

Antônio Guilherme Ferreira Viggiano
Fernando Fochi Silveira Araújo

**Desenvolvimento de uma biblioteca
computacional para sistemas de
recomendação de produtos de lojas de
comércio online**

São Paulo, Brasil
5 de novembro de 2014

Antônio Guilherme Ferreira Viggiano
Fernando Fochi Silveira Araújo

**Desenvolvimento de uma biblioteca computacional
para sistemas de recomendação de produtos de lojas
de comércio online**

Trabalho de Conclusão de Curso apresentado
ao Departamento de Engenharia Mecatrônica
da Escola Politécnica da Universidade de São
Paulo com requisito parcial para obtenção do
Grau de Engenheiro Mecatrônico.

Universidade de São Paulo
Escola Politécnica
Trabalho de Conclusão de Curso

Orientador: Prof. Dr. Fábio Gagliardi Cozman

São Paulo, Brasil
5 de novembro de 2014

Antônio Guilherme Ferreira Viggiano
Fernando Fochi Silveira Araújo

Desenvolvimento de uma biblioteca computacional para sistemas de recomendação de produtos de lojas de comércio online/ A.G.F. Viggiano; F.F.S. Araújo. – São Paulo, Brasil, 5 de novembro de 2014

77 p.

Orientador: Prof. Dr. Fábio Gagliardi Cozman

Trabalho de Formatura – Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia Mecatrônica e de Sistemas Mecânicos.

1. Inteligência artificial.
 2. Aprendizado computacional.
 3. Comercio eletrônico.
 4. Produtos I. Prof. Dr. Fábio Gagliardi Cozman. II. Universidade de São Paulo. Escola Politécnica. III. Departamento de Engenharia Mecatrônica e de Sistemas Mecânicos
-

Antônio Guilherme Ferreira Viggiano
Fernando Fochi Silveira Araújo

Desenvolvimento de uma biblioteca computacional para sistemas de recomendação de produtos de lojas de comércio online

Trabalho de Conclusão de Curso apresentado
ao Departamento de Engenharia Mecatrônica
da Escola Politécnica da Universidade de São
Paulo com requisito parcial para obtenção do
Grau de Engenheiro Mecatrônico.

Prof. Dr. Fábio Gagliardi Cozman
Orientador

Prof. Dr. Lucas Antonio Moscato
Convidado 1

Prof. Dr. Thiago de Castro Martins
Convidado 2

Prof. Dr. Arturo Forner Cordero
Convidado 3

Profa. Dra. Larissa Driemeier
Convidado 4

São Paulo, Brasil
5 de novembro de 2014

Agradecimentos

Agradecemos ao professor Fábio Cozman pela sua orientação e apoio durante todo o projeto. Agradecemos também ao professor Thiago Martins e aos demais orientadores das disciplinas PMR2500 e PMR2550 – Projeto de Conclusão do Curso I e II – por terem nos guiado na elaboração da monografia e por terem sempre exigido trabalhos de alta qualidade. Esse papel é fundamental na valorização do diploma de Engenharia Mecatrônica da Escola Politécnica.

Make things as simple as possible, but not simpler (Albert Einstein)

Resumo

Resumo em português

Palavras-chaves: este é o resumo em português

Abstract

This is the english abstract.

Key-words: latex, abntex, text editoration.

Lista de tabelas

Tabela 1 – Atributos a_{if}	25
Tabela 2 – Avaliações r_{ui}	25
Tabela 3 – Avaliações r_{ui}	27
Tabela 4 – Atributos a_{if}	27
Tabela 5 – Avaliação de sistemas de predição	40
Tabela 6 – Avaliações r_{ui}	41
Tabela 7 – Atributos a_{if}	41
Tabela 8 – d_{ij}^f	44
Tabela 9 – Medidas de distância entre alguns atributos	44
Tabela 10 – e_{ij}	44
Tabela 11 – w_f	44
Tabela 12 – s_{ij}	46
Tabela 13 – \hat{i}_u (FW)	46
Tabela 14 – TF _{uf}	46
Tabela 15 – IDF _f	46
Tabela 16 – w_{uf}	46
Tabela 17 – s_{uv}	46
Tabela 18 – f _{uf}	48
Tabela 19 – ω_{ui} (UP)	48
Tabela 20 – \hat{i}_u (UP)	48
Tabela 21 – ω_{ui} (UI)	48
Tabela 22 – \hat{i}_u (UI)	48
Tabela 23 – Parâmetros de influência no desempenho dos algoritmos de recomendação	57

Lista de símbolos

k	Número de vizinhos mais próximos
N	Tamanho da lista de recomendação
\mathcal{U}	Conjunto de todos os usuários
\mathcal{I}	Conjunto de todos os itens
\mathcal{F}	Conjunto de todos os atributos dos itens
u, v	Usuários
i, j	Itens
f	Atributos dos itens
$\mathbf{X}_{M \times N}, \mathbf{X}$	Matriz de elementos x_{mn}
\mathbf{x}_N, \mathbf{x}	Vetor de elementos x_n
\tilde{x}	Valor ótimo de x
\hat{x}	Valor estimado de x
$ \mathcal{X} $	Número de elementos do conjunto \mathcal{X}
\mathbf{R}, r_{ui}	Avaliação feita pelo usuário u do item i
\mathbf{A}, a_{if}	Atributo f presente no item i
$\mathbf{S}, s_{ij}, s_{uv}$	Similaridade entre itens i e j ou entre usuários u e v
\mathbf{W}, w_{uf}	Correlação ponderada entre usuário u e atributo f
\mathbf{w}, w_f	Peso do atributo f

Sumário

1	INTRODUÇÃO	19
1.1	Motivação	20
1.2	Objetivos	21
2	ESTADO DA ARTE	23
2.1	Estado da arte dos problemas	23
2.2	Estado da arte das soluções	26
2.3	Desafios científicos e tecnológicos	26
2.4	Soluções propostas	28
3	METODOLOGIA	31
3.1	Definição da Necessidade	31
3.2	Definição dos Parâmetros de Sucesso	31
3.3	Síntese de Soluções	31
3.4	Detalhamento da Solução	32
3.5	Modelamento e Simulação	32
3.6	Validação Cruzada	32
4	REQUISITOS	35
4.1	Diagrama de Casos de Uso	36
4.2	Diagrama de Atividades	37
4.3	Avaliação de Desempenho	37
5	DETALHAMENTO DE SOLUÇÕES	41
5.1	Algoritmo baseado na ponderação de atributos (FW)	41
5.2	Algoritmo baseado no perfil de usuários (UP)	43
5.3	Algoritmo baseado na correlação usuário-item (UI)	47
6	DESENVOLVIMENTO DA BIBLIOTECA	49
6.1	Recursos acadêmicos	49
6.2	Ferramentas utilizadas	50
6.3	Métodos computacionais	50
6.3.1	Estrutura da biblioteca	50
6.3.2	Algoritmo baseado na ponderação de atributos (FW)	50
6.3.3	Algoritmo baseado no perfil de usuários (UP)	50
6.3.4	Algoritmo baseado na correlação usuário-item (UI)	50
6.4	Ambiente de testes	50

7	RESULTADOS	57
7.1	Tamanho da lista de recomendações N	57
7.2	Percentual da base de aprendizado T	60
7.3	Percentual de avaliações “escondidas” dos usuários-teste na validação cruzada H	63
7.4	Valor mínimo para avaliações positivas M	63
7.5	Número de vizinhos mais próximos k	66
7.6	Conjunto de atributos dos itens \mathcal{F}	70
7.7	Medida de distância entre atributos d^f	72
7.8	Pesos dos atributos w_f	72
8	CONCLUSÃO	73
	Referências	75

1 Introdução

O comércio on-line se torna cada vez mais importante na vida das pessoas, de forma que a adoção deste método de compra é cada vez mais comum. Estima-se que em 2013 um bilhão de pessoas compraram online (1), gerando uma receita anual de 1,25 trilhão de dólares com expectativas de crescer 17% ao ano até 2017. (2). Para exemplificar, a gigante chinesa Alibaba está se preparando para abrir o seu capital na bolsa de valores americana. Esta oferta pública inicial poderá arrecadar até 25 bilhões de dólares, significando a maior oferta pública inicial do mercado acionário americano de todos os tempos. Isso transformaria a Alibaba o maior varejista online do mundo, com um valor avaliado em 158 bilhões de dólares (3).

Ao analisarmos o mercado brasileiro, percebemos que se trata de um comércio jovem e com bom potencial. Percebe-se ainda que, no Brasil, o hábito de se fazer compras pela internet não está consolidado. Apenas 19% dos compradores usam esse serviço semanalmente, enquanto em países mais desenvolvidos estes números chegam a 35% na Alemanha e a 39% no Reino Unido. Outro indicador de que o mercado brasileiro ainda é jovem é que 61% dos consumidores de varejo online utilizaram o serviço pela primeira vez nos últimos 4 anos (4).

Como o mercado de varejo online é novo como um todo, este ainda passará por algumas mudanças drásticas em um curto espaço de tempo. Um dos itens-chave destas mudanças é a capacidade de se analisar os dados gerados pelos consumidores. Com estes dados será possível segmentar os clientes mais facilmente e as empresas poderão direcionar suas investidas de forma mais eficiente, chegando ao ponto em que campanhas de marketing e precificação serão totalmente personalizadas (5). Uma das maneiras de se usar estes dados são os sistemas de recomendação.

“Sistemas de recomendação são ferramentas e técnicas de software destinadas a prover sugestões de itens para usuários” (6). O sistema tem o propósito de automatizar o processo de recomendação e auxiliar na tomada de decisão, podendo ser aplicado em diversas áreas da indústria, tais como na indicação de notícias, músicas, relações de amizade ou artigos científicos.

Estes sistemas são utilizados por diversos serviços online e geram um grande impacto quando utilizados corretamente. Em 2012, cerca de 75% dos vídeos assistidos através do site NetFlix foram acessados por meio de recomendações (7). Em 2006, as recomendações representaram 35% dos livros vendidos pela Amazon (8), enquanto em 2007 cerca de 38% das notícias lidas no Google News foram sugeridas por um sistema de recomendação (9).

De modo geral, um sistema de recomendação possui três etapas: a aquisição dos dados de entrada, a determinação das recomendações e finalmente a apresentação dos resultados ao usuário. A aquisição dos dados de entrada pode ser feita tanto de forma automática quanto manual, e em geral utiliza-se um banco de dados para armazenar essas informações. As sugestões são feitas segundo uma estratégia de recomendação determinada a priori, que pode ser fundamentada nas preferências do usuário, nas características dos itens ou em alguma formulação mista. Finalmente, os resultados são apresentados na interface sob variadas formas, como por exemplo em uma lista dos N itens mais relevantes para o usuário.

Conforme o tipo específico de itens recomendados, o design do sistema, a interface homem-máquina e o tipo de técnica de recomendação são construídos a fim de prover sugestões mais adequadas.

Os sistemas de recomendação são destinados primeiramente aos indivíduos que não possuem competência ou experiência suficiente para avaliar o grande número de opções do conjunto total de itens. Dessa forma, a interface homem-máquina é adaptada a cada um dos usuários, de maneira que eles recebam recomendações adequadas ao seu perfil. Essa ideia, amplamente divulgada por um antigo diretor executivo do e-commerce *Amazon.com*, se resume à sua fala de que “se você possui 2 milhões de clientes na web, você precisa ter 2 milhões de lojas na web” ([10](#)).

1.1 Motivação

Conforme apresentado, a quantidade de lojas de varejo online cresce em ritmo acelerado no Brasil e no mundo. Motivados pela importância econômica dos e-commerces, bem como pela possibilidade de criar um conjunto de ferramentas *open source* que possam ser utilizadas pela comunidade acadêmica e empresarial, propomos como Trabalho de Conclusão de Curso o desenvolvimento de uma biblioteca computacional para sistemas de recomendação de produtos de lojas de comércio online.

Esse pacote computacional é composto de métodos de leitura de dados de histórico de compras e de informações de clientes e produtos, de cálculo de sugestões de itens com base em algoritmos de recomendação e de análise de desempenho das recomendações.

A motivação de se criar uma biblioteca de software decorre principalmente da sua abrangência e capacidade de adaptação, visto que é possível atender a mais casos de uso que um sistema de recomendação completo. De um lado, um sistema de recomendação possui uma finalidade específica – como por exemplo de sugerir notícias para usuários de internet – e uma entrada e saída de dados específica – como por exemplo o fato de o usuário selecionar seus jornais preferidos antes de começar a leitura ou o fato de as notícias sempre estarem ordenada pelas mais recentes em uma tabela de sugestões. De outro lado,

uma biblioteca computacional pode receber qualquer tipo de dados e gerar qualquer saída de dados.

Caso uma empresa ou um acadêmico queira construir seu próprio sistema de recomendação, basta elaborar a conexão entre o pacote apresentado pela dupla, seu banco de dados e a interface gráfica de apresentação de resultados.

As contribuições científica e tecnológica deste trabalho para a Engenharia Mecatrônica estão sobretudo nos campos de inteligência artificial, de sistemas de informação e de automação de processos.

As competências acadêmicas necessárias para a execução desse trabalho envolvem algoritmos e estruturas de dados (abordados em PMR2300 – Computação para Automação), documentação e modelagem de sistemas computacionais (explicados em PMR2440 – Programação para Automação), sistemas de informação e banco de dados (tratados em PMR2490 – Sistemas de Informação) e inteligencia artificial, com enfase em aprendizado de máquina (aprofundados em PMR2728 – Teoria de Probabilidades em Inteligência Artificial e Robótica). As competências técnicas abrangem programação estatística e funcional, demonstradas através da linguagem R.

1.2 Objetivos

O objetivo do presente Trabalho de Conclusão de Curso é o desenvolvimento de uma biblioteca computacional para sistemas de recomendação de produtos de lojas de comércio online, e respectiva análise de desempenho das recomendações propostas.

O pacote de software é composto de três diferentes algoritmos de recomendação, além de funções para avaliar a qualidade das sugestões. Neste texto, será feita uma avaliação comparativa entre os três algoritmos. A explicação detalhada dos métodos se encontra no Capítulo 5.

A fim de se poder experimentar a influência de diversos parâmetros na qualidade das recomendações, todas as funções foram desenvolvidas integralmente pela dupla, e nenhuma biblioteca externa foi utilizada. O objetivo do trabalho não é, portanto, o uso de ferramentas de recomendação já disponíveis no mercado, mas sim a elaboração de uma biblioteca que possibilite a construção e análise de um sistema de recomendação próprio. Qualquer e-commerce interessado no assunto pode, portanto, apropriar-se do pacote de software e modificá-lo para atender a suas especificidades e melhorar as sugestões.

Essa ferramenta tem como finalidade a automatização do processo de recomendação de itens, e pode ser aplicada em diversas áreas da indústria, tais como na indicação de notícias, músicas, relações de amizade ou artigos científicos. No nosso trabalho, a biblioteca terá como foco a sugestão de produtos de lojas de comércio online que disponham de um

histórico de compras dos usuários e das características dos produtos.

A qualidade das recomendações será avaliada quanto a precisão, abrangência e tempo de execução. Uma descrição detalhada da avaliação do sistema de recomendação está descrita no Capítulo 4.3.

Por meio de uma validação cruzada, analisaremos a influência dos principais parâmetros do problema na qualidade das recomendações, como o tamanho do banco de dados, a quantidade de informações de itens e clientes utilizadas na recomendação e outros.

Será discutido o impacto dos principais desafios tecnológicos e científicos dos sistemas de recomendação na nossa proposta de solução, tais como a escalabilidade, a adaptação a novos usuários e a esparsidade dos dados (11).

Ao final, será possível extrair uma validação experimental das diretrizes fundamentais a serem seguidas por e-commerce que desejem desenvolver um sistema de recomendação próprio a partir da biblioteca desenvolvida neste trabalho.

2 Estado da Arte

As terminologias *cliente* e *usuário* neste texto serão intercambiáveis e sem distinção semântica, mesmo que na prática essas duas entidades possam ser diferentes. Da mesma forma, *item* e *produto* terão o mesmo significado neste trabalho.

A fim de tornar a formulação mais genérica, também não faremos distinção entre *avaliação positiva* de um item e *compra* de um item. Avaliação positiva é toda avaliação r_{ui} do item i feita pelo usuário u tal que $r_{ui} > M$, e avaliação negativa tal que $r_{ui} \leq M$, sendo M um valor mínimo escolhido a priori, indicador de que o usuário u “gostou” do item i . No caso de um banco de dados sem avaliações dos produtos, será levada em conta a compra dos itens e será admitida avaliação unitária e valor mínimo nulo. Desta forma, os bancos de dados que contenham informações do tipo “usuário u avaliou o item i em $r_{ui} = 3.54 > M$ ” e aqueles que contenham “usuário u comprou o item i , logo $r_{ui} = 1 > 0$ ” serão tratados equivalentemente. Vale observar que essa definição difere da Referência (12), em que avaliação positiva é aquela tal que $r_{ui} \geq M$.

2.1 Estado da arte dos problemas

O problema de recomendação pode ser formulado como se segue, adaptado da referência (13), com notação inspirada em (12):

“Seja \mathcal{U} o conjunto de todos os usuários e seja \mathcal{I} o conjunto de todos os itens que podem ser recomendados, tais como livros, filmes ou artigos científicos. Seja ℓ uma função de utilidade, que mede a relevância do produto i para usuário u . Em notação matemática, $\ell : \mathcal{U} \times \mathcal{I} \rightarrow \mathcal{R}$, onde \mathcal{R} é um conjunto totalmente ordenado – por exemplo, números inteiros ou números reais dentro de um determinado intervalo, em geral $\{-1, 0, +1\}$ ou $[1, 5]$. O objetivo do sistema de recomendação é determinar o item \tilde{i}_u que maximize a utilidade ℓ_{ui} do usuário u . ”

$$\forall u \in \mathcal{U}, \tilde{i}_u = \arg \max_{i \in \mathcal{I}} \ell_{ui} \quad (2.1)$$

O problema central da recomendação é que “em geral a função ℓ é desconhecida ou não é definida para todo o espaço $\mathcal{U} \times \mathcal{I}$ ”, e portanto determinar \tilde{i} através da equação 2.1 é inviável.

Em algumas formulações, “a utilidade é descrita pela avaliação r_{ui} do item i feita pelo usuário u ”. Neste caso, o sistema de recomendação busca determinar \hat{r}_{ui} que melhor se aproxime de r_{ui} , e a qualidade da recomendação é normalmente descrita pela distância

entre esses dois valores. Em outros sistemas, todavia, a utilidade é descrita diferentemente, de forma que o item com maior valor de \hat{r}_{ui} não é necessariamente recomendado.

Para lidar com o problema da recomendação, existem três grandes grupos de estratégias de sugestão de itens, segundo as referências (13, 14):

- Recomendações baseadas em conteúdo: o usuário recebe sugestões de itens similares àqueles pelos quais ele se interessou no passado;
- Recomendações colaborativas: o usuário recebe sugestões de itens que pessoas com preferências semelhantes gostaram no passado;
- Recomendações híbridas: esses métodos combinam características de sistemas colaborativos e baseados em conteúdo. O usuário recebe sugestões de itens compatíveis com seu perfil e de itens do interesse de usuários com perfil similar.

As estratégias de recomendação baseadas em conteúdo exploram os dados dos itens para calcular a sua relevância conforme o perfil do usuário. Suas técnicas de recomendação podem ser classificadas em dois grupos: aquelas baseadas em heurísticas ou memória – fazem a previsão com base em toda a coleção de itens anteriormente classificados pelos usuários – e aquelas baseadas em modelos – utilizam o conjunto de avaliações com o objetivo de descrever a interação entre usuários e itens, tal como em uma regressão linear ou em uma rede Bayesiana.

Na abordagem de sistemas baseados em conteúdo, a recomendação pode ser vista como um problema de aprendizado de máquina, em que o sistema adquire conhecimento sobre o usuário. Muitas vezes é recomendado que o aprendizado seja feito com base no perfil do usuário em uso contínuo, ao invés de forçá-lo a responder diversas perguntas demográficas (15) – idade, gênero, classe social, etc. O objetivo é categorizar novas informações baseadas em informações previamente adquiridas e rotuladas como interessantes ou não pelo usuário. Com estas informações em mão, é possível gerar modelos preditivos que evoluem conforme aparecem novas informações.

Em sistemas baseados em conteúdo, os itens a serem recomendados podem possuir diversos atributos e formas de classificação. Em documentos como e-mails, *websites* ou comentários de usuários, os textos não tem estrutura definida e a abordagem mais comum para escolher o melhor item é a mineração de informações. O usuário procura por uma lista de termos desejados e o sistema retorna os textos de maior relevância, tal como é feito em um motor de busca (16). Nesses casos, calcula-se a similaridade entre documentos a partir da importância das palavras ou termos similares, como a TF-IDF ou o classificador Bayesiano (17).

Em bancos de dados relacionais, os itens possuem uma categorização pré-definida, e sua relevância depende das suas características, descritas pela matriz de atributos **A**. Cada

feature pertence a um conjunto distinto, podendo ser booleano (possui ou não possui), inteiro ou real (preço, data, etc.), ou um coleção finita de valores (marca, modelo, gênero, etc.), como exemplifica a Tabela 1.

