matriz de similaridade de itens

A matriz de similaridade entre itens terá dimensão $m \times m$, sendo m o número de itens (colunas) da Matriz de Utilidade, onde a posição (j, l) é a similaridade entre os itens j e l . Veja a Tabela 5.

Tabela 5: Matriz de Similaridade entre itens

	Item 1	Item 2	Item 3	...	Item m
Item 1	1	$s_{1,2}^I$	$s_{1,3}^I$...	$s_{1,m}^I$
Item 2		1	$s_{2,3}^I$...	$s_{2,m}^I$
Item 3			1	...	$s_{3,m}^I$
⋮					⋮
Item m					1

Além disso, assim como a Matriz de Similaridade entre usuários, esta é uma matriz simétrica e com todos os elementos da diagonal principal iguais a 1, ou seja, $s_{j,j} = 1, j = 1, \dots, m$.

Exemplo 3.2.5 A Matriz de Similaridade entre itens para a matriz da Tabela 2 é apresentada abaixo:

Tabela 6: Matriz de Similaridade entre itens do exemplo

	Item 1	Item 2	Item 3	Item 4	Item 5
Item 1	1	0,0899	0,6747	0,8709	0,6068
Item 2		1	0,2099	0,1101	0,6363
Item 3			1	0,5403	0,7348
Item 4				1	0,3578
Item 5					1

Veja que a similaridade entre os itens 1 e 4, calculada no Exemplo 3.2.3 está apresentada na posição (1,4). O mesmo vale para a similaridade entre os itens 1 e 5, apresentada na posição (1,5).

3.2.4 Como Prever a Avaliação de um Item por um Usuário

Dado que alguma entrada $x_{i,j}$ da Matriz de Utilidade é nula, pode-se dizer que o usuário i não avaliou o item j . Com isso, a decisão de recomendar ou não este item é baseada na previsão da avaliação do usuário i ao item j . A seguir será apresentado como fazer previsão de avaliações com base na similaridade entre usuários e itens.

É importante salientar que o grupo de usuários (itens) semelhantes ao usuário i (item j) é criado a partir da Matriz de Similaridade entre usuários (itens) e usa-se algum critério pré-estabelecido para definir a escolha dos usuários semelhantes.

Por exemplo, escolhendo os N usuários (M itens) que apresentam maior semelhança ao usuário i (item j) específico que avaliaram o item j (foram avaliados pelo usuário i). Outra forma é definir uma constante de similaridade que a partir do cálculo das similaridades cosseno entre usuários (itens), ou seja, serão semelhantes ao usuário i (item j) todos os usuários (itens), cuja similaridade com o usuário i (item j) for maior que uma constante de similaridade pré-estabelecida.

Previsão com base na similaridade entre usuários

Quer-se prever a avaliação de um usuário i em relação ao item j . Primeiramente, se definirá um grupo de usuários semelhantes ao usuário i que avaliaram o item j , denotado por G_i^U . Após isto, a nota dada pelo usuário i ao item j será prevista a partir da média ponderada das notas dadas pelos usuários deste grupo ao item j em questão.

A previsão é dada pela fórmula:

$$\hat{A}_{i,j}^U = \frac{\sum_{\{k \in G_i^U | x_{k,j} \neq 0\}} x_{k,j} \times s_{i,k}^U}{\sum_{\{k \in G_i^U | x_{k,j} \neq 0\}} s_{i,k}^U} \quad (3.4)$$

Exemplo 3.2.6 *Suponha que, na Tabela 2 queremos prever a avaliação do usuário 1 ao item 2. Vamos considerar para um usuário k ser semelhante ao usuário 1 é necessário que a similaridade cosseno entre os dois seja maior ou igual a 0.5. Sendo assim, percebe-se através da Matriz de Similaridade apresentada na Tabela 4 que os usuários 2, 3 e 4 formam o grupo de usuários semelhantes ao usuário 1. Porém o usuário 4 não avaliou o item, logo a previsão da nota do usuário 1 ao item 2 pode ser calculada da seguinte forma:*

$$\hat{A}_{1,2}^U = \frac{(1 \times 0,5275) + (1 \times 0,5)}{0,5275 + 0,5} = 1$$

Portanto a previsão é que o usuário 1 teria dado nota 1 ao item 2, baseando-se nesta avaliação.

Previsão com base na similaridade entre itens

Da mesma forma, quer-se prever a avaliação de um usuário i em relação ao item j . Primeiramente, se definirá um grupo de itens semelhantes ao item j , denotado por G_j^I . Após isto, a nota dada pelo usuário i para o item j será prevista a partir da média ponderada das notas dadas aos itens semelhantes ao item j que o usuário i avaliou.

A previsão é dada pela fórmula:

$$\hat{A}_{i,j}^I = \frac{\sum_{\{k \in G_j^I | x_{i,k} \neq 0\}} x_{i,k} \times s_{j,k}^I}{\sum_{\{k \in G_j^I | x_{i,k} \neq 0\}} s_{j,k}^I} \quad (3.5)$$

Exemplo 3.2.7 *Suponha que queremos, agora, prever a avaliação do usuário 3 ao item 4 e que para um item k ser semelhante ao item 4, é necessário que a similaridade cosseno entre os dois também seja maior ou igual a 0.5. Sendo assim, percebe-se através da Matriz de Similaridade apresentada na Tabela 6 que os itens 1 e 3 formam o grupo de itens semelhantes ao item 4. Com isto, a previsão da nota do usuário 3 ao item 4 pode ser calculada da seguinte forma:*

$$\hat{A}_{3,4}^I = \frac{(2 \times 0,8704) + (4 \times 0,5403)}{0,8704 + 0,5403} = 2,766$$

Uma discussão que se faz presente é a questão das previsões contínuas. Elas não serão arredondadas pois se isto é feito, prejudica-se a análise em relação às recomendações. Vamos tomar como base o exemplo mostrado acima, se arredondamos $\hat{A}_{3,4}^I$ para 3, prevemos que o indivíduo 3 se mostrou indiferente ao item 4, mas a previsão é de que o indivíduo não deve gostar do item. Logo, um item irrelevante será recomendado ao usuário 3, prejudicando as análises quanto às recomendações. Os conceitos de relevância e irrelevância serão apresentados na Seção 3.5.

3.3 Diferentes abordagens para a Recomendação

Nesta Seção serão apresentadas 8 diferentes Abordagens usadas para realizar a recomendação dos itens. Cada Abordagem consiste na combinação das seguintes escolhas dentro da Filtragem Colaborativa, sendo estas:

Semelhança entre Usuários ou Itens

Esta escolha consiste no uso de semelhança entre usuários ou itens para avaliar se o item j deve ser recomendado ao usuário i .

Escala de Avaliação: Likert ou Escala -1 a 1

Esta escolha consiste no uso da escala Likert (CAZELLA, 2009) [7] ou da escala de -1 a 1 para avaliar se o item j deve ser recomendado ao usuário i . A escala de -1 a 1 será criada a partir de uma manipulação dos dados originalmente na escala Likert (CAZELLA, 2009) [7]. A idéia é classificar como -1 os itens avaliados como 1 ou 2 na escala Likert (CAZELLA, 2009) [7]. Os demais itens, avaliados como 3, 4 ou 5 na Escala Likert (CAZELLA, 2009) [7] serão classificados como 1. Veja que foi feita a escolha de colocar a avaliação 3 como 1 na escala de -1 a 1 , essa escolha será melhor discutida na Seção 3.4.

Definição dos grupos de semelhança: Número fixo ou constante de similaridade

Esta escolha consiste na definição de grupos semelhantes de usuários ou itens, sendo estes definidos a partir de um número fixo de N usuários (M itens) ou a partir de uma constante de similaridade para a comparação desta com a similaridade cosseno entre usuários (itens).

O organograma apresentado na Figura 1 ilustra bem as 8 possibilidades de Abordagens:

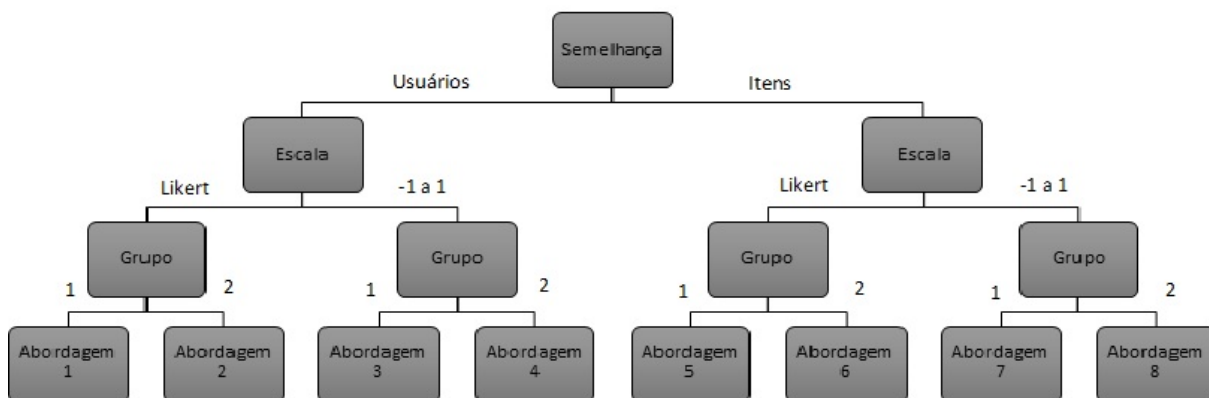


Figura 1: Abordagem para a recomendação de itens

Por exemplo, na Abordagem 3 será considerada a similaridade entre usuários, a escala $(-1$ a $1)$ e o critério de escolher os N usuários semelhantes ao usuário i para recomendar ou não um item j ao usuário i .¹

3.4 Diferentes formas de Realizar uma Recomendação

Neste trabalho o nosso objetivo é realizar recomendações a partir da Filtragem Colaborativa, isso será feito a partir da previsão de avaliação apresentada na Seção 3.3. Para que esse objetivo seja alcançado, deve-se saber quando é interessante recomendar um item para um usuário.

