

**UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA  
FILHO”  
FACULDADE DE CIÊNCIAS  
DEPARTAMENTO DE COMPUTAÇÃO  
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO**

**MATHEUS CESÁRIO MELLO DOS SANTOS**

**SISTEMA DE RECOMENDAÇÃO DE PRODUTOS ONLINE BASEADO  
EM PALVRAS-CHAVE**

**BAURU  
2014**

MATHEUS CESÁRIO MELLO DOS SANTOS

**SISTEMA DE RECOMENDAÇÃO DE PRODUTOS ONLINE BASEADO  
EM PALVRAS-CHAVE**

Orientadora: Profa. Dra. Simone das Graças Domingues do Prado

Trabalho de Conclusão de Curso  
apresentado ao Departamento de  
Computação da Faculdade de  
Ciências da Universidade Estadual  
Paulista “Júlio de Mesquita Filho”,  
para obtenção de título de Bacharel  
em Ciência da Computação.

BAURU  
2014



Santos, Matheus

Sistema de recomendação de produtos online baseado em palavras-chave / Matheus Cesário Mello dos Santos.- 2014.

55 f. :il.

Monografia (Bacharelado em Ciência da Computação) – Universidade Estadual Paulista “Júlio de Mesquita Filho”, Faculdade de Ciências, Bauru, 2014.

Orientação: Profa. Dra. Simone das Graças Domingues do Prado

1. Sistemas de recomendação. 2. Sistema baseado em conteúdo. 3. recomendação de produtos. Assunto. I. Santos, Matheus. II. Universidade Estadual Paulista “Júlio de Mesquita Filho”. III. Sistema de recomendação de produtos online baseado em palavras-chave.

MATHEUS CESÁRIO MELLO DOS SANTOS

**SISTEMA DE RECOMENDAÇÃO DE PRODUTOS ONLINE BASEADO  
EM PALVRAS-CHAVE**

Monografia apresentada junto à disciplina Projeto e Implementação de Sistemas II, do curso de Bacharelado em Ciência da Computação, Faculdade de Ciências, Universidade Estadual Paulista “Júlio de Mesquita Filho”, campus de Bauru, como parte do Trabalho de Conclusão de Curso.

**BANCA EXAMINADORA**

Profa. Dra. Simone das Graças Domingues do Prado  
Departamento de Computação – Faculdade de Ciências  
UNESP – Campus Bauru

Prof. Dr. Eduardo Martins Morgado  
Departamento de Computação – Faculdade de Ciências  
UNESP – Campus Bauru

Prof. Dr. Ivan Rizzo Guilherme  
Departamento de Estatística, Matemática Aplicada e Computação  
Instituto de Geociências e Ciências Exatas  
UNESP – campus de Rio Claro

Bauru, 17 de junho de 2014

Dedico este trabalho aos meus pais Josias e Josimara e minha irmã Barbara, que sempre acreditaram na minha capacidade e no meu potencial, à minha namorada Carolina pela motivação e apoio incondicional, à minha orientadora Profa. Dra. Simone das Graças Domingues do Prado e demais professores pelo apoio e incentivo ao longo de minha jornada na Universidade, à todos os meus amigos e àqueles que sentem paixão pelo que fazem.

Agradecimentos à minha orientadora Profa. Dra. Simone das Graças Domingues do Prado pelo apoio técnico, e à Felipe Cabral Minutti pelo auxílio na parte visual da aplicação.

## RESUMO

Este projeto tem como objetivo apresentar os diversos métodos utilizados para o desenvolvimento de sistemas de recomendação de itens para usuários e aplicar o método de recomendação baseado em conteúdo a um protótipo de sistema que tem por finalidade recomendar livros para usuários. Este trabalho expõe os métodos mais populares para a criação de sistemas capazes de oferecer itens (produtos) de acordo com o gosto do usuário, como filtragem colaborativa e o baseado em conteúdo. Também salienta técnicas que podem ser aplicadas para o cálculo da similaridade entre duas entidades, sendo estas, itens ou usuários, como o método de Pearson, cálculo do Cosseno de vetores e mais recentemente, uma proposta de utilização de um sistema bayesiano sob uma distribuição de Dirichlet. Este trabalho também possui o propósito de passar por diversos pontos sobre a concepção de uma aplicação *online*, ou seja, um *website*, tratando não só de questões orientadas à algoritmos, mas também a definição de ferramentas de desenvolvimento e técnicas de melhoramento da experiência do usuário. As ferramentas utilizadas para o desenvolvimento da página são listadas, e um tópico sobre *design* de *layout* também é discutido com a finalidade de enfatizar a importância do *layout* sobre a aplicação. Ao final, alguns exemplos de utilização de sistemas de recomendação são discutidos a fim de apresentar situações de aplicações.

**Palavras-chave:** sistemas de recomendação; sistema baseado em conteúdo; *website*; *user experience*; recomendação de produtos.

## ABSTRACT

This project aims to explore the many methods used for the development of recommendation systems to user's items and apply the content-based recommendation method on a prototype system whose purpose is to recommend books to users. This paper exposes the most popular methods for creating systems capable of providing items (products) according to user preferences, such as collaborative filtering and content-based. It also point different techniques that can be applied to calculate the similarity between two entities, for items or users, as the Pearson's method, calculating the cosine of vectors and more recently, a proposal to use a Bayesian system under a Dirichlet distribution. In addition, this work has the purpose to go through various points on the design of an online application, or a website, dealing not only oriented algorithms issues, but also the definition of development tools and techniques to improve the user's experience. The tools used for the development of the page are listed, and a topic about web design is also discussed in order to emphasize the importance of the layout of the application. At the end, some examples of recommender systems are presented for curiosity, learning and research purposes.

**Keywords:** recommender systems; system based on content; website; user experience; product recommendation.

## LISTA DE FIGURAS

Figura 1 - Relacionamento entre usuários e itens.....	13
Figura 2 - Recomendação baseado nas interações do usuário.....	14
Figura 3 - Distribuição cauda-longa Popularidade x Especificidade .....	16
Figura 4 - Arquitetura de alto-nível de um sistema de recomendação baseado em conteúdo..	20
Figura 5 – Técnicas de recomendação e suas fonte de dados.....	21
Figura 6 - Diagrama que ilustra o fluxo do processo no modelo MVC .....	28
Figura 7 – Relacionamento entre um item e sua característica. ....	29
Figura 8 - Conjunto de características e seus relacionamentos com itens da base de dados....	30
Figura 9 - Tipos de interações que um usuário pode executar em uma aplicação .....	31
Figura 10 - Interação de um usuário com outro usuário, um item ou item de um usuário.....	32
Figura 11 - Exemplo de grafo.....	32
Figura 12 - Grafo direcionado. O valor da aresta (A,B) pode ser diferente de (B,A) .....	33
Figura 13 - Página inicial. ....	35
Figura 14 - Página de visualização de um item .....	36
Figura 15 - Página pessoal do usuário. ....	36
Figura 16 - Perfil do usuário.....	37
Figura 17 - Uma aplicação responsiva adequa seu layout para diversos tipos de dispositivos. .....	38
Figura 18 - Visão do protótipo em resolução reduzida (um smartphone por exemplo).....	39
Figura 19 - Relacionamento entre itens sendo estabelecidos ao longo do tempo, pela interação dos usuários. ....	42
Figura 20 - Grafo exemplificando problema do item mais popular e os itens mais específicos (difíceis de encontrar).....	43

## **LISTA DE TABELAS**

Tabela 1 - Etapas do projeto .....	11
Tabela 2 - Desafios de um sistema de recomendação .....	16
Tabela 3 - Itens, suas características e relevâncias.....	33

## SUMÁRIO

1. Introdução.....	9
2. Fundamentação teórica.....	13
2.1. O sistema de recomendação.....	13
2.1.1. Recomendações colaborativas.....	18
2.1.2. Baseado em conteúdo.....	18
2.1.3. Métodos híbridos.....	20
2.1.4. Modelo de cluster.....	21
2.2. Cálculo de similaridade entre itens.....	22
2.2.1. Coeficiente de similaridade de Jaccard.....	23
2.2.2. Distância Euclidiana.....	24
2.2.3. Método do cosseno.....	24
2.2.4. Método de pearson.....	24
2.2.5. Similaridade baseado em uma distribuição de diriclhett.....	25
2.3. Ferramentas computacionais.....	26
2.3.1. MySQL (Structured Query Language).....	26
2.3.2. PHP (Hypertext Preprocessor).....	27
2.3.3. Javascript.....	27
2.3.4. CSS (Cascading Style Sheets).....	27
2.3.5. HTML (HyperText Markup Language).....	28
2.3.6. MVC (Model-View-Controller).....	28
3. Sobre a aplicação.....	29
3.1. Modelo.....	29
3.2. Controle.....	33
3.3. Visão.....	35
4. Estudo de caso.....	40
4.1.1. Amazon.....	40
4.1.2. Goodreads.....	41
5. Conclusões.....	42
REFERÊNCIAS BIBLIOGRÁFICAS.....	48

## 1. INTRODUÇÃO

Diariamente é gerada na internet uma gigantesca quantidade de informações, dos mais variados tipos. Esta produção exagerada torna a mineração de dados extremamente laborosa, e conseqüentemente, a busca por informações mais concisas ou específicas têm se tornado cada vez mais complicada.

“I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A "Semantic Web", which makes this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The "intelligent agents" people have touted for ages will finally materialize.”<sup>1</sup> (BERNERS-LEE e FISCHETTI, 1999, cap. 12).

Os mecanismos de busca atuais logo não serão mais capazes de varrer tantas informações que não possuem significados ou conexões com outras informações, e isso prejudicará a qualidade com que os resultados serão exibidos para o usuário. Partindo deste princípio, algumas iniciativas passaram a buscar uma nova abordagem para análise de dados na internet. Estes esforços se tornaram, o que hoje é chamado de, Web 3.0.

Dentro deste volume crescente de informações, pode-se observar o crescimento de plataformas com bases de dados gigantescas, tanto de produtos quanto de usuários, como por exemplo as páginas de comércio eletrônico, ou *e-commerces*. Grandes *e-commerces* possuem bases enormes de produtos e suas descrições, e este volume acarreta problemas na hora de oferecer conteúdo a seus clientes. Foi pensando nisto que diversas soluções foram propostas, com a finalidade de extrair informações precisas e apresentá-las de forma estratégica visando aumentar as oportunidades de lucro. Assim, surgiram sistemas capazes de coletar informações sobre clientes e oferecer um conteúdo correspondente aos seus interesses sem que ele necessite buscar por estas informações. Um sistema de recomendação de produtos é um conjunto de métodos que visam analisar um perfil de determinado usuário e a partir destas informações, buscar conteúdo que tenha alguma semelhança com as características compartilhadas. Além da oportunidade de aumento de vendas, estes sistemas também podem ser aplicados em diversas outras áreas, com o intuito de auxiliar as buscas dos usuários, oferecendo produtos antes mesmo dele procurar. Assim, este tipo de sistema age como um suporte à tomada de decisão.