Tabela 1 – Atributos a_{if}

	f_1	f_2	f_3	f_4
i_1	1	50	0.8	P
i_2	0	75	0.3	M
i_3	1	30	0.4	G

As recomendações colaborativas, por sua vez, tentam prever a utilidade dos itens para cada cliente com base em itens previamente avaliados por outros usuários. Elas podem ser baseadas em usuários, isto é, na escolha de clientes que possuam avaliações similares de produtos, quanto baseadas em itens, na escolha de produtos avaliados similarmente (18).

Mais formalmente, quando a filtragem colaborativa é baseada em usuários, a utilidade ℓ_{ui} de um item i para um usuário u é estimada com base nas utilidades $\ell_{v_k^u}$ dos usuários $v_k^u \in \mathcal{U}$ que são “similares” ao usuário u . De maneira análoga, quando baseada em itens, a utilidade ℓ_{ui} é prevista com base nas utilidades $\ell_{uj_k^u}$, dado itens $j_k^u \in \mathcal{I}$ que são “similares” aos itens i .

Na prática, o cálculo das recomendações para sistemas colaborativos é feito a partir da matriz de avaliações \mathbf{R} . Isso pode ser exemplificado pela Tabela 2, que possui avaliações de 1 a 5, sendo $M = 2$. Em um sistema usuário-usuário, o cliente u_1 receberia recomendação do item i_4 , pois para os itens i_2 e i_3 suas avaliações foram similares às do cliente u_2 . Já para um sistema item-item, o usuário u_3 receberia recomendação do item i_3 , pois este tem avaliações similares às do item i_2 , avaliado positivamente pelo usuário u_3 .

Tabela 2 – Avaliações r_{ui}

	i_1	i_2	i_3	i_4
u_1	-	4	3	-
u_2	-	4	3	5
u_3	2	5	-	1

Por fim, as recomendações híbridas combinam aspectos tanto da filtragem colaborativa (baseada em usuários ou em itens) quanto da filtragem baseada em conteúdo, com o objetivo de atingir uma melhor recomendação ou de superar problemas recorrentes nas técnicas individuais, como a esparsidade (*sparsity*) dos dados ou o *cold start* (19).

2.2 Estado da arte das soluções

Do ponto de vista do estado da arte das soluções, as variáveis de interesse estão ligadas ao número de usuários no sistema, ao número de itens, à medida de qualidade da recomendação e ao custo computacional (20).

No que se refere à dependência do número de usuários, a filtragem colaborativa baseada em usuários é extremamente efetiva para um baixo número de usuários. A filtragem colaborativa a base de itens é consideravelmente pior para um baixo número de usuários, mas supera todos os outros métodos baseados em memória conforme o número de clientes aumenta.

A dependência do número de itens é, de certa forma, oposta à de usuários: a filtragem colaborativa baseada em itens é extremamente efetiva para poucos itens, enquanto aquela baseada em usuários supera todos os outros métodos baseados em memória para grandes quantidades de itens.

Com relação à medida de qualidade, avaliada a partir da acurácia dos dados, a filtragem baseada em usuários e a baseada em itens mostram uma dependência semelhante. Na análise de menor erro quadrático médio entre o item sugerido e o item efetivamente comprado, todos os métodos de recomendação variam não-linearmente com o número de usuários, itens e acurácia, e de modo geral há um compromisso (*trade-off*) entre a esparsidade (*sparsity*) dos dados e o tempo de processamento.

2.3 Desafios científicos e tecnológicos

Um dos maiores desafios tecnológicos dos sistemas de recomendação é, atualmente, o da escalabilidade (15). O sistema de recomendação deve ser flexível no sentido de poder operar igualmente bem tanto em pequenas quanto em grandes bases de dados, que podem chegar até centenas de milhões de clientes (21) e de produtos (22). Isso significa que as recomendações devem ser suficientemente rápidas e ainda assim prover sugestões valiosas aos consumidores.

Um problema muito comum nos sistemas de recomendação é o do *cold start*, que atinge principalmente os sistemas de filtragem colaborativa, grandemente dependentes da matriz de avaliações \mathbf{R} . Quando itens ou usuários são inicialmente introduzidos no sistema, existe pouca ou nenhuma informação sobre eles. O sistema é incapaz de realizar inferências sobre quais itens recomendar ao novo usuário ou sobre quais produtos são similares ao novo item. Na Tabela 3, por exemplo, o item i_{100} não possui nenhuma avaliação, e nunca seria recomendado em sistemas puramente baseados em itens. Analogamente, o usuário u_3 também não teria nenhuma sugestão de itens para sistemas puramente baseados em usuários.

Tabela 3 – Avaliações r_{ui}

	i_1	i_2	...	i_{100}
u_1	5	4	...	-
u_2	-	2	...	-
u_3	-	-	...	-

Também é uma grande dificuldade dos algoritmos baseados em filtragem colaborativa a esparsidade dos dados. Como a maioria dos clientes interage com uma pequena quantidade de itens, a matriz de avaliações tem em geral menos de 1% dos valores preenchidos, e o sistema deve prever os outros valores (23).

Outro desafio científico é referente à diversidade das recomendações realizadas, também chamado de excesso de especialização (*over-specialization*) (13). Ao mesmo tempo que o sistema deve apresentar itens similares ao que o usuário está procurando, ele também deve sugerir itens que o usuário desconheça ou que nem saiba que poderiam interessá-lo. Esse problema afeta principalmente os sistemas baseados em conteúdo, pois itens com características similares tendem a ser sempre recomendados. Na Tabela 4, o item i_2 seria sugerido para um usuário que tenha avaliado i_1 , apesar de esse não apresentar nenhuma característica diferente do item previamente comprado. Para contornar essa dificuldade, costuma-se introduzir elementos de aleatoriedade na recomendação, por exemplo a partir de algoritmos genéticos (14).

Tabela 4 – Atributos a_{if}

	f_1	f_2	f_3
i_1	1	50	0.8
i_2	1	50	0.8
i_3	0	75	0.3

Além desse desafio, existe também o da análise rasa do conteúdo (*shallow content analysis*). O sistema, ao avaliar a característica dos itens, não consegue extraír importantes aspectos para o usuário caso eles não estejam explicitamente descritos na categorização do banco de dados. Isso pode ocorrer, por exemplo, com fatores externos ao produto que influenciem na compra do usuário, como em datas comemorativas, em compras induzidas por propaganda, em compras “impulsivas”, etc. Se um usuário comprou um arranjo de flores no dia das mães, não é necessariamente verdade que ele se interessa por flores. Da mesma maneira, sistemas que ignorem a sazonalidade de certos produtos não recomendariam arranjos de flores para clientes que não tenham comprado itens parecidos, mesmo no dia das mães.

Por fim, um desafio científico que este trabalho enfrentará é a execução de um sistema híbrido do ponto de vista de efemeridade e persistência, ao construir um modelo

de recomendação que integre as preferências de curto e longo termo dos usuários (10). A análise dos dados de compras anteriores, bem como de dados demográficos, deverá portanto ser incorporada à análise de característica dos produtos, a fim de enriquecer a acurácia do sistema (15).

Esse tópico de pesquisa inclui ainda diversos desafios científicos e tecnológicos que não serão tratados no nosso projeto, tais como a preservação da privacidade dos usuários, a criação de modelos de recomendação inter-domínios, o desenvolvimento de sistemas descentralizados operando em redes computacionais distribuídas, a otimização de sistemas para sequências de recomendações, a otimização de sistemas para dispositivos móveis e outros. Um sistema de recomendação inteligente também deveria prever quando enviar uma determinada recomendação, e não agir apenas mediante requisição dos clientes (24).

2.4 Soluções propostas

Este Trabalho de Conclusão de Curso aborda três propostas de solução para o problema da recomendação, sendo duas delas retiradas de referências bibliográficas (12, 25), e uma outra apresentada pela dupla. O objetivo é realizar uma análise comparativa entre cada um dos métodos e estabelecer diretrizes para sua aplicação em e-commerce. Os algoritmos propostos estão descritos com maior detalhe no Capítulo 5.

Todas as soluções são algoritmos híbridos, por utilizarem na recomendação tanto a matriz de avaliações **R** quanto a matriz de atributos **A**. Optou-se por dar importância aos algoritmos híbridos em razão de os e-commerce estruturarem seus bancos de dados em torno da descrição dos itens à venda. De modo geral, as tabelas de itens possuem dezenas de atributos, dependendo do ramo de negócios da loja, e pouco detalhe é dado à interação entre o grupo de usuários e itens. A tabela de histórico de compras se limita a informações como data e método de pagamento, e detém pouca informação adicional que possa ser utilizada na recomendação de produtos. Dessa forma supusemos que métodos puramente colaborativos, fundamentados na avaliação dos itens por parte dos usuários, teriam pior desempenho que métodos baseados em conteúdo, que exploram as características dos itens na recomendação.

A solução *FW* determina a similaridade de dois itens a partir de medidas de distância para cada um dos atributos dos itens, ponderadas por pesos determinados na regressão linear de uma equação descrita pelo interesse dos usuários em cada *feature*.

O método *UP* parte do princípio que os usuários estão interessados nos atributos dos itens, e traça correlações entre esses dois elementos para obter pesos que servirão de base para o cálculo da similaridade inter-usuários, utilizada na recomendação pelo método da vizinhança (*nearest neighbors*).

A variante *UI*, elaborada pela dupla, recomenda o melhor item a partir das matrizes de correlação usuário-atributo e atributo-item, a fim de obter a matriz de correlação usuário-item. Espera-se que essa solução tenha desempenho similar ao método de base *UP*, pois ambos buscam explorar as características dos itens para determinar a preferência do usuário.

3 Metodologia

A metodologia de projeto deste Trabalho de Conclusão de Curso foi fundamentada principalmente na Referência 26. Por se tratar de um projeto de Engenharia de Software, foi necessário dar ênfase às etapas iterativas de desenvolvimento dos algoritmos. Esse processo cíclico, com fases de especificação, desenvolvimento e validação, permitiu obter resultados preliminares e os modificar os algoritmos ao longo da disciplina, ajustando detalhes e melhorando o sistema gradativamente (27).

A metodologia de execução do projeto, assim como a de avaliação dos resultados, pode ser consolidada da seguinte maneira:

3.1 Definição da Necessidade

Com o crescente número de lojas de comércio online, tornou-se necessário a criação de sistemas que pudessem entender e prever o comportamento de consumidores, a fim de oferecer produtos específicos para cada um deles, aumentando o número de vendas e a satisfação do cliente. Observa-se atualmente que o número de sistemas de recomendação gratuitos, de fácil integração e de código aberto (*open source*) são limitados e não correspondem às necessidades do mercado. Existe, pois, a necessidade da criação de um sistema que possa ser utilizado por e-commerces que desejem estabelecer seu próprio sistema de recomendação ou mesmo por indivíduos interessados na temática da recomendação de itens.

3.2 Definição dos Parâmetros de Sucesso

O sucesso do projeto pode ser medido em duas frentes: a primeira, quantitativa, mede a precisão e a abrangência das recomendações; a segunda, qualitativa, avalia se o sistema responde bem aos problemas recorrentes desse tópico de pesquisa, tais como a escalabilidade, o excesso de especialização e outros.

3.3 Síntese de Soluções

Nesta fase do projeto, foram propostas possíveis soluções para o desafio da recomendação. Decidiu-se avaliar dois métodos híbridos do meio acadêmico e um outro elaborado pela dupla.

3.4 Detalhamento da Solução

Após a escolha dos métodos de recomendação, as soluções foram detalhadas matematicamente segundo uma mesma notação, e a estrutura dos algoritmos foi descrita e exemplificada. Neste ponto, escolheu-se também a linguagem de programação (R) e a forma de entrada e saída de dados (arquivos .csv).

A fim de facilitar o pré-processamento dos dados, estabelecemos que seriam necessários dois arquivos. Um deles deve conter a matriz de atributos **A** e o outro, a matriz de avaliações **R**.

$$\mathbf{A} = \begin{bmatrix} a_{i_1 f_1} & a_{i_1 f_2} & a_{i_1 f_3} & \dots \\ a_{i_2 f_1} & a_{i_2 f_2} & a_{i_2 f_3} & \dots \\ a_{i_3 f_1} & a_{i_3 f_2} & a_{i_3 f_3} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (3.1)$$

$$\mathbf{R} = \begin{bmatrix} r_{u_1 i_1} & r_{u_1 i_2} & r_{u_1 i_3} & \dots \\ r_{u_2 i_1} & r_{u_2 i_2} & r_{u_2 i_3} & \dots \\ r_{u_3 i_1} & r_{u_3 i_2} & r_{u_3 i_3} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (3.2)$$

3.5 Modelamento e Simulação

Uma vez determinada a forma de entrada de informações, definiram-se os conjuntos de dados a serem utilizados. O primeiro conjunto de dados abertos é proveniente do sistema de recomendações de filmes MovieLens (<http://movielens.umn.edu>). Nessa base de dados, o catálogo de filme faz o papel de catálogo de produtos, e o histórico de compras se refere à avaliação dos filmes feita por cada usuário. Outro conjunto de dados abertos é do website Internet Movie Database (IMDB). Na nossa análise, esses dois bancos foram utilizados complementarmente. TODO falar como juntamos

Os métodos escolhidos foram codificados em R e testados com inicialmente com o banco de dados 100k. Posteriormente, testamos os algoritmos no banco IMDB, a fim de avaliar a qualidade das recomendações mediante a mudanças na base de dados.

Ainda na etapa de implementação, confirmamos a validade de cada um dos métodos aplicando-os nas matrizes-referência (Tabelas 6 e 7).

3.6 Validação Cruzada

A fim de realizar um estudo comparativo (*benchmarking*) com os artigos de referência, mantivemos a mesma metodologia de avaliação de qualidade do artigo 12.

Em particular, implementamos uma validação cruzada considerando $T = 75\%$ do banco de dados como base de treinamento ou aprendizado e os 25% restantes como base de testes. Em seguida, mascaramos $H = 75\%$ das avaliações dos usuários-teste, de modo a medir a qualidade do sistema de recomendação em prever os itens positivamente avaliados. Cerca de uma dezena de parâmetros de interesse foram avaliados para cada um dos métodos (Tabela 23).

Com foco na reproduzibilidade do trabalho, realizamos todas as amostragens em R utilizando o número 2 como semente aleatória (*state seed*). Além disso, não fizemos distinção entre valores não observados (*NA value/NULL value*) e avaliações nulas ($r_{ui} = 0$), pois na maioria dos casos essa simplificação é válida.

Sabe-se que a extração de um modelo por meio de uma validação cruzada sobre uma mesma base de dados pode gerar *overfitting* (Referência 28). Para não cair nesse erro, utilizamos dois bancos de dados, sendo um deles para avaliação da qualidade das recomendações mediante mudanças em parâmetros do problema e o outro para avaliação dos algoritmos em uma base totalmente diferente, sem modelagem *a priori*.

Como a complexidade dos algoritmos excedia o limite dos computadores pessoais da dupla, foi necessário contratar o serviço de computação nas nuvens Amazon Web Services.

Alugamos duas máquinas virtuais do tipo `r3.large`, otimizadas para memória. As máquinas, de especificação 2 vCPU, 15 GB de memória RAM e sistema operacional Amazon Linux AMI release 2014.09 x86_64, baseado em RHEL Fedora, custaram USD 0,175 por hora de uso. Todos os testes foram realizados em aproximadamente 12 horas, custando apenas USD 4,20 (menos de R\$ 12,00).

4 Requisitos

A partir dos objetivos deste Trabalho de Conclusão de Curso, é possível extrair os requisitos funcionais do sistema de recomendação. Esses requisitos ditam principalmente sobre a escalabilidade e o desempenho das recomendações do sistema.

Como as sugestões serão calculadas com antecedência, não há necessidade para uma elevada taxa de recomendações por período de tempo (*throughput*). Deseja-se contudo que o sistema possa gerar todas as recomendações para um banco de dados de cem mil clientes em uma hora, isto é, que tenha *throughput* mínimo de 28 recomendação por segundo. Os sistemas de recomendação tradicionais possuem *throughput* de cerca de 500 recomendações por segundo, mas operam em servidores dedicados de maior potência computacional (29).

A fim de poder estabelecer uma base comparativa entre o sistema proposto *UI* e os sistemas de referência *FW* e *UP*, serão utilizados os mesmos indicadores de desempenho dos artigos-base: precisão, abrangência e medida F_1 (12, 25). Precisão é a porcentagem de casos corretamente preditos em relação ao tamanho da lista de recomendações. Abrangência é a razão entre o número de itens corretamente preditos e daqueles que foram efetivamente avaliados pelo usuário. A medida F_1 , por sua vez, é a média harmônica entre precisão e abrangência.

Todas essas métricas são dependentes dos diversos parâmetros do problema, como do tamanho da lista de recomendações N , da quantidade de vizinhos mais próximos k , e principalmente do banco de dados de teste. Como os artigos de referência não os disponibilizaram integralmente, serão estimados os valores de precisão, abrangência e medida F_1 para o banco de dados da dupla.

Espera-se que a precisão, abrangência e consequentemente a medida F_1 sejam maiores que 20%. Esses valores foram escolhidos por serem superiores aos de algoritmos puramente baseados em conteúdo ou em filtragem colaborativa (12, 25). Na prática, o resultado mais importante é a comparação entre os três métodos para um banco de dados de referência.

Neste trabalho o *benchmarking* é feito por meio da união de dois bancos amplamente utilizados na comunidade científica de Sistemas de Recomendação. O primeiro, denominado MovieLens 100k, é composto de 100 000 avaliações (valores inteiros de 1 a 5) de 943 usuários para 1682 filmes (30). Além disso, cada usuário (idade, sexo, profissão, logradouro) avaliou pelo menos 20 filmes (categoria, ano de publicação). O segundo banco de dados é extraído do Internet Movie Database (IMDB), e possui 28 819 filmes. Esse banco está presente na biblioteca *ggplot2* da linguagem de programação R (31). A união desses dois bancos é denominada 100k-IMDB, e a metodologia para essa união está descrita na Seção 3.5.

Os requisitos funcionais são suportados por requisitos não-funcionais, e estes são determinados pelas restrições sobre o projeto ou execução, tais como desenvolvimento e confiabilidade.

O sistema de recomendação deverá poder ser utilizado por qualquer e-commerce que disponha de um banco de dados de clientes, produtos e histórico de compras, desde que o formato de entrada, a ser especificado no Capítulo 7, seja seguido.

Além disso o sistema deverá ser desenvolvido em tecnologias abertas (*open source*) que tenham um alto número de colaboradores, como o sistema de gestão de banco de dados MySQL ou a linguagem de programação estatística R, a fim de torná-lo reutilizável por alunos ou e-commerces interessados.

Por fim, o sistema de recomendação deverá ser escalável e flexível no sentido de poder operar igualmente bem tanto em pequenas quanto em grandes bases de dados.

Apesar serem importantes parâmetros de um sistema de recomendação, a taxa de recomendações por período de tempo e a escalabilidade estão intimamente relacionados ao orçamento do projeto. Pode-se obter virtualmente qualquer *throughput* desejado, contanto que haja investimento equivalente em infra-estrutura computacional. O mesmo não é válido para os parâmetros de qualidade da recomendação, que dependem tão somente dos algoritmos de sugestão. Neste trabalho, assumimos que o sistema de operará em microcomputadores pessoais, e por isso o requisito funcional *throughput* se faz necessário.

Com os requisitos do sistema de recomendação definidos, devemos estruturar o seu relacionamento com o administrador do sistema. Para isto determinamos seus casos de uso, classes e atividades.

4.1 Diagrama de Casos de Uso

Os casos de uso se dividem em *Avaliar Performance*, *Configurar Banco de Dados*, *Recomendar UI*, *Recomendar UP*, *Recomendar FW*.

O caso de Uso *Avaliar Performance*, para avaliar a performance do sistema de recomendação. *Avaliar Performance* se relaciona com outros 3 casos de uso. O primeiro deles, *Mascarar Dados*, serve para mascarar dados de alguns usuários-teste na matriz de avaliações. Para, após, compará-los às recomendações calculadas pelo sistema. O segundo caso, *Dividir Banco de Treino*, tem o objetivo de dividir o banco de dados em dois, um para o aprendizado do sistema e o segundo para testes. O terceiro caso, *Devolver Indicadores*, torna os indicadores de performance acessíveis ao administrador do sistema.

O caso de uso *Configurar Banco de Dados* tem a utilidade de converter o banco de dados de entrada em matrizes. O caso *Ler Itens*, é o encarregado de ler o arquivo de itens fornecido pelo banco de dados, assim como *Ler Usuários* e *Ler Histórico* são

para usuários e histórico. Os casos de uso *Gerar Matriz de Atributos* e *Gerar Matriz de Avaliação* constroem as matrizes de acordo com dados lidos pelos casos de uso de leitura.

Já os casos de uso *Recomendar UI*, *Recomendar UP* e *Recomendar FW*, são os casos de uso em que se geram as recomendações pelos métodos baseados na correlação usuário-item, no perfil de usuários e na ponderação de atributos respectivamente. O método de recomendação baseado na ponderação de atributos necessita de outros 3 casos de uso, o *Normalizar Matriz*, onde se normaliza as colunas da matriz de distância entre atributos. Esta distância entre atributos é calculada pelos dois outros casos de uso, *Fazer Delta de Kronecker* e *Fazer Índice Jaccard*.

Estes casos de uso foram representados no diagrama de casos de uso (Figura 1).

4.2 Diagrama de Atividades

Para representar o fluxo de informação foi considerado o diagrama de atividades. Assim é possível visualizar os processos que vão desde a informação fornecida pelo usuário à geração de recomendações.