Suponha que a avaliação prevista para o usuário i em relação ao item j foi menor que 3 (ou negativa na escala -1 a 1); neste caso não iremos recomendar o item j ao usuário i . Mas se a avaliação prevista foi maior que 3 (ou positiva na escala -1 a 1), a recomendação do item j ao usuário i será feita.

E se a avaliação prevista foi 3? Deve-se ou não recomendar o item j ao usuário i ? Isso depende de cada problema. Para tomar uma decisão diante desta situação temos de avaliar se preferimos deixar de recomendar um item para um usuário interessado nele ou preferimos recomendar um item que o usuário pode não gostar.

¹1 - Grupos de similaridades formados através de um número específico de elementos similares;
2 - Grupos de similaridades formados através de uma constante de similaridade pré-estabelecida

Neste trabalho, iremos recomendar os itens para os usuários que tiveram como previsão uma nota 3, pois entendemos que para o nosso problema é mais preocupante deixar de recomendar itens (no caso, filmes) para usuários que possivelmente irão se interessar por estes itens do que recomendar itens para usuários que possivelmente possam não gostar destes itens.

3.5 Medidas de Comparação para uma Recomendação

A seguir serão apresentadas as medidas de comparação entre as 8 Abordagens apresentadas na Seção 3.3. Para realizar as previsões e recomendações de todas as entradas não-nulas da Matriz de Utilidade seguiremos os seguintes passos:

1. Para uma entrada selecionada, substituir $x_{i,j}$ por 0 na Matriz de Utilidade.
2. Para cada Abordagem decidir entre recomendar ou não o item j para o usuário i .
3. Contabilizar se a recomendação foi correta ou não, usando tabelas de dupla entrada para as 8 abordagens.
4. Repetir este processo para todas as entradas não-nulas da Matriz de Utilidade.

Assim após realizar os 4 passos acima em cada entrada não nula, para cada Abordagem será montada uma tabela de dupla entrada com as seguintes classificações: (Recomendado ou Não Recomendado) e (Relevante ou Irrelevante). O item j será recomendado para o usuário i se $\hat{A}_{i,j} \geq 3$. Caso contrário, o item j não será recomendado ao usuário i . Agora vamos dizer que o item j será “relevante” para o usuário i se $x_{i,j} \geq 3$. Caso contrário o item j será “irrelevante” para o usuário i .

O conceito de relevância vem do fato de um item j ser bem avaliado pelo usuário i . Portanto, este item deve ser recomendado para ao usuário i , baseado nas Abordagens já citadas na Seção 3.3.

Abaixo será apresentado um exemplo para mostrar de forma clara estes conceitos. Vamos apresentá-los para a avaliação do usuário 1 referente ao item 1:

Exemplo 3.5.1 Como mostrado na matriz 2, o item 1 será relevante em relação ao usuário j pois $x_{1,1} = 5 \geq 3$.

Agora vamos anular esta nota e calcular a previsão desta avaliação baseada nas similaridade entre usuários e itens, usando os conceitos mostrados na seção 3.2.4. Primeiramente, vamos fazer os cálculos considerando a similaridade entre usuários.

De acordo com a matriz 4, sabe-se que os usuários semelhantes ao usuário 1 são os usuários 2,3 e 4. Porém, o usuário 2 não avaliou o item 1 e, portanto, será retirado deste cálculo. Com isso, a previsão desta entrada será dada por:

$$\hat{A}_{1,1}^U = \frac{(2 \times 0,5) + (4 \times 0,9328)}{0,5 + 0,9328} \approx 3,3021$$

A partir deste resultado, recomenda-se o item 1 ao usuário 1 pois $\hat{A}_{1,1}^U \approx 3.3021 > 3$ se considerarmos a similaridade entre usuários.

Agora, vamos fazer os cálculos considerando a similaridade entre itens. De acordo com a Matriz 6, sabe-se que os itens semelhantes ao item 1 são os itens 3,4 e 5. Com isso, a previsão desta entrada será dada por:

$$\hat{A}_{1,1}^I = \frac{(2 \times 0,6747) + (4 \times 0,8709) + (1 \times 0,6068)}{0,6747 + 0,8709 + 0,6068} \approx 2,5273$$

A partir deste resultado, não se recomenda o item 1 ao usuário 1 pois $\hat{A}_{1,1}^I \approx 2,5273 < 3$ se considerarmos a similaridade entre itens.

A partir dessas classificações podemos montar a tabela de classificação a seguir:

Tabela 7: Tabela de classificação

	Relevante	Irrelevante	Total
Recomendado	$N_{rel,rec}$	$N_{irr,rec}$	N_{rec}
Não Recomendado	$N_{rel,n}$	$N_{irr,n}$	N_n
Total	N_{rel}	N_{irr}	N

onde,

- $N_{rel,rec}$: total de pares ordenados relevantes e recomendados;
- $N_{irr,rec}$: total de pares ordenados irrelevantes e recomendados;
- $N_{rel,n}$: total de pares ordenados relevantes e não recomendados;
- $N_{irr,n}$: total de pares ordenados irrelevantes e não recomendados;

- N_{rec} : total de pares ordenados recomendados = $N_{irr,rec} + N_{rel,rec}$
- N_n : total de pares ordenados não recomendados = $N_{irr,n} + N_{rel,n}$
- N_{rel} : total de pares ordenados relevantes = $N_{rel,n} + N_{rel,rec}$
- N_{irr} : total de pares ordenados irrelevantes = $N_{irr,n} + N_{irr,rec}$

A partir desta tabela será possível calcular 5 medidas que serão levadas em consideração para comparar as 8 diferentes Abordagens, sendo estas: *Recall*, *Precision*, Medida F, Sensibilidade e Especificidade. Para mais detalhes veja Chaves [6].

3.5.1 *Recall*, *Precision* e Medida F

A métrica de classificação *Recall* é definida como a proporção de itens relevantes e recomendados com relação ao número total de itens relevantes existentes, ou seja, é a probabilidade de um item relevante ser recomendado. Sua fórmula é dada por:

$$Recall = R = \frac{N_{rel,rec}}{N_{rel}} \quad (3.6)$$

A métrica *Precision* é definido como a proporção de itens relevantes e recomendados com relação ao número total de itens recomendados, ou seja, é a probabilidade de um item recomendado ser relevante. Sua fórmula é dada por:

$$Precision = P = \frac{N_{rel,rec}}{N_{rec}} \quad (3.7)$$

Estas métricas são as mais utilizadas para avaliar a capacidade de recomendação de sistemas. Quanto maior o *Recall* e quanto maior o *Precision*, melhor é a recomendação. Para se obter uma métrica única foi criada a medida F, que é uma combinação das métricas *Recall* e *Precision*. Sua fórmula é dada por:

$$F = \frac{2RP}{P + R} \quad (3.8)$$

A medida F $\in [0, 1]$, e quanto maior for esta medida melhor a capacidade de recomendação do método.

3.5.2 Sensibilidade e Especificidade

A partir da tabela de classificação também é possível definir a Sensibilidade e Especificidade do sistema de recomendação (MARTINEZ, 2003) [12].

A Sensibilidade é definida como a proporção de itens relevantes e recomendados com relação ao número total de itens recomendados, ou seja, é probabilidade de um item ser recomendado dado que ele é relevante:

$$\text{Sensibilidade} = \mathbb{P}(\text{item foi recomendado} \mid \text{item é relevante}) = \frac{N_{rel,rec}}{N_{rel}} \quad (3.9)$$

Veja que a Expressão 3.9 é definida da mesma forma que a expressão do *Recall* 3.6, ou seja, ambas as métricas têm a mesma definição.

Por fim, a Especificidade é definida como a proporção de itens irrelevantes e não recomendados com relação ao número total de itens não recomendados, ou seja, é probabilidade de um item ser não recomendado dado que ele é irrelevante:

$$\text{Especificidade} = \mathbb{P}(\text{item não foi recomendado} \mid \text{item é irrelevante}) = \frac{N_{irr,n}}{N_{irr}} \quad (3.10)$$

Por se tratarem de medidas que aferem a capacidade do sistema acertar, esperamos que estes valores sejam altos, próximos de 1.

4 Análise dos Resultados

4.1 Descrição da Amostra

Todos os telespectadores foram analisados quanto ao número de filmes que os mesmos assistiram. Na tabela abaixo, pode-se perceber que a média de filmes assistidos foi de 13,73 filmes, o que representa 25 % dos filmes oferecidos, e que 25 % dos telespectadores assistiram mais da metade dos filmes, ou seja, mais de 25 filmes na lista de 50.

Tabela 8: Medidas Descritivas para o Número de Filmes assistidos por Usuário

Estatística	Min	1º quartil	Mediana	Média	3º quartil	Max	Desvio-Padrão
Nº de Filmes	2	8	13	13,73	18	41	7,3627

Da mesma forma, procurou-se analisar o número de usuários que havia assistido cada filme disponibilizado para a avaliação. Pode-se perceber que em média, cada um dos 50 filmes listados no questionário foi visto por 146 usuários entrevistados, a maioria dos filmes não chegou a ser assistidos por sequer 200 usuários e apenas um filme foi assistido mais que 300 vezes, sendo este o filme de comédia nacional “Minha mãe é uma peça”. Ele foi assistido 361 vezes. Já o filme menos assistido foi a animação “Caminhando com Dinossauros”, que foi visto por apenas 20 usuários.

