Escola Politécnica da Universidade de São Paulo

PMR2500 – Projeto de Conclusão do Curso I

Desenvolvimento de um Sistema de Recomendação para e-Commerces

Nome	*Número USP*
Antônio Guilherme Ferreira Viggiano	6846450
Fernando Fochi Silveira Araújo	5894546

Orientador

Prof. Dr. Fábio Gagliardi Cozman

1 de junho de 2014

Sumário

1 Introdução

"Sistemas de recomendação são ferramentas e técnicas de software destinadas a prover sugestões de itens para usuários" (1). 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.

De modo geral, um sistema de recomendação possui três etapas: a aquisição dos dados de entrada do usuário e dos itens, 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. A determinação das recomendações é feita 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 a 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 possibilidades do conjunto total de itens. Dessa forma, a interface homem-máquina é personalizada diferentemente para cada um dos usuários, de maneira que eles recebam recomendações adequadas ao seu perfil. Essa ideia, amplamente divulgada por um antigo diretor executivo da empresa *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" (2).

Motivados pela importância econômica crescente de lojas de varejo online, bem como pela possibilidade de criar um conjunto de ferramentas *open source* que possam ser utilizadas abertamente pela comunidade, propomos como Trabalho de Conclusão de Curso o desenvolvimento de um sistema de recomendação de produtos de e-commerces.

A contribuição científica e tecnológica do trabalho para a Engenharia Mecatrônica estão sobretudo nos campos de sistemas de informação, de automação de processos e de inteligência artificial. As competências acadêmicas necessárias para a sua execução envolvem algoritmos e estruturas de dados, aprendizado de máquina e modelagem de bancos de dados. As competências técnicas abrangem programação estatística e orientada a objetos (R ou Java, por exemplo) e em linguagem de consulta estruturada (SQL).

2 Estado da Arte

2.1 Estado da arte dos problemas

O problema de recomendação pode ser formulado como se segue, adaptado da referência (3), com notação inspirada em (4):

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, ou seja, $\ell : \mathcal{U} \times \mathcal{I} \to \mathrm{R}$, onde R é um conjunto totalmente ordenado (por exemplo, números inteiros não-negativos ou números reais dentro de um determinado intervalo, em geral $[1, 5]$). O objetivo do sistema de recomendação é determinar o item $\hat{\imath}$ que maximize a utilidade ℓ_{ui} do usuário u.

$$\forall u \in \mathcal{U},\ \hat{\imath}_u = \arg\max_{i \in \mathcal{I}} \ell_{ui} \tag{2.1}$$

O problema central da recomendação é que a função ℓ é em geral desconhecida, e portanto determinar $\hat{\imath}$ 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.

Para lidar com esse problema, existem três grandes grupos de estratégias de sugestão de itens, conforme apresenta a referência (5) (TODO refazer essa parte):

- Recomendações baseadas em conteúdo: o usuário recebe recomendações com base nas descrições dos atributos dos itens;
- Recomendações colaborativas
  - Baseada em usuários: o usuário recebe recomendações de itens que pessoas com gostos e preferências semelhantes gostaram no passado;
  - Baseada em itens: o usuário recebe recomendações de itens semelhantes aos que ele gostou no passado;
- Recomendações híbridas: esses métodos combinam métodos colaborativos e métodos baseados em conteúdo.

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 – essencialmente 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 um modelo, como em uma regressão linear ou em uma rede Bayesiana.

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 reviews de usuários, os itens são textos sem estrutura definida e a abordagem mais comum é a de recuperação de informação – o usuário procura por uma lista de termos desejados e o sistema retorna os textos que contém aqueles termos com maior relevância, tal como é feito em um motor de busca (6). Nesses casos, calcula-se a similaridade entre documentos a partir de formulações que levam em conta as palavras ou termos escritos, como a TF-IDF ou o classificador Bayesiano (7).

Na abordagem de sistemas baseados em conteúdo, a recomendação pode ser vista como um problema de aprendizado que explora os conhecimentos sobre o usuário. Muitas vezes é recomendado que o aprendizado seja feito com base no perfil do usuário conforme o uso continuo, ao invés de forçá-lo a responder diversas perguntas demográficas (8). Também chamado de aprendizado de máquina, o objetivo é aprender a 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.