O primeiro diagrama de atividade (Figura 2) representa o registro de uma avaliação por parte do cliente, onde o cliente solicita o item para visualizá-lo, a plataforma Web faz o pedido de informações ao banco de dados. O banco de dados devolve as informações pedidas e a plataforma o exibe no dispositivo do cliente. Após a visualização o cliente avalia o item, a plataforma web informa ao banco de dados sobre a avaliação que a registra e a plataforma Web confirma a avaliação finalizando o processo.

O segundo diagrama (Figura 3), representa o fluxo de informação desde o pedido de uma recomendação pela plataforma web até a finalização da atividade pelo cliente. O primeiro passo é o pedido de uma recomendação, ao banco de dados, pela plataforma web. O banco de dados envia as informações necessárias para o sistema de recomendação que de acordo com as regras pré-estabelecidas as calcula e retorna para o banco de dados. O banco de dados registra estas recomendações e informa a plataforma web, que envia a recomendação ao usuário. O usuário avalia o item recomendado e finaliza o processo.

4.3 Avaliação de Desempenho

De modo geral os sistemas de recomendação tem o objetivo de apresentar ao usuário itens pelos quais ele possa se interessar. O desempenho de um sistema de recomendação se mede, portanto, na qualidade com a qual ele executa essa tarefa.

Essa qualidade pode ser medida de diferentes maneiras, tal como pela medida de distância entre os produtos recomendados e aqueles que seriam efetivamente comprados

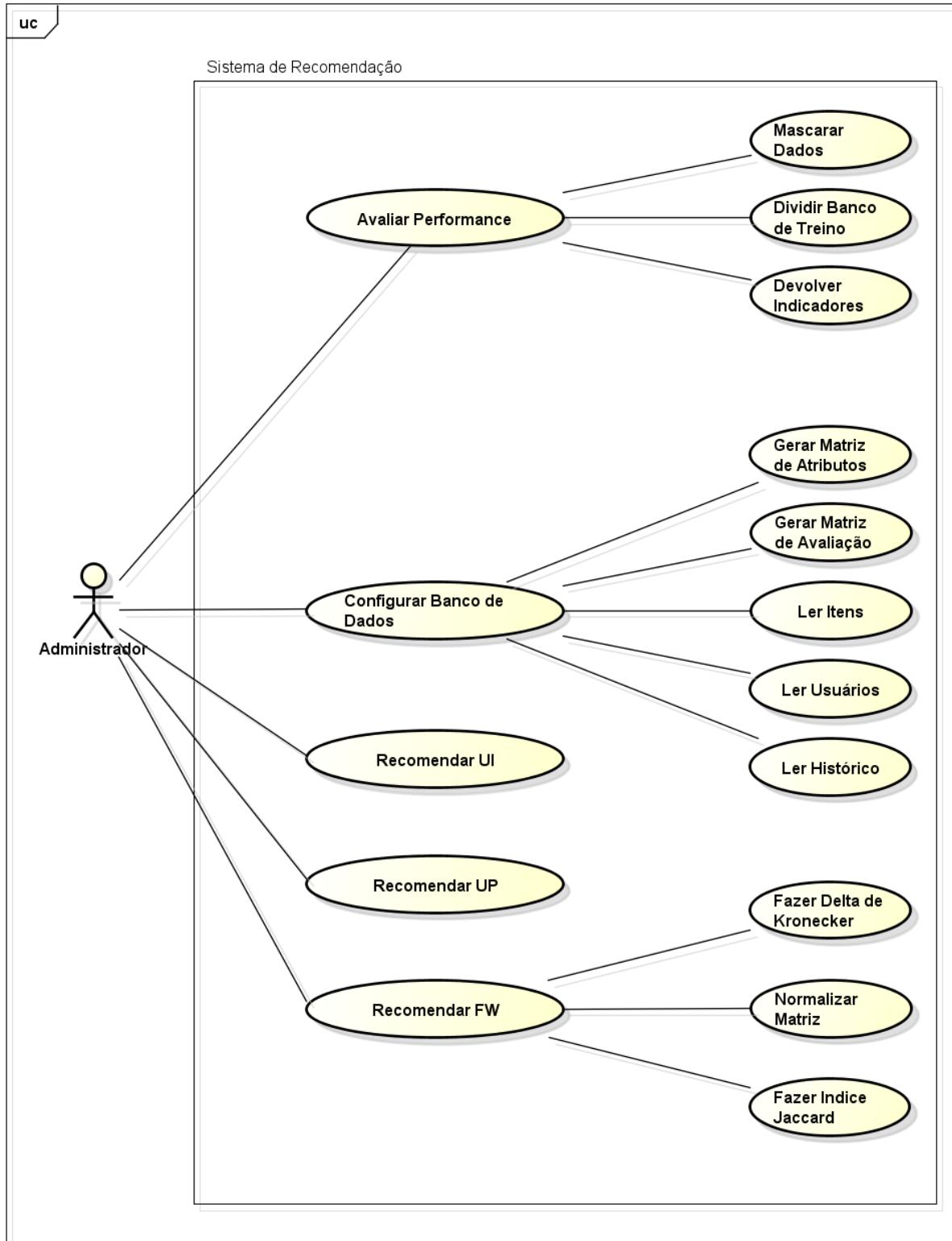


Figura 1 – Diagrama de casos de uso representando os relacionamentos entre o administrador e o sistema de recomendações.

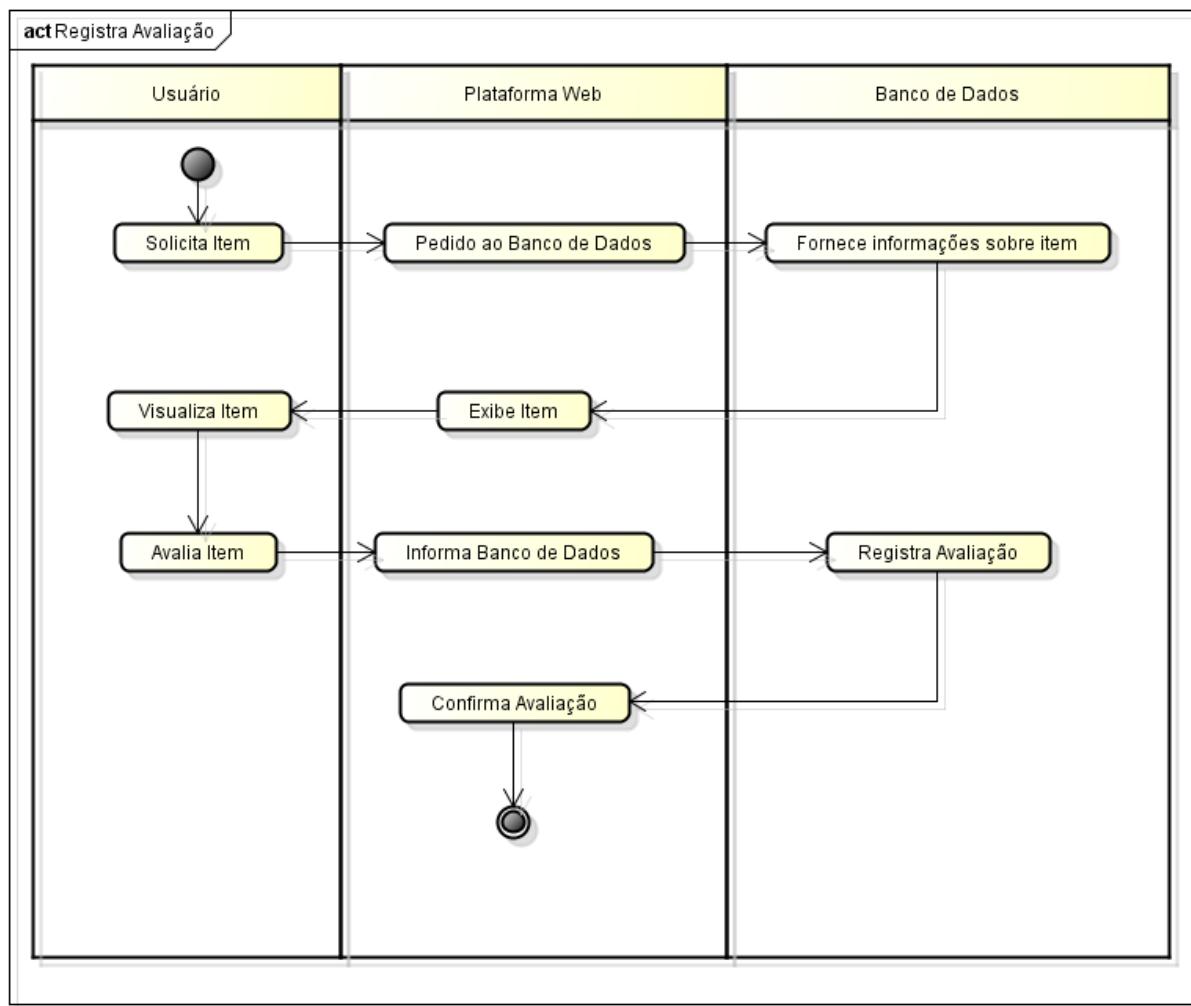


Figura 2 – Diagrama Atividades - Registro de Avaliação

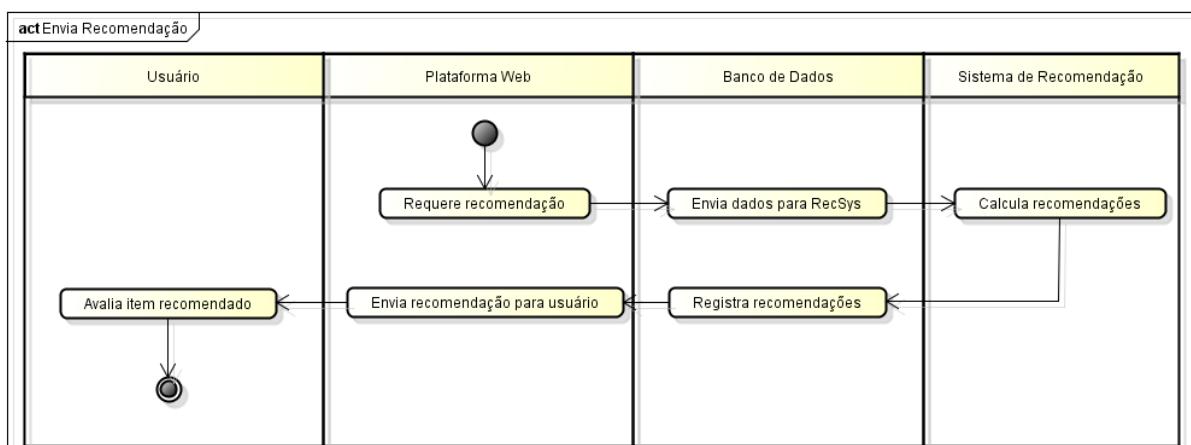


Figura 3 – Diagrama Atividades - Gerar Recomendação

i pelo cliente em uma validação cruzada (*cross validation*). Outras medidas de predição também podem ser utilizadas, a exemplo de trabalhos de recuperação de informação, tais como acurácia (*accuracy*), especificidade (*specificity*), precisão (*precision*), abrangência (*recall*), medida F_1 (F_1 -*score*), e outras.

No nosso Trabalho de Conclusão de Curso, serão utilizados precisão, abrangência, e medida F_1 . Essas medidas foram escolhidas a fim de se poder estabelecer uma base comparativa com os textos de referência, que também as utilizam. Elas estão sumarizadas na Tabela 5. As quantidades VP , FP , VN e FN significam o número de verdadeiro e falso positivos e o número de verdadeiro e falso negativos.

Tabela 5 – Avaliação de sistemas de predição

Medida	Fórmula	Significado
Precisão	$\frac{VP}{VP+FP}$	Porcentagem de casos positivos corretamente preditos.
Abrangência	$\frac{VP}{VP+FN}$	Porcentagem de casos positivos sobre aqueles que foram marcados como positivos.
F_1	$2 \cdot \frac{\text{Precisão} \cdot \text{Abrangência}}{\text{Precisão} + \text{Abrangência}}$	Média harmônica entre precisão e abrangência.

Por fim, avaliaremos o desempenho do sistema mediante a mudança nas variáveis de importância do problema, como por exemplo na quantidade de atributos utilizados na recomendação. O tempo de execução também será avaliado em função do algoritmo utilizado e do tamanho do banco de dados.

5 Detalhamento de Soluções

A fim de facilitar a compreensão dos métodos propostos neste trabalho, serão utilizadas as matrizes de avaliações \mathbf{R} e de atributos \mathbf{A} abaixo, adaptadas da Referência 25. Em todos os exemplos, considera-se valor mínimo $M = 2$. Os logaritmos são expressos em base 10 e todos os pesos w_f (descritos a seguir) são utilizados.

Tabela 6 – Avaliações r_{ui}

	i_1	i_2	i_3	i_4	i_5	i_6
u_1	-	4	-	-	5	-
u_2	-	3	-	4	-	-
u_3	-	-	-	-	-	4
u_4	5	-	3	-	-	-

Tabela 7 – Atributos a_{if}

	f_1	f_2	f_3	f_4
i_1	0	1	0	0
i_2	1	1	0	0
i_3	0	1	1	0
i_4	0	1	0	0
i_5	1	1	1	0
i_6	0	0	0	1

5.1 Algoritmo baseado na ponderação de atributos (FW)

O primeiro algoritmo que utilizaremos no sistema de recomendação, adaptado da Referência 12 e denominado ponderação de atributos, *feature weighting* ou *FW*, trata-se de um híbrido entre filtragem colaborativa e filtragem baseada em conteúdo. A partir da regressão linear de dados de uma rede social (*Internet Movie Database, IMDB*), extraem-se os pesos que determinam a importância de cada atributo dos itens, e é onde ocorre a filtragem colaborativa dos usuários. Após obtenção dos pesos, realiza-se a filtragem baseada em conteúdo para determinar os itens com maior similaridade, que são finalmente recomendados.

Na filtragem baseada em conteúdo, “cada item é representado por um vetor de atributos ou *features*”. A similaridade s_{ij} entre dois itens i e j é dada pela média ponderada

das distâncias entre as *features* dos itens:

$$s_{ij} = \sum_f w_f (1 - d_{fij}) \quad (5.1)$$

As distâncias entre os atributos d_f são determinadas conforme o tipo de dado avaliado e seu domínio, normalizadas no intervalo $[0, 1]$.

Para atributos literais, como categoria, marca, cor, etc., uma possível medida de distância é o delta de Kronecker descrito em 5.2. A similaridade entre as cores “azul” e “vermelho” é, nesse caso, 0, e sua distância é 1. O valor da distância é nulo se e somente se os atributos são idênticos.

Para atributos pertencentes a uma coleção finita de itens, tais como os atores participantes de um filme, é possível estabelecer a similaridade entre dois conjuntos a partir do índice Jaccard, descrito em 5.3. Neste caso, a similaridade entre os conjuntos $\{\text{Al Pacino, Tom Hanks}\}$ e $\{\text{Tom Hanks, Marlon Brando}\}$ é $1/3$, e a sua distância é $2/3$.

$$\delta_{mn} = \begin{cases} 1, & \text{se } m = n \\ 0, & \text{se } m \neq n \end{cases} \quad (5.2)$$

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (5.3)$$

Vale considerar a correlação entre atributos no cálculo das distâncias: a similaridade de duas marcas de calçado, por exemplo, é maior que a de duas marcas de produtos de categorias diferentes, mesmo que as marcas sejam distintas nos dois casos. Em uma primeira análise, todavia, utilizaremos para a maior parte das *features* as medidas de distância do delta de Kronecker 5.4 (Tabela 8) e do índice Jaccard 5.5. Isso significa que se os atributos de dois itens são idênticos, a distância é nula e portanto a similaridade é máxima. O sumário de algumas medidas de distância que podem ser utilizadas estão na Tabela 9.

$$\begin{aligned} d_{fij} &= 1 - \delta_{ij}^f \\ &= 1 - \delta_{a_{if} a_{jf}} \end{aligned} \quad (5.4)$$

$$\begin{aligned} d_{fij} &= 1 - J^f(i, j) \\ &= 1 - J(a_{if}, a_{jf}) \end{aligned} \quad (5.5)$$

Os pesos w_f são a priori desconhecidos. A Referência 12 os determina a partir de uma regressão linear do tipo 5.6, onde e_{ij} é o número de usuários que se interessam

tanto por i quanto por j . Esses valores permitem determinar “o julgamento humano de similaridade entre itens”, e pode ser calculado a partir da matriz de avaliações, conforme a equação 5.7 (Tabela 10). O operador booleano b_M , descrito pela Equação 5.8, nada mais é que uma ferramenta matemática para se poder extrair o número de usuários que avaliaram *positivamente* tanto i quanto j a partir de \mathbf{R} .

$$e_{ij} = w_0 + \sum_f w_f (1 - d_{fij}) \quad (5.6)$$

$$e_{ij} = \sum_u b_M (r_{ui} r_{uj}) \quad (5.7)$$

$$b_M (x) = \begin{cases} 1, & \text{se } x > M \\ 0, & \text{se } x \leq M \end{cases} \quad (5.8)$$

Desta forma, os pesos w_f são determinados a partir resolução do sistema de equações lineares 5.9 (Tabela 11). Apenas os pesos positivos e com valor absoluto expressivo (maior que um piso arbitrariamente escolhido a posteriori) são utilizados na recomendação.

$$w_0 + \sum_f w_f (1 - d_{fij}) = \sum_u b_0 (r_{ui} r_{uj}), \quad \forall i \neq j \quad (5.9)$$

Calcula-se a matriz de similaridade \mathbf{S} pela Equação 5.1 (Tabela 12) e recomendam-se os itens similares àqueles já comprados, segundo 5.10 (Tabela 13).

$$\hat{i}_u = \arg \max_{i \in \{i \mid r_{ui} > 0\}, j} s_{ij} \quad (5.10)$$

5.2 Algoritmo baseado no perfil de usuários (UP)

O segundo algoritmo, adaptado da Referência 25, é um híbrido entre filtragem colaborativa e filtragem baseada em conteúdo. Os atributos dos itens são ponderados no cálculo de similaridade, com pesos extraídos de um modelo de perfil de usuários, denominado *user profile* ou *UP*. Esse perfil leva em consideração o interesse dos usuários por *features*, indiretamente calculado a partir de seu interesse pelos itens.

Para se determinar a relevância de f para u , deve-se levar em conta não somente a frequência com a qual uma característica aparece, mas também o fato de algumas características estarem contidas na maioria dos itens. Determina-se, então, os pesos w_{uf} , que mostram a relevância de f para u , a partir da medida estatística TF-IDF (*term frequency-inverse document frequency*), presente em formulações de recuperação de informação e mineração de dados (Equação 5.13).