Tabela 9: Medidas Descritivas para o Número de Usuários que assistiram cada filme

Estatística	Min	1º quartil	Mediana	Média	3º quartil	Max	Desvio-Padrão
Nº de Filmes	20	66,75	152,5	146	202,5	361	81,78

Dentro deste contexto, é importante analisar descritivamente quantos filmes foram avaliados por cada usuário. Na tabela abaixo, vale ressaltar que a maioria dos itens não foi avaliada por cada usuário e que a matriz de utilidade tem uma grande quantidade de entradas vazias, como já foi dito na Seção 3.2.1.

Neste grupo de entradas avaliadas, também se torna importante descrever a amostra quanto à opinião de cada telespectador em relação ao filme por ele assistido. Vale ressaltar

Tabela 10: Comparação entre número de entradas avaliadas e não avaliadas

Tipo de Entrada	Avaliada	Não Avaliada
Quantidade (Porcentagem)	19224 (72,54%)	7276 (27,46%)

que a tabela abaixo mostra como a maioria das avaliações foi positiva em relação aos filmes já que menos de 10 % destas avaliações estiveram entre “Detestar o filme” ou “Não gostar do filme” e aproximadamente 75 % das avaliações estiveram entre “Gostar do filme” ou “Gostar muito do filme”.

Tabela 11: Opinião dos usuários em relação aos filmes disponíveis no questionário

Avaliação	Quantidade (Porcentagem)
Detestou	203 (2,79%)
Não gostou	516 (7,09%)
Ficou indiferente	1001 (13,76%)
Gostou	2959 (40,67%)
Gostou muito	2597 (35,69%)

4.2 Detalhes da Implementação

Nesta seção serão apresentados os parâmetros utilizados durante a implementação das abordagens já mostradas na Seção 3.3. Entre eles estão os valores adotados para a definição dos grupos de semelhança e o critério usado para recomendar os filmes para os telespectadores.

Para as Abordagens 1 e 3, que usam a semelhança entre usuários (telespectadores) e consideram o tamanho do grupo de usuários semelhantes fixo igual a N , de acordo com a Equação 3.4, foi considerado $N = 10$. Já para as Abordagens 2 e 4, que também usam a semelhança entre usuários (telespectadores) porém consideram que um usuário faz parte do grupo de usuários semelhantes se a similaridade for igual ou superior a uma constante de similaridade específica, foi considerada a constante 0,2.

Para as Abordagens 5 e 7, que usam a semelhança entre itens (filmes) e consideram o tamanho do grupo de itens semelhantes fixo igual a M , de acordo com a Equação 3.5, foi considerado $M = 2$. Já para as Abordagens 6 e 8, que também usam a semelhança entre itens (filmes) porém consideram que um item faz parte do grupo de itens semelhantes se a similaridade for igual ou superior a uma constante de similaridade específica, também foi considerada a constante 0,2.

O critério usado para recomendar um filme a algum telespectador será o seguinte: Para

as 4 abordagens que consideram a escala de avaliação Likert, um filme será recomendado à algum usuário se a previsão da nota, calculada com base em uma das 4 abordagens descritas, for maior ou igual a 3. Para as demais abordagens que consideram a escala -1 a 1 de avaliação, um filme será recomendado à algum usuário se a previsão da nota for maior que 0.

O pseudocódigo para o cálculo das previsões das notas, a realização das recomendações e a formação das tabelas de classificação é mostrado abaixo:

1. Monte uma cópia da Matriz de Utilidade, denominada MU.
2. Para cada (i, j) tal que $MU_{i,j} \neq 0$, faça os passos 3-9 a seguir.
3. Guarde o valor de $MU_{i,j}$ em uma variável *nota*.
4. Faça $MU_{i,j} = 0$.
5. Calcule a matriz de similaridade, a partir da matriz MU modificada.
6. Determine o grupo de usuários (itens) semelhantes a i (j).
7. Com base nestes grupos, calcule $\hat{A}_{i,j}$ de acordo com as Equações 3.4 ou 3.5.
8. Incremente a tabela de dupla entrada de acordo com os valores das notas (Relevante ou Não Relevante) e das previsões (Recomendado ou Não Recomendado).
9. Faça $MU_{i,j} = nota$.

4.3 Resultados

Seguem abaixo os resultados referentes às recomendações dos filmes para as 8 abordagens já descritas na Seção 3.3.

Abordagem 1

Tabela 12: Tabela de classificação para a Abordagem 1

	Relevante	Irrelevante	Total
Recomendado	6324	655	6979
Não Recomendado	215	63	278

Pode-se perceber na tabela acima que, para $6324 + 63 = 6387$ entradas, os filmes foram classificados da maneira correta. Porém, observa-se também que $655 + 215 = 870$ entradas, a Abordagem 1 não se mostrou eficiente. Isto acontece pois, para estas entradas, o método recomenda um filme que não deve ser recomendado a algum telespectador (Item Recomendado e Irrelevante) ou não recomenda um filme que deve ser recomendado a algum telespectador (Item Não Recomendado e Relevante). Vale lembrar também que apenas 63 entre os $655 + 63 = 718$ irrelevantes acabaram não sendo recomendados.

Abordagem 2

Tabela 13: Tabela de classificação para a Abordagem 2

	Relevante	Irrelevante	Total
Recomendado	6526	715	7241
Não Recomendado	13	3	16

Para esta Abordagem, o método recomenda os filmes aos telespectadores em quase todas as entradas. Novamente pode-se perceber um alto número de entradas que foram classificadas como “Recomendadas e Irrelevantes”. A Especificidade, que idealmente deveria ter um valor alto, está próxima a zero, ver tabela 20. Assim, pode-se afirmar que a abordagem apresentou resultados de má qualidade e, com isso, não foi feita uma recomendação eficiente dos filmes.

Abordagem 3

Tabela 14: Tabela de classificação para a Abordagem 3

	Relevante	Irrelevante	Total
Recomendado	6531	711	7242
Não Recomendado	8	7	15

Assim como na Abordagem 2, o método recomendou os filmes aos telespectadores em quase todas as entradas e a Especificidade também apresentou um valor muito baixo, ver tabela 20.

Abordagem 4

Tabela 15: Tabela de classificação para a Abordagem 4

	Relevante	Irrelevante	Total
Recomendado	6535	717	7252
Não Recomendado	4	1	5

Nesta Abordagem, pode-se verificar que, os filmes não foram recomendados aos telespectadores em apenas 5 entradas. Com isso, muitos filmes estão sendo mal-recomendados, prejudicando a aplicação do método de recomendação.

Abordagem 5

Tabela 16: Tabela de classificação para a Abordagem 5

	Relevante	Irrelevante	Total
Recomendado	6090	556	6646
Não Recomendado	449	162	611

Esta Abordagem classificou mais de 1000 entradas de forma equivocada. Com isso, as métricas *Recall* e *Precision* acabam sendo prejudicadas, pois à medida que recomendam-se filmes que não deveriam ser recomendados e não se recomendam filmes que deveriam ser recomendados, os valores destas métricas caem, ver tabela 20. Em compensação, comparando com os resultados das abordagens mostradas até então, esta Abordagem foi a que teve a maior Especificidade, visto que 162 entradas foram classificadas como “Não Recomendadas e Irrelevantes”.

Abordagem 6

Tabela 17: Tabela de classificação para a Abordagem 6

	Relevante	Irrelevante	Total
Recomendado	6440	678	7118
Não Recomendado	99	40	139

Em relação a Abordagem 5, houve uma queda no número de filmes não recomendados e relevantes, porém o número de filmes que foram recomendados e irrelevantes aumentou.

Abordagem 7

Tabela 18: Tabela de classificação para a Abordagem 7

	Relevante	Irrelevante	Total
Recomendado	6158	582	6740
Não Recomendado	381	136	517

Para esta abordagem, ainda tem-se uma grande quantidade de entradas que foram classificadas como “Recomendadas e Irrelevantes” ou “Não Recomendadas e Relevantes”, porém, com o alto número de entradas classificadas como “Não Recomendadas e Irrelevantes”, a métrica de Especificidade aumenta em relação às demais abordagens.

Abordagem 8

Tabela 19: Tabela de classificação para a Abordagem 8

	Relevante	Irrelevante	Total
Recomendado	6531	717	7248
Não Recomendado	8	1	9

O método para a Abordagem 8 recomendou os filmes aos telespectadores em quase todas as entradas. Também vale ressaltar o fato que 717 entradas foram referentes a filmes que foram recomendados mas que não deveriam ter sido. Portanto, pode-se afirmar que a abordagem apresentou resultados de má qualidade e, com isso, não foi feita uma recomendação eficiente dos filmes.

Para ilustrar melhor a qualidade de cada método e das previsões calculadas, segue abaixo a tabela referente às 4 métricas de classificação mostradas nas Seções 3.5.1 e 3.5.2.

Tabela 20: Métricas de classificação

	Recall(Sensibilidade)	Precision	Medida F	Especificidade
Abordagem 1	0,9671	0,9061	0,9356	0,0877
Abordagem 2	0,9980	0,9013	0,9472	0,0042
Abordagem 3	0,9988	0,9018	0,9478	0,0097
Abordagem 4	0,9994	0,9011	0,9477	0,0014
Abordagem 5	0,9313	0,9163	0,9237	0,2256
Abordagem 6	0,9849	0,9047	0,9431	0,0557
Abordagem 7	0,9417	0,9136	0,9275	0,1894
Abordagem 8	0,9988	0,9011	0,9474	0,0014

De forma geral, as medidas *Recall*, *Precision* e *F* tiveram valores bons, enquanto que a *Especificidade* foi ruim para todas as Abordagens. Isso se deve ao fato de que para as 8 abordagens quase todas as entradas foram recomendadas. Em casos extremos, como as Abordagens 2, 3, 4 e 8, menos de 10 entre as 718 entradas irrelevantes não foram recomendadas.