Desta forma, o objetivo deste trabalho é propor um sistema de recomendação de produtos baseado em conteúdo. O sistema deverá ser capaz de analisar o perfil de um usuário e auxiliá-lo na busca por produtos dentro de um *website*, oferecendo-lhe itens de possível interesse.

Os sistemas de recomendação de produtos podem utilizar diversas técnicas, dentre elas a filtragem colaborativa, recomendação baseada em conteúdo e a recomendação baseada em *clusters*. Este trabalho discutirá a ideia principal de cada técnica, e demonstrará, por meio de uma aplicação, como a técnica de recomendação baseada em conteúdo funciona dentro de um ambiente, em prol da obtenção do melhor conteúdo possível para os usuários. Além disso, o

---

<sup>1</sup> “Eu tenho um sonho onde a Web [e os computadores conectados a ela] serão capazes de analisar todos os dados da Web – conteúdo, links, e transações entre pessoas e computadores. A “Web semântica”, a qual torna tudo isso possível, ainda não emergiu, mas quando acontecer, todos os mecanismos de troca, burocracia e nossas atividades diárias serão todas gerenciadas por máquinas conversando com máquinas. Os tão prometidos “agentes inteligentes” se tornará realidade”, (Tradução nossa).

projeto também abrange outras questões, como o *web design* e a engenharia de *software*, e visita também diversas ferramentas e métodos para o desenvolvimento de um *website* moderno de caráter comercial.

O objeto utilizado como item de recomendação neste trabalho gira em torno do universo dos livros. A ideia de criar um sistema capaz de recomendar livros, autores e gêneros partiu da dificuldade que os usuário têm para encontrar informações precisas devido ao grande volume de dados vigente na internet. Este projeto vem com o propósito de oferecer uma ferramenta de suporte à busca de informações, tendo como intuito apresentar itens baseados nas ações anteriores e nos interesses vindos do perfil de um usuário.

De acordo com o Google<sup>2</sup>, existem cerca de 130 milhões de livros espalhados pelo mundo. Isto quer dizer que há muita informação a ser varrida, analisada e apresentada (PARR, 2010). Assim, a preocupação com o desenvolvimento de ferramentas que possam auxiliar as pessoas a buscarem as melhores opções de acordo com seus interesses se faz evidente. Em contrapartida, uma ferramenta como essa também oferece informações valiosas sobre quais são os títulos mais buscados e qual a relação que um objeto (livro, autor, gênero, etc) têm com outro objeto. Este cálculo que determina a similaridade entre dois objetos ajuda na classificação das obras e seus autores, e pode vir a se tornar objeto de pesquisa e aplicação tanto no âmbito comercial, por meio do desenvolvimento de uma página de comércio eletrônico como a Amazon<sup>3</sup>, quanto no âmbito científico, por meio da catalogação das obras, como o Goodreads<sup>4</sup> e o Google Books<sup>5</sup>.

A motivação para este trabalho parte da dificuldade que muitos usuários têm em decidir quais obras eles devem ler. Esta dificuldade advém das mais variadas causas, entre elas a dificuldade de encontrar obras que tenham características que os interessem, o preconceito de iniciar uma nova leitura ou até mesmo a falta de informações sobre outras obras.

A partir daí, surgiu a vontade de criar um sistema que fosse capaz de suprir essa necessidade e solucionar estes problemas ao oferecer itens similares aos interesses das pessoas.

A opção pelo universo da leitura vem da paixão que o autor deste trabalho tem para com os livros, e pela curiosidade em desenvolver um sistema que fosse capaz de auxiliar usuários na escolha dos mesmos.

Devido ao grande volume de informações ao qual somos bombardeados constantemente, um sistema que ofereça suporte a essas buscas e seja capaz de recomendar livros, autores e gêneros antes mesmo de iniciar a procura pode ser de vital importância para a aplicação, pois evita a evasão de usuários devido à exaustão da pesquisa por produtos específicos. Além do mais, a propaganda direcionada e personalizada ajuda a manter os clientes ativos, apresentando sempre novos conteúdos cujo usuário possa acessar, e em casos de aplicações comerciais, representa um ponto vital pois pode garantir um aumento na venda de produtos.

---

<sup>2</sup> Companhia americana de tecnologia cujo um dos projetos, o Google Books, tem o propósito de digitalizar todas as obras já escritas no mundo. Site: <http://www.google.com.br/googlebooks/about/>

<sup>3</sup> Companhia americana especializada em comércio eletrônico. Site: <http://www.amazon.com/>

<sup>4</sup> Site que oferece um serviço de classificação e avaliação de livros. Site: <http://www.goodreads.com/>

<sup>5</sup> Projeto que oferece livros digitalizados para o público. Site: <http://books.google.com/>

Por isso, unindo o interesse pela inteligência artificial, mineração de dados e desenvolvimento de projetos, algoritmos e a paixão pelos livros, este trabalho vem como um protótipo de solução para este problema, abordando cada tópico de forma isolada.

Inicialmente este trabalho foi dividido em 3 etapas diferentes, que são as seguintes: estudo sobre a teoria a ser aplicada e embasamento científico; estudo sobre as ferramentas para o desenvolvimento e prova do método; desenvolvimento da aplicação (união das duas primeiras etapas).

A primeira etapa ocorreu de forma breve no início, tomando maior força nas etapas críticas do projeto, ou seja, durante a implementação do algoritmo e o desenvolvimento da estrutura de dados. Este período compreende o tempo destinado às pesquisas sobre o funcionamento de um sistema convencional de recomendação e sua forma de atuação sobre os interesses do usuário. Estudos sobre estruturas de dados, ontologias, redes neurais e mineração de dados também foram necessários para melhor entendimento e domínio do assunto principal tratado.

A etapa dois se deu ao longo de todo o projeto e envolveu estudos sobre as linguagens de programação PHP, Javascript, CSS e HTML para a produção da página de internet. O modelo de arquitetura de *software* aplicado durante o processo de produção foi o MVC (Modelo Controle e Visão). O modelo MVVM (Modelo Visão Visão Modelo) também foi considerado durante a concepção da arquitetura do sistema.

Ao final, a etapa três envolveu a união de todos os conceitos estudados durante as etapas anteriores para a criação de um sistema online de recomendação de produtos baseado em palavras-chave. Abaixo, a tabela 1 mostra como as etapas foram divididas ao longo do tempo de concepção e produção do projeto.

Tabela 1 - Etapas do projeto

- a) Fundamentação e pesquisas sobre a área;
- b) Pesquisa sobre conceitos de comércio eletrônico e dados mercadológicos;
- c) Desenvolvimento da arquitetura a ser aplicada;
- d) Desenvolvimento da arquitetura da aplicação em conjunto baseado na arquitetura definida;
- e) Desenvolvimento da aplicação (4 meses);
- f) Exposição de dados e informações;
- g) Entrega do projeto;

Durante as pesquisas, a vontade inicial era a criação de uma base de conhecimento consistente que oferecesse recomendações precisas de acordo com o perfil de um determinado usuário. Ao longo do tempo, após a construção da arquitetura de software inicial, optou-se por restringir certas etapas do projeto e focar somente em um módulo. Assim, o plano escolhido foi o de desenvolver um módulo de recomendação baseado em um conjunto de características (ou interesses) de um usuário. A escolha do universo literário como objeto a ser recomendado (produtos) se deu devido à grandes possibilidades que este universo permitiria para o projeto, como a alta gama de informações sobre diversas obras literárias na internet; o interesse de diversas pessoas em livros, autores, gêneros, entre outros; o valor educacional que esta aplicação agrega, ao manipular informações que podem proporcionar curiosidade e vontade de aprender; o grande número de projetos que têm como objetivo a divulgação e difusão de informações por meio de livros em formatos digitais; a liberdade para extrair e apresentar

informações sobre estes itens; além do fato dos livros representarem uma grande paixão do autor deste trabalho.

“Uma boa parte da ação dos organismos educativos e de fomento científico e tecnológico deveria estar encaminhada a cobrir a necessidade de disseminar em toda a população o conhecimento disponível, através de programas tanto de formação como de informação de longo alcance.” (LAUFER).

## 2. FUNDAMENTAÇÃO TEÓRICA

Como mencionado no capítulo 1, as duas primeiras etapas do desenvolvimento deste trabalho foram a pesquisa sobre os métodos a serem utilizados e o estudo sobre as ferramentas adequadas para a implementação da ideia. Uma ênfase maior para cada ferramenta pesquisada e utilizada será dada ao longo deste capítulo, explicando também o porquê da escolha de um instrumento em detrimento de outro.

### 2.1. O SISTEMA DE RECOMENDAÇÃO

Durante o desenvolvimento da aplicação, convenhou-se utilizar para este trabalho a denominação “objeto” como toda entidade que está inserida no sistema. Uma entidade é um objeto que possui informações e é descrito por meio de palavras-chave. Essas palavras-chave são as características da entidade e por meio de um peso pré-definido, que representa o valor que a característica correspondente tem em relação ao objeto que caracteriza, é possível determinar uma relação entre dois objetos. Isso possibilita a exibição de itens semelhantes aos que o usuário busca e, possivelmente, a apresentação de um conjunto de itens que possam interessá-lo. A figura 1 ilustra esse relacionamento entre produtos e usuários:

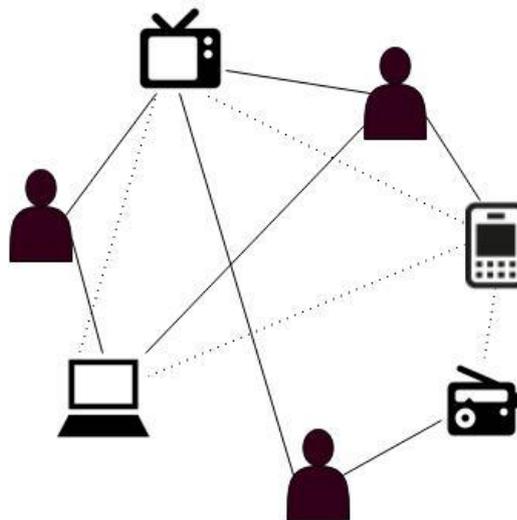


Figura 1 - Relacionamento entre usuários e itens

O usuário também é descrito como uma entidade, ou ainda, um objeto, porém suas características se dão de forma implícita. Isto quer dizer que uma entidade do tipo usuário tem como características os atributos dos itens o qual acessa e atua sobre. Estes itens podem ser páginas que o descrevem por completo ou marcações (*links*) presentes nos perfis de outros usuários e suas formas de atuação podem ser avaliações, comentários, adições à sua própria lista de marcações, entre outros. Assim, se baseando nestas interações, novas opções podem surgir para o usuário, como ilustra a figura 2:

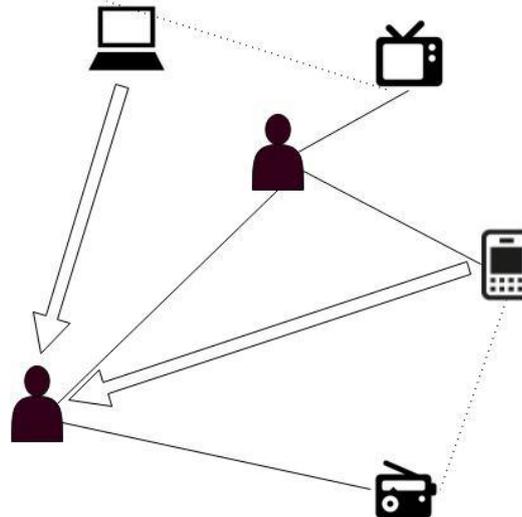


Figura 2 - Recomendação baseado nas interações do usuário

Na aplicação, os objetos inseridos foram divididos em dois grupos: atuantes e não-atuantes. Um objeto atuante é aquele que pratica alguma ação dentro do ambiente, ou seja, é uma entidade que atua diretamente sobre o cenário ao qual está inserido. Baseado nesta classificação tem-se o usuário. O usuário atua sobre a aplicação de forma direta e suas ações e decisões alteram a estrutura de classificação do sistema, que em resposta, exhibe itens que tenham certa possibilidade de serem acessados por ele. Ao longo da interação entre o utilizador e o sistema, as características que definem o usuário serão refinadas, o que pode auxiliar o sistema a determinar qual item é o mais adequado para um usuário, baseado em seu perfil de preferências.