As recomendações colaborativas baseadas em usuários, por sua vez, tentam prever a utilidade dos itens para cada usuário baseado em itens previamente avaliados por outros usuários. Mais formalmente, a utilidade $u(c, s)$ de um item s para um usuário c é estimada com base nas utilidades $u(c_j, s)$ propostas por usuários $c_j \in C$ que são "similares" ao usuário c. Por exemplo, em um sistema de recomendação de filmes, a fim de recomendar um título para um usuário c, o sistema tenta identificar "avaliadores" com gostos similares ao do usuário c, e então indica-se os filmes que os usuários c_j recomendariam. De maneira análoga, as recomendações colaborativas baseadas em itens, tentam prever a utilidade $u(c, s)$ com base nas utilidades $u(c, s_j)$, dado itens $s_j \in S$ que são "similares" aos itens s (9).

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 dispersão de dados ou o *cold start* (10).

2.2 Estado da arte das soluções

Do ponto de vista do estado da arte das soluções, as variáveis de interesse estão ligadas do número de usuários no sistema, ao número de itens, ao nível de dispersão, à medida de qualidade da recomendação e ao custo computacional (11).

No que se refere à dependência do número de usuários, a filtragem colaborativa a base de usuários é extremamente efetiva para um baixo número de usuários, mas tem uma dependência quase constante em relação a essa quantidade. 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 para quantidades maiores.

A dependência do número de itens é, de certa forma, oposta à de usuários: a filtragem colaborativa a base de itens é extremamente efetiva para poucos itens, mas tem uma dependência quase constante no número de itens. A filtragem colaborativa baseada em usuários tem performance consideravelmente pior de início, mas supera todos os outros métodos baseados em memória para maiores quantidades de usuários.

Com relação ao nível de dispersão dos dados, a filtragem baseada em usuários e a baseada em itens mostram uma dependência semelhante. Na medida de qualidade de recomendação (menor erro quadrático médio), todos os métodos de recomendação variam não-linearmente com o número de usuários, itens e nível de dispersão, e de modo geral há um *trade-off* entre a acurácia e o tempo de processamento da sugestão de produtos.

2.3 Desafios científicos e tecnológicos

Um dos maiores desafios tecnológicos dos sistemas de recomendação é, atualmente, o da escalabilidade (8). O sistema de recomendação deverá ser flexível no sentido de poder operar igualmente bem tanto em conjuntos pequenas quanto em grandes bases de dados, que podem chegar até centenas de milhões de clientes (12) e de produtos (13). Isso significa que as recomendações devem ser suficientemente rápidas e ainda assim prover sugestões valiosas aos consumidores.

Um sistema de recomendação inteligente também deve prever quando enviar uma determinada recomendação, e não agir apenas mediante requisição do cliente (14). É interessante, por exemplo, enviar recomendações de produtos com descontos a usuários que estão há algum tempo inativos no site, para que eles retornem a comprar. Da mesma forma, um sistema inteligente poderia sugerir produtos do lar a um usuário detectado como recém-casado.

Outro desafio científico ainda em estágio inicial de pesquisa é referente à diversidade das recomendações realizadas, também chamado de excesso de especialização (3). 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.

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 (2). 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 (8).

Esse tópico de pesquisa inclui ainda diversos desafios científicos e tecnológicos que não foram aqui detalhados, 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. Entretanto, esses desafios são menos relevantes porque não se aplicam diretamente aos objetivos do nosso projeto, que serão especificados no Capítulo 3.

3 Objetivos

O objetivo do presente Trabalho de Conclusão de Curso é o desenvolvimento de um Sistema de Recomendação de produtos para lojas de comércio online, baseado em algoritmos de filtragem colaborativa item a item e a respectiva análise de desempenho das recomendações propostas.

Esse sistema será, do ponto de vista da taxonomia tradicional dos sistemas de recomendação (2, 8), automático e persistente. Isso significa que as sugestões serão dadas sem a interação do usuário e que as compras anteriores serão levadas em conta. Essas características aproximam o entregável das ferramentas de marketing via e-mail, que sugerem produtos com uma determinada frequência aos usuários com base em seu histórico de compras.