As Abordagens 1, 5 e 7, pelos resultados apresentados acima, foram as abordagens com desempenho menos ruim, uma vez que tiveram os maiores valores para a *Especificidade*. Isso se deve ao fato destas abordagens terem apresentado o maior número de entradas classificadas como “Não Recomendadas” entre as irrelevantes, o que, conseqüentemente, também aumenta a taxa de *Especificidade*.

Outra consequência da grande quantidade de itens irrelevantes que foram recomendados foi o fato que isto prejudicou a qualidade dos métodos adotados quanto as métricas *Precision* e *F*. Estas medidas poderiam ter valores ainda maiores se uma quantidade maior de entradas fossem classificadas de maneira correta, isto é, os filmes que deveriam ser recomendados sendo, de fato, recomendados e os filmes que não deveriam ser recomendados não sendo, de fato, recomendados.

Apesar dos resultados ruins, podemos comparar os resultados das 8 abordagens a fim de avaliar a influência das escolhas feitas na Seção 3.3 que define cada abordagem. A começar pela escolha de semelhança entre usuários ou itens, comparando as abordagens 1 com 5, 2 com 6, 3 com 7 e 4 com 8, veja figura 1. A partir desta comparação, não é possível afirmar que uma escolha recomende itens a usuários melhor do que a outra escolha, porém observou-se valores levemente melhores, para as abordagens que usaram a semelhança entre itens.

Para as escolhas relacionadas a escala de avaliação, comparou-se as abordagens 1 com 3, 2 com 4, 5 com 7 e 6 com 8, respectivamente, veja figura 1. Após esta avaliação,

foi verificado que as abordagens que usaram a escala *Likert* de avaliação tiveram um melhor desempenho em relação às abordagens que usaram a escala modificada de -1 a 1 , comparando os valores da tabela de dupla entrada e das medidas de classificação. Com isso, pode-se dizer que estas abordagens fariam recomendações mais eficientes dos filmes aos telespectadores.

Por fim, analisando a definição dos grupos de semelhança, comparou-se as abordagens 1 com 2, 3 com 4, 5 com 6 e 7 com 8, respectivamente, veja figura 1. Se percebeu que as abordagens que tinham grupos de semelhança definidos por um número específico de N usuários (M itens) semelhantes a um usuário i (item j) apresentaram métricas de classificação melhores que as abordagens que usaram a comparação da similaridade cosseno entre usuários (itens) com uma constante de similaridade pré-estabelecida para formar os grupos de semelhança. Vale destacar a melhora acentuada da taxa de Especificidade quando as medidas de classificação são comparadas através da definição dos grupos de semelhança.

Baseando-se na análise dos resultados para cada escolha, para as tabelas de dupla entrada e para as métricas de classificação, pode-se concluir que a Abordagem 5 produziu os melhores resultados para a recomendação de itens para usuários.

4.4 Resultados com critérios ajustados

Diante dos resultados ruins, principalmente para a métrica Especificidade, devido ao baixo número de itens recomendados, decidiu-se realizar uma análise de sensibilidade nos parâmetros do método. Com isso, espera-se estudar se é possível melhorar os resultados usando diferentes valores para os parâmetros. Para a realização desta análise, as 8 abordagens serão avaliadas novamente, agora com diferentes valores para os parâmetros: Constante de similaridade, Número de usuários (itens) semelhantes e Nota de corte para o critério entre recomendar e não recomendar. Devido ao grande consumo de tempo para rodar as 8 abordagens, em vez de considerarmos todas as 7257 entradas não nulas da Matriz de Utilidade, as abordagens serão avaliadas em 100 entradas selecionadas de forma aleatória.

Falando primeiramente da escolha das 100 avaliações, vale lembrar que o método de escolha destas entradas que será mostrado a seguir teve como base o trabalho de conclusão de curso de Chavez [6].

Vale lembrar que na avaliação anterior, os valores usados para os parâmetros foram:

10 para o número de usuários semelhantes; 2 para o número de itens semelhantes; 0,2 para a constante de similaridade e 3 para a nota de corte usada para recomendar ou não um item a algum usuário.

Na nova avaliação, os valores usados para o número de usuários semelhantes serão, respectivamente, 5, 8 e 10 usuários semelhantes; para os itens semelhantes serão usados, respectivamente, 2, 4 e 6 itens semelhantes. Já para a constante de similaridade, as constantes usadas serão, respectivamente, 0,2, 0,3 e 0,4. Por fim, a nota de corte para o critério entre recomendar ou não um item a algum usuário vai variar entre 3, 4 e 5.

4.4.1 Escolha das 100 avaliações que serão retiradas da amostra

Para a escolha das 100 avaliações, será recolhida uma amostra estratificada a fim de garantir que pares (Usuários, Itens) com diferentes características farão parte da amostra selecionada.

Antes de assumirmos estas 100 avaliações como desconhecidas definiremos as seguintes variáveis:

- z_i : Quantidade de itens avaliados pelo usuário i . Este valor irá variar de 2 à m , dado que não nos interessará o caso em que o usuário avaliou apenas o item em questão.
- y_j : Quantidade de usuários que avaliaram o item j . Este valor irá variar de 2 à n , dado que não nos interessará o caso em que o item foi avaliado apenas pelo usuário em questão.
- $x_{i,j}$: a avaliação do usuário i para o item j .

Em seguida determinamos faixas de valores para as variáveis z_i e y_j . Por exemplo, as faixas para z_i podem ser $[2, 10]$, $(10, 20]$, $(20, 30]$, \dots , $(m - 10, m]$. Já as faixas para y_j podem ser $[2, 100]$, $(100, 200]$, \dots , $(n - 100, n]$. Os exemplos de faixas para z_i possuem amplitude menor que as faixas para y_j pois existem menos itens avaliados por um único usuário do que usuários que avaliaram um mesmo item. Assim definiremos diferentes extratos como uma combinação de uma faixa de z_i , uma faixa de y_j e um valor de $x_{i,j}$. Por exemplo, um extrato pode ser z_i em $(40, 50]$, y_j em $(100, 200]$ e $x_{i,j} = 4$, que representa os pares (Usuário, Item) para os quais o usuário avaliou entre 40 e 50 itens, o item foi avaliado por 100 à 200 usuários e a nota do usuário para esse item foi 4.

Determinaremos o número de extratos a serem definidos pelo número de faixas para z_i , o número de faixas para y_j e o número de valores que $x_{i,j}$ pode assumir. Sendo a variável z_i dividida em 5 faixas, a variável y_j em 4 e com $x_{i,j}$ assumindo 5 valores, o número de extratos será $5 \times 4 \times 5 = 100$. Uma vez que definidos os extratos, selecionaremos as 100 entradas não nulas a serem retiradas por sorteio aleatório dentro de cada extrato. É importante ressaltar que estas entradas serão selecionadas uma a uma. Com isso, não deixamos uma matriz que já é esparsa ter ainda mais entradas nulas. (CHAVES, 2014) [6].

4.4.2 Apresentação dos novos resultados

As tabelas que serão apresentadas nesta seção mostram os novos resultados das métricas de classificação das seções 3.5.1 e 3.5.2, avaliadas em relação a combinação de parâmetros mostrada anteriormente no começo desta seção.

Analisando primeiramente as notas de corte, pode-se perceber que, como esperado, os desempenhos dos métodos para a nota 5 foram muito ruins, visto que as medidas *Recall*, *Precision* e *F* apresentaram valores muito baixos em relação a esta nota. Pode-se observar que, de uma forma geral, que os resultados para a nota de corte 3 apresentaram desequilíbrio entre as medidas *Recall*, *Precision* e *F*, que geralmente apresentaram altos valores, em relação a taxa de Especificidade, que apresentaram baixos valores. Já os resultados para a nota de corte 4 apresentaram medidas de valor intermediário e equilíbrio para os valores das 4 métricas de classificação.

Verificou-se, também, que a medida *Recall* caiu e a taxa de Especificidade aumentou com aumento da nota de 3 para 4. Uma vez que o problema dos resultados anteriores era o baixo valor da Especificidade e que as medidas se apresentaram mais equilibradas para a nota 4, esta foi a que apresentou os melhores resultados nesta análise. Por estas razões, a nota de corte 4 é a única que será considerada nas próximas análises a partir de então.

Quanto às escalas de avaliação, nota-se a má qualidade do método quando a escala de -1 a 1 é adotada. Esta escala apresentou valores muito ruins para a taxa de Especificidade e a queda nos valores das métricas *Precision* e *F* ao passo que a nota de corte ia aumentando. Optou-se, então, por desconsiderar as abordagens que usam esta escala na análise de sensibilidade dos parâmetros.