Tabela 8 – d_{ij}^f

f_1	i_1	i_2	i_3	i_4	i_5	i_6	f_2	i_1	i_2	i_3	i_4	i_5	i_6
i_1	-	0	1	1	0	1	i_1	-	1	1	1	1	0
i_2	0	-	0	0	1	0	i_2	1	-	1	1	1	0
i_3	1	0	-	1	0	1	i_3	1	1	-	1	1	0
i_4	1	0	1	-	0	1	i_4	1	1	1	-	1	0
i_5	0	1	0	0	-	0	i_5	1	1	1	1	-	0
i_6	1	0	1	1	0	-	i_6	0	0	0	0	0	-
f_3	i_1	i_2	i_3	i_4	i_5	i_6	f_4	i_1	i_2	i_3	i_4	i_5	i_6
i_1	-	1	0	1	0	1	i_1	-	1	1	1	1	0
i_2	1	-	0	1	0	1	i_2	1	-	1	1	1	0
i_3	0	0	-	0	1	0	i_3	1	1	-	1	1	0
i_4	1	1	0	-	0	1	i_4	1	1	1	-	1	0
i_5	0	0	1	0	-	0	i_5	1	1	1	1	-	0
i_6	1	1	0	1	0	-	i_6	0	0	0	0	0	-

Tabela 9 – Medidas de distância entre alguns atributos

Atributo f	Domínio F	Distância d_f
Marca	Literal	$1 - \delta_{ij}^f$
Esporte	Literal	$1 - \delta_{ij}^f$
Gênero	Literal	$1 - \delta_{ij}^f$
Categoria	Conjunto Literal	$1 - J^f(i, j)$
Preço	\mathbb{R}	$\frac{ a_{if} - a_{jf} }{\max_{i,j} a_{if} - a_{jf} }$
Data	\mathbb{R} milissegundos a partir de epoch (32)	$\frac{ a_{if} - a_{jf} }{\max_{i,j} a_{if} - a_{jf} }$

Tabela 10 – e_{ij}

	i_1	i_2	i_3	i_4	i_5	i_6
i_1	-	0	1	0	0	0
i_2	0	-	0	1	1	0
i_3	1	0	-	0	0	0
i_4	0	1	0	-	0	0
i_5	0	1	0	0	-	0
i_6	0	0	0	0	0	-

Tabela 11 – w_f

w_0	w_1	w_2	w_3	w_4
0.41	-0.22	-0.34	-0.03	-

Em nosso caso, TF ou *feature frequency* é a “similaridade intra-usuários”, igual ao número de vezes em que a *feature* f aparece no perfil do usuário u (Equação 5.11, Tabela 14). Se o usuário avaliou *positivamente* algum item r_{ui} , tal que r_{ui} é superior a um valor mínimo M , considera-se que u tem interesse TF_{uf} nos atributos f dos itens i , representados por a_{if} .

$$\text{TF}_{uf} = \sum_i b_M(r_{ui} | a_{if}) \quad (5.11)$$

O termo IDF ou *inverse user frequency* é a “dissimilaridade inter-usuários”, relacionada com o inverso da frequência de um atributo f dentro de todos os usuários (Equação 5.12, Tabela 15).

$$\text{IDF}_f = \log \left(\frac{|\mathcal{U}|}{\sum_u b_0(\text{TF}_{uf})} \right) \quad (5.12)$$

Os pesos w_{uf} , obtidos na TF-IDF 5.13 (Tabela 16), são utilizados para calcular a similaridade s_{uv} entre dois usuários u e v , conforme as Equações 5.14 e 5.15 (Tabela 17).

$$w_{uf} = \text{TF}_{uf} \text{IDF}_f \quad (5.13)$$

$$s_{uv} = \frac{\sum_{f \in \mathcal{F}_{uv}} w_{uf} w_{vf}}{\sqrt{\sum_{f \in \mathcal{F}_{uv}} w_{uf}^2} \sqrt{\sum_{f \in \mathcal{F}_{uv}} w_{vf}^2}} \quad (5.14)$$

$$\begin{aligned} \mathcal{F}_{uv} &= \mathcal{F}_u \cap \mathcal{F}_v \\ \mathcal{F}_u &= \{f \in \mathcal{F} \mid t_{uf} > 0\} \end{aligned} \quad (5.15)$$

Dispondo-se de \mathbf{S} , selecionam-se os k vizinhos mais próximos v_k^u com maior similaridade s_{uv} , dentre todos $v \neq u$. Posteriormente, determina-se o conjunto $\mathcal{I}_{v_k^u} = \{i \mid r_{v_k^u i} > M\}$ de itens i avaliados positivamente por v_k^u . Em 5.16 avalia-se a frequência total f_{uf} dos atributos f para os itens de $\mathcal{I}_{v_k^u}$ (Tabela 18).

$$f_{uf} = \sum_{i \in \mathcal{I}_{v_k^u}} b_0(a_{if}) \quad (5.16)$$

Por fim, a partir da Equação 5.17 calcula-se o peso ω_{ui} (Tabela 19) de cada item e gera-se a lista dos *top-N* produtos a serem recomendados para o usuário u , conforme 5.18 (Tabela 20).

$$\omega_{ui} = \sum_f a_{if} f_{uf} \quad (5.17)$$

Tabela 12 – s_{ij}

	i_1	i_2	i_3	i_4	i_5	i_6
i_1	-	0.44	1.00	0.93	0.51	0.17
i_2	0.44	-	0.51	0.44	1	-0.32
i_3	1.00	0.51	-	1.00	0.44	0.24
i_4	0.93	0.44	1.00	-	0.51	0.17
i_5	0.51	1.00	0.44	0.51	-	-0.25
i_6	0.17	-0.33	0.24	0.17	-0.25	-

Tabela 13 – \hat{u}_u (FW)

u_1	u_2	u_3	u_4
3	5	3	4

Tabela 14 – TF_{uf}

	f_1	f_2	f_3	f_4
u_1	2	2	1	0
u_2	1	2	0	0
u_3	0	0	0	1
u_4	0	2	1	0

Tabela 15 – IDF_f

f_1	f_2	f_3	f_4
0.30	0.12	0.30	0.60

Tabela 16 – w_{uf}

	f_1	f_2	f_3	f_4
u_1	0.60	0.25	0.30	0
u_2	0.30	0.25	0	0
u_3	0	0	0	0.60
u_4	0	0.25	0.30	0

Tabela 17 – s_{uv}

	u_1	u_2	u_3	u_4
u_1	-	0.96	0	1
u_2	0.96	-	0	1
u_3	0	0	-	0
u_4	1	1	0	-

$$\hat{i}_u = \arg \max_{i \in \{i \mid r_{ui} = 0\}} \omega_{ui} \quad (5.18)$$

5.3 Algoritmo baseado na correlação usuário-item (UI)

Este método se trata de uma variante da solução *UP*, e também está embasado no cálculo da preferência do usuário por *features*, medida através do seu interesse pelos itens. O algoritmo *UI* utiliza as matrizes de correlação ponderada entre usuários e atributos **W** e a matriz de atributos dos itens **A** no cálculo da correlação usuário-item.

A lista dos N produtos a serem recomendados decorre portanto do cálculo de ω_{ui} (Equação 5.19, Tabela 21) e da escolha dos itens que maximizem essa variável para cada usuário (Equação 5.18, Tabela 22).

$$\omega_{ui} = \sum_f w_{uf} a_{if} \quad (5.19)$$

Ao passo que o método *UP* recomenda itens a partir dos k vizinhos mais próximos, o algoritmo *UI* busca os itens com *features* mais similares aos atributos pelos quais u se interessa, diretamente através da matriz de atributos.

Espera-se que esse tipo de recomendação forneça sugestões de qualidade similar ao algoritmo original, pois os dois tem a mesma fundamentação inicial. Pode-se observar que, para o exemplo-base, ambos algoritmos forneceram a mesma recomendação para três de quatro usuários.

Tabela 18 – f_{uf}

	f_1	f_2	f_3	f_4
u_1	0	2	1	0
u_2	1	3	2	0
u_3	1	2	0	0
u_4	2	3	1	0

Tabela 19 – ω_{ui} (UP)

	i_1	i_2	i_3	i_4	i_5	i_6
u_1	2	0	3	0	0	0
u_2	3	0	5	0	6	0
u_3	0	3	0	2	0	0
u_4	0	5	0	3	6	0

Tabela 20 – \hat{i}_u (UP)

u_1	u_2	u_3	u_4
3	5	2	5

Tabela 21 – ω_{ui} (UI)

	i_1	i_2	i_3	i_4	i_5	i_6
u_1	0.25	0.85	0.55	0.25	1.15	0
u_2	0.25	0.55	0.25	0.25	0.55	0
u_3	0	0	0	0	0	0.60
u_4	0.25	0.25	0.55	0.25	0.55	0

Tabela 22 – \hat{i}_u (UI)

u_1	u_2	u_3	u_4
3	5	-	5

6 Desenvolvimento da biblioteca

6.1 Recursos acadêmicos

Diversos recursos extra-curriculares foram de fundamental importância para o sucesso deste trabalho. Foram aplicados ensinamentos práticos de quatro cursos da plataforma online de *e-learning* Coursera (<https://www.coursera.org/>), seja relacionados a teoria dos sistemas de recomendação, seja relacionados a configuração de máquinas virtuais dos servidores da Amazon Web Services.

O curso “Redes: Amigos, Dinheiro e Bytes” (Networks: Friends, Money, and Bytes – <https://www.coursera.org/course/friendsmoneybytes>), teve papel importante na introdução a temas ligados à rede mundial de computadores. Mais especificadamente, a aula 4 aborda, de maneira simples mas repleta de exemplos, a temática de sugestão de itens através da pergunta “Como o Netflix recomenda filmes?”. Essa aula ajudou-nos a compreender a teoria por trás do algoritmo de recomendação do Netflix detalhado na Referência 33.

Outro curso que influenciou diretamente o nosso Trabalho de Conclusão de Curso foi “Computação para Análise de Dados” (Computing for Data Analysis – <https://www.coursera.org/course/compdata>). As quatro semanas de aula ensinaram a leitura de dados formatados em R, o tratamento de dados, a definição de métodos estatísticos, como por exemplo de regressão linear, a aplicação de cálculos vetorizados e a construção de gráficos e tabelas.

Aliado a essas aulas, aprendemos também o paradigma funcional, amplamente utilizado em R, durante as sete semanas de “Princípios de Programação Funcional em Scala” (Functional Programming Principles in Scala – <https://www.coursera.org/course/progfun>).

Por fim, o curso de doze semanas de duração “Engenharia de Startup” (Startup Engineering – <https://www.coursera.org/course/startup>) nos ensinou a trabalhar com diversas ferramentas de software necessárias para a realização dos testes de desempenho dos algoritmos. Utilizamos máquinas virtuais, linha de comando Unix, versionamento de código em git e editores de texto sem interface gráfica (tais como vi e Emacs). Além disso, o *setup* de máquinas virtuais na Amazon Web Services também era abordada no curso, facilitando a configuração do ambiente de testes e a automatização desse processo.

6.2 Ferramentas utilizadas

A programação da biblioteca computacional se deu por meio do ambiente de desenvolvimento integrado RStudio versão 0.98.953 ([<http://www.rstudio.com/>](http://www.rstudio.com/)). Esse IDE inclui um console, um editor de texto e um corretor de sintaxe que suporta a execução de código direta, bem como ferramentas para traçar gráficos, histórico de comandos, depuração de erros e gerenciamento de espaço de trabalho. Além disso, o RStudio está disponível via licença de código aberto GPLv3 (Afferro General Public License version 3) para os principais sistemas operacionais (Windows, Mac e Linux).

6.3 Métodos computacionais

6.3.1 Estrutura da biblioteca

A biblioteca está estruturada em quatro principais

6.3.2 Algoritmo baseado na ponderação de atributos (FW)

6.3.3 Algoritmo baseado no perfil de usuários (UP)

6.3.4 Algoritmo baseado na correlação usuário-item (UI)

6.4 Ambiente de testes

Inicialmente, realizamos os testes de qualidade nos nossos próprios computadores pessoais. Todavia, a execução de testes sucessivos exigia muita capacidade computacional, principalmente quanto a memória virtual.

A alocação de objetos e matrizes na memória RAM é muito custosa, principalmente na etapa de determinação de medidas de distância d_{ij}^f para o algoritmo FW (Equação 5.1). Uma matriz de dimensão $|\mathcal{I}| \times |\mathcal{I}| \times |\mathcal{F}|$ possui $1682 \times 1682 \times 25$ elementos (cerca de 71 milhões), e ocupa aproximadamente 500 MB de memória. No total, 4 GB de memória RAM são utilizadas durante todos os testes, fazendo-se necessário o uso máquinas dedicadas.

Por essa razão, realizamos todas as etapas de recomendação e avaliação de qualidade em máquinas *memory-optimized* nos servidores da Amazon Web Services ([<http://aws.amazon.com/>](http://aws.amazon.com/)). Visto que o serviço é cobrado por hora-máquina, desenvolvemos um *script* de inicialização para instalar todos os pacotes de programação e execução imediata dos testes, permitindo assim reduzir os custos da análise.

As etapas de configuração do ambiente de testes envolvem o cadastro na Amazon Web Services, a criação de uma máquina virtual, a instalação das ferramentas de programação e o descarregamento do código de testes.

Após o cadastro na AWS, deve-se seguir os seguintes passos para a criação de uma máquina virtual:

1. Login na plataforma (Figura 4);
2. Acesso ao serviço de máquinas virtuais Elastic Compute Cloud ou EC2. Para criar uma nova máquina, basta clicar em *Launch Instance* (Figura 5);
3. Escolha da configuração do software da máquina. É possível escolher entre diversos sistemas operacionais, versões e distribuições. Para avançar, deve-se clicar em *Select*. No nosso caso, escolhemos a configuração *Amazon Linux AMI 2014.09.1 (HVM)*, em virtude da facilidade de se instalar pacotes adicionais (Figura 6);
4. Escolha do tipo de máquina. No nosso caso, escolhemos uma máquina otimizada para memória RAM. Em seguida, avançamos em *Next: Configure Instance Details*. Nos *Steps 3, 4 e 5* do serviço, não modificamos nenhuma opção pré-configurada (Figura 7);
5. Criação da chave privada a ser utilizada para conexão com a máquina. Depois da criação de uma nova chave, pode-se descarregá-la através de *Download Key Pair* e finalmente iniciar a máquina em *Launch Instances* (Figura 8);
6. Inicialização da máquina e do DNS público. Esse é o endereço de conexão com a máquina, e pode ser obtido na seção *Description → Public DNS* (Figura 9).

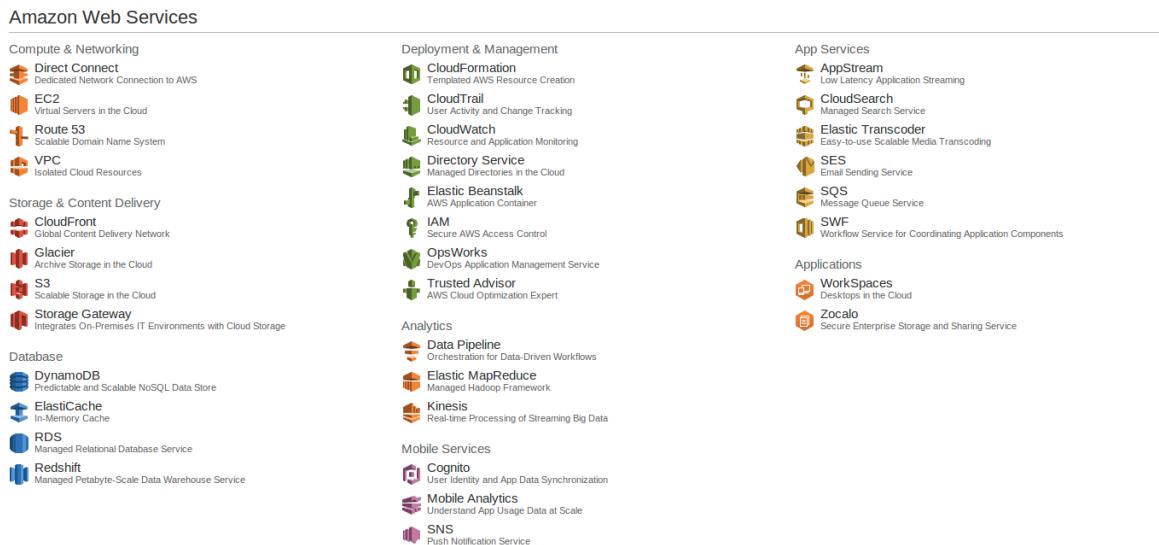


Figura 4 – Serviços da Amazon Web Services

The screenshot shows the AWS EC2 Dashboard. On the left, a sidebar lists various services: EC2 Dashboard, Events, Tags, Reports, Instances (with sub-options Instances, Spot Requests, Reserved Instances), AMIs (with sub-options AMIs, Bundle Tasks), Elastic Block Store (with sub-options Volumes, Snapshots), Network & Security (with sub-options Security Groups, Elastic IPs, Placement Groups, Load Balancers, Key Pairs, Network Interfaces), and Auto Scaling (with sub-options Launch Configurations, Auto Scaling Groups). The main content area is titled "Resources" and shows the following statistics for the US East (N. Virginia) region:

- 0 Running Instances
- 0 Volumes
- 1 Key Pair
- 0 Placement Groups
- 0 Elastic IPs
- 0 Snapshots
- 0 Load Balancers
- 5 Security Groups

A note at the bottom says "Easily deploy Ruby, PHP, Java, .NET, Python, Node.js & Docker applications with [Elastic Beanstalk](#)." A "Hide" link is also present.

The "Create Instance" section follows, with a "Launch Instance" button and a note about launching instances in the US East (N. Virginia) region.

The "Service Health" section shows the status of various services across different availability zones:

- US East (N. Virginia): This service is operating normally.
- us-east-1a: Availability zone is operating normally.
- us-east-1b: Availability zone is operating normally.
- us-east-1c: Availability zone is operating normally.
- us-east-1e: Availability zone is operating normally.

A "Service Health Dashboard" link is provided.

Figura 5 – Serviços de Elastic Compute Cloud (EC2)

This screenshot shows the "Step 1: Choose an Amazon Machine Image (AMI)" page. It lists several AMI options:

- Amazon Linux AMI 2014.09.1 (HVM) - ami-b66ed3de**: Free tier eligible. Root device type: ebs. Virtualization type: hvm. Description: The Amazon Linux AMI is an EBS backed image. It includes the 3.14 kernel, Ruby 2.1, PHP 5.5, PostgreSQL 9.3, Docker 1.2, the AWS command line tools, and repository access to many other packages. A "Select" button and "64-bit" link are available.
- Red Hat Enterprise Linux 7.0 (HVM), SSD Volume Type - ami-a8d369c0**: Free tier eligible. Root device type: ebs. Virtualization type: hvm. Description: Red Hat Enterprise Linux version 7.0 (HVM), EBS General Purpose (SSD) Volume Type. A "Select" button and "64-bit" link are available.
- SuSE Linux Enterprise Server 12 (HVM), SSD Volume Type - ami-aeb532c6**: Free tier eligible. Root device type: ebs. Virtualization type: hvm. Description: SuSE Linux Enterprise Server 12 (HVM), EBS General Purpose (SSD) Volume Type. Public Cloud, Advanced Systems Management, Web and Scripting, and Legacy modules enabled. A "Select" button and "64-bit" link are available.
- Ubuntu Server 14.04 LTS (HVM), SSD Volume Type - ami-9ea11cf6**: Free tier eligible. Root device type: ebs. Virtualization type: hvm. Description: Ubuntu Server 14.04 LTS (HVM), EBS General Purpose (SSD) Volume Type. Support available from Canonical (<http://www.ubuntu.com/cloud/services>). A "Select" button and "64-bit" link are available.
- Microsoft Windows Server 2012 R2 Base - ami-ba13ab2**: Free tier eligible. Root device type: ebs. Virtualization type: hvm. Description: Microsoft Windows 2012 R2 Standard edition with 64-bit architecture. [English]. A "Select" button and "64-bit" link are available.
- Microsoft Windows Server 2012 R2 with SQL Server Web - ami-9c0bb3f4**: Windows. Root device type: ebs. Virtualization type: hvm. Description: Microsoft Windows Server 2012 R2 Standard edition, 64-bit architecture, Microsoft SQL Server 2014 Web edition. [English]. A "Select" button and "64-bit" link are available.
- Microsoft Windows Server 2012 R2 with SQL Server Standard - ami-a416aecc**: Windows. Root device type: ebs. Virtualization type: hvm. Description: Microsoft Windows Server 2012 R2 Standard edition, 64-bit architecture, Microsoft SQL Server 2014 Standard edition. [English]. A "Select" button and "64-bit" link are available.
- Microsoft Windows Server 2012 Base - ami-3214ac5a**: Windows. Root device type: ebs. Virtualization type: hvm. Description: Microsoft Windows Server 2012 Standard edition with 64-bit architecture. [English]. A "Select" button and "64-bit" link are available.

A "Cancel and Exit" link is at the top right, and a "1 to 22 of 22 AMIs" link is at the bottom right.