A qualidade das recomendações será avaliada tanto em termos da similaridade entre os itens efetivamente comprados pelo cliente com aqueles previstos pelo sistema de recomendação, quanto em termos de indicadores de erro tipo I e erro tipo II, como a medida F (15).

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 ou a quantidade de informações de itens e clientes utilizadas na recomendação.

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 dispersão dos dados (8). Também serão avaliadas as diferentes medidas de similaridade e modelos de predição na qualidade das recomendações.

Ao final, será possível extrair uma validação experimental das diretrizes fundamentais a serem seguidas por e-commerces que desejem desenvolver um sistema de recomendação próprio, a partir de um banco de dados de clientes, produtos e histórico de compras.

4 Metodologia

O presente Trabalho de Conclusão de Curso se fundamenta na metodologia de um projeto de engenharia. Por se tratar de um projeto de Engenharia de Software, alguns desses passos são adaptados a fim de levar em conta o desenvolvimento do código computacional. Como o projeto de um software é um processo cíclico com etapas de especificação, desenvolvimento, validação e manutenção, a criação do produto ocorre de maneira incremental, diferentemente de certos projetos de outras áreas da engenharia (16).

A metodologia de trabalho proposta pode ser, então, consolidada da seguinte maneira:

4.1 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 e aumentar 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 ou da academia. 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.

4.2 Definição dos Parâmetros de Sucesso

O sucesso do projeto poderá ser medido em duas frentes, a primeira sendo a verificação entre as sugestões do sistema e as compras feitas pelos consumidores e a segunda é a escalabilidade do sistema de recomendação. Visto que a tendência é o aumento da base de consumidores e de itens, há um aumento no custo computacional para gerar recomendações, e o sistema deve responder sem grande demora ou perda de qualidade.

4.3 Síntese de Soluções

Nesta fase do projeto serão propostas possíveis soluções para o problema proposto. Aqui o problema principal deverá ser dividido em partes menores, que idealmente são mutuamente exclusivas e coletivamente exaustivas (cobrem todos os pontos uma só vez), que serão individualmente resolvidas. Por exemplo, o método de se fazer o cálculo de

medidas de similaridade entre dois itens influencia na taxa de sucesso da recomendação, e o método de se expandir este cálculo para os outros itens influencia na escalabilidade do sistema.

4.4 Processo de escolha

O processo de escolha da solução deverá levar em conta três pontos. O primeiro, eliminatório, é a viabilidade técnica da solução – não será levado em conta soluções de execução inviável. Os outros dois, classificatórios, levam em conta a os parâmetros de sucesso do projeto, devendo assim maximizar a escalabilidade e a taxa de recomendações bem sucedidas.

4.5 Detalhamento da Solução

No detalhamento da solução, serão levantados os pontos que serão comparados entre os itens e a estrutura dos algoritmos que gerarão as recomendações.

4.6 Projeto Básico

Aqui serão codificados os métodos escolhidos para o cálculo da recomendação para uma item qualquer e o método de aplicação deste cálculo para todos os outros casos. Etapa de projeto é incremental e ocorre em ciclos, acompanhada ela própria de testes unitários e testes de integração.

4.7 Modelamento e Simulação

Para o modelamento e simulação utilizaremos partes dos bancos de dados que temos disponíveis e serão feitas simulações até que tenhamos resultados satisfatórios.

4.8 Projeto Executivo

O Projeto Executivo conterá os métodos escolhidos e a forma de implementá-los em um e-commerce, a fim que seja possível aplicá-lo independentemente da área de atuação desta empresa.

4.9 Protótipos/Testes

A realização de testes será feita com os bancos de dados de centenas de milhares de itens ou de avaliações. Visto que será feita uma validação cruzada, será necessário descartar os dados e reformular a solução caso as recomendações não atinjam os requisitos funcionais. Isso evita que o projeto seja moldado para operar somente com aquele banco de dados específico.

4.10 Produto

Assim que a fase de testes for concluída com êxito, o Projeto Executivo se torna o Produto, já que este é um projeto voltado à programação e aplicação em novos negócios.

5 Requisitos

A partir dos casos de uso propostos e do projeto do sistema para o sistema de recomendação, é possível extrair os requisitos funcionais do software. Esses requisitos ditam principalmente sobre a escalabilidade e acurácia do sistema.