Tabela 21: Resultados ajustados para a Abordagem 1

n° de usuários semelhantes	Medidas de comparação	≥ 3	≥ 4	$= 5$
5	Recall (Sensibilidade)	0,9344	0,4634	0,05
	Precision	0,6627	0,7308	1
	Medida F	0,7755	0,5672	0,0952
	Especificidade	0,2564	0,8814	1
8	Recall (Sensibilidade)	0,9672	0,3902	0
	Precision	0,6344	0,64	0
	Medida F	0,7662	0,4848	0
10	Especificidade	0,1282	0,8474	1
	Recall (Sensibilidade)	0,9836	0,4146	0
	Precision	0,6383	0,5667	0
	Medida F	0,7742	0,4789	0
	Especificidade	0,1282	0,7797	1

Tabela 22: Resultados ajustados para a Abordagem 2

Constante de similaridade	Medidas de comparação	≥ 3	≥ 4	$= 5$
0,2	Recall (Sensibilidade)	1	0,6342	0
	Precision	0,6162	0,6667	0
	Medida F	0,7625	0,65	0
	Especificidade	0,0256	0,7797	1
0,3	Recall (Sensibilidade)	0,9836	0,5854	0
	Precision	0,6122	0,6486	0
	Medida F	0,7547	0,6154	0
0,4	Especificidade	0,0256	0,7797	1
	Recall (Sensibilidade)	0,9836	0,5854	0,05
	Precision	0,6186	0,6486	1
	Medida F	0,7595	0,6154	0,0952
	Especificidade	0,0512	0,7797	1

Tabela 23: Resultados ajustados para a Abordagem 3

n° de usuários semelhantes	Medidas de comparação	$\geq 3(1)$	$\geq 4(1)$	$= 5(1)$
5	Recall (Sensibilidade)	1	1	1
	Precision	0,6162	0,4141	0,2020
	Medida F	0,7625	0,5857	0,3361
	Especificidade	0,0256	0,0164	0,0125
8	Recall (Sensibilidade)	1	1	1
	Precision	0,6162	0,4141	0,2020
	Medida F	0,7625	0,5857	0,3361
	Especificidade	0,0256	0,0164	0,0125
	Recall (Sensibilidade)	1	1	1
10	Precision	0,61	0,41	0,2
	Medida F	0,7578	0,5816	0,3333
	Especificidade	0	0	0

Tabela 24: Resultados ajustados para a Abordagem 4

Constante de similaridade	Medidas de comparação	$\geq 3(1)$	$\geq 4(1)$	$= 5(1)$
0,2	Recall (Sensibilidade)	1	1	1
	Precision	0,61	0,41	0,2
	Medida F	0,7578	0,5816	0,3333
	Especificidade	0	0	0
0,3	Recall (Sensibilidade)	1	1	1
	Precision	0,61	0,41	0,2
	Medida F	0,7578	0,5816	0,3333
	Especificidade	0	0	0
	Recall (Sensibilidade)	1	0,9756	1
0,4	Precision	0,1	0,4598	0,2299
	Medida F	0,7578	0,6250	0,3738
	Especificidade	0	0,2034	0,1625

Tabela 26: Resultados ajustados para a Abordagem 6

Constante de similaridade	Medidas de comparação	≥ 3	≥ 4	$= 5$
0,2	Recall (Sensibilidade)	0,9836	0,6585	0
	Precision	0,6122	0,4821	0
	Medida F	0,7547	0,5567	0
	Especificidade	0,0256	0,5085	1
0,3	Recall (Sensibilidade)	0,8525	0,6585	0,1
	Precision	0,6047	0,5294	1
	Medida F	0,7075	0,5869	0,1818
	Especificidade	0,1282	0,5932	1
0,4	Recall (Sensibilidade)	0,6885	0,4878	0,05
	Precision	0,6087	0,5263	0,2857
	Medida F	0,6462	0,5063	0,1481
	Especificidade	0,3077	0,6949	0,9375

Tabela 25: Resultados ajustados para a Abordagem 5

n° de itens semelhantes	Medidas de comparação	≥ 3	≥ 4	$= 5$
2	Recall (Sensibilidade)	0,9508	0,8537	0,3
	Precision	0,6237	0,4667	0,5455
	Medida F	0,7532	0,6034	0,3871
	Especificidade	0,1025	0,3220	0,9375
4	Recall (Sensibilidade)	0,9836	0,6098	0,05
	Precision	0,6316	0,4464	1
	Medida F	0,7692	0,5155	0,0952
	Especificidade	0,1026	0,4746	1
6	Recall (Sensibilidade)	0,9836	0,6585	0
	Precision	0,6186	0,4655	0
	Medida F	0,7595	0,5455	0
	Especificidade	0,0513	0,4746	1

Tabela 28: Resultados ajustados para a Abordagem 8

Constante de similaridade	Medidas de comparação	$\geq 3(1)$	$\geq 4(1)$	$= 5(1)$
0,2	Recall (Sensibilidade)	0,9836	0,9756	1
	Precision	0,6061	0,4040	0,2020
	Medida F	0,75	0,5714	0,3361
	Especificidade	0	0	0,0125
0,3	Recall (Sensibilidade)	0,8852	0,9268	1
	Precision	0,6067	0,4270	0,2247
	Medida F	0,72	0,5846	0,3670
	Especificidade	0,1026	0,1356	0,1375
0,4	Recall (Sensibilidade)	0,7541	0,7805	0,8
	Precision	0,6216	0,4324	0,2162
	Medida F	0,6815	0,5565	0,3404
	Especificidade	0,2821	0,2881	0,2750

Tabela 27: Resultados ajustados para a Abordagem 7

n° de itens semelhantes	Medidas de comparação	$\geq 3(1)$	$\geq 4(1)$	$= 5(1)$
2	Recall (Sensibilidade)	0,9508	0,9756	0,95
	Precision	0,6237	0,4301	0,2043
	Medida F	0,7532	0,5970	0,3363
	Especificidade	0,1026	0,1017	0,0750
4	Recall (Sensibilidade)	0,9836	1	1
	Precision	0,6316	0,4316	0,2105
	Medida F	0,7692	0,6029	0,3478
	Especificidade	0,1026	0,0847	0,0625
6	Recall (Sensibilidade)	1	1	1
	Precision	0,6224	0,4184	0,2041
	Medida F	0,7673	0,5899	0,3390
	Especificidade	0,0513	0,0339	0,0250

Para o número de usuários semelhantes, o melhor desempenho aconteceu quando os grupos de semelhança eram formados por 5 usuários. Este parâmetro apresentou valores mais altos para as métricas de classificação, comparado aos grupos de semelhança formados por 8 e 10 usuários, principalmente para a métrica *Precision* e a taxa de Especificidade. Já o número de itens semelhantes apresentou valores decrescentes para medida *Recall* e crescentes para a taxa de Especificidade, a medida que o número de itens semelhantes para a formação dos grupos aumentava. Por apresentar valores mais equilibrados e ligeiramente superiores para todas as métricas, a escolha de 6 itens semelhantes foi a que apresentou melhor desempenho em relação às outras escolhas deste parâmetro.

Os resultados na análise das constantes de similaridades variaram de acordo com cada abordagem. Na Abordagem 2, a escolha da constante 0,2 se mostrou mais eficaz, haja vista que obteve valores levemente superiores para as métricas de classificação, comparando com as constantes de similaridade 0,3 e 0,4. Na abordagem 6, as métricas se comportaram de forma diferente com o aumento da constante de similaridade. Ao passo que a medida *Precision* e a taxa de Especificidade cresciam com o aumento da constante de similaridade, a medida *Recall* caía e a *F* oscilava entre valores intermediários. Com isso, a escolha da constante 0,3 foi a que apresentou os melhores resultados para esta abordagem, pois apresentou valores equilibrados e acima de 0,5 para todas as métricas, diferentemente das constantes 0,2 e 0,4.

Com o objetivo de analisar mais detalhadamente os resultados para a Abordagem 5, abordagem que teve o melhor desempenho na avaliação dos resultados para todas as entradas não nulas da Matriz de Utilidade, uma nova análise de sensibilidade será formulado. Este estudo será baseado na comparação dos resultados referentes às métricas *Recall*, *Precision*, *F* e a taxa de Especificidade usando os mesmos valores definidos na Seção 4.4 para o número de itens semelhantes e para a nota de corte. A grande diferença é que ao invés de selecionar 100 entradas não nulas da Matriz de Utilidade, todas as 7257 entradas avaliadas serão selecionadas, como foi feito na Seção 4.3.

De uma forma geral, pode-se perceber que as métricas da Tabela 29 tinham valores mais altos e equilibrados, na comparação com os valores da Tabela 25 e com a primeira análise dos resultados na Seção 4.3.

Tabela 29: Análise de sensibilidade detalhada - Abordagem 5

n° de itens semelhantes	Medidas de comparação	≥ 3	≥ 4	$= 5$
2	Recall (Sensibilidade)	0,9313	0,7569	0,2625
	Precision	0,9162	0,8230	0,5372
	Medida F	0,9237	0,7885	0,3526
	Especificidade	0,2242	0,4747	0,8747
4	Recall (Sensibilidade)	0,9630	0,6857	0,0912
	Precision	0,9110	0,8338	0,6036
	Medida F	0,9363	0,7526	0,1585
	Especificidade	0,1435	0,5591	0,9668
6	Recall (Sensibilidade)	0,9723	0,6681	0,0387
	Precision	0,9095	0,8330	0,7042
	Medida F	0,9398	0,7415	0,0733
	Especificidade	0,1184	0,5679	0,9910

As métricas apresentaram resultados divergentes quanto às notas de corte; a medida que a nota de corte aumentava, as medidas *Recall*, *Precision* e *F* diminuavam enquanto que a taxa de Especificidade aumentava. A nota 3 apresentou métricas de valores elevados, exceto para a taxa de Especificidade, que se mostrou baixa. Os resultados para a nota 5 foram ineficientes. Exceto para os grupos formados por 2 itens semelhantes a algum item j , as medidas *Recall* e *F* tiveram valores muito baixos, evidenciando o fato que o método não recomendou muitas entradas que deveriam ter sido recomendadas. Já a taxa de Especificidade esteve bastante elevada e em crescimento. Finalmente, para a nota de corte 4, as métricas se mostraram bem mais equilibradas, com uma grande melhora das taxas de Especificidade, que obtiveram valores intermediários sem causar uma grande queda nos valores das outras medidas. Por isso, a nota de corte 4 é a única que será considerada na análise do número de itens semelhantes a partir de então.