Por conseguinte, objetos não-atuantes são aqueles que não alteram o ambiente de forma explícita. Nesta classificação estão inseridos os itens de escolha do usuário, no caso deste trabalho foram utilizados como tema livros, autores, gêneros, etc. Os itens não-atuantes são produtos os quais os usuários estarão buscando. A escolha da terminologia não-atuante decorre do fato de que estes objetos não podem agir sobre outros itens diretamente, como os usuários fazem ao avaliar um livro, comentar sobre uma página ou ainda adicionar outros itens em suas listas de escolhas. Um objeto não-atuante possui características explícitas nomeadas palavras-chave que foram pré-definidas e representam os atributos que descrevem o objeto em questão. A forma de atuação destes objetos é de forma implícita: as características do objeto atraem a atenção ou a busca de um usuário, que então atuará sobre ele, gerando informações que serão utilizadas pelo sistema de classificação.

Assim, ao longo do trabalho, fica definido que a distinção entre os objetos do sistema seriam denominados usuários (objetos atuantes) e itens (objetos não atuantes, ou ainda, produtos) para evidenciar de forma mais enfática a diferença de papéis de cada entidade envolvida na aplicação.

Um sistema de recomendação é um sistema que utiliza informações do usuário, comparando com as informações de produtos de uma base de dados para oferecer produtos que se encaixem com seus interesses ou seu histórico de itens já acessados de alguma forma. No

âmbito dos *e-commerces*<sup>6</sup>, este sistema pode recomendar itens semelhantes aos que o cliente busca e também oferecer opções baseado nas compras efetuadas anteriormente (análise do carrinho de compras). No âmbito de um site de recomendação de filmes, o sistema analisa as características do usuário para oferecer opções de filme cujo gênero o interesse. De forma análoga, o aplicativo desenvolvido neste trabalho tem como funcionalidade principal analisar as características de cada item, neste caso livros, e, oferecer novas opções de leitura observando as informações extraídas de seu perfil.

Um dos benefícios que um sistema de recomendação oferece é a possibilidade de apresentar itens com maior pertinência em relação à sua busca. Vale destacar também que devido à infinidade de informações dentro da base de dados, o direcionamento de objetos que ofereçam maior semelhança com os gostos do usuário é de suma importância. Uma vez que a busca se torna mais personalizada, a chance do usuário optar por mais itens (no caso de uma página *e-commerce* que oferece produtos para venda) representa um possível ganho financeiro para a empresa, enquanto que o próprio cliente, no caso o usuário, tem sua permanência na busca diminuída, poupando-lhe tempo. Outro motivo que torna estes sistemas necessários é o alcance que eles podem oferecer à busca do usuário. Uma loja de produtos pode oferecer milhares de itens, porém nem todos estarão à vista dos clientes. Basicamente, os itens mais populares da loja física serão os mais visitados e possivelmente, comprados, enquanto os menos conhecidos serão relegados ao anonimato. Ao contrário do que acontece com a loja física, uma loja virtual pode oferecer uma vasta gama de itens, e milhares de itens podem ser visualizados pelos clientes em um período de tempo muito menor. Esta visibilidade que as lojas virtuais oferecem permite aos clientes a descoberta de novos itens e interesses. O compromisso do sistema de recomendações se faz imprescindível neste momento, de forma a executar o melhor serviço, oferecendo itens de interesse ao usuário dentro do sistema.

Este fenômeno pode ser ilustrado pela chamada teoria da cauda-longa (*long-tail*). Cauda-longa é um termo utilizado em estatística para identificar distribuições de dados de forma decrescente e foi popularizada por Chris Anderson<sup>7</sup>, partindo de um artigo de Clay Shirky, “*Power Laws, Weblogs and Inequality*” (SHIRKY, 2003). Aplicando o diagrama ao conceito de exibição de itens por lojas físicas e virtuais, tem-se que as lojas físicas oferecem em sua maioria os itens que são populares, no caso dos livros, os *best-sellers* e obras famosas. Em contrapartida, lojas virtuais conseguem oferecer toda uma variedade de produtos; além dos *best-sellers*, apresentam diversos outros itens. A figura 3 é um gráfico que elucida este efeito:

---

<sup>6</sup> Página de internet que realiza transações comerciais efetuadas de forma eletrônica.

<sup>7</sup> Escritor Americano. Publicou o livro “*The Long Tail: Why the Future of Business Is Selling Less of More*” (“Cauda Longa: por que o futuro dos negócios é vender pouco de muito”, Tradução nossa), que trata de relações de comércio na Internet (ANDERSON, 2014).



Figura 3 - Distribuição cauda-longa Popularidade x Especificidade

Porém um empecilho quanto a estes sistemas é a grande dificuldade de mantê-los, uma vez que, na maioria das vezes, a estrutura de dados é representada por uma matriz esparsa, ou seja, uma matriz que possui poucos valores calculados (diferentes de zero). Desta forma a dificuldade aumenta cada vez mais, tanto para manter dados sobre novos usuários e itens quanto para usuários mais experientes, que possuem muitos interesses já registrados. O esquema a seguir (tabela 2) mostra as fontes de extração de dados que um sistema pode utilizar para análise, ilustrando os pontos que devem ser observados quando na determinação do perfil usuário.

Tabela 2 - Desafios de um sistema de recomendação

Desafios de um sistema de recomendação	
Capturar comportamento em um site de vendas	Item visitado, Item visitado e recomendado Item avaliado, Item adicionado ao carrinho de compras, Item adicionado aos favoritos, Item adicionado a lista de presentes, Item comprado
Capturar comportamento em um site de conteúdo	Conteúdo visitado, Conteúdo visitado recomendado, Conteúdo revisado, Conteúdo buscado, Conteúdo baixado
Adquirindo recomendações	Adquirindo recomendação de um item Adquirindo recomendação de um usuário Justificando a recomendação
Início e fim de uma sessão do usuário	Início da sessão, Fim da sessão
Avaliar itens	Avaliar um item, Exibir avaliação do item
Classificar usuários e itens	Classificar usuário, Selecionar classificação atribuída ao usuário, Classificar item, Selecionar classificação atribuída ao item

Além do grande volume de dados que devem ser analisados para oferecer um bom serviço, é necessário ponderar sobre gerar recomendações a partir de um item-frio, ou seja, um usuário recém-chegado, o qual não possui nenhuma característica registrada, ou um item adicionado recentemente, o qual não possui nenhuma nova avaliação feita. É necessário lidar com o outro extremo desta situação, ou seja, um usuário experiente que possui milhares de

recomendações, ou um item detentor de diversas recomendações e características. Devido a estes cenários, sistemas complementares ou alternativas mais robustas como inteligência artificial e mineração de dados são necessários para oferecer um serviço de qualidade e não-redundante.

Em (LINDEN, SMITH e YORK, 2003), são discutidas dificuldades que os *e-commerces*, páginas de internet de vendas de produtos, têm ao manipular um volume grande e disperso de dados:

“A large retailer might have huge amounts of data, tens of millions of customers and millions of distinct catalog items. Many applications require the results set to be returned in realtime, in no more than half a second, while still producing high-quality recommendations; new customers typically have extremely limited information, based on only a few purchases or product ratings; older customers can have a glut of information, based on thousands of purchases and ratings; customer data is volatile: each interaction provides valuable customer data, and the algorithm must respond immediately to new information.”<sup>8</sup>, (LINDEN, SMITH e YORK, 2003, p. 76).

A não-redundância é ato de sempre apresentar ao usuário itens novos. Geralmente isso é feito analisando os itens acessados (comprados, adicionados, etc), removendo-os do conjunto ou substituindo-os por seus itens semelhantes, em uma situação chamada semelhança item-para-item.

Xavier Amatriain destaca que três aspectos devem ser considerados ao desenhar um sistema de recomendação: “Similarity or distance measures; issue of sampling as a way to reduce the number of items in very large collections while preserving its main characteristics and reduce dimensionality”<sup>9</sup> (AMATRIAIN, *et al.*, 2011, p. 41). Estes aspectos definem em linhas gerais os principais desafios e pontos a serem considerados durante o *design* do sistema de recomendação. Determinar o algoritmo de cálculo de semelhança que melhor se adequa à base de dados; reduzir a base de dados, excluindo características ou itens redundantes ou desnecessários, restringindo a base de dados às características essenciais de cada item, reduzindo a dimensão do conjunto de similaridades e determinar o melhor método para seleção de amostras de itens do conjunto para a recomendação (apresentação do item ao usuário).

Também outro ponto importante a salientar é o fato de os sistemas de recomendação terem o âmago de sua funcionalidade baseada no fato de utilizarem dados pessoais do usuário. De acordo com Paul Resnik e Val R. Varian, a violação de privacidade é um ponto crucial a se analisar quando se pensa em um sistema de recomendação:

“People often use recommender systems to make decision. Based on recommendations by other individuals or authorities, choices can be made even without adequate first-hand knowledge of the

---

<sup>8</sup> “Um grande varejista pode ter enormes quantidades de dados, dezenas de milhões de clientes e milhões de itens de catálogo distintas. Muitas aplicações requerem o conjunto de resultados a serem retornados em tempo real, em não mais do que meio segundo, enquanto continua produzindo recomendações de alta qualidade; novos clientes normalmente têm informações extremamente limitadas, com base em apenas algumas compras ou classificações de produtos; clientes mais velhos podem ter um excesso de informações, baseado em milhares de compras e avaliações; os dados do cliente são voláteis: cada interação fornece dados de clientes valiosos, e o algoritmo deve responder imediatamente às essas novas informações”, (Tradução nossa).