Como as recomendações serão calculadas com antecedência e dadas de forma automática, não há necessidade para um elevado *throughput* ou taxa de transferência (quantidade de recomendações feitas por período de tempo). 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 (17).

O sistema também deve ser suficientemente acurado para prover recomendações úteis para os clientes. Espera-se que o desvio médio entre todas as previsões de qualidade de itens e os produtos efetivamente avaliados pelos clientes, ou seja, o erro absoluto médio, seja de no máximo $1,00$, para avaliações que variam de $1,00$ a $5,00$. No caso de bancos de dados que não contém a avaliação dos produtos por parte dos clientes, esse requisito pode ser substituído pelo desvio médio entre as similaridades dos produtos sugeridos e aqueles verdadeiramente comprados. Para os sistemas de recomendação tradicionais, esse valor é de cerca de $0,85$ (18).

Os requisitos funcionais são suportados por requisitos não-funcionais, e estes são determinados pelas restrições sobre o projeto ou execução, tais como requisitos de desempenho, de segurança ou 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, 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 estrutural C, a fim de torná-lo genérico e reutilizável.

Por fim, o sistema de recomendação deverá 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 (12) e de produtos (13).

6 Cronograma

O cronograma de atividades da dupla busca seguir o cronograma proposto pela banca avaliadora dos trabalhos de conclusão de curso, estando sempre à frente das entregas em pelo menos uma semana. Dessa maneira, é possível apresentar a entrega antecipadamente ao orientador e falar sobre possíveis mudanças ou correções.

Além disso, semanalmente os alunos se reunem com o orientador a fim de conversar sobre o andamento do projeto, apresentar-lhe o esboço dos relatórios e discutir a implementação dos algoritmos.

Para o segundo semestre, trabalharemos na implementação do sistema de recomendação já no período de férias escolares, para poder ter uma amostra funcional no início das aulas. Em seguida, daremos início ao relatório final em paralelo com os testes de performance do sistema de recomendação, e esperamos finalizar o projeto dentro do prazo estipulado.

O cronograma detalhado da dupla está descrito a seguir:

09/04 Análise do banco de dados e determinação das medidas de similaridade

16/04 Esboço do relatório final

23/04 Validação I do relatório final

07/05 Validação II do relatório final

14/05 Desenvolvimento dos Algoritmos de Recomendação

28/05 Validação III do relatório final

04/06 Esboço do resumo final e da apresentação

09/06 Validação do resumo final

11/06 Validação da apresentação

13/06 Ensaio da apresentação

23/06 Apresentação para o orientador

09/07 Desenvolvimento do sistema de recomendação

16/07 Desenvolvimento do sistema de recomendação

23/07 Desenvolvimento do sistema de recomendação

30/07 Desenvolvimento do sistema de recomendação

13/08 Relatório de atividades de implementação

27/08 Primeiros testes com o sistema (desvio de similaridade para uma base teste)

03/09 Testes com o sistema (validação cruzada)

24/09 Melhorias incrementais e relatório de atividades

15/10 Relatório aprofundado de atividades

05/11 Elaboração da apresentação e finalização dos relatórios

12/11 Melhorias incrementais

7 Resultados

Até o presente momento, os resultados deste Trabalho de Conclusão de Curso concentram-se na definição de necessidades, de parâmetros de sucesso e de síntese de possíveis soluções.

Visto que a primeira etapa de um sistema de recomendação é a coleta e manipulação de dados, definimos que o sistema partirá de um banco de dados relacional MySQL. Essa escolha decorre do fato de este ser o sistema de gestão de banco de dados *open source* mais utilizado no mundo. O *input* virá sob a forma de três tabelas genéricas de informações de clientes, C, itens, S, e histórico de compras, H. Os dados de cliente c e item s devem vir, cada um deles, sob a forma de um identificador e eventuais dados complementares. Os dados de histórico h, sob a forma de um identificador e da relação de identificadores cliente-item daquela compra, assim como de eventuais dados complementares.