Analisando agora o número de itens semelhantes, pode-se observar que para grupos de semelhança formado por 2 itens, as métricas *Recall*, *Precision* e *F* tiveram valores altos, porém a taxa de Especificidade ainda se encontrou abaixo de 0,5. Para grupos formados por 4 e, posteriormente, 6 itens, a taxa de Especificidade aumentou enquanto que as outras medidas se mantiveram estáveis e altas. Com isso, pode-se concluir que os resultados para a nota de corte 4 e para grupos de itens semelhantes formados por 4 ou 6 itens foram os melhores para esta análise de sensibilidade e, portanto, seriam bons parâmetros na formulação de um novo método para a recomendação destes filmes aos telespectadores.

5 Conclusões

Neste trabalho foi usada a técnica de Filtragem Colaborativa a fim de comparar, dentre as combinações de escolhas da Filtragem Colaborativa que formavam 8 abordagens, quais destas abordagens produziam as melhores recomendações de filmes para os telespectadores. A partir desta comparação, o objetivo seria recomendar os filmes aos usuários. As informações do banco de dados foram coletadas pela aluna Juliana Giglio para a sua monografia [6] através de um questionário na Internet, onde 525 usuários avaliaram um lista de 50 filmes. Os filmes eram avaliados com base em uma escala de 5 pontos, denominada *Escala Likert*, onde os usuários davam notas de 1 a 5, que representavam respectivamente às classificações: “Gostei muito”, “Gostei”, “Achei indiferente”, “Não gostei” e “Detestei”.

A fim de fazer esta análise, foram calculadas previsões de avaliações para todas as 7257 entradas não nulas do banco de dados, aqui denominado Matriz de Utilidade. É importante ressaltar que estas abordagens são formadas a partir de escolhas ligadas ao tipo de semelhança, à escala de avaliação e a maneira que os grupos de semelhança são criados.

Para as abordagens 1 e 3, que usaram a semelhança entre usuários (telespectadores) e consideraram o tamanho do grupo de usuários semelhantes fixo igual a N , foi considerado $N = 10$. As abordagens 5 e 7, que usaram a semelhança entre itens (filmes) e consideraram o tamanho do grupo de itens semelhantes fixo igual a M , foi considerado $M = 2$. Finalmente, para as demais abordagens, que consideraram um usuário (um item) pertencente ao grupo de usuários (itens) semelhantes se a similaridade fosse igual ou superior a uma nota de corte específica, a nota de corte considerada foi 0,2.

Após a aplicação da Filtragem Colaborativa para todas as abordagens, percebeu-se que muitas entradas eram classificadas de maneira errada, com um alto índice de filmes sendo recomendados, mesmo quando não deveriam. Os casos mais extremos aconteceram para as Abordagens 2, 3, 4 e 8, que deixaram de recomendar menos de 20 filmes em 7257 avaliações. Isto influenciou no cálculo das métricas de classificação, principalmente nas

taxas de Especificidade, que apresentou taxas muito baixas para todas as Abordagens; apenas para as Abordagens 5 e 7, esta taxa ultrapassou 0,15.

Na análise dos resultados a partir das escolhas da Filtragem Colaborativa, foi possível afirmar que as escolhas da semelhança entre itens, da escala *Likert* de avaliação e do número específico de usuários ou itens para a formação dos grupos de semelhança apresentaram os melhores resultados para as métricas de classificação e, portanto, a Abordagem 5 seria a que melhor recomendaria os itens aos usuários.

Motivado pelos resultados ruins, resolveu-se fazer uma análise de sensibilidade nos parâmetros, calculando as métricas de classificação citadas nas seções 3.5.1 e 3.5.2 para diferentes valores dos parâmetros. Nesta análise, os valores usados para a formação de um grupo de usuários semelhantes foram, respectivamente, 5, 8 e 10 usuários; para a formação de um grupo de itens semelhantes foram, respectivamente, 2, 4 e 6 itens. Para a comparação baseada em uma constante de similaridade, foram usadas as constantes 0,2, 0,3 e 0,4. Além disso, a nota de corte para recomendar ou não um item a algum usuário variou entre 3, 4 ou 5.

Em relação às notas de corte, as métricas relacionadas a nota 4 apresentaram resultados melhores na comparação com as outras notas de corte, apresentando valores intermediários para quase todas as métricas, diferentemente das notas 3 e 5, que mostraram métricas desequilibradas, principalmente para as medidas *Recall* e a taxa de Especificidade, para quase todas as abordagens.

Os resultados para as abordagens que usaram a Escala Likert se mostraram bem mais eficientes que os resultados da escala -1 a 1. Esta apresentou métricas *Recall* e taxas de Especificidade com bastante discrepância, com as medidas *Recall* muito altas e as taxas de Especificidade muito baixas para todas as abordagens que utilizaram esta escala.

Os grupos de 5 usuários semelhantes apresentaram resultados melhores em relação aos grupos formados por 8 ou 10 usuários semelhantes, por apresentarem medidas *Precision* e Especificidade de valor ligeiramente mais alto na comparação com os grupos citados logo acima. Os grupos de 6 itens semelhantes apresentaram resultados mais eficazes que os grupos formados por 2 ou 4 itens semelhantes. Por mostrar métricas mais equilibradas e levemente superiores a 0,5, esta escolha obteve vantagem sobre às outras.

Em relação às constantes de similaridade, os desempenhos variaram de acordo com cada abordagem. Para a Abordagem 2, a constante 0,2 obteve valores um pouco maiores para as medidas de classificação, em comparação com as outras constantes. Na Abordagem

6, a constante 0,3 foi a que mostrou os melhores resultados, apesar da grande oscilação das métricas nesta abordagem.

Após esta avaliação, foi feita uma análise de sensibilidade detalhada para a Abordagem 5, produziu os melhores resultados para a recomendação de itens para usuários. Se percebeu que esta análise produziu resultados mais eficazes para as métricas de classificação que os resultados apresentados na Tabela 20. Também se verificou que a recomendação dos itens aos usuários seria feita com mais qualidade se os parâmetros da nota de corte e do número de itens semelhantes passassem por mudanças. Mais precisamente, se a nota de corte subisse de 3 para 4 e se os grupos de itens semelhantes fossem formados por 4 ou 6 itens, ao invés de 2.

Referências

- [1] RESNICK, P.; VARIAN, H. R. Recommender systems. *Communications of the ACM*, ACM, v. 40, n. 3, p. 56–58, 1997.
- [2] CAZELLA, S. C.; NUNES, M.; REATEGUI, E. B. A ciência da opinião: Estado da arte em sistemas de recomendação. In: *XXX Congresso da Sociedade Brasileira de Computação—Jornada de Atualização em Informática (JAI)*. [S.l.: s.n.], 2010. p. 161–216.
- [3] SHARDANAND, U.; MAES, P. Social information filtering: algorithms for automating “word of mouth”. In: ACM PRESS/ADDISON-WESLEY PUBLISHING CO. *Proceedings of the SIGCHI conference on Human factors in computing systems*. [S.l.], 1995. p. 210–217.
- [4] BALABANOVIĆ, M.; SHOHAM, Y. Fab: content-based, collaborative recommendation. *Communications of the ACM*, ACM, v. 40, n. 3, p. 66–72, 1997.
- [5] LINDEN, G.; SMITH, B.; YORK, J. Amazon. com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, IEEE, v. 7, n. 1, p. 76–80, 2003.
- [6] CHAVEZ, J. G. *Recomendação de Filmes Utilizando Filtragem Colaborativa*. Niterói, 2014.
- [7] CAZELLA, S. C. et al. Recomendação de objetos de aprendizagem empregando filtragem colaborativa e competências. *Simpósio Brasileiro de Informática na Educação (SBIE)*, 2009.
- [8] ADOMAVICIUS, G.; TUZHILIN, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, IEEE, v. 17, n. 6, p. 734–749, 2005.
- [9] REATEGUI, E. B.; CAZELLA, S. C. Sistemas de recomendação. In: CITESEER. *XXV Congresso da Sociedade Brasileira de Computação. Universidade do Vale do Rio dos Sinos (UNISINOS)*. São Leopoldo. [S.l.], 2005.
- [10] MCJONES, P. Eachmovie collaborative filtering data set. *DEC Systems Research Center*, v. 249, 1997.
- [11] RAJARAMAN, A. et al. *Mining of massive datasets*. [S.l.]: Cambridge University Press Cambridge, 2012.
- [12] MARTINEZ, E. Z.; LOUZADA-NETO, F.; PEREIRA, d. B. B. A curva roc para testes diagnósticos. *Cad Saúde Coletiva*, v. 11, n. 1, p. 7–31, 2003.