<sup>9</sup> “Medidas de similaridade ou distância; amostragem, visando reduzir o número de itens em grandes coleções, preservando suas características principais, e a dimensionalidade destas coleções”, (Tradução nossa).

alternatives. However, recommender systems raise certain social problems, including those that relate to the issues of incentives for generating recommendations and of personal privacy. Another consideration is the high cost of maintaining such a system.”<sup>10</sup> (RESNICK e VARIAN, 1997, p. 1).

Um mecanismo de recomendação pode ser construído seguindo diversas diretrizes, dentre elas os sistemas baseados em conteúdo, filtragem colaborativa ou ainda, uma mistura das anteriores em um modelo híbrido. Outra estratégia também utilizada é o modelo de *clusters*<sup>11</sup>, que pode ser aplicado tanto para os usuários quanto para os itens.

### 2.1.1. RECOMENDAÇÕES COLABORATIVAS

A filtragem colaborativa filtra informações baseadas nas avaliações anteriores dos usuários sobre outros itens. Este método parte do princípio de se um primeiro usuário avaliou um item, um segundo usuário que tenha avaliado os mesmos itens que o primeiro, exceto este, terá uma probabilidade de se interessar por este item, e conseqüentemente comprá-lo ou avaliá-lo. A abordagem deste método difere do baseado em conteúdo na questão da análise de similaridades: enquanto o método de filtragem colaborativa busca comparar o perfil de usuários, sendo avaliações, comentários ou compras, o método de filtragem baseado em conteúdo busca classificar os itens, oferecendo ao usuário itens semelhantes aos que ele já se interessou. Uma vantagem da filtragem colaborativa é o fato de o sistema não precisar de um método de análise e classificação de conteúdo. Em linhas gerais, isto quer dizer que o sistema é capaz de recomendar um item sem precisar conhecer seus detalhes. Uma vez que ele mede as informações do perfil do usuário com a de outros usuários do sistema, atribuir características para calcular a semelhança torna-se uma tarefa dispensável. Esta aproximação de recomendação por filtragem colaborativa é amplamente utilizada atualmente por mídias sociais<sup>12</sup>, como Facebook<sup>13</sup>, LinkedIn<sup>14</sup> e MySpace<sup>15</sup> para recomendar amigos, páginas ou grupos. Para recomendação de produtos, tem-se como exemplo o site Netflix, de aluguel de vídeos como séries, filmes e documentários; e a Amazon, site de venda de produtos diversos.

### 2.1.2. BASEADO EM CONTEÚDO

Um sistema de recomendações baseado em conteúdo tem como foco definir as características de cada item dentro do conjunto e comparar com o perfil do usuário. Cada item possui uma série de aspectos que podem classificá-lo dentro de determinados gêneros, preços, tipos, cores

<sup>10</sup> “As pessoas costumam usar sistemas de recomendação para tomar uma decisão. Com base nas recomendações de outras pessoas ou do sistema, as escolhas podem ser feitas mesmo sem adequado conhecimento em primeira mão das alternativas. No entanto, sistemas de recomendação levantam certos problemas sociais, incluindo aqueles que se relacionam com as questões de incentivos para a geração de recomendações e de privacidade pessoal. Outra consideração é o elevado custo de manutenção de um tal sistema”, (Tradução nossa).

<sup>11</sup> “Um grupo de coisas semelhantes ou pessoas posicionados ou que ocorrem em conjunto”, (OXFORD DICTIONARIES, Tradução nossa).

<sup>12</sup> O conceito de mídias sociais (social media) precede a Internet e as ferramentas tecnológicas. Trata-se da produção de conteúdos de forma descentralizada e sem o controle editorial de grandes grupos (SCHIVINSKI e DABROWSKI, 2014).

<sup>13</sup> Facebook é um site de rede social livre que permite aos utilizadores registrados criar perfis, fazer upload de fotos e vídeo, enviar mensagens e manter contato com os amigos. Site: <http://www.facebook.com>

<sup>14</sup> LinkedIn é uma rede social projetada especificamente para a comunidade empresarial. Site: <https://www.linkedin.com/>

<sup>15</sup> MySpace é um site que oferece e-mail, fórum, comunidades, vídeos e espaço weblog. Site: <https://myspace.com/>

entre outros. Estes aspectos podem ter um peso, ou seja, um valor representando a relevância deste atributo em relação ao item. A finalidade deste sistema é reunir todas as características e compará-las com as presentes no perfil do usuário. Acerca deste tipo de procedimento, existem algumas considerações a serem analisadas durante seu desenvolvimento. O primeiro ponto que deve ser analisado é sobre a extração dos dados. É necessário definir qual método fará a extração de informações para classificar cada item da forma mais precisa possível, e isso pode ser feito de forma automática ou manual. Na forma automática, torna-se imperativo a escolha de um algoritmo de recuperação de informações ou de extração de informações que seja capaz de classificar corretamente os atributos dos itens do sistema. Na forma manual, o sistema permite ao usuário, comum ou administrador, classificar os atributos dos itens do sistema. Ainda é possível trabalhar das duas formas, ou seja, permitindo que usuários façam suas avaliações enquanto um sistema complementar, paralelamente, realiza análises sobre os objetos do conjunto (itens).

O segundo ponto a ser considerado é de que as características do item e do usuário devem ter a mesma equivalência, para garantir a integridade das decisões do sistema.

Por fim, o terceiro e último ponto centra-se na extração de informações do perfil do usuário. Um algoritmo de aprendizado deve ser aplicado para que este possa analisar o perfil e apresentar quais características, e as respectivas relevâncias de cada uma, o usuário possui, comparando com a base existente sobre itens, podendo assim executar o processo de diagnóstico, classificação e recomendação item-usuário (item para o usuário). Dentre as possíveis aplicações para esta tarefa, estão a utilização de algoritmos genéticos, redes neurais (RUSSO, 2006), realimentação de relevâncias, sistemas Bayesianos (JIE, GUIBING e YORKE-SMITH, 2013) e técnicas de aprendizado de máquina (AGARWAL e CHEN, 2011).

A figura 4 descreve o processo de um sistema recomendação baseado em conteúdo. Partindo da etapa de coleta de informações (*information source*), o processo de análise de dados (*content analyzer*) irá tratar as informações adquiridas, adequando à estrutura de dados utilizada pelo sistema de recomendação (*profile learner*). Por sua vez, o sistema (*profil learner*) irá comparar as informações coletadas dos usuários com as informações sobre os itens, e então fará a classificação dos usuários para cada perfil, enviando então para o componente de filtragem (*filtering component*), que gerará a lista de recomendações de itens para os usuários, de acordo com o perfil definido na etapa anterior (análise do *profile learner*). O usuário  $u_a$  atuará sobre esta lista de recomendações, e o resultado disso estará na forma *feedback*. Este *feedback* é a resposta de  $u_a$  em relação aos itens sugeridos a ele (itens provenientes de *List of recommendations*). Esta resposta pode ser negativa ou positiva, e estas informações serão enviadas de volta ao sistema de recomendação (*Profile Learner*) que irá tratar estas informações, a fim de refinar o perfil do usuário com o intuito de oferecer itens de forma mais personalizada e precisa, observando também um aumento nos acertos das previsões. Abaixo, pode-se observar o fluxo do processo descrito anteriormente:

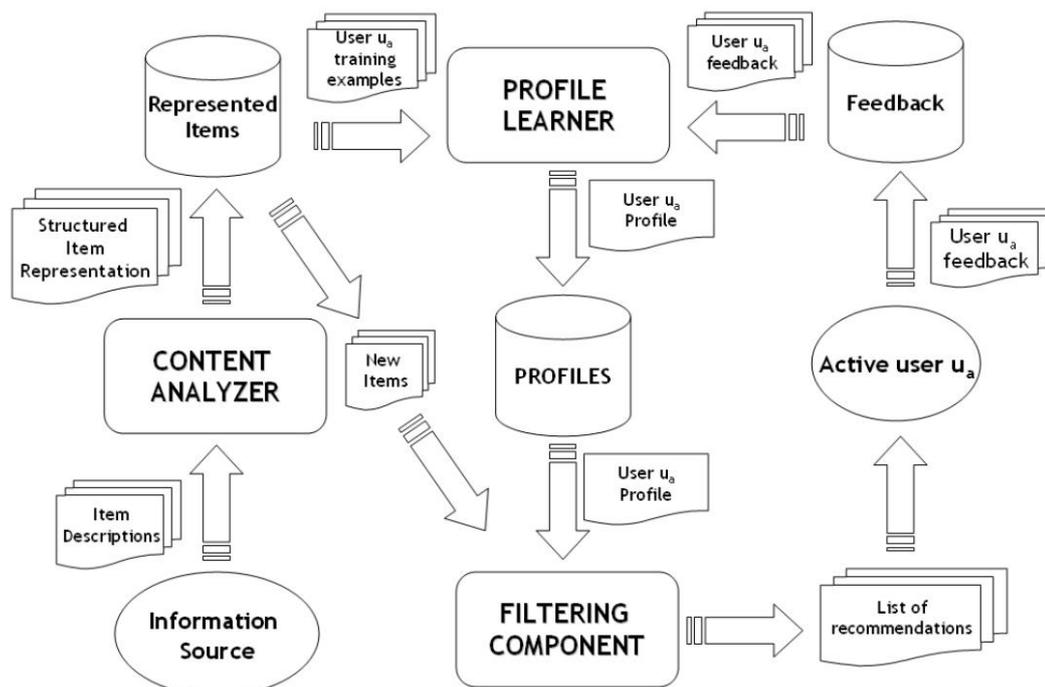


Figura 4 - Arquitetura de alto-nível de um sistema de recomendação baseado em conteúdo.

Fonte: (LOPS, DE GEMMIS e SEMERA, 2012, p. 76).

O modelo escolhido para a aplicação foi o baseado em conteúdo. Baseado em um método de extração e classificação de atributos para cada item, o algoritmo é capaz de criar uma rede de semelhanças entre cada item, e baseado nas experiências anteriores do usuário, apresentar itens que possuam características parecidas com as visitadas anteriormente (ou adicionadas, vistas, avaliadas, etc), apresentando então opções que venham a interessar o usuário.

### 2.1.3. MÉTODOS HÍBRIDOS

Ainda é possível misturar conceitos que utilizam as duas formas de aproximação: classificação de itens e filtragem colaborativa baseada nas avaliações dos usuários. Robin Burke define um sistema híbrido como “[...] any recommender system that combines multiple recommendation techniques together to produce its output”<sup>16</sup> (BURKE, 2010, p. 380). Netflix é um exemplo de aplicação que utiliza ambos os métodos para criar recomendações para seus usuários: além de classificar os itens em diversos atributos e compará-los com os gostos do usuário, o sistema também compara as avaliações do usuário com outros. Dentro desta abordagem, pode-se dividi-la em estratégias diferentes, todas representando o uso de múltiplas formas de recomendação: incremental, ou seja, quando as recomendações geradas pelos sistemas são unidas e calculadas para exibir um único resultado; de sistema alternante, quando as recomendações são utilizadas ora por um sistema, ora por outro; misto, quando os resultados são apresentados juntos (mas não são somados como na estratégia incremental); combinação de resultados, quando os resultados são combinados e enviados para um único

<sup>16</sup> “[...] qualquer sistema de recomendação que combina várias técnicas de recomendação em conjunto para produzir sua saída”, (Tradução nossa).

sistema de recomendação; de acréscimo, quando um resultado de um sistema interfere na relevância (valor) do resultado de outro sistema; sistema de cascata, onde os algoritmos geram recomendações baseado em prioridades pré-definidas organizando a apresentação dos resultados; e de metadados, onde um sistema gera um certo modelo de dados que será então utilizado por outro sistema de recomendação.