$$\begin{aligned}
\mathbf{c} \in \mathbf{C}, \mathbf{c} &= \begin{bmatrix} \text{id}_c & \text{atributo}_1 & \text{atributo}_2 & \dots & \text{atributo}_n \end{bmatrix}^T \\
\mathbf{s} \in \mathbf{S}, \mathbf{s} &= \begin{bmatrix} \text{id}_s & \text{atributo}_1 & \text{atributo}_2 & \dots & \text{atributo}_m \end{bmatrix}^T \\
\mathbf{h} \in \mathbf{H}, \mathbf{h} &= \begin{bmatrix} \text{id}_h & \text{id}_c & \text{id}_s & \text{atributo}_1 & \text{atributo}_2 & \dots & \text{atributo}_p \end{bmatrix}^T
\end{aligned} \tag{7.1}$$

Uma vez determinada a forma de entrada de dados, definiu-se a escolha do conjunto de dados a serem utilizados. O primeiro conjunto de dados abertos é proveniente do website de recomendações de filmes MovieLens (<http://movielens.umn.edu>). Nessa base de dados, o catálogo de filme faz o papel de catálogo de produtos pelos quais os usuários possam se interessar, e o histórico de compras se refere à avaliação dos filmes feita por cada usuário. Outros conjuntos de dados também serão explorados pela dupla, tais como os dados de classificação de músicas do serviço Yahoo! Music (<http://webscope.sandbox.yahoo.com>) ou de dados anônimos de e-commerces.

Por fim, as possíveis soluções do projeto abrangem o cálculo das medidas de similaridade entre itens, para os conjuntos de dados que não foram previamente tratados. Determinar um valor numérico entre dois produtos distintos, tais como uma camiseta e uma prancha de *surf*, é uma tarefa complexa e sujeita a erros humanos. Após reflexão e leitura de referências, definimos possíveis maneiras de realizar esse cálculo:

- Grupos de similaridade: a similaridade dos itens seria definida por pelas características do item (como “esporte radical”, “corrida”, “filme de aventura”, etc). Seria necessário classificar manualmente esses atributos (por exemplo, determinar que “esporte radical” tem similaridade de 3/5 com “corrida” e 1/5 com “produtos de limpeza”).

- Histórico de compra da comunidade: quanto mais usuários comprarem a mesma dupla de itens, maior a similaridade entre estes itens. A classificação se faz pelo histórico mas o índice de similaridade pertence aos itens e não entre os usuários.

- Ranking por arestas: adaptado do sistema *edge rank*, de classificação de posts no Facebook, este sistema levaria em conta a multiplicação de diversas características do item. Características como: popularidade da marca, número de compras por visualização do item, popularidade da marca para o usuário em questão (quantos itens ele comprou desta marca), há quanto tempo o item foi lançado e a relação do item com as compras anteriores do usuário (se este tipo de item já fez sucesso com este usuário).

Para os itens que já foram avaliados, como no caso dos conjuntos de dados de classificação de filmes ou músicas, essa etapa não é necessária, pois a similaridade entre dois itens provém diretamente da avaliação do usuário. Dizer que um filme A tem avaliação $4/5$ e um filme B tem avaliação $5/5$ equivale a dizer que os dois são interessantes para aquele usuários, e por isso vale recomendar filmes similares a A e B. Nesse caso, passa-se diretamente para a etapa das recomendações.

Os resultados práticos de cálculo de similaridade ou descrição completa do banco de dados serão apresentados no relatório final do trabalho, em conjunto com os demais resultados da evolução do projeto.

8 Andamento do Projeto

De início pensamos fazer um sistema de recomendação utilizando algoritmos de filtragem colaborativa baseada em itens, principalmente motivados pela leitura inicial de (9), que mostrava as vantagens desse método comparado à filtragem colaborativa baseada em usuários.

Todavia, percebemos grande parte dos e-commerces estruturam seus bancos de dados em torno da descrição dos itens vendidos e das informações dos clientes. Pouco detalhe é dado à interação entre esses dois grupos, com exceção da tabela de compras, que se limita a informações como data e método de pagamento. Dessa forma concluímos que os métodos de filtragem colaborativa, fundamentados na avaliação dos itens por parte dos usuários, teriam pior desempenho que métodos baseados em conteúdo.

Métodos de recomendação baseados em conteúdo podem explorar a classificação dos produtos no banco de dados a fim de determinar as sugestões. Os atributos podem ser diversos, dependendo do ramo de negócios do e-commerce, tais como marca, esporte, categoria, sexo, idade, cor, preço, etc.