ANEXO A – Código do trabalho no software R

```

require(car)
require(BioPhysConnectoR)
A = read.csv(file = "resposta_filmes.csv",header=T,fill=T)
D = as.matrix(A) #D = matriz de utilidade
###-----
##Selecao das entradas nao nulas de D
par_i = NULL
par_j = NULL
for(i in 1:nrow(D)){
  for(j in 1:ncol(D)){
    if((D[i,j] != 0) && (sum(D[i,]!=0) > 2)){
      par_i = c(par_i,i)
      par_j = c(par_j,j)
    }
  }
}
length(par_j) #total de entradas nao nulas = 7257
###-----
##Resultados - Abordagem 1
N = 10 #numero de usuarios semelhantes levados em consideracao
n = length(par_i)
prev = NULL #Vetor das previsoes
nota = NULL #Construcao do vetor das notas
for(k in 1:n){
  i = par_i[k] #telespectador do par selecionado
  j = par_j[k] #filme do par selecionado

```

```

nota[k] = D[i,j] #nota do telespectador i para o filme j guardada no vetor nota
D[i,j] = 0 #zerando essa entrada para fazer previsao
#Similaridade com o telespectador i
M = matrix(NA,nrow(D),nrow(D))
l=i
for(u in 1:nrow(D)){
  M[l,u] =
    (sum(D[l,] * D[u,])) / (sqrt(sum(D[l,]^2)) * sqrt(sum(D[u,]^2)))
  if(M[l,u] == "NaN"){
    M[l,u] = 0
  }
}
#Quem sao os usuarios semelhantes ao usuario i que assistiram o filme j
si = M[i,]
usuarios = c(1:nrow(M))
clas = mat.sort(cbind(usuarios,si),2) #dentro do pacote BioPhysConnectoR
l = nrow(M) #a posicao (indice) em clas do usuario avaliado
Gi = NULL #grupo de usuarios semelhantes a i que assistiram j
while(length(Gi)<N){
  u = clas[l,1]
  if(D[u,j]!=0){
    Gi = c(Gi,u)
  }
  l = l - 1
}
#Previsoes para Likert
prev[k] = (sum(D[Gi,j]*si[Gi]))/(sum(si[Gi]))
D[i,j] = nota[k] #retornar valor original para matriz D
}
##Tabela de classificacao
RECM = NULL
RELV = NULL
for(t in 1:n){
  if(prev[t] >= 3){
    RECM[t] = "Rec"
  }
}

```

```

}else{
  RECM[t] = "Rec_N"
}
if(nota[t] >= 3){
  RELV[t] = "Rel"
}else{
  RELV[t] = "Rel_N"
}
}
tab = table(RECM,RELV)
##Metricas de Classificacao
Rc = tab[1,1] / sum(tab[,1]) #Recall / Sensibilidade
Ps = tab[1,1] / sum(tab[1,]) #Precision
MF = (2*Rc*Ps)/(Rc+Ps)      #Medida F
Ep = tab[2,2] / sum(tab[,2]) #Especificidade
###-----
#Resultados - Abordagem 2
prev2 = NULL #Vetor das previsoes
nota = NULL #vetor das notas
#Construcao do vetor das notas
for(k in 1:n){
  i = par_i[k] #telespectador do par selecionado
  j = par_j[k] #filme do par selecionado
  nota[k] = D[i,j] #nota do telespectador i para o filme j guardada no vetor nota
  D[i,j] = 0 #zerando essa entrada para fazer previsao
  #Similaridade com o telespectador i
  M = matrix(NA,nrow(D),nrow(D))
  l=i
  for(u in 1:nrow(D)){
    M[l,u] =
      (sum(D[l,] * D[u,])) / (sqrt(sum(D[l,]^2)) * sqrt(sum(D[u,]^2)))
    if(M[l,u] == "NaN"){
      M[l,u] = 0
    }
  }
}

```

```

#Quem sao os usuarios semelhantes ao usuario i que assistiram o filme j
si = M[i,]
usuarios = c(1:nrow(M))
clas = mat.sort(cbind(usuarios,si),2) #dentro do pacote BioPhysConnectoR
l = nrow(M) #a posicao (indice) em clas do usuario avaliado
Gi = NULL #grupo de usuarios semelhantes a i que assistiram j
while(clas[l,2] > 0.2){ #0,2 - Constante de Similaridade adotada
  u = clas[l,1]
  if(D[u,j]!=0){
    Gi = c(Gi,u)
  }
  l = l - 1
}
#Previsoes para Likert
prev2[k] = (sum(D[Gi,j]*si[Gi]))/(sum(si[Gi]))
D[i,j] = nota[k] #E depois do calculo da previsao, a entrada voltada a ser colocada
}
##Tabela de classificacao
RECM2 = NULL
RELV2 = NULL
for(t in 1:n){
  if(prev2[t] >= 3){
    RECM2[t] = "Rec"
  }else{
    RECM2[t] = "Rec_N"
  }
  if(nota[t] >= 3){
    RELV2[t] = "Rel"
  }else{
    RELV2[t] = "Rel_N"
  }
}
tab2 = table(RECM2,RELV2)
##Metricas de classificacao
Rc2 = tab2[1,1] / sum(tab2[,1]) #Recall e Sensibilidade

```

```

Ps2 = tab2[1,1] / sum(tab2[1,]) #Precision
MF2 = (2*Rc2*Ps2)/(Rc2+Ps2)      #Medida F
Ep2 = tab2[2,2] / sum(tab2[,2]) #Especificidade
###_-----
#Resultados - Abordagem 3
G = recode(D,"0=0;1:2=-1;3:5=1")
H = matrix(G,nrow(D),ncol(D))
prev3 = NULL #vETOR DAS PREVISOES
nota2 = NULL #VETOR DAS NOTAS
#Construcao do vetor das notas
for(k in 1:n){
  i = par_i[k] #telespectador do par selecionado
  j = par_j[k] #filme do par selecionado
  nota2[k] = H[i,j] #Entrada nao nula e guardada no vetor nota
  H[i,j] = 0 #Apos isto, esta entrada e zerada na Matriz de Utilidade
#####
##Matriz de Similaridade - Usuarios
M = matrix(NA,nrow(D),nrow(D))
l=i
for(u in 1:nrow(D)){
  M[l,u] =
    (sum(D[l,] * D[u,])) / (sqrt(sum(D[l,]^2)) * sqrt(sum(D[u,]^2)))
  if(M[l,u] == "NaN"){
    M[l,u] = 0
  }
}
}
#Quem sao os usuarios semelhantes ao usuario i que assistiram o filme j
si = M[i,]
usuarios = c(1:nrow(M))
clas = mat.sort(cbind(usuarios,si),2) #dentro do pacote BioPhysConnectoR
l = nrow(M) #a posicao (indice) em clas do usuario avaliado
Gi = NULL #grupo de usuarios semelhantes a i que assistiram j
while(length(Gi)<N){
  u = clas[l,1]
  if(H[u,j]!=0){

```

```

    Gi = c(Gi,u)
  }
  l = l - 1
}
#Previsoes para -1 A 1
prev3[k] = (sum(H[Gi,j]*si[Gi]))/(sum(si[Gi]))
H[i,j] = nota2[k] #E depois do calculo da previsao, a entrada voltada a ser col
}
##Tabela de classificacao
RECM3 = NULL
RELV3 = NULL
for(t in 1:n){
  if(prev3[t] > 0){
    RECM3[t] = "Rec"
  }else{
    RECM3[t] = "Rec_N"
  }
  if(nota2[t] == 1){
    RELV3[t] = "Rel"
  }else{
    RELV3[t] = "Rel_N"
  }
}
}
tab3 = table(RECM3,RELV3)
##Metricas de classificacao
Rc3 = tab3[1,1] / sum(tab3[,1]) #Recall e Sensibilidade
Ps3 = tab3[1,1] / sum(tab3[1,]) #Precision
MF3 = (2*Rc3*Ps3)/(Rc3+Ps3) #Medida F
Ep3 = tab3[2,2] / sum(tab3[,2]) #Especificidade
###-----
#Resultados - Abordagem 4
prev4 = NULL #vETOR DAS PREVISOES
nota2 = NULL #VETOR DAS NOTAS
#Construcao do vetor das notas
for(k in 1:n){

```

```

i = par_i[k] #telespectador do par selecionado
j = par_j[k] #filme do par selecionado
nota2[k] = H[i,j] #Entrada nao nula e guardada no vetor nota
H[i,j] = 0 #Apos isto, esta entrada e zerada na Matriz de Utilidade
#####
##Matriz de Similaridade - Usuarios
M = matrix(NA,nrow(D),nrow(D))
l=i
for(u in 1:nrow(D)){
  M[l,u] = (sum(D[l,] * D[u,])) / (sqrt(sum(D[l,]^2)) * sqrt(sum(D[u,]^2)))
  if(M[l,u] == "NaN"){
    M[l,u] = 0
  }
}
#Quem sao os usuarios semelhantes ao usuario i que assistiram o filme j
si = M[i,]
usuarios = c(1:nrow(M))
clas = mat.sort(cbind(usuarios,si),2) #dentro do pacote BioPhysConnectoR
l = nrow(M) #a posicao (indice) em clas do usuario avaliado
Gi = NULL #grupo de usuarios semelhantes a i que assistiram j
while(clas[l,2] > 0.2){ #0,2 - Constante de Similaridade adotada
  u = clas[l,1]
  if(H[u,j]!=0){
    Gi = c(Gi,u)
  }
  l = l - 1
}
#previsoes para -1 A 1
prev4[k] = (sum(H[Gi,j]*si[Gi]))/(sum(si[Gi]))
H[i,j] = nota2[k] #E depois do calculo da previsao, a entrada voltada a ser col
print(paste("iteracao",k,"fim -",Sys.time()))
}
##Tabela de classificacao
RECM4 = NULL
RELV4 = NULL

```



```

for(t in 1:n){
  if(prev4[t] > 0){
    RECM4[t] = "Rec"
  }else{
    RECM4[t] = "Rec_N"
  }
  if(nota2[t] == 1){
    RELV4[t] = "Rel"
  }else{
    RELV4[t] = "Rel_N"
  }
}
tab4 = table(RECM4,RELV4)
##Metricas de classificacao
Rc4 = tab4[1,1] / sum(tab4[,1]) #Recall e Sensibilidade
Ps4 = tab4[1,1] / sum(tab4[1,]) #Precision
MF4 = (2*Rc4*Ps4)/(Rc4+Ps4) #Medida F
Ep4 = tab4[2,2] / sum(tab4[,2]) #Especificidade
###-----
#Resultados - Abordagem 5
N = 2 # num de itens semelhantes levados em consideracao
prev5 = NULL #vETOR DAS PREVISOES
nota = NULL #VETOR DAS NOTAS
#Construcao do vetor das notas
for(k in 1:n){
  i = par_i[k] #telespectador do par selecionado
  j = par_j[k] #filme do par selecionado
  nota[k] = D[i,j] #Entrada n?o nula e guardada no vetor nota
  D[i,j] = 0 #Apos isto, esta entrada e zerada na Matriz de Utilidade
##Matriz de Similaridade - Itens
M = matrix(NA,ncol(D),ncol(D))
u=j
for(l in 1:ncol(D)){
  M[l,u] = (sum(D[,l] * D[,u])) / (sqrt(sum(D[,l]^2)) * sqrt(sum(D[,u]^2)))
  if(M[l,u] == "NaN"){