Figura 6 – Escolha do tipo de configuração do software da máquina

Step 2: Choose an Instance Type							
	Family	Type	vCPUs	Memory (GB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance
Currently selected: r3.large (6.5 ECUs, 2 vCPUs, 2.5 GHz, Intel Xeon E5-2670v2, 15 GiB memory, 1 x 32 GiB Storage Capacity)							
<input type="checkbox"/>	General purpose	t2.micro <small>Free tier eligible</small>	1	1	EBS only	-	Low to Moderate
<input type="checkbox"/>	General purpose	t2.small	1	2	EBS only	-	Low to Moderate
<input type="checkbox"/>	General purpose	t2.medium	2	4	EBS only	-	Low to Moderate
<input type="checkbox"/>	General purpose	m3.medium	1	3.75	1 x 4 (SSD)	-	Moderate
<input type="checkbox"/>	General purpose	m3.large	2	7.5	1 x 32 (SSD)	-	Moderate
<input type="checkbox"/>	General purpose	m3.xlarge	4	15	2 x 40 (SSD)	Yes	High
<input type="checkbox"/>	General purpose	m3.2xlarge	8	30	2 x 80 (SSD)	Yes	High
<input type="checkbox"/>	Compute optimized	c3.large	2	3.75	2 x 16 (SSD)	-	Moderate
<input type="checkbox"/>	Compute optimized	c3.xlarge	4	7.5	2 x 40 (SSD)	Yes	Moderate
<input type="checkbox"/>	Compute optimized	c3.2xlarge	8	15	2 x 80 (SSD)	Yes	High
<input type="checkbox"/>	Compute optimized	c3.4xlarge	16	30	2 x 160 (SSD)	Yes	High
<input type="checkbox"/>	Compute optimized	c3.8xlarge	32	60	2 x 320 (SSD)	-	10 Gigabit
<input type="checkbox"/>	GPU instances	g2.2xlarge	8	15	1 x 60 (SSD)	Yes	High
<input checked="" type="checkbox"/>	Memory optimized	r3.large	2	15	1 x 32 (SSD)	-	Moderate
<input type="checkbox"/>	Memory optimized	r3.xlarge	4	30.5	1 x 80 (SSD)	Yes	Moderate

[Cancel](#) [Previous](#) [Review and Launch](#) [Next: Configure Instance Details](#)

Figura 7 – Escolha do tipo máquina

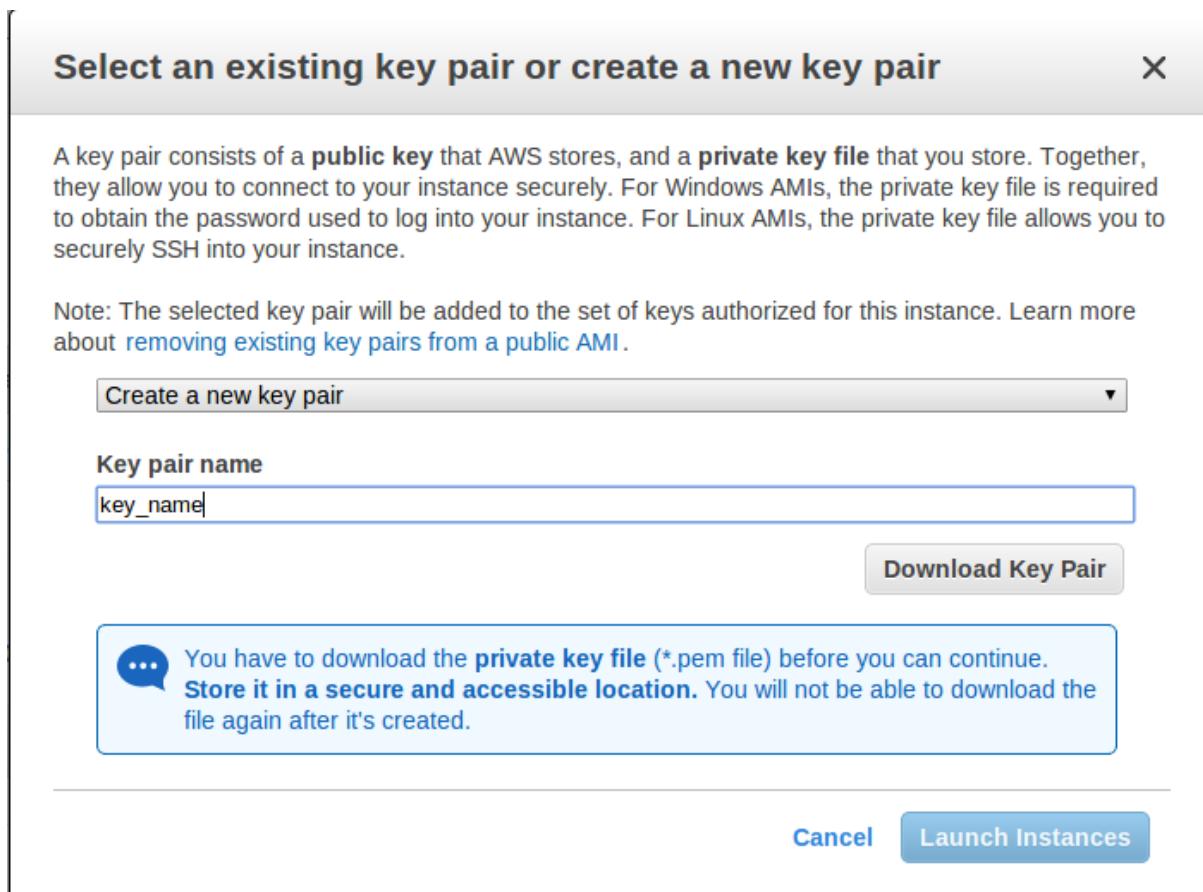


Figura 8 – Criação da chave privada

The screenshot shows two windows from the AWS Management Console. The top window is a list of instances, showing one instance named 'i-129954f3' which is 'running'. The bottom window is a detailed view of the same instance, showing its configuration. Key details include:

- Instance ID:** i-129954f3
- Public DNS:** ec2-54-88-186-106.compute-1.amazonaws.com
- Public IP:** 54.88.186.106
- Key Name:** vaio
- Launch Time:** November 5, 2014 5:38:29 AM
- AMI ID:** ami-b66ed0de
- Platform:** -
- IAM role:** -
- Owner:** 010719389704
- Launch time:** November 5, 2014 5:38:29 AM UTC-2 (less than one hour)

Figura 9 – Criação da máquina e do DNS público (*Public DNS*)

Uma vez criada a máquina, deve-se utilizar a chave privada para realizar um *secure shell* e conectar-se remotamente ao EC2. A partir de um computador pessoal dotado de um interpretador de comandos `bash`, utilizamos a seguinte instrução:

```
ssh -i ~/Downloads/key_name.pem
ec2-user@ec2-54-88-186-106.compute-1.amazonaws.com
```

Em seguida, para a automatização do ambiente de testes e rápida configuração de novas máquinas, criamos um arquivo `script.sh` na linguagem de programação `bash`. Esse *script* instala os pacotes `R` e `git` e cria a chave pública necessária para acessar o servidor onde o código da biblioteca está hospedado. Em virtude de sua popularidade, utilizamos o serviço de hospedagem de códigos abertos GitHub (<<https://github.com/aviggiano/tcc>>).

```
#!/usr/bin/env bash
sudo su          # login como administrador. necessario para instalar
                  # pacotes no sistema operacional
yes | yum install R  # instala o pacote R no linux
yes | yum install git # instala o git para descarregar os metodos de
                      # recomendacao
ssh-keygen -t rsa    # gera a chave para conectar-se ao GitHub
cat ~/.ssh/id_rsa.pub # imprime a chave publica, que deve ser adicionada nas
                      # configuracoes do GitHub
```

A saída do *script* é uma chave no seguinte formato:

```
ssh-rsa AAAAB3NzaC1yc2EAAAQABAAQCB/bxcszoyHzyqtTXp4f1lQ3OT58Lsb7QLx+7nQ6
y0OIoWhK+r5ynSVi0BpTC+2hMrlg1rZTC1ED7Nb+SI9bRvf+1UYW0iVUXtwAVColMNBDlfE7QCWbJm
TmmBLcv9PIoCAvCfrxBh+f1W3hXG388/LIEjZJckPYogho0jPAnFv3IXAGtVniV6cBcTTfKPUnX+np
6xiqnf4tYQpmPW/mnxk9s3bbEmcE1eYJkrE2IWdzy6EBnR9D4cBW5D8/VMM54xMJzugWZ0//sIjLLT
OoFTTrroiwr+OX2DxqFdgCy8Agx1WZTeGhBAW1nvIVr5WVcWVBSzBCZfg8mYe+zYnbwl ec2-user@
ip-10-168-40-38
```

Após a obtenção da chave, deve-se cadastrá-la na página correspondente do GitHub (<https://github.com/settings/ssh>), como mostra a Figura 10.

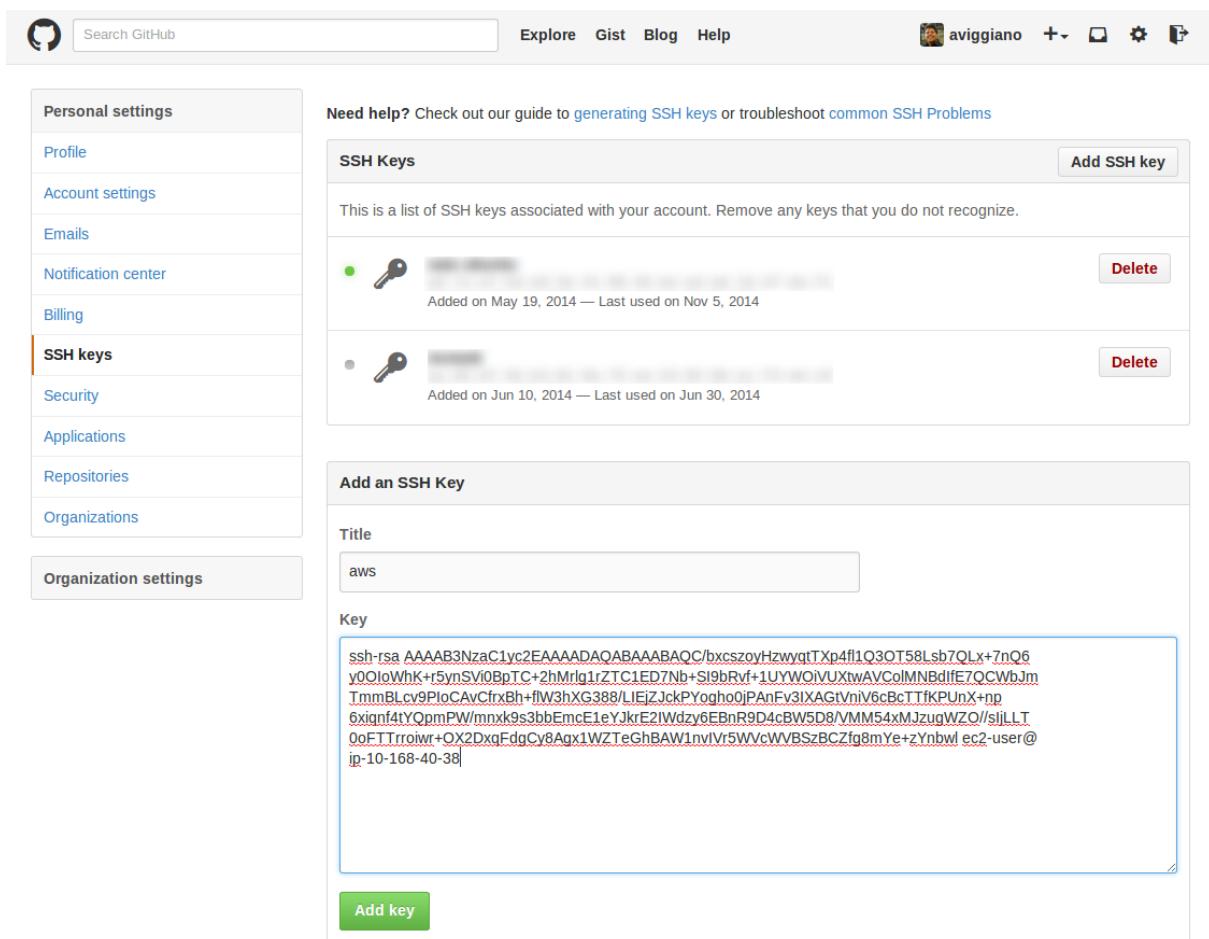


Figura 10 – Cadastro da chave pública no GitHub

Uma vez tendo habilitado a máquina virtual da AWS para manipulação do repositório da biblioteca, pode-se descarregar o código e executar o *script* de testes de qualidade:

```
#!/usr/bin/env bash
git clone git@github.com:aviggiano/tcc # clona o repositorio
cd tcc && Rscript recsys/run_tests.R # executa o script de testes
```

7 Resultados

Os resultados deste trabalho são a análise de desempenho dos algoritmos propostos, em termos de precisão, abrangência e tempo computacional, mediante a mudanças em suas variáveis de importância (Tabela 23).

Além disso, as metodologias de solução de cada um dos sistemas serão debatidas, de modo a explorar casos de uso particulares e a propor melhorias nos métodos computacionais. Serão respondidas perguntas como “O que acontece com itens ou usuários sem nenhuma avaliação?” e “Qual o desempenho dos métodos para outros bancos de dados?”.

Tabela 23 – Parâmetros de influência no desempenho dos algoritmos de recomendação

Variável	Descrição	Valor padrão
N	Tamanho da lista de recomendação	20
T	Percentual da base de aprendizado na validação cruzada	75%
H	Percentual de avaliações “escondidas” dos usuários-teste na validação cruzada	75%
M	Valor mínimo para avaliações positivas	2
k	Número de vizinhos mais próximos	10
\mathcal{F}	Conjunto de atributos dos itens	Todos atributos
d^f	Medida de distância entre atributos	$\ \cdot\ ^f$
w_f	Pesos dos atributos	$w_f > 0$

7.1 Tamanho da lista de recomendações N

Assim como mostra a literatura, a medida que o tamanho da lista de recomendações aumenta, a precisão cai e a abrangência cresce (Figuras 11 e 12). A primeira decresce com N porque a quantidade de itens sugeridos se torna excessivamente maior que a quantidade de itens positivamente avaliados pelos usuários-teste. A segunda, por sua vez, cresce com N porque a probabilidade de sugerirmos itens relevantes para o usuário aumenta quando sugerimos mais itens. Para $N = |\mathcal{I}|$, a abrangência atinge 100%, pois todos os itens teriam sido recomendados.

O método UP supera os dois outros algoritmos para todos os valores de N , tanto em precisão quanto em abrangência, como se observa pelo gráfico das medidas F_1 .

Contrariamente ao esperado, a qualidade de recomendação do algoritmo UI é sensivelmente inferior à do algoritmo UP. Isso se deve ao fato de a correlação usuário-item daquele método colocar ênfase no valor do atributo a_{if} , mesmo que esses atributos não

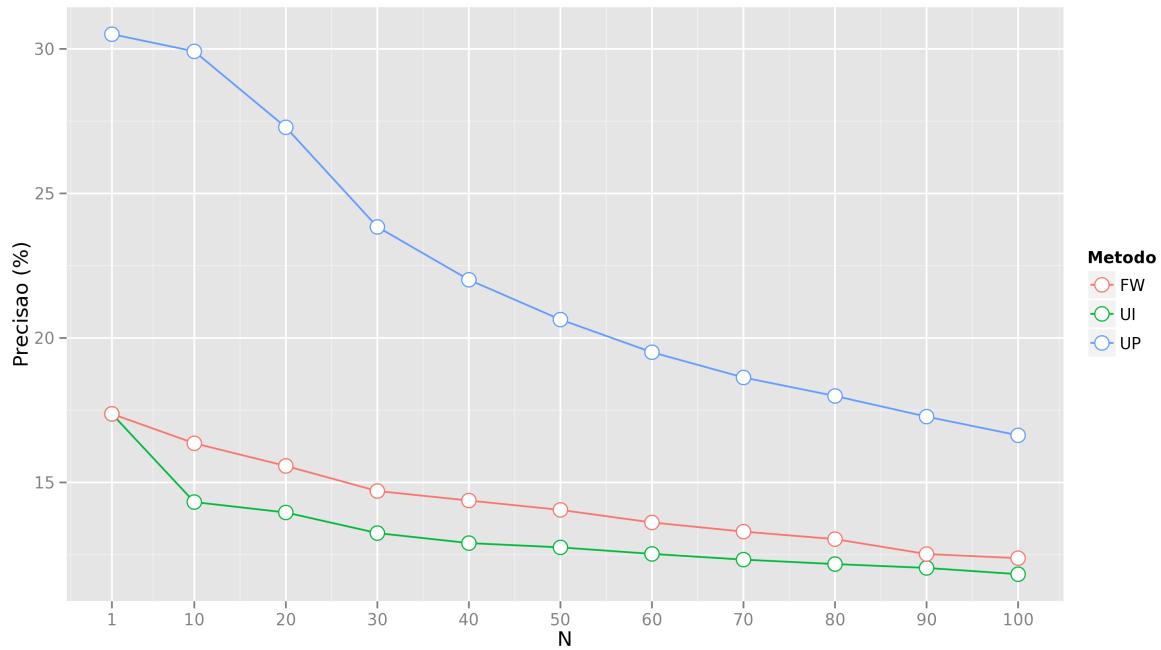


Figura 11 – Precisão em função do tamanho da lista de recomendações N

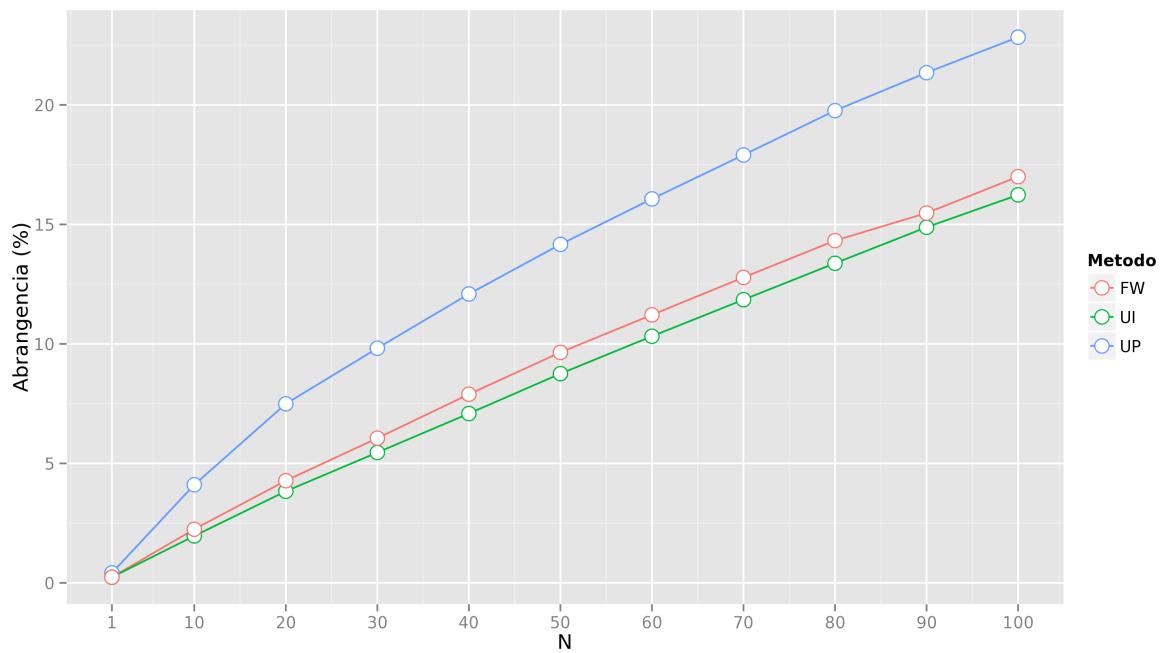
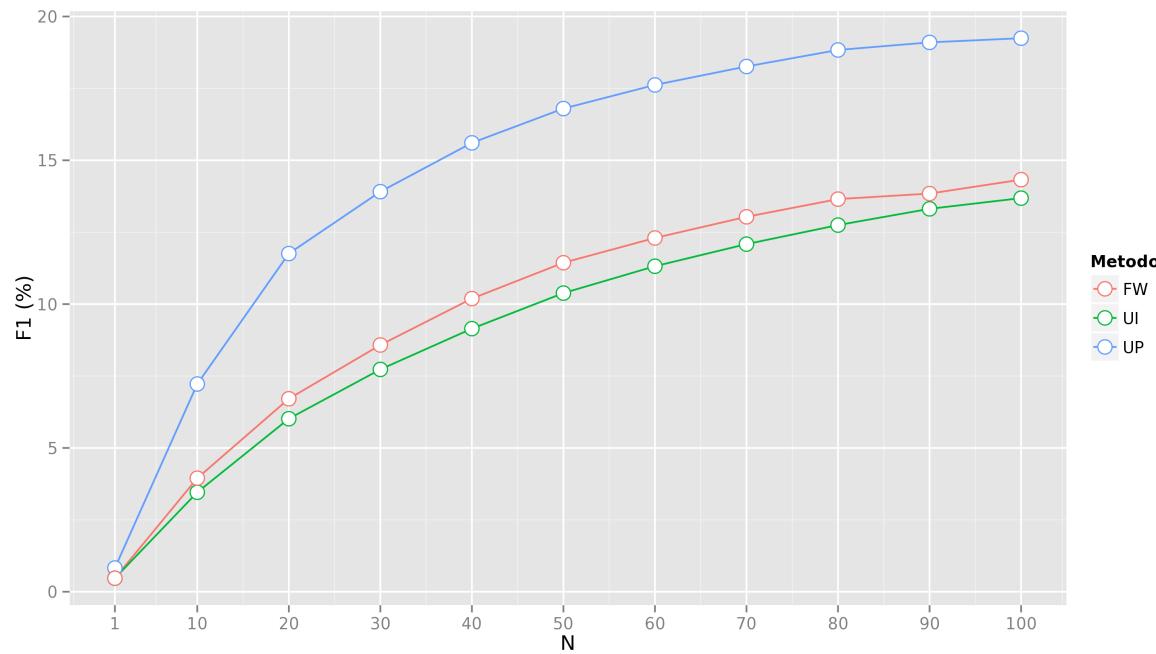
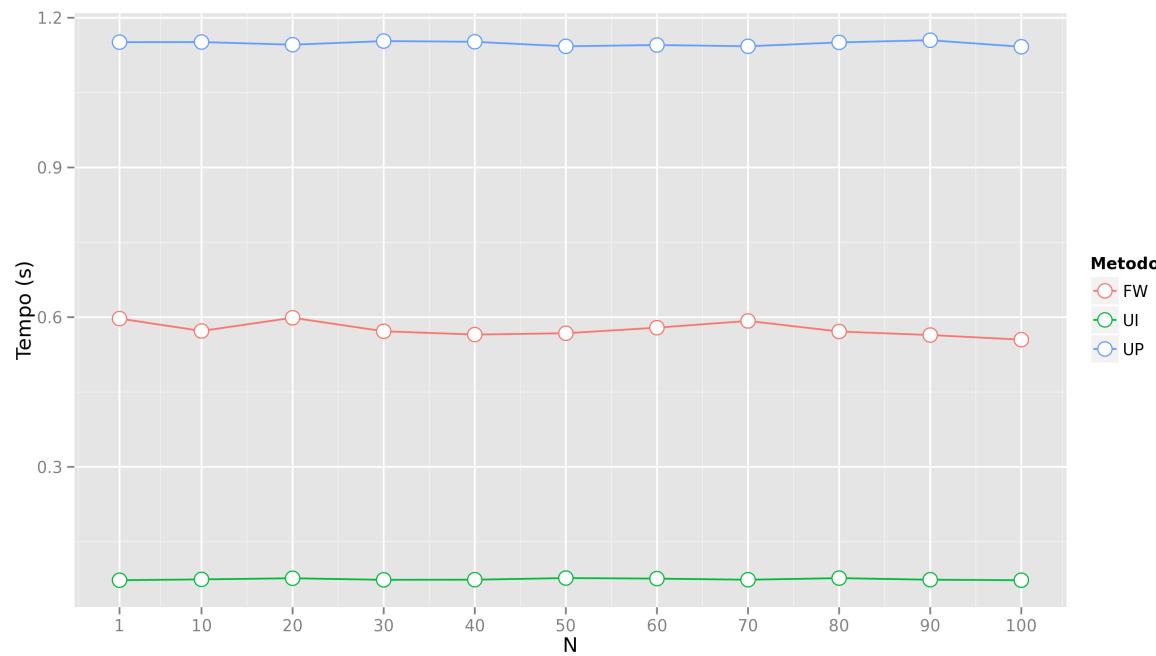


Figura 12 – Abrangência em função do tamanho da lista de recomendações N

Figura 13 – Medida F_1 em função do tamanho da lista de recomendações N Figura 14 – Tempo de execução em função do tamanho da lista de recomendações N

sejam diretamente proporcionais à preferência do usuário. Esse cálculo é incoerente, por exemplo, para atributos $f = \text{data}$: mesmo que o usuário tenha um elevado interesse w_{uf} por filmes antigos, o valor de a_{if} não leva em conta se sua preferência é por filmes da década de 1970 ou 1990. Nesse caso, o algoritmo indicaria incorretamente que filmes mais recentes são mais adequados para aquele usuário, porque possuem maior a_{if} .