...

Utilizaremos um banco de dados de e-commerce fornecido anonimamente. Para o segundo semestre, utilizaremos também outros bancos, como o TODO MovieLens, TODO ClickBus. Ele está estruturado da seguinte maneira: TODO

...

Para o segundo semestre deste ano, desenvolveremos um sistema de recomendação a partir de diferentes algoritmos e faremos uma análise de desempenho para cada um deles. Utilizaremos algoritmos inspirados em (19) e (4). O primeiro artigo 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 segundo texto parte do princípio que os usuários estão interessados nos atributos dos itens, traçando correlações entre esses dois elementos até chegar nos pesos que servirão de base para a matriz de similaridade de usuários, utilizada na recomendação pelo método da vizinhança (*nearest neighbors*).

...

A aquisição de dados será feita a partir de um banco de dados genérico, que deverá alimentar o sistema por meio de arquivos de texto com valores separados por vírgulas (`.csv`). A fim de facilitar o pré-processamento dos dados, exigem-se três arquivos, cada um com uma tabela de itens, clientes e histórico de compras. Caso existam outras tabelas

no banco de dados, o sistema deverá ser alterado para levar em conta o processamento dos arquivos suplementares.

...

Os resultados das recomendações serão entregues, da mesma forma, por meio de um arquivo `.csv` contendo o identificador de cada usuário com as *top-N* recomendações de produtos, assim como o valor numérico associado à recomendação. Esse resultado é o mais importante do ponto de vista do e-commerce, que o utilizará como estratégia de marketing na sugestão de produtos.

...

Uma das maiores dificuldades do sistema de recomendação é a escala, isto é, o fato de os sistemas lidarem com quantidades de itens e clientes da ordem de centenas de milhares, exigindo algoritmos eficientes e inviabilizando implementações computacionalmente complexas. Outra grande dificuldade é a esparsidade dos dados, ou seja, o fato de a maioria dos clientes nunca ter interagido com mais de algumas unidades de itens, fazendo com que a matriz de relação usuário-item tenha apenas uma quantidade muito pequena de valores preenchidos, da ordem de 1% (20).

...

De modo geral os sistemas de recomendação tem o objetivo de apresentar ao usuário itens pelos quais ele possa se interessar e, no caso de um e-commerce, que ele vá adquirir. 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 ($\tilde{\mathbf{x}}$) e aqueles que seriam efetivamente comprados ($\mathbf{x}$) pelo cliente em uma validação cruzada (*cross validation*). Essa medida pode ser, por exemplo, a distância L_1 (erro médio absoluto, $|\tilde{\mathbf{x}} - \mathbf{x}|$) ou a distância L_2 (erro quadrático médio, $\sqrt{|\tilde{\mathbf{x}} - \mathbf{x}|^2}$)

Outras medidas de predição também serão utilizadas, tais como acurácia (*accuracy*), especificidade (*specificity*), precisão (*precision*), abrangência (*recall*) e a medida F_1 (F_1-*score*). Elas estão sumarizadas na Tabela 1.

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 e na capacidade de lidar com problemas como o *cold start*. O tempo de execução também será avaliado em função do tamanho do banco de dados.

...

Problemas que o recsys enfrentará: cold start

...

Algoritmos

O primeiro algoritmo que utilizamos no sistema de recomendação, adaptado de (4) e doravante denominado de *feature weighting*, ponderação de atributos ou *FW*, se trata 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. Essa rede social permite determinar o julgamento humano de similaridade entre itens, e por isso se trata de uma filtragem colaborativa dos usuários. Após obtenção dos pesos, realiza-se a filtragem baseada em conteúdo e posteriormente os itens com maior similaridade são 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 das *features* dos itens:

$$s_{ij} = \sum_f w_f \ (1 - d_{fij}) \tag{8.1}$$

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 8.2. É possível considerar a correlação entre atributos f, g (a similaridade de duas marcas de calçado é maior que a de duas marcas de produtos de categorias distintas, mesmo que as marcas sejam diferentes), mas em uma primeira análise utilizaremos para a maior parte dos atributos a medida de distância $d_{fij} = 1 - \delta_{ij}^f$. Isso significa que se as *features* de dois itens são idênticas, a distância é nula e logo a similaridade é máxima. O sumário das medidas de distância estão na Tabela 2.