```

```

    M[l,u] = 0
  }
}
#Quem sao os itens semelhantes ao item j que foram assistidos pelo usuario i
si = M[,j]
itens = c(1:ncol(M))
clas = mat.sort(cbind(itens,si),2) #dentro do pacote BioPhysConnectoR
l = ncol(M) #a posicao (indice) em clas do item avaliado
Gi = NULL #grupo de itens semelhantes ao item j que foram assistidos pelo usuario
while(length(Gi)<N && l>0){
  u = clas[l,1]
  if(D[i,u]!=0){
    Gi = c(Gi,u)
  }
  l = l - 1
}

#Previsoes para likert
prev5[k] = (sum(D[i,Gi]*si[Gi]))/(sum(si[Gi]))
D[i,j] = nota[k] #E depois do calculo da previsao, a entrada voltada a ser colocada
print(paste("iteracao",k,"fim -",Sys.time()))
}

##Tabela de classificacao
RECM5 = NULL
RELV5 = NULL
for(t in 1:n){
  if(prev5[t] >= 3){
    RECM5[t] = "Rec"
  }else{
    RECM5[t] = "Rec_N"
  }
  if(nota[t] >= 3){
    RELV5[t] = "Rel"
  }else{
    RELV5[t] = "Rel_N"
  }
}

```

```

    }
}
tab5 = table(RECM5,RELV5)
##Metricas de classificacao
Rc5 = tab5[1,1] / sum(tab5[,1]) #Recall e Sensibilidade
Ps5 = tab5[1,1] / sum(tab5[1,]) #Precision
MF5 = (2*Rc5*Ps5)/(Rc5+Ps5)      #Medida F
Ep5 = tab5[2,2] / sum(tab5[,2]) #Especificidade
###-----
##Resultados - Abordagem 6
prev6 = NULL #vETOR DAS PREVISOES
nota = NULL #VETOR DAS NOTAS
#Construcao do vetor das notas
for(k in 1:n){
  i = par_i[k] #telespectador do par selecionado
  j = par_j[k] #filme do par selecionado
  nota[k] = D[i,j] #Entrada n?o nula e guardada no vetor nota
  D[i,j] = 0 #Apos isto, esta entrada e zerada na Matriz de Utilidade
  ##Matriz de Similaridade - Itens
  M = matrix(NA,ncol(D),ncol(D))
  u=j
  for(l in 1:ncol(D)){
    M[l,u] = (sum(D[,l] * D[,u])) / (sqrt(sum(D[,l]^2)) * sqrt(sum(D[,u]^2)))
    if(M[l,u] == "NaN"){
      M[l,u] = 0
    }
  }
}
#Quem sao os itens semelhantes ao item j que foram assistidos pelo usuario i
si = M[,j]
itens = c(1:ncol(M))
clas = mat.sort(cbind(itens,si),2) #dentro do pacote BioPhysConnector
l = ncol(M) #a posicao (indice) em clas do item avaliado
Gi = NULL #grupo de itens semelhantes ao item j que foram assistidos pelo usuario
while(clas[l,2] > 0.2){
  u = clas[l,1]

```

```

    if(D[i,u]!=0){
      Gi = c(Gi,u)
    }
    l = l - 1
  } #0,2 - Constante de Similaridade
  #Previsoes para likert
  prev6[k] = (sum(D[i,Gi]*si[Gi]))/(sum(si[Gi]))
  D[i,j] = nota[k] #E depois do calculo da previsao, a entrada voltada a ser colocada
}
#Tabela de classificacao
RECM6 = NULL
RELV6 = NULL
for(t in 1:n){
  if(prev6[t] >= 3){
    RECM6[t] = "Rec"
  }else{
    RECM6[t] = "Rec_N"
  }
  if(nota[t] >= 3){
    RELV6[t] = "Rel"
  }else{
    RELV6[t] = "Rel_N"
  }
}
}
tab6 = table(RECM6,RELV6)
##Metricas de classificacao
Rc6 = tab6[1,1] / sum(tab6[,1]) #Recall e Sensibilidade
Ps6 = tab6[1,1] / sum(tab6[1,]) #Precision
MF6 = (2*Rc6*Ps6)/(Rc6+Ps6) #Medida F
Ep6 = tab6[2,2] / sum(tab6[,2]) #Especificidade
tab6;Rc6;Ps6;MF6;Ep6
###_-----
#Resultados - Abordagem 7
N = 2 # num de itens semelhantes levados em consideracao
n = length(par_i)

```

```

prev7 = NULL #vETOR DAS PREVISOES
nota2 = NULL #VETOR DAS NOTAS
#Construcao do vetor das notas
for(k in 1:n){
  i = par_i[k] #telespectador do par selecionado
  j = par_j[k] #filme do par selecionado
  nota2[k] = H[i,j] #Entrada nao nula e guardada no vetor nota
  H[i,j] = 0 #Apos isto, esta entrada e zerada na Matriz de Utilidade
  ##Matriz de Similaridade - Itens
  M = matrix(NA,ncol(D),ncol(D))
  u=j
  for(l in 1:ncol(D)){
    M[l,u] = (sum(D[,l] * D[,u])) / (sqrt(sum(D[,l]^2)) * sqrt(sum(D[,u]^2)))
    if(M[l,u] == "NaN"){
      M[l,u] = 0
    }
  }
  #Quem sao os itens semelhantes ao item j que foram assistidos pelo usuario i
  si = M[,j]
  itens = c(1:ncol(M))
  clas = mat.sort(cbind(itens,si),2) #dentro do pacote BioPhysConnector
  l = ncol(M) #a posicao (indice) em clas do item avaliado
  Gi = NULL #grupo de itens semelhantes ao item j que foram assistidos pelo usuario i
  while(length(Gi)<N && l>0){
    u = clas[l,1]
    if(H[i,u]!=0){
      Gi = c(Gi,u)
    }
    l = l - 1
  }
  #previsoes para likert
  prev7[k] = (sum(H[i,Gi]*si[Gi]))/(sum(si[Gi]))
  H[i,j] = nota2[k] #E depois do calculo da previsao, a entrada voltada a ser colada
}
#Tabela de classificacao

```

```

RECM7 = NULL
RELV7 = NULL
for(t in 1:n){
  if(prev7[t] > 0){
    RECM7[t] = "Rec"
  }else{
    RECM7[t] = "Rec_N"
  }
  if(nota2[t] == 1){
    RELV7[t] = "Rel"
  }else{
    RELV7[t] = "Rel_N"
  }
}
}
tab7 = table(RECM7,RELV7)
##Metricas de classificacao
Rc7 = tab7[1,1] / sum(tab7[,1]) #Recall e Sensibilidade
Ps7 = tab7[1,1] / sum(tab7[1,]) #Precision
MF7 = (2*Rc7*Ps7)/(Rc7+Ps7)      #Medida F
Ep7 = tab7[2,2] / sum(tab7[,2]) #Especificidade
###_-----
#Abordagem 8
prev8 = NULL #vETOR DAS PREVISOES
nota2 = NULL #VETOR DAS NOTAS
#Construcao do vetor das notas
for(k in 1:n){
  i = par_i[k] #telespectador do par selecionado
  j = par_j[k] #filme do par selecionado
  nota2[k] = H[i,j] #Entrada n?o nula e guardada no vetor nota
  H[i,j] = 0 #Apos isto, esta entrada e zerada na Matriz de Utilidade
##Matriz de Similaridade - Itens
M = matrix(NA,ncol(D),ncol(D))
u=j
for(l in 1:ncol(D)){
  M[l,u] = (sum(D[,l] * D[,u])) / (sqrt(sum(D[,l]^2)) * sqrt(sum(D[,u]^2)))
}
}

```

```

    if(M[l,u] == "NaN"){
      M[l,u] = 0
    }
  }
}
#Quem sao os itens semelhantes ao item j que foram assistidos pelo usuario i
si = M[,j]
itens = c(1:ncol(M))
clas = mat.sort(cbind(itens,si),2) #dentro do pacote BioPhysConnector
l = ncol(M) #a posicao (indice) em clas do item avaliado
Gi = NULL #grupo de itens semelhantes ao item j que foram assistidos pelo usuario i
while(clas[l,2] > 0.2){
  u = clas[l,1]
  if(H[i,u]!=0){
    Gi = c(Gi,u)
  }
  l = l - 1
} #0,2 - Constante de Similaridade

#Previsoes para likert
prev8[k] = (sum(H[i,Gi]*si[Gi]))/(sum(si[Gi]))
H[i,j] = nota2[k] #E depois do calculo da previsao, a entrada voltada a ser colada
}
##Tabela de classificacao
RECM8 = NULL
RELV8 = NULL
for(t in 1:n){
  if(prev8[t] > 0){
    RECM8[t] = "Rec"
  }else{
    RECM8[t] = "Rec_N"
  }
  if(nota2[t] == 1){
    RELV8[t] = "Rel"
  }else{
    RELV8[t] = "Rel_N"
  }
}

```

```
    }  
  }  
  tab8 = table(RECM8,RELV8)  
  ##Metricas de classificacao  
  Rc8 = tab8[1,1] / sum(tab8[,1]) #Recall e Sensibilidade  
  Ps8 = tab8[1,1] / sum(tab8[1,]) #Precision  
  MF8 = (2*Rc8*Ps8)/(Rc8+Ps8)      #Medida F  
  Ep8 = tab8[2,2] / sum(tab8[,2]) #Especificidade
```