A fim de corrigir essa falha no algoritmo UI, seria necessário, por exemplo, aplicar nos atributos a_{if} uma função g_f que crescesse no mesmo sentido do interesse do usuário por aquela *feature*. Dessa forma, o cálculo $\sum_f w_{uf} g_f(a_{if})$ significaria de fato a similaridade entre o usuário u e o item i medida através de seu interesse $g(a_{if})$ pelas *features* f .

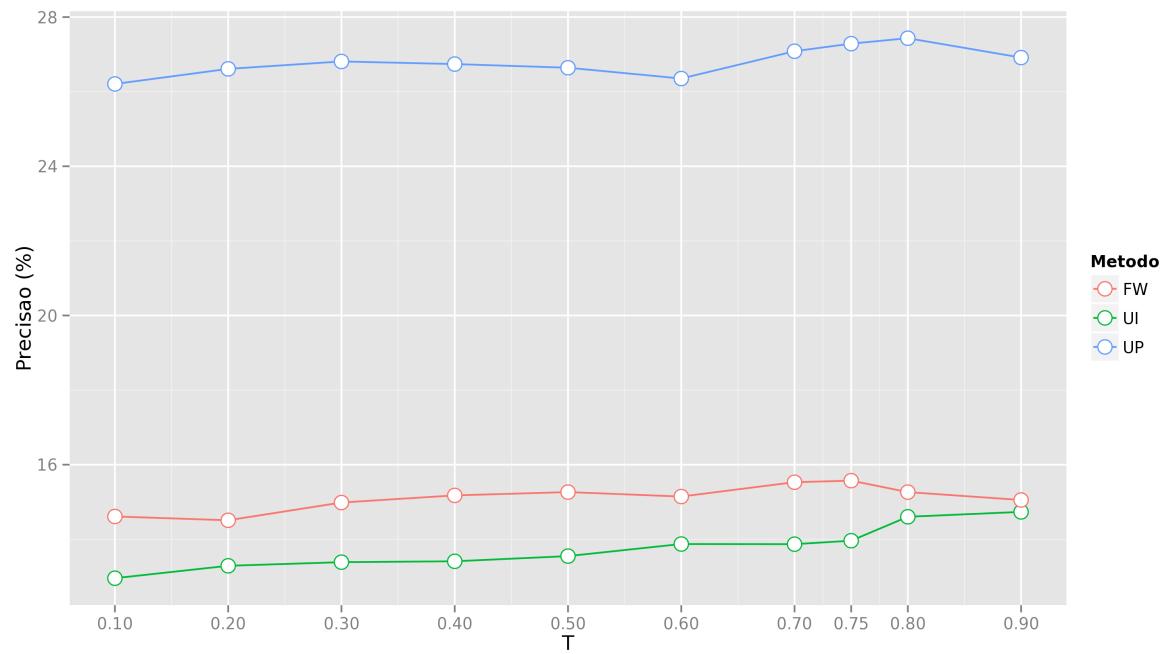
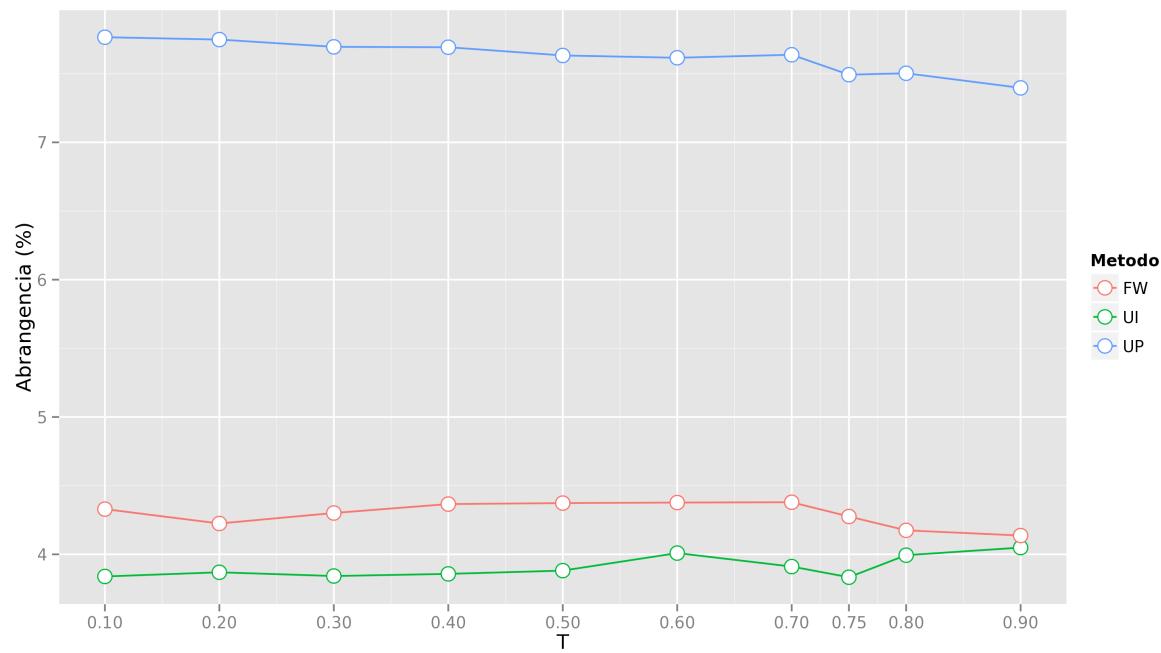
Apesar de alta qualidade das recomendações do método UP, este possui também a maior complexidade computacional. Seu tempo de execução é 2 vezes maior que o do método FW e 4 vezes maior que o do método UI. Todavia, nenhum desses tempos de execução é crítico, tendo em vista que o sistema não seria colocado diretamente à disposição dos clientes, mas que as recomendações seriam enviadas via email, por exemplo.

Apenas o método UI atende ao requisito de *throughput* mínimo de 28 recomendações para cada usuário por segundo. Dado que a base de testes possui 25% do total de usuários, correspondente a 236 clientes para o banco 100k, o tempo de execução máximo dos métodos deveria ser de 0.15 min. A fim de melhorar a velocidade das recomendações, a solução mais eficiente é a mudança da linguagem de programação. O uso de linguagens C, C++ ou Python pode melhorar o desempenho computacional em até 500 vezes (34).

7.2 Percentual da base de aprendizado T

A medida que o percentual da base de aprendizados aumenta, a precisão de todos os métodos cresce ligeiramente. Isso é consequência do caráter colaborativo dos algoritmos, já que a qualidade da recomendação depende da quantidade total de dados. Entretanto, pode-se observar que a abrangência e a medida F_1 são praticamente constantes para valores crescentes de T de modo que esse parâmetro não tem grande relevância para o sucesso do sistema de recomendação.

O parâmetro T não exerce nenhuma influência sobre o tempo de execução dos métodos UP e UI, mas apenas sobre o método FW. Isso ocorre porque a etapa de maior custo computacional (Equação 5.9) é linearmente dependente da quantidade de usuários $|\mathcal{U}|$. Quanto menos usuários-teste, mais veloz é o algoritmo.

Figura 15 – Precisão em função do percentual da base de aprendizado T Figura 16 – Abrangência em função do percentual da base de aprendizado T

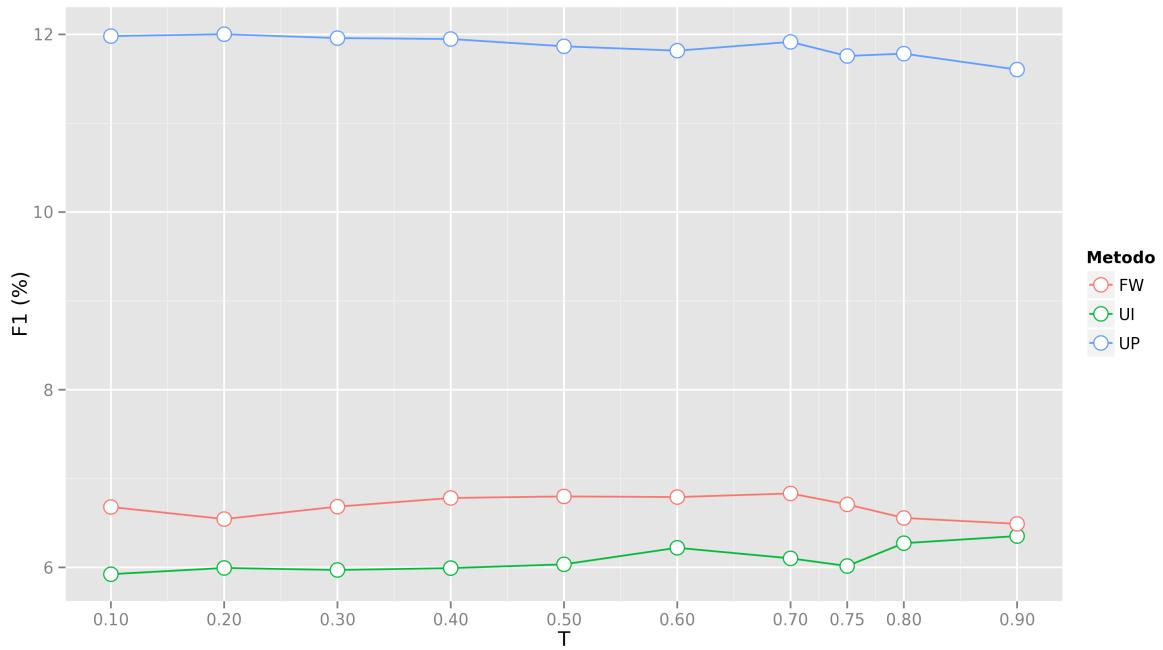


Figura 17 – Medida F_1 em função do percentual da base de aprendizado T

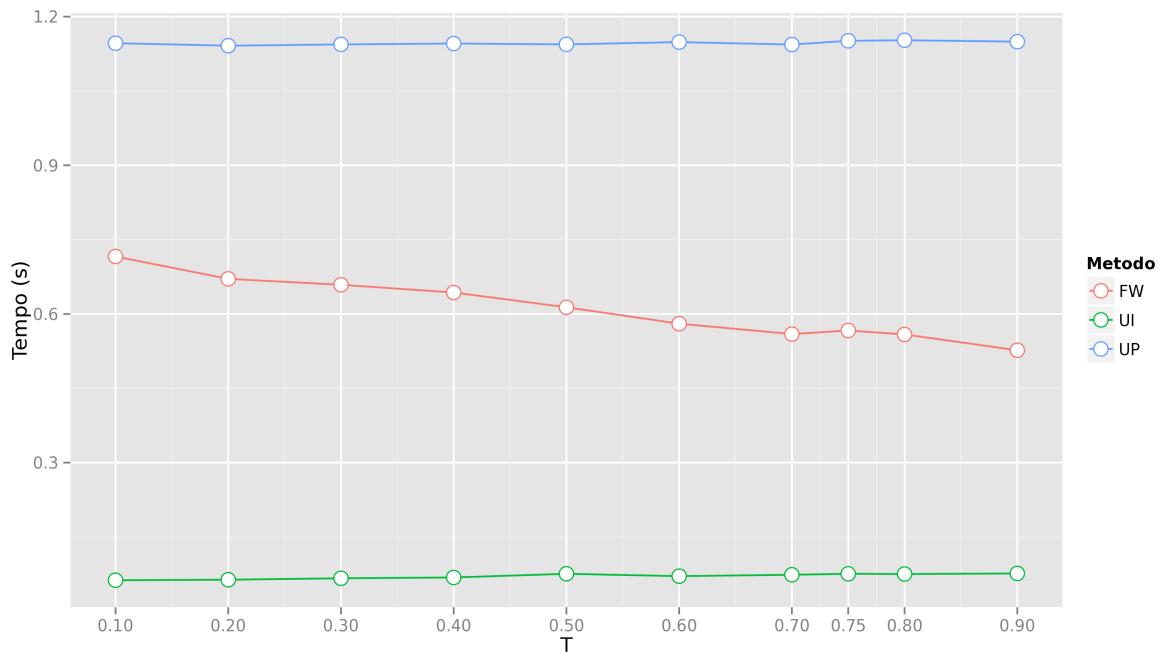


Figura 18 – Tempo de execução em função do percentual da base de aprendizado T

7.3 Percentual de avaliações “escondidas” dos usuários-teste na validação cruzada H

Quanto maior o número de avaliações “escondidas”, mais fácil é acertar os itens dos usuários-teste, pois a lista de recomendação é pequena em relação ao total de itens positivamente avaliados pelo usuário. Por esse motivo, a precisão cresce com H para todos os métodos.

Para o algoritmo FW, a precisão atinge seu máximo em $H = 75\%$ e depois decresce ligeiramente (Figura 19). Isso ocorre porque o cálculo dos pesos w_f depende da quantidade de avaliações r_{ui} . Existe, pois, um compromisso (*tradeoff*) entre facilidade de se acertar itens avaliados quando há muitas avaliações escondidas e a dificuldade de se estimar w_f quando não há muitos dados de avaliações.

Ao passo que a precisão dos métodos aumenta com H , a abrangência diminui. Visto que a quantidade de itens da lista *top-N* é fixa, quanto maior o número de itens “escondidos”, mais difícil é de se retornar todos os itens relevantes.

O resultado de uma precisão crescente em função de H e uma acurácia decrescente é que a medida F_1 possui um ponto de máximo. Para todos os métodos, o valor máximo é tal que $H = 50\%$.

Quanto ao tempo de execução, a influência é a mesma do parâmetro T : para o método FW, a etapa de maior custo computacional (Equação 5.9) é linearmente dependente da quantidade de itens $|\mathcal{I}|$. Quanto menos avaliações de itens dos usuários-teste, mais veloz é o algoritmo.

7.4 Valor mínimo para avaliações positivas M

Contrariamente ao que esperávamos, tornar o algoritmo mais “seletivo” não melhora sua precisão. Apesar de o valor mínimo M estar intimamente ligado com a noção de “avaliação positiva” e de entrar no cálculo de parâmetros importantes dos métodos (Equações 5.7 e 5.11), esse parâmetro pouco influencia a precisão para $0 \leq M \leq 2$.

Esse resultado pode ser explicado pelo fato de que a maioria das avaliações são positivas (Figura 24), e portanto b_M tem quase o mesmo efeito de b_0 . Isso não ocorre somente pelo fato de os clientes comprarem itens similares a seus gostos, e portanto de raramente se decepcionarem, mas também pelo fato de que os usuários tem menos disposição para darem avaliações negativas. Esse fenômeno se chama *hidden feedback*, e se caracteriza pelo fato de que os itens avaliados não são escolhidos ao acaso, mas sim por despertarem aspectos de interesse das preferências do usuário, indo além dos valores numéricos das avaliações (33).