A figura 5 é um diagrama que apresenta diversos métodos de recomendação e as bases de informação que utilizam.

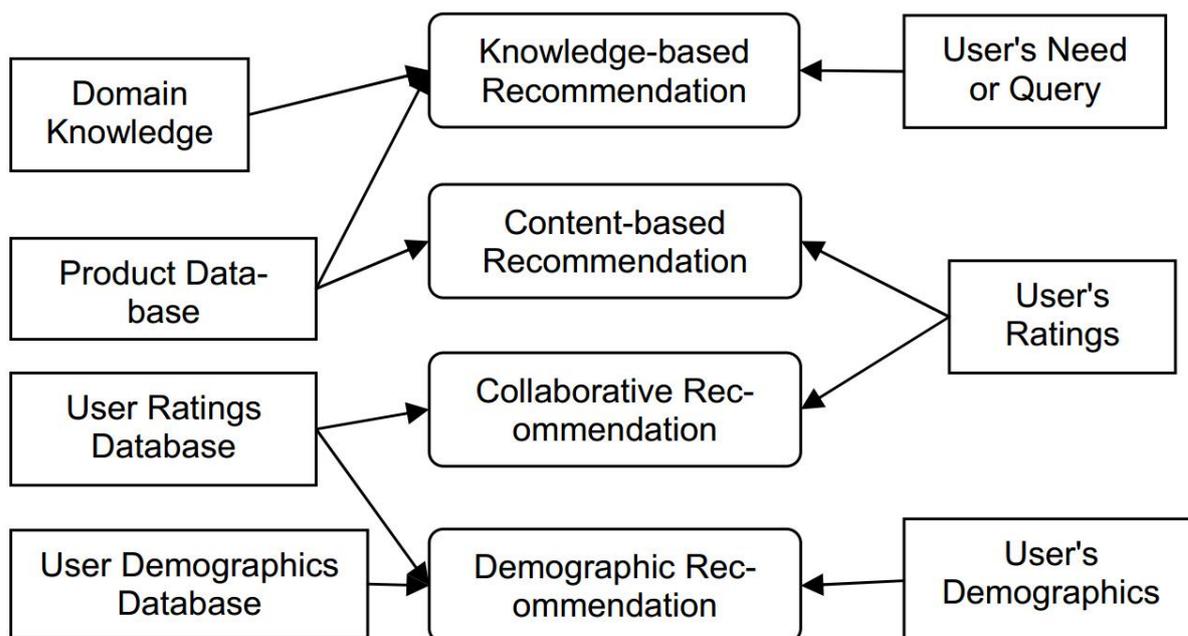


Figura 5 – Técnicas de recomendação e suas fonte de dados.

Fonte: (BURKE, 2010, p. 379, Tradução nossa).

#### 2.1.4. MODELO DE CLUSTER

Há também uma outra técnica chamada de análise por modelo de *cluster*, ou *clustering*. O dicionário *Oxford* define *cluster* como “A group of similar things or people positioned or occurring closely together”<sup>17</sup> (OXFORD DICTIONARIES). Isto quer dizer que a estratégia empregada neste modelo trata das análises de dados discriminados por grupos. O modelo de *cluster* faz uma divisão dos usuários levando em conta uma determinada característica, como tipo de usuário, ou seja, se é novo ou já possui muitas contribuições; geográfica, agrupando os usuários de acordo com a região onde mora; idade, separando os usuário de acordo com a faixa etária.

“O problema de clusterização possui aplicações nas mais variadas áreas de pesquisa incluindo, por exemplo: computação visual e gráfica, computação médica, biologia computacional, redes de comunicações, engenharia de transportes, redes de computadores, sistemas de manufatura, entre outras.” (OCHI, DIAS e STÊNIO, 2004, p. 1).

<sup>17</sup> “Um grupo de coisas semelhantes ou pessoas posicionados ou que ocorrem em conjunto”, (Tradução nossa).

A diferença desta abordagem está no tratamento do comportamento dos grupos de usuário, e não no indivíduo (usuário) em si. Luiz Satoru Ochi define que um problema de *clusterização* “[...] consiste em dado uma base de dados X, agrupar (clusterizar) os objetos (elementos) de X de modo que objetos mais similares fiquem no mesmo *cluster* e objetos menos similares sejam alocados para *clusters* distintos” (OCHI, DIAS e STÊNIO, 2004, p. 1).

A vantagem deste modelo é conseguir identificar grupos de usuários que possuam comportamentos ou atributos semelhantes, para então recomendar produtos específicos para este conjunto. Porém, uma desvantagem deste modelo é a não personalização das recomendações, pois o perfil do usuário não é analisado por completo, mas sim algumas características que o classificam dentro de um conjunto. A dificuldade na utilização de *cluster* está na escolha da característica que irá segmentar os usuários em diversos grupos, e como o sistema, no caso o algoritmo, irá trabalhar para realizar esta separação. Manuel Eduardo Ares Brea destaca quatro pontos importantes a serem considerados ao implementar um modelo de *clustering*, que são “[...] *the extraction of the constraints, the robustness of the algorithms, the feasibility of the constraints and the utility of the constraints.*”<sup>18</sup> (BREAS, 2013, p. 25). Portanto, quanto à classificação de itens e perfis, a utilização desta técnica se mostra uma ótima complementação à construção de um sistema de recomendação, uma vez que ele é capaz de segregar os itens em grupos, reduzindo o universo de atributos e permitindo que o sistema de recomendação possa atacar (extrair informações e classificar) cada grupo de forma mais direta ao invés de tratar de toda a base de dados, uma vez que o algoritmo irá atuar sobre um grupo que possua uma ou mais características em comum. “*Grouping people into cluster based on the items they have purchased allows accurate recommendations of new items for purchase: if you and I have liked many of the same movies, then I will probably enjoy other movies that you like.*”<sup>19</sup> (UNGAR e FOSTER, 1998, p. 114).

## 2.2. CÁLCULO DE SIMILARIDADE ENTRE ITENS

Como dito anteriormente no capítulo de introdução deste trabalho, item é definido como um objeto que está inserido no conjunto de dados do sistema. Por sua vez, este item é composto por características, ou atributos, e seus respectivos pesos, onde este peso representa uma medida que define a relevância da característica para o item. Como também mencionado anteriormente (capítulo 2.1), um sistema de recomendação utiliza essas características para aproximar itens de usuários, a fim de recomendar itens semelhantes. Para calcular esta semelhança, diversos algoritmos podem ser aplicados, dependendo do modelo de recomendação a ser utilizado e a proposta da aplicação.

Vale ressaltar que dois algoritmos de cálculo de semelhanças, método da similaridade por Cosseno e método de Pearson, usam em sua maioria conjuntos de atributos com valores todos positivos. Em cenários onde é permitido atributos com pesos negativos, estes algoritmos se mostram desvantajosos.

---

<sup>18</sup> “[...] a extração dos restrições, a robustez dos algoritmos, a viabilidade das limitações e a utilidade dos restrições”, (Tradução nossa).

<sup>19</sup> “O agrupamento de pessoas em conjunto com base nos itens que eles tenham adquirido permite recomendações precisas de novos itens para compra: se eu e você gostamos muito dos mesmos filmes, então eu provavelmente irei desfrutar de outros filmes que você goste.”, (Tradução nossa).

Há ainda outras alternativas, dentre elas a de utilizar um sistema Bayesiano para o cálculo de similaridades, o qual permite conjuntos de características com pesos negativos. (JIE, GUIBING e YORKE-SMITH, 2013).

A seguir será discutido alguns métodos considerados neste trabalho, os quais podem ser utilizados para calcular as semelhanças entre itens.

### 2.2.1. COEFICIENTE DE SIMILARIDADE DE JACCARD

Índice de Jaccard ou coeficiente de Jaccard (LEVANDOWSKY e WINTER, 1971) é um método estatístico que determina a similaridade entre dois itens a partir da intersecção de seus conjuntos. Seu coeficiente é definido como a razão do número de atributos resultantes da intersecção de dois conjuntos pelo número de atributos resultantes da união entre estes dois conjuntos. Sua fórmula é definida como:

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

Este método calcula a similaridade de maneira binária, ou seja, leva em conta somente os atributos que os itens têm em comum, e seu alcance varia entre zero (quando os dois itens não tem similaridade alguma) e um (quando os dois itens são exatamente iguais).

$$0 \leq J(A, B) \leq 1$$

Este método é frequentemente utilizado na área de pesquisas ecológicas, para calcular a presença ou falta de características em diferentes situações, como por exemplo, caracterizar locais cuja determinada espécie possua atributos suficientes para se estabelecer.

Uma variação análoga a este método, chamado método de Tanimoto, consiste em transformar estes conjuntos de atributos em um mapa de bits e então calcular sua similaridade da mesma forma que o coeficiente de Jaccard (TANIMOTO, 1957).

$$T_s = \frac{\sum_i (x_i \cap y_i)}{\sum_i (x_i \cup y_i)}$$

Embora sejam métodos que calculam a similaridade entre conjuntos, são geralmente descartados quando se quer implementar um sistema de recomendação de produtos, devido ao fato de se utilizar a intersecção entre os conjuntos, o que o torna um sistema binário. Este aspecto compromete o cálculo da similaridade, uma vez que os atributos envolvidos não possuem uma relevância (ou pertencem ou não pertencem no conjunto do item) definida.

### 2.2.2. DISTÂNCIA EUCLIDIANA

O método de cálculo de similaridade por distância euclidiana (DEZA e DEZA, 1994) considera dois conjuntos (itens) como vetores, e seus atributos representam as coordenadas dos vetores. Por meio do cálculo de distância euclidiana, é possível definir a similaridade dos itens baseando-se no resultado da distância: quanto mais próximos, mais semelhantes estes itens são. A equação a seguir mostra a fórmula utilizada:

$$d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + \dots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

### 2.2.3. MÉTODO DO COSSENO

Este método considera os itens como vetores, cujos atributos representam os valores de cada coordenada. O resultado é obtido calculando o cosseno destes dois conjuntos da seguinte forma:

$$\cos(\theta) = \frac{A \cdot B}{|A||B|} = \frac{\sum_1^n A_i \times B_i}{\sqrt{\sum_1^n (A_i)^2} \times \sqrt{\sum_1^n (B_i)^2}}$$

O método do cosseno leva em conta a orientação dos vetores, e não sua magnitude. Desta forma, os resultados que se seguem estão dentro de um intervalo, onde 0 representa a perpendicularidade dos vetores (90°), 1 representa total semelhança (0°) e -1 representa vetores diametralmente opostos.