$$\delta_{ij}^f = \begin{cases} 1, & \text{se } i = j \\ 0, & \text{se } i \neq j \end{cases} \tag{8.2}$$

Os pesos w_f são a priori desconhecidos. A referência (4) os determina a partir de um conjunto de equações do tipo 8.3, onde e_{ij} é o número de usuários que se interessam tanto por i quanto por j.

$$e_{ij} = w_0 + \sum_f w_f \ d_{fij} \tag{8.3}$$

A partir da matriz de avaliações $\mathbf{R}$, pode-se determinar e_{ij}, conforme a equação 8.4, onde r_{ui} é a avaliação do item i feita pelo usuário u e b_0 é o operador booleano descrito

por 8.5.

$$e_{ij} = \sum_{u} \mathrm{b}_0 \left(r_{ui} \; r_{uj} \right) \tag{8.4}$$

$$\mathrm{b}_y \left(x \right) = \begin{cases} 1, & \text{se } x > y \\ 0, & \text{se } x \leq y \end{cases} \tag{8.5}$$

Desta forma, os pesos w_f são determinados a partir resolução do sistema de equações lineares 8.6. Apenas os pesos positivos e com valor absoluto expressivo (maior que um piso arbitrariamente escolhido a posteriori) são utilizados na recomendação. Calcula-se a matriz de similaridade $\mathbf{S}$ pela equação 8.1 e recomenda-se os itens mais similares àqueles já comprados.

$$w_0 + \sum_{f} w_f \; d_{fij} = \sum_{u} \mathrm{b}_0 \left(r_{ui} \; r_{uj} \right), \; \forall i \neq j \tag{8.6}$$

...

O segundo algoritmo, adaptado de (19), é um hibrido 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. 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 a_{if} nos atributos f do item i. A correlação t_{uf} entre usuários e *features* é descrita por 8.7.

$$t_{uf} = \sum_{i} \mathrm{b}_M \left(r_{ui} \; a_{if} \right) \tag{8.7}$$

Os pesos w_{uf} que mostram a relevância de f para u são determinados a partir da estatística TF-IDF (*term frequency–inverse document frequency*), presente em formulações de recuperação de informação e mineração de dados. Em nosso caso, TF ou *feature frequency* é a similaridade intra-usuários p_{uf} – número de vezes em que a *feature* f aparece no perfil do usuário u (equação 8.8). IDF ou *inverse user frequency* é a dissimilaridade inter-usuários q_f – relacionada com o inverso da frequência $\hat{q}_f$ de um atributo f dentro de todos os usuários (equações 8.9 e 8.10).

$$p_{uf} = t_{uf} \tag{8.8}$$

$$\hat{q}_f = \sum_{u} \mathrm{b}_0 \left(t_{uf} \right) \tag{8.9}$$

$$q_f = \log\left(\frac{|\,\mathcal{U}\,|}{\hat{q}_f}\right) \tag{8.10}$$

Os pesos, descritos por 8.11, são utilizados para calcular a similaridade s_{uv} entre dois usuários u e v, conforme 8.12.

$$w_{uf} = p_{uf}\ q_f \tag{8.11}$$

$$\begin{aligned} s_{uv} &= \frac{\sum\limits_{f\in\mathcal{F}_{uv}} w_{uf}\ w_{vf}}{\sqrt{\sum\limits_{f\in\mathcal{F}_{uv}} w_{uf}^2}\sqrt{\sum\limits_{f\in\mathcal{F}_{uv}} w_{vf}^2}} \\ \mathcal{F}_{uv} &= \mathcal{F}_u \cap \mathcal{F}_v \\ \mathcal{F}_u &= \{f \in \mathcal{F} \mid t_{uf} > 0\} \end{aligned} \tag{8.12}$$

Dispondo-se de $\mathbf{S}$, selecionam-se os k vizinhos mais próximos v_k de u com maior similaridade s_{uv}. Posteriormente, determina-se o conjunto $I_{v_k} = \{i \mid r_{v_k i} > 0\}$ de itens i avaliados por v_k. Em 8.13 avalia-se a frequência total fr_f dos atributos f para os itens de I_{v_k}. Por fim, a partir da equação 8.14 calcula-se o peso ω_i de cada item e gera-se a lista dos *top-N* itens a serem recomendados para o usuário u.

$$\mathrm{fr}_f = \sum_{i\in I_{v_k}} \mathrm{b}_0\left(a_{if}\right) \tag{8.13}$$

$$\omega_i = \sum_f a_{if}\ \mathrm{fr}_f \tag{8.14}$$

...