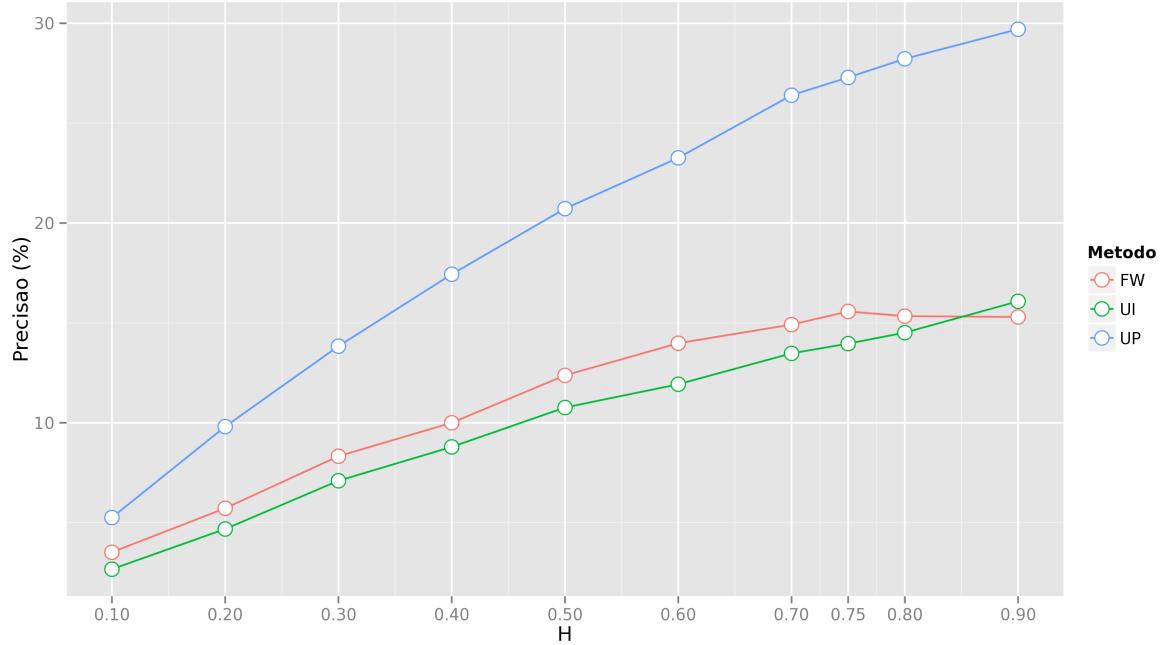


Figura 19 – Precisão em função do percentual de avaliações “escondidas” H

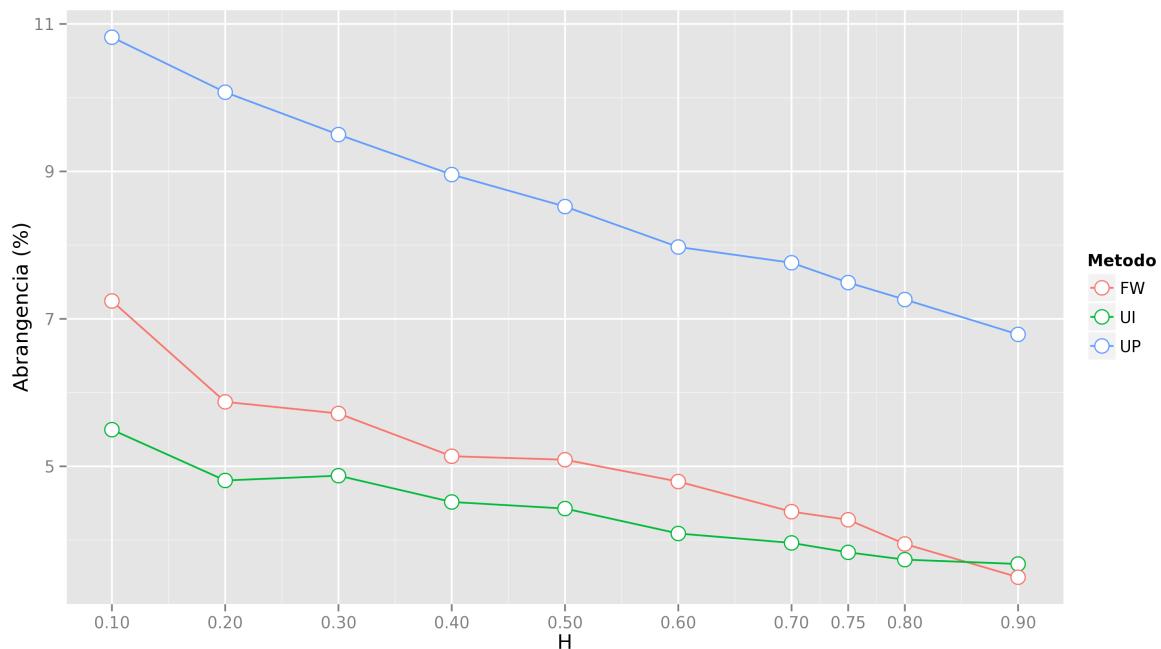
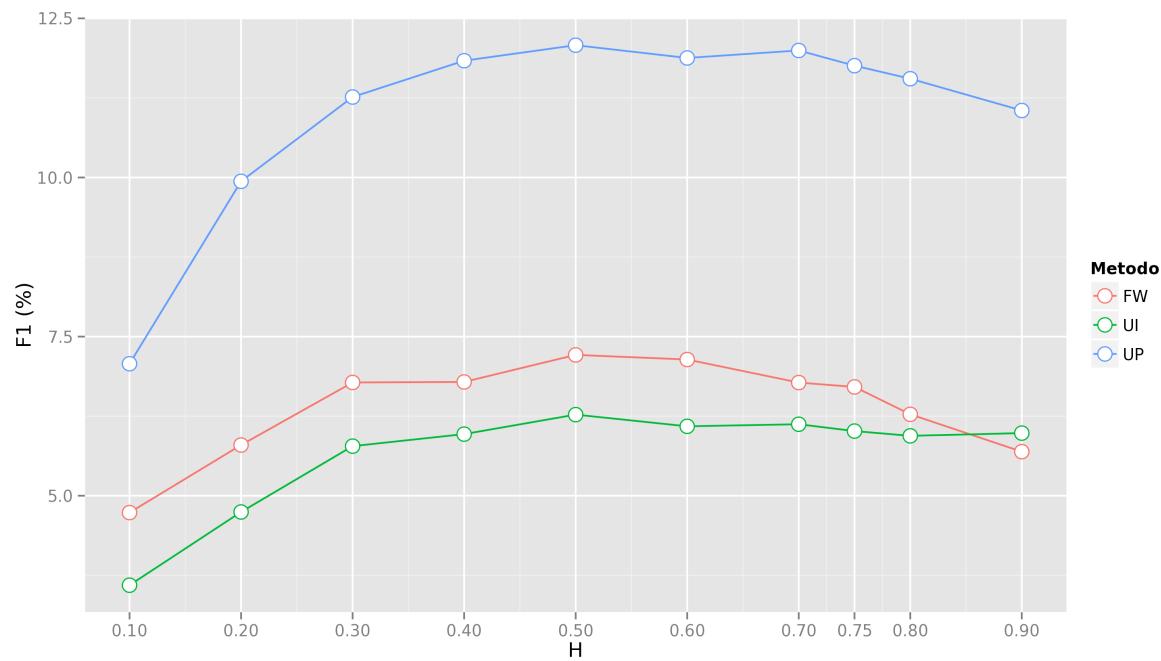
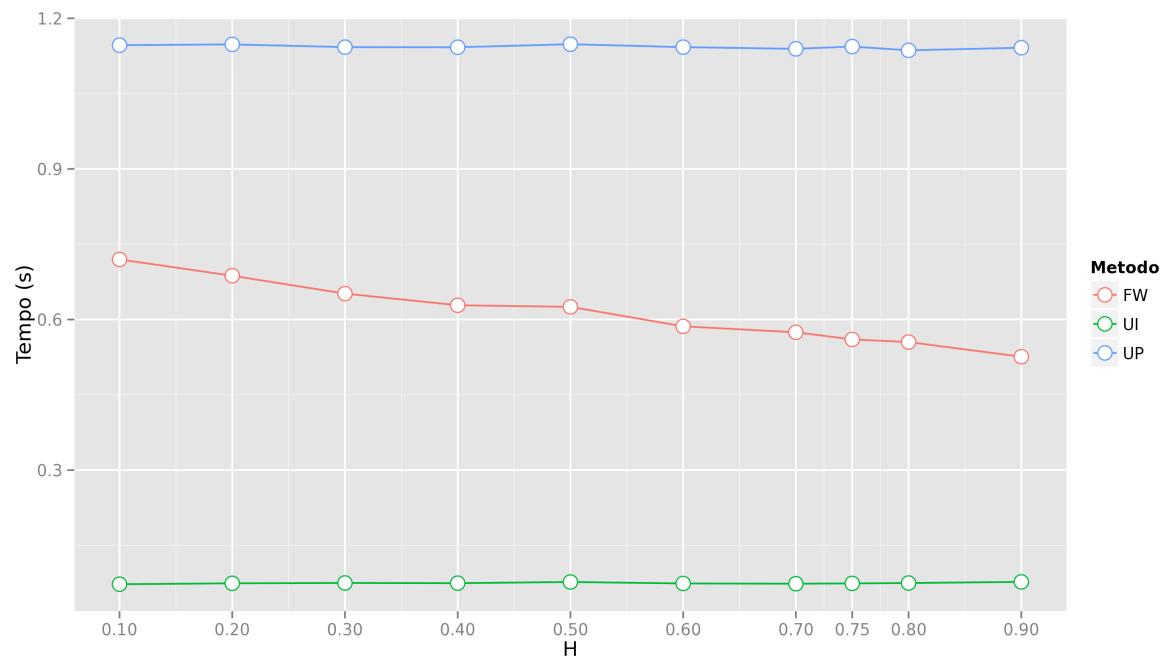


Figura 20 – Abrangência em função do percentual de avaliações “escondidas” H

Figura 21 – Medida F_1 em função do percentual de avaliações “escondidas” H Figura 22 – Tempo de execução em função do percentual de avaliações “escondidas” H

Ao se analisar a abrangência dos métodos, a seletividade influencia na recomendação. Quanto maior M , menor é a quantidade de itens muito bem avaliados. Estes possuem elevada similaridade-correlação e são facilmente escolhidos pelos algoritmos. Por esse motivo, melhor é o desempenho do sistema.

A complexidade computacional dos algoritmos também depende de M , já que mais ou menos parâmetros são analisados no cálculo da TF-IDF (métodos UI e UP) e dos pesos dos atributos (método FW).

Um detalhe a se observar é que a precisão é nula e a abrangência é inexistente para $M = 5$, já que todas as avaliações r_{ui} pertencem ao conjunto $\{1, 2, 3, 4, 5\}$. Para o algoritmo UI, tanto a precisão quanto a abrangência são nulas para $M = 4$.

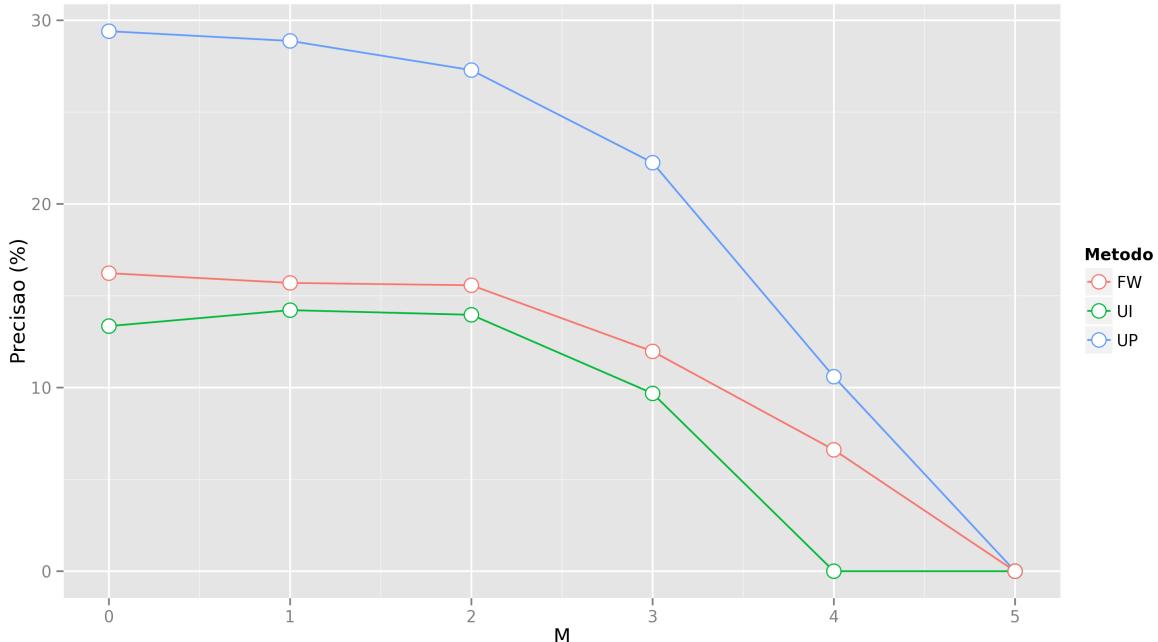
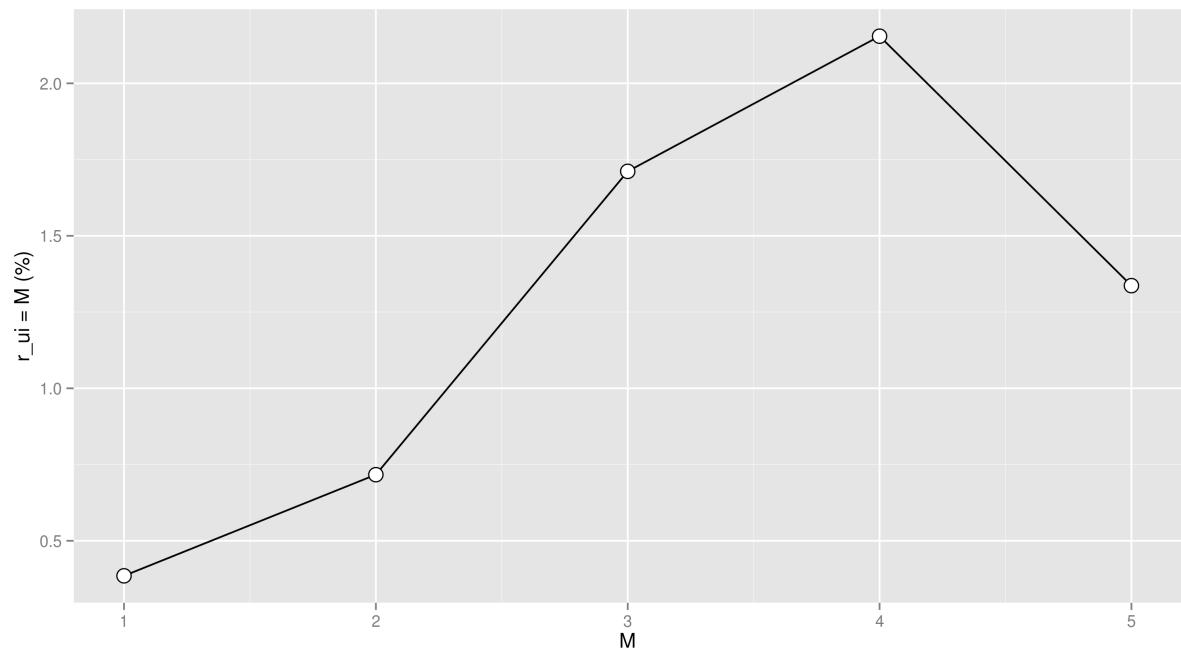
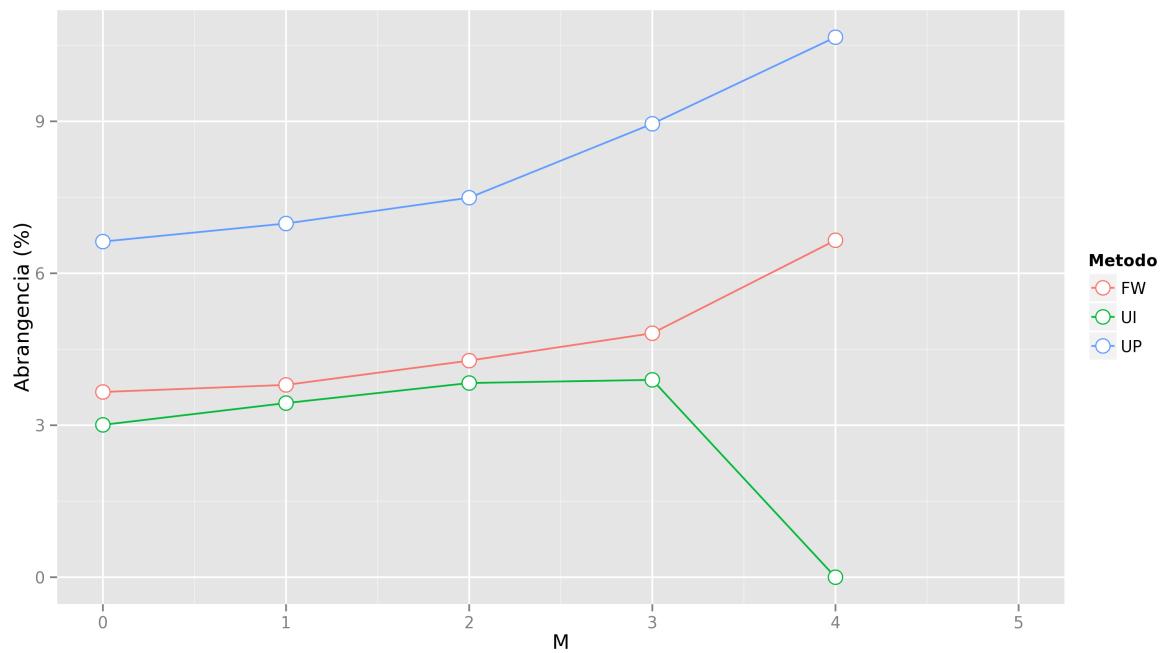


Figura 23 – Precisão em função do valor mínimo para avaliações positivas M

7.5 Número de vizinhos mais próximos k

O único método que recomenda itens com base nos vizinhos mais próximos é o UP. Percebe-se que com o aumento de k , a precisão e a abrangência caem, pois a vizinhança se torna excessivamente grande e repleta de usuários sem muita similaridade com o usuário-teste. Pode-se observar que o valor máximo de precisão e acurácia ocorre para $k = 20$ (Figura 30).

Figura 24 – Percentual de avaliações por valor de M Figura 25 – Abrangência em função do valor mínimo para avaliações positivas M

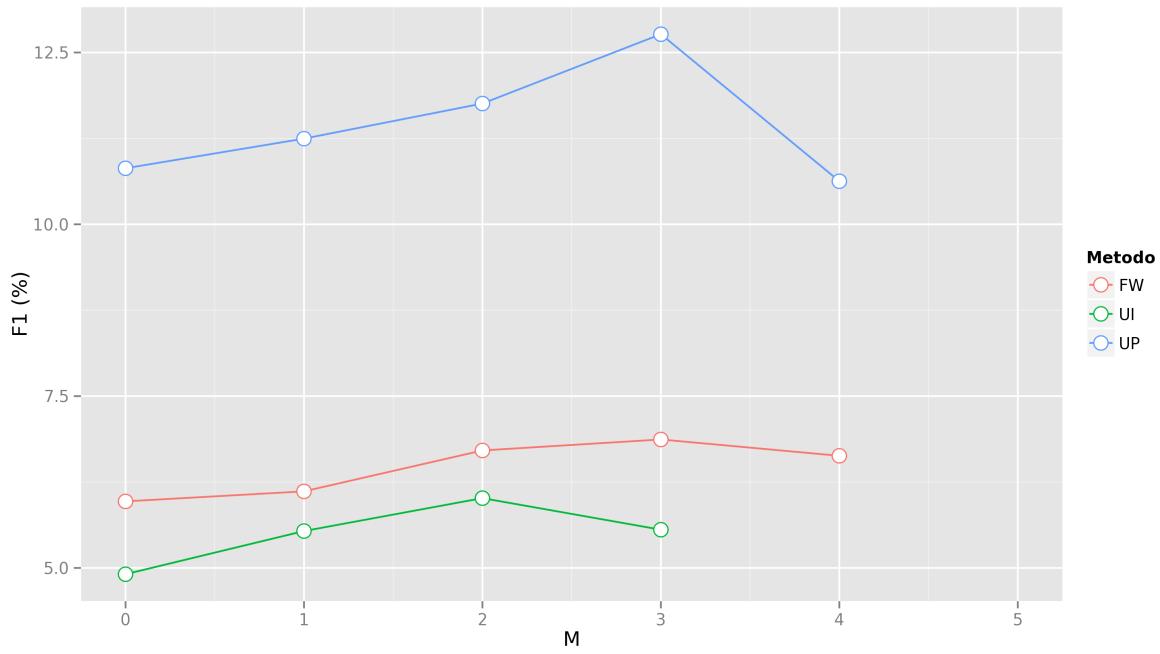


Figura 26 – Medida F_1 em função do valor mínimo para avaliações positivas M

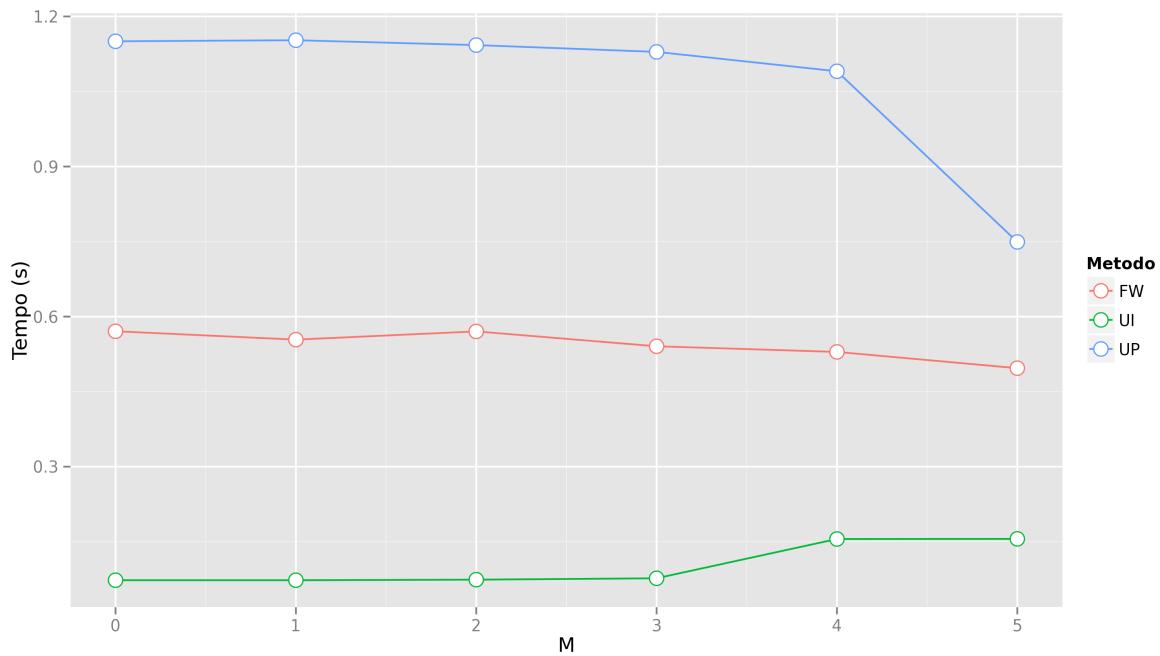
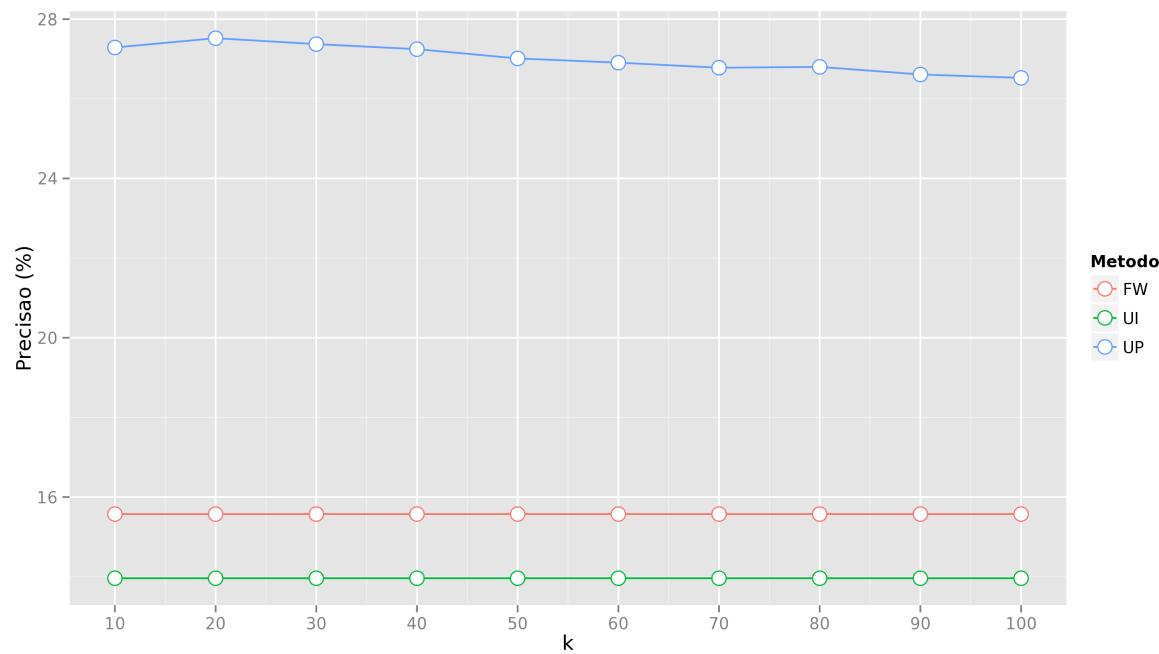
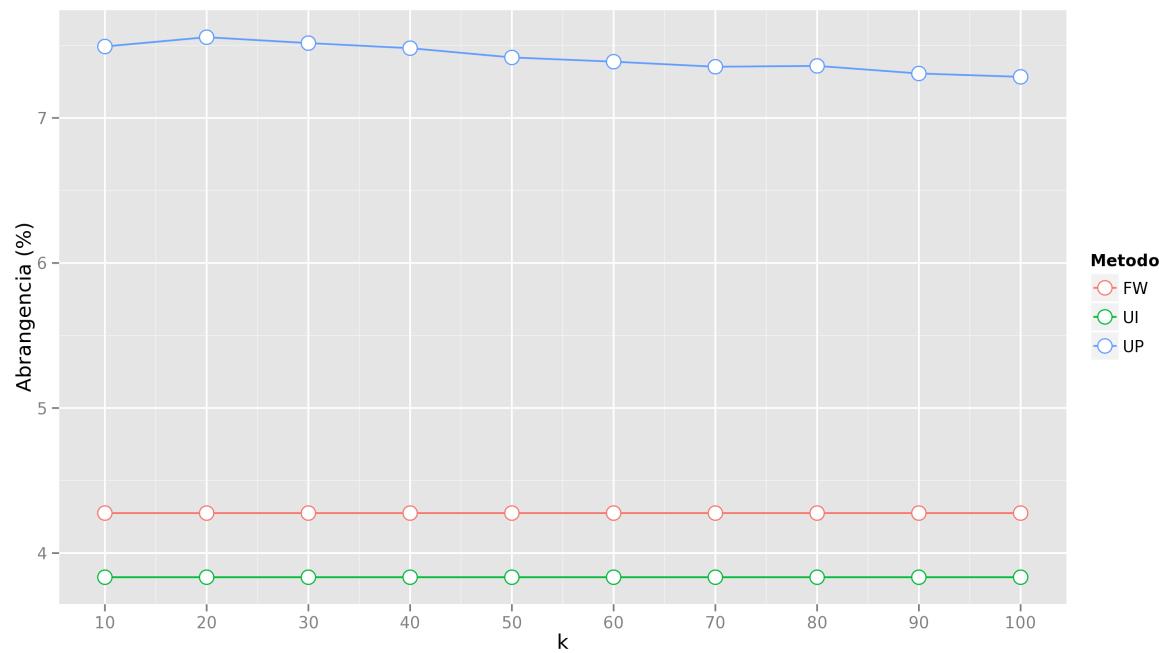