Este procedimento é geralmente aplicado em casos onde o espaço amostral apresenta somente valores positivos, ou seja, os vetores são todos positivos. Assim, o intervalo dos resultados fica restrito a 0 e 1. Estes resultados podem então ser utilizados para calcular a semelhança entre itens e usuários.

### 2.2.4. MÉTODO DE PEARSON

O método de cálculo do coeficiente de correlação de Pearson (WIKIPEDIA, 2003) mede a correlação linear entre duas variáveis aleatórias X e Y, onde o intervalo pode variar de +1 à -1, onde 1 é a total correlação entre as variáveis (igualdade), 0 significa que as variáveis são totalmente diferentes e -1 significa que as variáveis são totalmente opostas entre si.

Sua fórmula é definida como sendo a covariância de duas variáveis dividido pelo produto de seus desvios padrão.

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{cov(x, y)}{\sqrt{var(x) \times var(y)}}$$

A covariância entre os vetores é um cálculo definido como o somatório do produto de cada variável menos sua média aritmética:

$$\bar{x} = \frac{1}{n} \times \sum_{i=1}^n x_i$$

$$\bar{y} = \frac{1}{n} \times \sum_{i=1}^n y_i$$

Assim, ao calcular a similaridade entre dois itens, basta tratar estes itens como dois vetores distintos e seus respectivos atributos como suas coordenadas, e então aplicar o método de Pearson para calcular a correlação (similaridade). O método de correlação de Pearson mede o relacionamento entre duas amostras de dados ao longo do tempo. Isso quer dizer que o resultado indicará quanto uma variável muda em detrimento da variação da outra ao longo da distribuição. Este relacionamento pode ser utilizado para calcular previsões sobre itens que o usuário possa se interessar, pois se baseia na correlação encontrada ao longo do histórico de mudança das variáveis, podendo ser por exemplo, a média das avaliações dos usuários.

### **2.2.5. SIMILARIDADE BASEADO EM UMA DISTRIBUIÇÃO DE DIRICLHET**

Guibing Guo, Jie Zhang e Neil Yorke-Smith (JIE, GUIBING e YORKE-SMITH, 2013) propõe ainda uma outra estratégia para o cálculo de similaridade entre dois itens, utilizando o sistema Bayesiano e distribuições de Dirichlet. Este método, aplicado para sistemas de recomendação por filtragem colaborativa, levam em consideração ambos os valores das avaliações e o número de avaliações. Logo, estes dados são levados em consideração quando na modelagem pelo método de distribuição de Dirichlet (JIE, GUIBING e YORKE-SMITH, 2013).

Em seguida, o conjunto de similaridade entre usuários é modelado como a média ponderada da distância encontrada, de acordo com seus pesos de importância, o que corresponde à quantidade de novas evidências de queda na distância. É considerado o cenário onde os usuários que são semelhantes são calculados por meio do comprimento destes vetores, denominado como chance de correlação (JIE, GUIBING e YORKE-SMITH, 2013).

Uma distribuição Dirichlet, é uma distribuição discreta multivariada com um parâmetro (vetorial) não-negativo e real. É definida como uma distribuição de Bayes por que as variáveis envolvidas tratam de incertezas dentro do sistema. “*The Dirichlet distribution is a model of how proportions vary.*”<sup>20</sup> (MINKA, 2000, p. 1). Abaixo, a fórmula que define a medição baseado em uma distribuição de Dirichlet segundo (JIE, GUIBING e YORKE-SMITH, 2013):

$$p(x|\alpha) = \frac{\tau(a_0)}{\prod_{i=1}^n \tau(a_i)} \times \prod_{i=1}^n x_i^{a_i-1}$$

### 2.3. FERRAMENTAS COMPUTACIONAIS

Para desenvolver a aplicação, foram realizadas uma série de pesquisas para selecionar as ferramentas ideais. Sendo a aplicação uma página de internet (*website*), as linguagens de programação selecionadas são orientadas para este tipo de propósito e são bastante populares em se tratando de *websites*. Para melhor entendimento, as linguagens foram divididas em subcapítulos, sendo cada capítulo um resumo sobre sua função dentro do projeto, além de uma visão geral sobre suas finalidades. Neste capítulo também será discutido o modelo de arquitetura de software chamado MVC (*model-view-control*). Este modelo é comumente utilizado para a construção de interfaces com o usuário, e foi adotado neste trabalho com o propósito de separar funções distintas, agrupando-as em três grupos interconectados: o modelo, o controle e a visão. Seguindo este mesmo padrão, os capítulos subsequentes que tratam de etapas específicas do projeto também estão divididos de acordo com o MVC, com o objetivo de segmentar cada assunto relacionado ao funcionamento da aplicação.

#### 2.3.1. MYSQL (STRUCTURED QUERY LANGUAGE)

MySQL é um sistema de gerenciamento de banco de dados (SGBD), que utiliza a linguagem SQL (*Structured Query Language*) como interface. É considerado atualmente um dos bancos de dados mais populares, com mais de 10 milhões de instalações pelo mundo (MYSQL, 2014). É um sistema de licença livre e que apresenta uma ótima performance com bancos de dados de tamanho médio.

Um SGBD é o sistema responsável pela organização e armazenamento dos dados que são utilizados pela aplicação. Dentro dele, as entidades relacionais definidas pelo projeto são traduzidas em tabelas, que são então acessados pelos programas. Basicamente, é um mecanismo que armazena dados em forma de tabelas, porém estes sistemas possuem recursos adicionais que facilitam a manipulação destas informações e garantem a integridade das mesmas.

Pelo seu fácil manuseio, recursos oferecidos e licença livre para uso, o MySQL foi selecionado para esta aplicação. Como o protótipo não utiliza uma base de dados gigante, o desempenho não é comprometido, garantindo uma ótima performance deste SGBD.

---

<sup>20</sup> “A distribuição Dirichlet é um modelo de como estas proporções variam.”, (Tradução nossa).

### 2.3.2. PHP (HYPERTEXT PREPROCESSOR)

É uma linguagem interpretada livre, usada originalmente apenas para o desenvolvimento de aplicações presentes e atuantes no lado do servidor, capazes de gerar conteúdo dinâmico na internet.

É uma linguagem de licença livre que pode ser embutida ao HTML para gerar conteúdo de uma página de internet. Atua do lado do servidor, e seu papel geralmente é o de controlar o fluxo de informações que saem do servidor e vão para o lado do cliente, e vice-versa. É uma linguagem interpretada de tipagem fraca e orientada a objetos.

### 2.3.3. JAVASCRIPT

JavaScript é uma linguagem de programação interpretada originalmente implementada como parte dos navegadores *web* para que scripts pudessem ser executados do lado do cliente e assim pudessem interagir com o usuário sem a necessidade de acessar o servidor, controlando o navegador, realizando comunicação assíncrona e alterando o conteúdo do documento exibido.

Por meio de um conjunto de técnicas denominadas Ajax (*Asynchronous Javascript and XML*<sup>21</sup>), é possível requisitar dados do servidor sem a necessidade de atualizar a página, tornando a aplicação mais dinâmica e interativa. Assim, também é possível alterar os elementos de uma página, além de requisitar e exibir informações vindas do banco de dados sem ter de recarregar toda a página novamente.

Atualmente surgiram também diversos projetos (*plugins*<sup>22</sup>) que incorporaram as funcionalidades do Javascript, criando abstrações e encapsulando métodos, tornando-os mais simples de implementar e garantindo segurança quanto à consistência do código. Desta forma o tempo de produção também é reduzido, tendo em vista que funcionalidades essenciais, como a comunicação dinâmica entre o cliente e o banco de dados, já estão incluídas e testadas à exaustão, evitando o retrabalho.

### 2.3.4. CSS (CASCADING STYLE SHEETS)

Cascading Style Sheets (ou simplesmente CSS) é uma linguagem de folhas de estilo utilizada para definir a apresentação de documentos escritos em uma linguagem de marcações, como HTML ou XML. Seu principal benefício é prover a separação entre o formato e o conteúdo de um documento.

Com o propósito de manter as marcações separadas de seu estilo, o CSS é utilizado para organizar os atributos de cada elemento da página em classes ou pelos tipos das marcações. Desta forma, é possível definir o caráter visual de cada elemento dentro da página, bem como

---

<sup>21</sup> “Javascript e XML assíncronos”, (Tradução nossa).

<sup>22</sup> Componente que adiciona recursos à um software ou código.

alterá-las utilizando o Javascript, à fim de apresentar uma aplicação visualmente atraente e de fácil manuseio para o usuário.

### 2.3.5. HTML (HYPERTEXT MARKUP LANGUAGE)

Hypertext Markup Language (HTML) é uma linguagem de marcações e representa a base da internet. Sua estrutura é definida por marcações chamadas *tags*. Estas *tags* podem exercer funções diversas dentro do documento, e são responsáveis pela organização e apresentação do conteúdo de forma clara, seguindo uma estrutura pré-definida.

### 2.3.6. MVC (MODEL-VIEW-CONTROLLER)

O modelo de três camadas MVC divide um aplicativo de modo que a lógica de negócio resida no meio destas três camadas. A arquitetura MVC – (Modelo Visualização Controle) fornece uma maneira de dividir a funcionalidade envolvida na manutenção e apresentação dos dados de uma aplicação. A arquitetura MVC foi originalmente desenvolvida para mapear as tarefas tradicionais de entrada, processamento e saída para o modelo de interação com o usuário. Usando o padrão MVC fica fácil mapear esses conceitos no domínio de aplicações Web multicamadas. A figura 6 ilustra o fluxo do processo MVC:

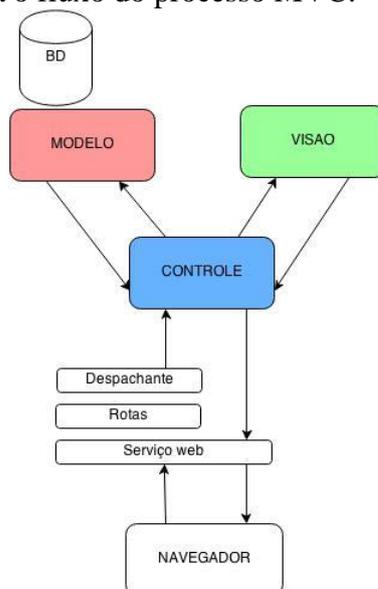


Figura 6 - Diagrama que ilustra o fluxo do processo no modelo MVC

Este modelo foi adotado para proporcionar uma estrutura e organização mais limpa para o projeto. Suas etapas bem definidas ajudam à separar as funções do projeto de forma a otimizar a manutenibilidade e clareza do código. Como este projeto requiriu conhecimento em diversas linguagens (Javascript, PHP, HTML, SQL), sendo possível interconectar todas elas por meio da mesma página, ou seja, escrever trechos de códigos de linguagens diferentes no mesmo arquivo, a opção pela arquitetura MVC se deu pela necessidade de separar estas funções em módulos diferentes, a fim de manter a legibilidade do código e evitar confusões quanto ao papel de cada linguagem dentro do projeto.