A simbologia utilizada neste presente trabalho de conclusão de curso é adaptada de (4), e está descrita na Tabela 3. As terminologias *cliente* e *usuário* 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 texto.

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} 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 admitido *rating* unitário. Desta forma os bancos de dados que contenham informações do tipo "usuário u avaliou o item i em $r_{ui} = 3.54 > 3.00$" e aqueles que contenham "usuário u comprou o item i, e logo $r_{ui} = 1$" são tratados equivalentemente.

Tabela 1 – 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.
Especificidade	$\frac{VN}{VN+FP}$	Porcentagem de casos negativos sobre aqueles que foram marcados como negativos.
Acurácia	$\frac{VP+VN}{VP+VN+FP+FN}$	Porcentagem de predições corretas.
Medida F_1	$2 \cdot \frac{\text{Precisão} \cdot \text{Abrangência}}{\text{Precisão} + \text{Abrangência}}$	Média harmônica entre precisão e abrangência.

Tabela 2 – Medidas de distância entre atributos

Medida	**Fórmula**	**Significado**

Tabela 3 – Simbologia

Símbolo	**Definição**
k	Número de vizinhos mais próximos
N	Tamanho da lista de recomendação
$\mathcal{U}$	Conjunto de todos os usuários
$\mathcal{F}$	Conjunto de todos os atributos
$\mathcal{I}$	Conjunto de todos os itens
u, v	Usuários
i, j	Itens
f, g	Atributos
$\mathbf{X}_{M\times N}$, $\mathbf{X}$	Matriz de elementos x_{mn}
$\mathbf{x}_N$, $\mathbf{x}$	Vetor de elementos x_n
$\lvert\mathcal{X}\rvert$	Número de elementos do conjunto $\mathcal{X}$
$\mathbf{R}, r_{ui}$	Avaliação do item i pelo usuário u
$\mathbf{A}, a_{if}$	Descrição numérica do atributo f presente no item i
$\mathbf{T}, t_{uf}$	Correlação entre usuário u e atributo f
$\mathbf{w}, w_f$	Peso do atributo f
$\mathbf{W}, w_{uf}$	Correlação ponderada entre usuário u e atributo f
$\mathbf{S}, s_{ij}$	Similaridade entre itens i e j

Referências

1 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 3.

2 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 3 vezes nas páginas 3, 7 e 8.

3 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 2 vezes nas páginas 4 e 6.

4 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 4 vezes nas páginas 4, 17, 19 e 21.

5 BALABANOVIC, M.; SHOHAM, Y. Fab: Content-based, collaborative recommendation. *Communications of the ACM*, v. 40, p. 66–72, 1997. Citado na página 4.

6 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 5.

7 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 5.

8 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 4 vezes nas páginas 5, 6, 7 e 8.

9 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 2 vezes nas páginas 5 e 17.

10 BURKE, R. Hybrid web recommender systems. In: *The adaptive web*. [S.l.]: Springer, 2007. p. 377–408. Citado na página 5.

11 LEE, J.; SUN, M.; LEBANON, G. A comparative study of collaborative filtering algorithms. *arXiv preprint arXiv:1205.3193*, 2012. Citado na página 6.

12 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-Amazon-for-a-Seamless-Buying-Experience>. Citado 2 vezes nas páginas 6 e 12.

13 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 2 vezes nas páginas 6 e 12.

14 LOPS, P.; GEMMIS, M. de; SEMERARO, G. In: *Recommender Systems Handbook*. [S.l.]: Springer, 2011. Citado na página 6.

15 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 8.

16 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 9.

17 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 12.

18 SARWAR, B. M. et al. Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering. In: CITESEER. [S.l.], 2002. Citado na página 12.

19 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 2 vezes nas páginas 17 e 20.

20 FENNELL, J. Collaborative filtering on sparse rating data for yelp. com. 2009. Citado na página 18.