Figura 27 – Tempo de execução em função do valor mínimo para avaliações positivas M

Figura 28 – Precisão em função do número de vizinhos mais próximos k Figura 29 – Abrangência em função do número de vizinhos mais próximos k

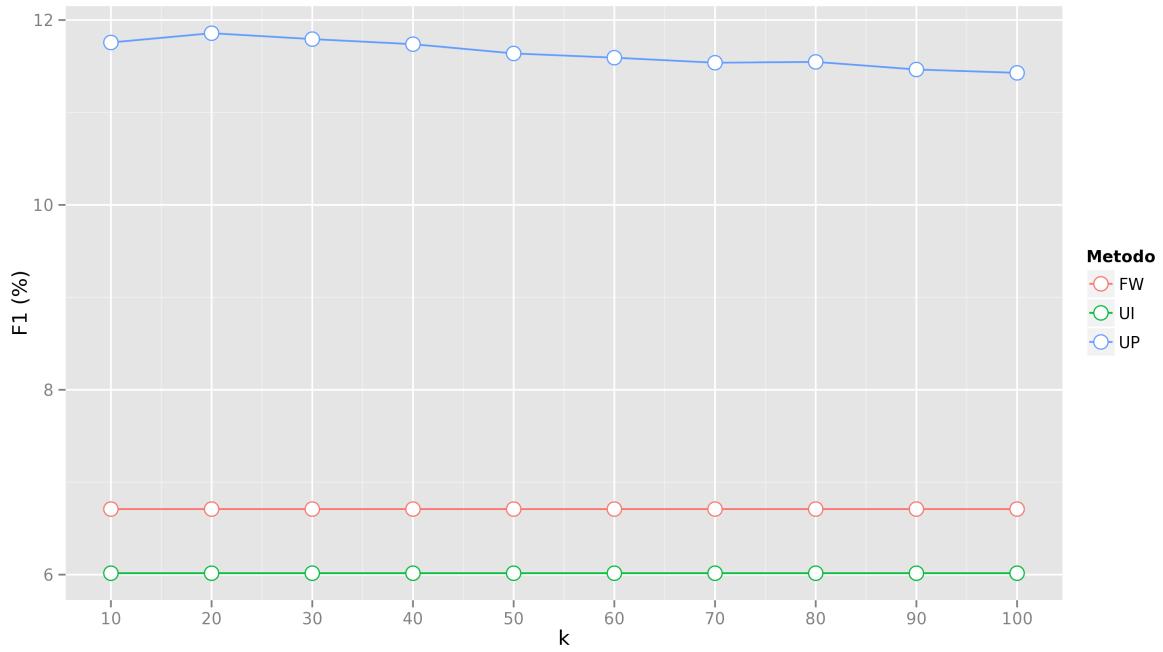


Figura 30 – Medida F_1 em função do número de vizinhos mais próximos k

7.6 Conjunto de atributos dos itens \mathcal{F}

Para o banco de dados 100k-IMDB, o conjunto de atributos dos itens é $\mathcal{F} = \{\text{data de lançamento, gênero, duração, orçamento, avaliação, votos}\}$.

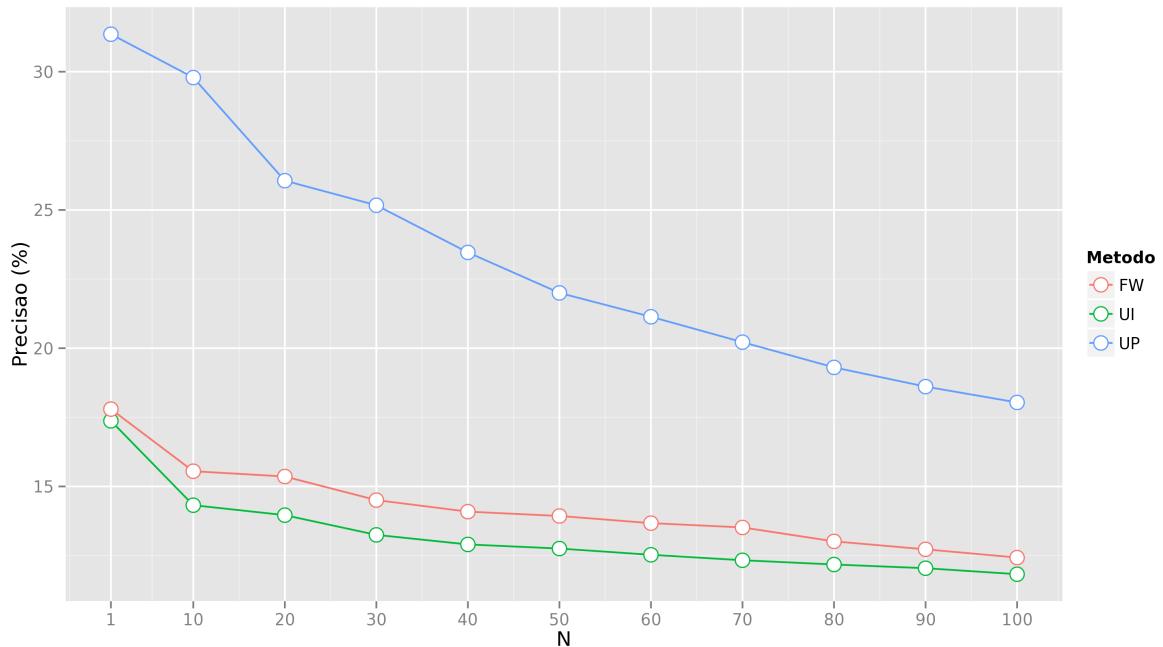
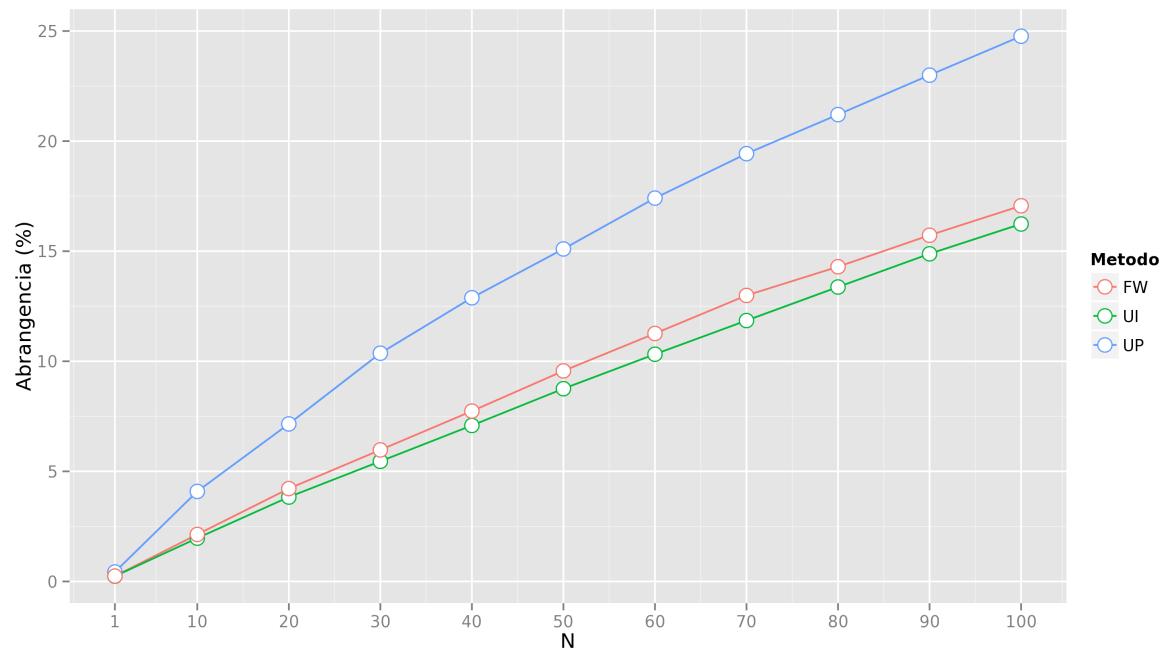
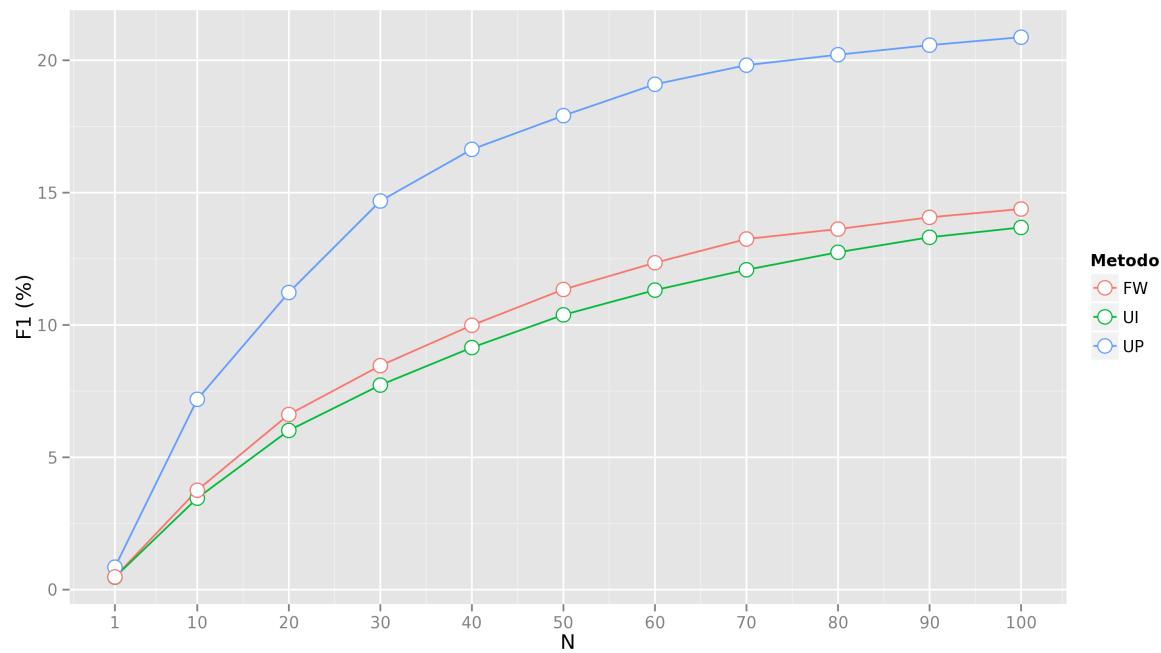


Figura 31 – Precisão em função do tamanho da lista de recomendações N

Figura 32 – Abrangência em função do tamanho da lista de recomendações N Figura 33 – Medida F_1 em função do tamanho da lista de recomendações N

7.7 Medida de distância entre atributos d^f

7.8 Pesos dos atributos w_f

8 Conclusão

TODO arrumar

Para os trabalhos futuros, iremos realizar a validação cruzada e avaliar se os requisitos funcionais foram estabelecidos. Em seguida, procuraremos melhorar o sistema de recomendação a fim de torná-lo mais genérico. Buscaremos eliminar restrições quanto a entrada e saída de dados, de forma que elas sejam completamente arbitrárias. O objetivo é que o usuário possa informar ao sistema como é formado sua base, e que todo o tratamento preliminar seja feito automaticamente.

Caso haja tempo, trabalharemos também na construção de um *driver* que possibilite a conexão entre o sistema de recomendação e um banco de dados SQL, sem que seja necessária a etapa intermediária de arquivos `csv` para aquisição de dados. Planejamos elaborar um *website* para o sistema de recomendação e exportar toda a lógica para um servidor dedicado. Outra melhoria desejada é a reconstrução dos métodos na linguagem de programação C, a fim de melhorar a performance computacional. Dessa forma, o serviço de “sistema de recomendação nas nuvens” estaria completo e poderia ser utilizado por e-commerce reais.

Referências

- 1 EMARKETER. *B2C Ecommerce Climbs Worldwide, as Emerging Markets Drive Sales Higher*. 2013. Disponível em: <<http://www.emarketer.com/Article/B2C-Ecommerce-Climbs-Worldwide-Emerging-Markets-Drive-Sales-Higher/1010004>>. Citado na página 19.
- 2 EMARKETER. *Global B2C Ecommerce Sales to Hit \$1.5 Trillion This Year Driven by Growth in Emerging Markets*. 2014. Disponível em: <<http://www.emarketer.com/Article/Global-B2C-Ecommerce-Sales-Hit-15-Trillion-This-Year-Driven-by-Growth-Emerging-Markets/1010575>>. Citado na página 19.
- 3 MAC, R.; SOLOMON, B. *Alibaba Boosts IPO Price Range, Could Raise Up To \$25 Billion*. 2014. Disponível em: <<http://www.forbes.com/sites/ryanmac/2014/09/15/alibaba-raises-ipo-price-range-could-raise-up-to-25-billion/>>. Citado na página 19.
- 4 COOPERS, P. W. *Total Retail Global Survey of Online Shoppers*. 2014. Disponível em: <<http://www.pwc.com/gx/en/retail-consumer/retail-consumer-publications/global-multi-channel-consumer-survey/index.jhtml>>. Citado na página 19.
- 5 COOPERS, P. W. *The Go-to-Market Revolution - Igniting Growth with Marketing, Sales, and Pricing*. 2014. Disponível em: <https://www.bcgperspectives.com/content/articles/go_to_market_strategy_growth_go_to_market_revolution_igniting_growth_marketing_sales_pricing>. Citado na página 19.
- 6 RICCI, L. R. F.; SHAPIRA, B. Introduction to recommender systems handbook. In: *Recommender Systems Handbook*. [S.l.]: Springer, 2011. p. 1–35. Citado na página 19.
- 7 AMATRIAIN, X. *Netflix Recommendations: Beyond the 5 stars*. 2012. Disponível em: <<http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html>>. Citado na página 19.
- 8 MARSHALL, M. *Aggregate Knowledge raises \$5M from Kleiner, on a roll*. 2006. Disponível em: <<http://venturebeat.com/2006/12/10/aggregate-knowledge-raises-5m-from-kleiner-on-a-roll/>>. Citado na página 19.
- 9 DAS, A. S. et al. Google news personalization: scalable online collaborative filtering. In: ACM. *Proceedings of the 16th international conference on World Wide Web*. [S.l.], 2007. p. 271–280. Citado na página 19.
- 10 SCHAFER, J. B.; KONSTAN, J.; RIEDL, J. Recommender systems in e-commerce. In: ACM. *Proceedings of the 1st ACM conference on Electronic commerce*. [S.l.], 1999. p. 158–166. Citado 2 vezes nas páginas 20 e 28.
- 11 SARWAR, B. et al. Analysis of recommendation algorithms for e-commerce. In: ACM. *Proceedings of the 2nd ACM conference on Electronic commerce*. [S.l.], 2000. p. 158–167. Citado na página 22.

- 12 SYMEONIDIS, P.; NANOPoulos, A.; MANOLOPOULOS, Y. Feature-weighted user model for recommender systems. In: *User Modeling 2007*. [S.l.]: Springer, 2007. p. 97–106. Citado 6 vezes nas páginas 23, 28, 32, 35, 41 e 42.
- 13 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. Citado 3 vezes nas páginas 23, 24 e 27.
- 14 BALABANOVIC, M.; SHOHAM, Y. Fab: Content-based, collaborative recommendation. *Communications of the ACM*, v. 40, p. 66–72, 1997. Citado 2 vezes nas páginas 24 e 27.
- 15 WEI, K.; HUANG, J.; FU, S. A survey of e-commerce recommender systems. In: IEEE. *Service Systems and Service Management, 2007 International Conference on*. [S.l.], 2007. p. 1–5. Citado 3 vezes nas páginas 24, 26 e 28.
- 16 SCHAFER, J. B.; KONSTAN, J. A.; RIEDL, J. E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, v. 5, p. 115–153, 2001. Citado na página 24.
- 17 LOPS, P.; GEMMIS, M. de; SEMERARO, G. Content-based recommender systems: State of the art and trends. In: *Recommender Systems Handbook*. [S.l.]: Springer, 2011. p. 73–105. Citado na página 24.
- 18 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. Citado na página 25.
- 19 BURKE, R. Hybrid web recommender systems. In: *The adaptive web*. [S.l.]: Springer, 2007. p. 377–408. Citado na página 25.
- 20 LEE, J.; SUN, M.; LEBANON, G. A comparative study of collaborative filtering algorithms. *arXiv preprint arXiv:1205.3193*, 2012. Citado na página 26.
- 21 TUTOL, L. *Amazon Launches ‘Login and Pay with Amazon’ for a Seamless Buying Experience*. 2013. Disponível em: <<http://services.amazon.com/post/Tx2A98P3EKP62O2/Amazon-Launches-Login-and-Pay-with-Account-for-a-Seamless-Buying-Experience>>. Citado na página 26.
- 22 PALLADINO, V. *Amazon sold 426 items per second in run-up to Christmas*. 2013. Disponível em: <<http://www.theverge.com/2013/12/26/5245008/amazon-sees-prime-spike-in-2013-holiday-season>>. Citado na página 26.
- 23 FENNELL, J. Collaborative filtering on sparse rating data for yelp. com. 2009. Citado na página 27.
- 24 LOPS, P.; GEMMIS, M. de; SEMERARO, G. In: *Recommender Systems Handbook*. [S.l.]: Springer, 2011. Citado na página 28.
- 25 DEBNATH, S.; GANGULY, N.; MITRA, P. Feature weighting in content based recommendation system using social network analysis. In: ACM. *Proceedings of the 17th international conference on World Wide Web*. [S.l.], 2008. p. 1041–1042. Citado 4 vezes nas páginas 28, 35, 41 e 43.

- 26 A Guide To The Project Management Body Of Knowledge (PMBOK Guides). [S.l.]: Project Management Institute, 2004. ISBN 193069945X, 9781933890517. Citado na página 31.
- 27 LARMAN, C.; BASILI, V. R. Iterative and incremental development: A brief history. *Computer*, IEEE Computer Society, Los Alamitos, CA, USA, v. 36, n. 6, p. 47–56, 2003. ISSN 0018-9162. Citado na página 31.
- 28 NG, A. Y. Preventing "overfitting" of cross-validation data. In: . [S.l.: s.n.]. Citado na página 33.
- 29 SARWAR, B. et al. Item-based collaborative filtering recommendation algorithms. In: ACM. *Proceedings of the 10th international conference on World Wide Web*. [S.l.], 2001. p. 285–295. Citado na página 35.
- 30 MOVIELENS. *MovieLens 100k Dataset*. 1998. Disponível em: <<http://files.grouplens.org/datasets/movielens/ml-100k-README.txt>>. Citado na página 35.
- 31 WICKHAM, H. *Movies dataset*. 2006. Disponível em: <<http://docs.ggplot2.org/0.9.3/movies.html>>. Citado na página 35.
- 32 HOPE, C. *Epoch*. 2014. Disponível em: <<http://www.computerhope.com/jargon/e/epoch.htm>>. Citado na página 44.
- 33 LOPS, P.; GEMMIS, M. de; SEMERARO, G. Advances in collaborative filtering. In: *Recommender Systems Handbook*. [S.l.]: Springer, 2011. p. 145–184. Citado 2 vezes nas páginas 49 e 63.
- 34 COOK, J. D. *Benchmarking C++, Python, R, etc.* 2014. Disponível em: <<http://www.johndcook.com/blog/2014/06/20/benchmarking-c-python-r-etc/>>. Citado na página 60.