### 3. SOBRE A APLICAÇÃO

Baseado no modelo MVC (modelo-visão-controle) (POPE e KRASNER, 1988), a apresentação da construção do trabalho foi dividida em três etapas. A primeira etapa trata do modelo de dados que foi definido para a aplicação, sua arquitetura e apresentação de como as informações são armazenadas. A segunda etapa, Controle, apresenta os algoritmos aplicados e trechos de código relacionados ao tratamento das informações e a geração de recomendações aos usuários. A última etapa, Visão, trata da exibição destas informações para os usuário, ou seja, como as recomendações são exibidas, as tecnologias aplicadas para a construção do *front-end*<sup>23</sup> da aplicação, bem como conceitos de *layout* e apresentação de ideias futuras.

Esta forma de separação foi definida para segmentar diferentes pontos que foram tratados neste trabalho. Além das pesquisas realizadas sobre diferentes métodos que poderiam ser aplicados em um sistema de recomendação, também foram pesquisadas formas de armazenar a informação e diferentes tecnologias foram selecionadas e estudadas para implementar a arquitetura escolhida.

Desta forma, um dos objetivos deste trabalho é apresentar uma estrutura completa, partindo de sua concepção, definição de arquitetura e tecnologias, desenvolvimento, até seu produto final, abrindo também novos horizontes para futuras implementações ou pesquisas sobre os assuntos tratados.

#### 3.1. MODELO

O modelo de recomendação escolhido foi o baseado em conteúdo. Como comentando anteriormente (capítulo 2.1.2), o método de recomendação baseado em conteúdo tem como premissa caracterizar cada item, baseando-se em informações pré-definidas ou determinadas pelo usuário.

Assim, o modelo de entidade que ilustra o relacionamento entre itens e suas características é descrito na figura 7:

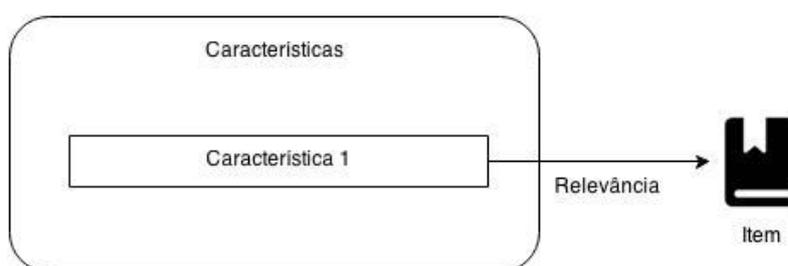


Figura 7 – Relacionamento entre um item e sua característica.

Onde os diversos objetos representam os itens do conjunto de itens, cada um possuindo informações únicas e Características representa o conjunto de palavras-chave utilizadas na aplicação, bem como suas identificações. As palavras-chave utilizadas neste conjunto são

<sup>23</sup> *Front-end* é uma etapa do processo responsável pela coleta de dados vindos do usuário, as processando a fim de adequá-la de acordo com as especificações necessárias para que o sistema possa utilizar.

informações que descrevem um objeto e podem ser compostas por qualquer descrição pertinente ao objeto, por exemplo, a descrição do conteúdo do livro, descrições sobre seu gênero, estilo literário do autor, características de seus personagens ou o mundo no qual a obra está ambientada, entre outros. A relevância desta característica em relação ao objeto que descreve é dado por meio de um valor numérico entre 0 e 1. Quanto maior a importância deste atributo para o objeto, maior será este peso, e se uma determinada característica for capaz de descrever o objeto porém não possua tanta relevância, este valor estará próximo de 0. Este conjunto de características deverá então ser capaz de representar o objeto por meio de etiquetas, ou atributos, ponderados de acordo com sua recorrência ou relevância. Por fim, tem-se o relacionamento e a sua relevância entre essas entidades, definido pelo conjunto das características, como é mostrado na figura 8:

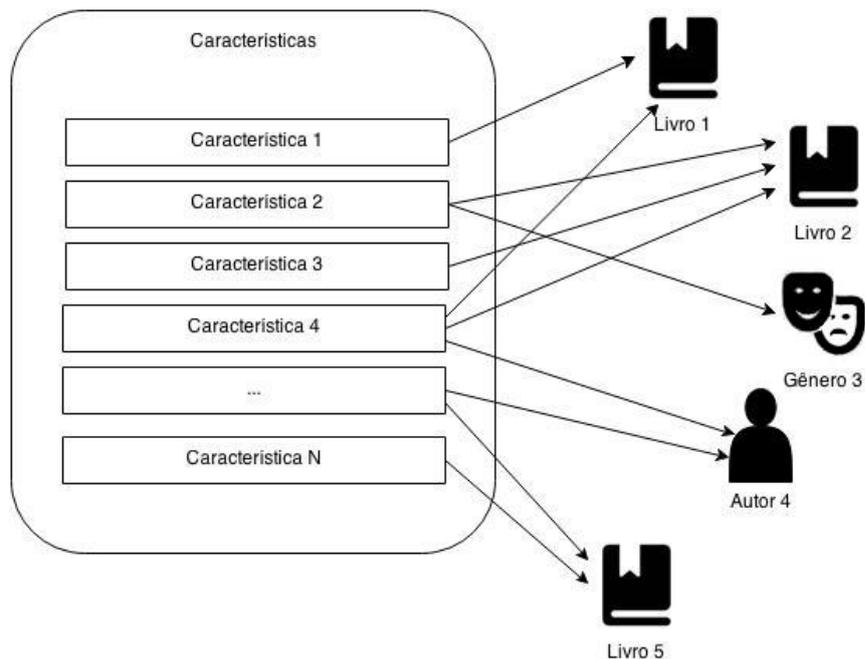


Figura 8 - Conjunto de características e seus relacionamentos com itens da base de dados

Por sua vez, os dados referentes aos usuários são definidos como uma entidade chamada *User*. Nesta entidade são armazenadas informações importantes sobre o usuário. As características de um usuário são determinadas de acordo com as ações que ele executa. As ações que um usuário pode executar são diversas: visualizar um item, adicioná-lo à sua lista, realizar um comentário sobre e aprová-lo. Posteriormente, as informações sobre as ações do usuário serão utilizadas pelo sistema de recomendação para reunir as características de todos os itens cujo usuário teve contato. Desta forma, uma lista de características pode ser gerada, informando os atributos que interessam o usuário. A figura 9 ilustra as ações que um usuário pode executar:

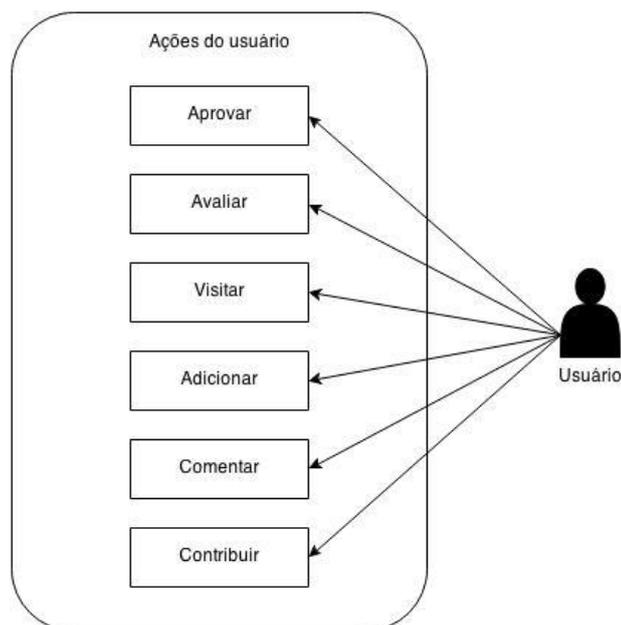


Figura 9 - Tipos de interações que um usuário pode executar em uma aplicação

Onde:

- Aprovar significa que o usuário aprovou o item.
- Avaliar significa que o usuário avaliou (deu uma nota) ao item.
- Ver é o ato de visualizar uma página.
- Converter é o ato de adicionar o item à sua lista.
- Comentar significa realizar um comentário sobre um item.
- Enviar significa enviar uma contribuição para uma página, como por exemplo, sugerir uma mudança, ou recomendar este item para outro usuário.

Como descrito anteriormente (capítulo 1), o usuário é um objeto atuante dentro do sistema, e pode interagir com dois tipos de itens diferentes: os originais, ou seja, os itens que são exibidos em páginas próprias, no caso do *website*, páginas que descrevem os itens, como uma página de exibição de produto de um *website* de vendas; e os itens pertencentes à listas de outros usuários, que são marcações que tem ligação direta com o original (links para as páginas dos produtos). Este segundo tipo são chamados de marcações e fazem parte da lista de itens adicionados do usuário, em analogia ao carrinho de compras presente nos *websites* de vendas. Um usuário também pode interagir com outro usuário de forma direta, como por exemplo, visitar seu perfil ou escrever um comentário para ele.

Assim, tem-se que um usuário pode atuar em três áreas diferentes: sobre a lista de outro usuário, sobre as páginas de itens que o sistema oferece e sobre outros usuários. A figura 10 ilustra as formas de interação possíveis, que geram informações para o sistema de recomendação:

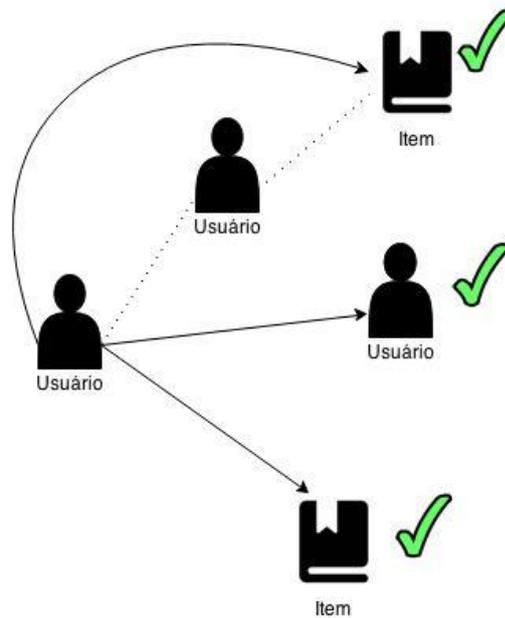


Figura 10 - Interação de um usuário com outro usuário, um item ou item de um usuário

A estrutura de dados utilizada para ilustrar a base de relacionamentos entre os itens foi o grafo. Isto se deve ao fato da necessidade detectada de se criar uma lista que representasse o relacionamento entre dois itens, sendo este relacionamento composto por um peso que determina a similaridade entre eles. Um grafo é um conjunto não-vazio  $V$  cujos elementos são chamados vértices, e um conjunto  $A$  de arestas. Uma aresta é um par não-ordenado  $(v_i, v_j)$ , onde  $v_i$  e  $v_j$  são elementos de  $V$ . Assim, podemos representar os relacionamentos entre os itens por meio de um grafo, onde cada vértice representa um item presente no sistema e a ligação entre eles, ou seja, a aresta entre dois vértices representa a similaridade entre os dois itens. A similaridade entre dois itens é um valor que mensura a semelhança entre dois objetos, e no caso deste trabalho, é calculado a partir dos conjuntos de características destes dois objetos. Por meio do cálculo do cosseno de vetores (mencionado no capítulo 2.2.3), é possível então chegar a um valor numérico entre 0 e 1 que descreve a semelhança entre estes dois itens. Normalmente, utiliza-se uma representação gráfica de um grafo, como ilustra a figura 11.

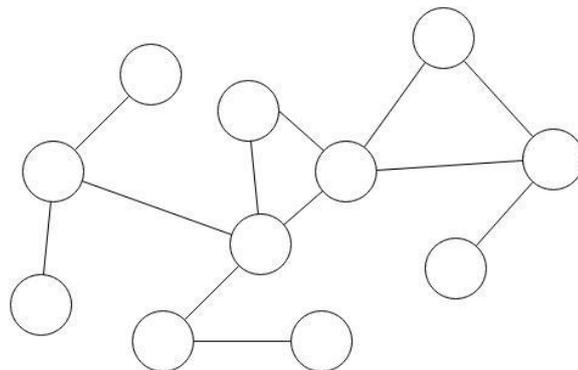


Figura 11 - Exemplo de grafo

Cada vértice representa um item e as arestas que ligam os vértices representam a similaridade entre estes itens. Vale a pena destacar que o grafo gerado é orientado. Isso quer dizer que a similaridade entre os itens  $A \rightarrow B$  será diferente da similaridade  $B \rightarrow A$ . Isto se deve ao fato de os itens não possuírem exatamente as mesmas características, gerando divergências nos resultados pelo método adotado para o cálculo. A figura 12 mostra a interação entre dois vértices conectados por duas arestas direcionadas:

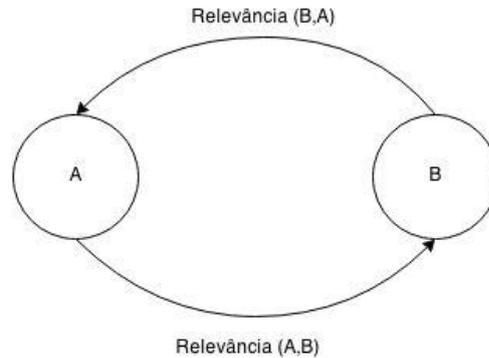


Figura 12 - Grafo direcionado. O valor da aresta (A,B) pode ser diferente de (B,A)

### 3.2. CONTROLE

Quanto aos métodos para calcular a similaridade entre dois itens que foram pesquisados, o implementado foi o método do cosseno, por ser, junto do método de correlação de Pearson, um dos mais populares e por se encaixar na proposta do projeto, quer dizer, a criação de um sistema de recomendação baseado em conteúdo cuja similaridade seja resultado do cálculo das características dos itens correspondentes.

Sendo um item representado por uma lista de características e seus pesos, podemos converter estes itens em vetores, como mostra a tabela 3. Assim, podemos utilizar a fórmula geométrica de cálculo do cosseno entre dois vetores para determinar a similaridade entre eles:

Tabela 3 - Itens, suas características e relevâncias.

Item	Características			
	1	2	3	4
A	0.2	0.7	0.5	1
B	0.4	0.5	1	-

Sendo  $A = (0.2, 0.7, 0.5, 1)$  e  $B = (0.4, 0.5, 1, 0)$ , sua similaridade pode ser calculada pelo cálculo de cosseno de vetores:

$$\cos(\theta) = \frac{A \cdot B}{|A||B|} = \frac{\sum_1^n A_i \times B_i}{\sqrt{\sum_1^n (A_i)^2} \times \sqrt{\sum_1^n (B_i)^2}}$$

O uso deste processo torna a classificação de itens automática, trazendo informações relevantes de diversas fontes, nutrindo o sistema com dados importantes que irão caracterizar cada item da base de dados. Dado um sistema de recomendação baseado em conteúdo, um sistema complementar de classificação de itens se faz muito útil, tanto para atribuir características quanto para analisar as características que são adicionadas posteriormente, comparando sua relevância, visando manter a consistência dos dados.

Christopher D. Manning, Prabhakar Raghavan e Hinrich Schütze definem o processo de extração de informações como “[...] *finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)*”<sup>24</sup> (MANNING, RAGHAVAN e SCHÜTZE, 2009).

Durante os testes com os algoritmos implementados, os atributos gerados foram todos artificiais. Porém, foi projetado um método que realiza a extração de informações da internet. A função deste método é, dado o *link* da página em questão, extrair palavras-chave para caracterizar o item. Palavras-chave, neste contexto, significam palavras mais recorrentes dentro de todo o conteúdo. Realizando a extração de informações de forma rústica, é possível criar uma descrição inicial sobre os itens que compõe a base de dados. Com o tempo, esta descrição poderia ser revisada e reavaliada tanto pelo sistema, como por exemplo, a criação de outros métodos de extração de informações, quanto pelos usuários, votando se aquele atributo seria ou não condizente com o item, além da possibilidade do usuário poder adicionar novas palavras-chave, contribuindo para a base de informações.

Quando o usuário acessa uma página, como descrito no capítulo sobre o Modelo, este pode atuar sobre um item, uma marcação de um usuário ou sobre o usuário. Para cada tipo de ação, é gerada uma informação. Estas informações são então reunidas e uma lista de itens que o usuário acessou, direta ou indiretamente (por meio de marcações) é gerada. Para cada item, são extraídas suas características, gerando então uma lista de atributos. Ao final, estes atributos são utilizados para obter itens que tenham características aproximadas à lista gerada.

Tendo uma base de dados pré-definida, a similaridade entre os itens é atualizada toda vez que um usuário acessa uma página. Para evitar sobrecarga, cada acesso do usuário extrai uma amostra de itens do grafo, e então atualiza seus pesos (similaridade). Desta forma o grafo mantém uma atualização constante, pois, eventualmente, um item pode ser removido da base de dados, ou uma característica de um item pode ser alterada.

No procedimento de recomendação de produtos, a inteligência artificial auxilia no refino e precisão das informações que são requisitadas e retornadas, aumentando a eficácia do processo. Diversas estratégias podem ser adotadas num sistema de recomendação. Estes sistemas têm a característica de extrair informações sobre os usuários e adequá-las de acordo com as necessidades e dados disponíveis na base dados, a fim de oferecer produtos mais relevantes para o usuário. De forma análoga, o algoritmo implementado neste trabalho tem esta característica. Cada acesso de um usuário, além de gerar informações pertinentes ao seu perfil, aciona um mecanismo de atualização e criação de arestas (similaridade) entre os vértices (produtos) que compartilham pelo menos uma característica. Isso quer dizer que o

---

<sup>24</sup> “[...] a descoberta de materiais (geralmente documentos) de natureza não estruturada (geralmente texto) que satisfaz uma necessidade de informação de dentro de grandes coleções (geralmente armazenadas em computadores)”, (Tradução nossa).

grafo de itens será constantemente atualizado, baseado no número de acessos e nos itens acessados. Além disso, o algoritmo poderia também ser pré-processado antes de entrar em produção, ou seja, antes do sistema ser efetivamente lançado como produto (colocado online). A ideia é que cada usuário acione o mecanismo de cálculo de similaridade, de forma não-supervisionada.

### 3.3. VISÃO

Esta etapa do projeto trata da apresentação das informações para o usuário, como os itens estão dispostos e quais recursos o usuário dispõe para navegar na aplicação. Como mencionado no capítulo anterior (capítulo 3.2), toda vez que o usuário visita uma página, uma série de informações são armazenadas para depois serem analisadas pelo algoritmo de recomendação. Também, toda vez que o usuário acessa uma página de item ou interage com algum item (página ou item que esteja no perfil do usuário), informações sobre essa interação (ação) são também armazenados. Feito o cálculo de similaridades e retornada a lista baseada no perfil (gostos) do usuário, os itens são distribuídos na página para seleção, como mostra a figura 13:

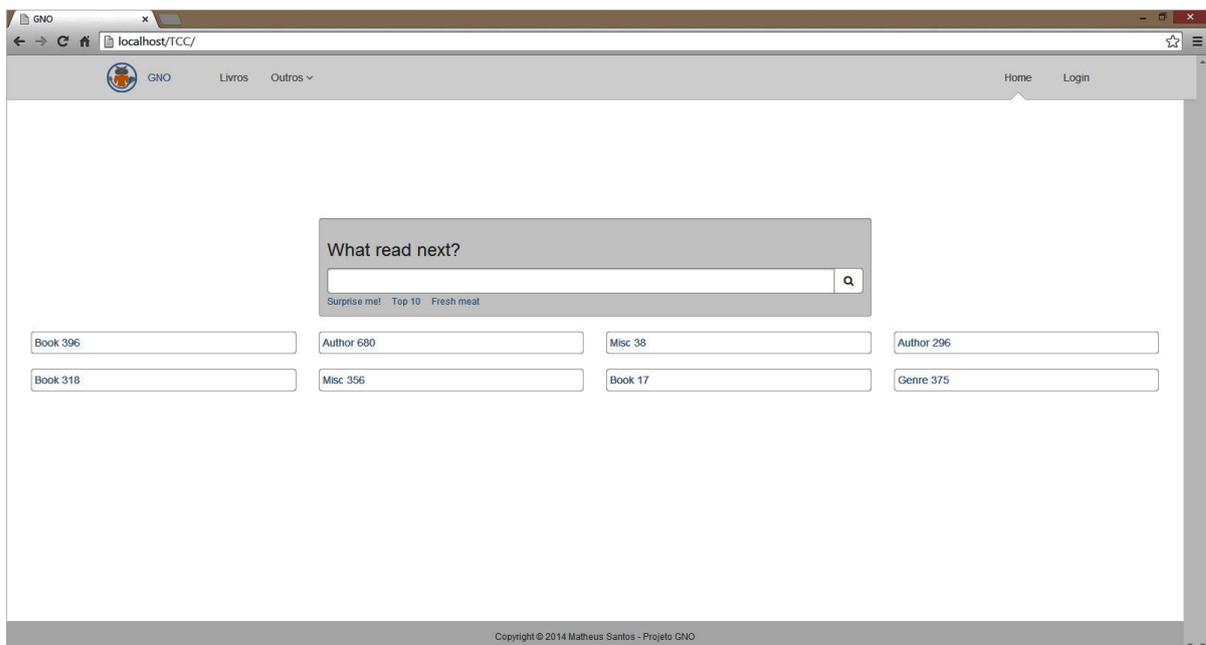


Figura 13 - Página inicial.

Ao selecionar um item, o usuário é redirecionado para a página que contém as informações pertinentes (figura 14), exibindo dados específicos sobre este item, como suas características, comentários de outros usuários, descrição, nome do autor, etc.

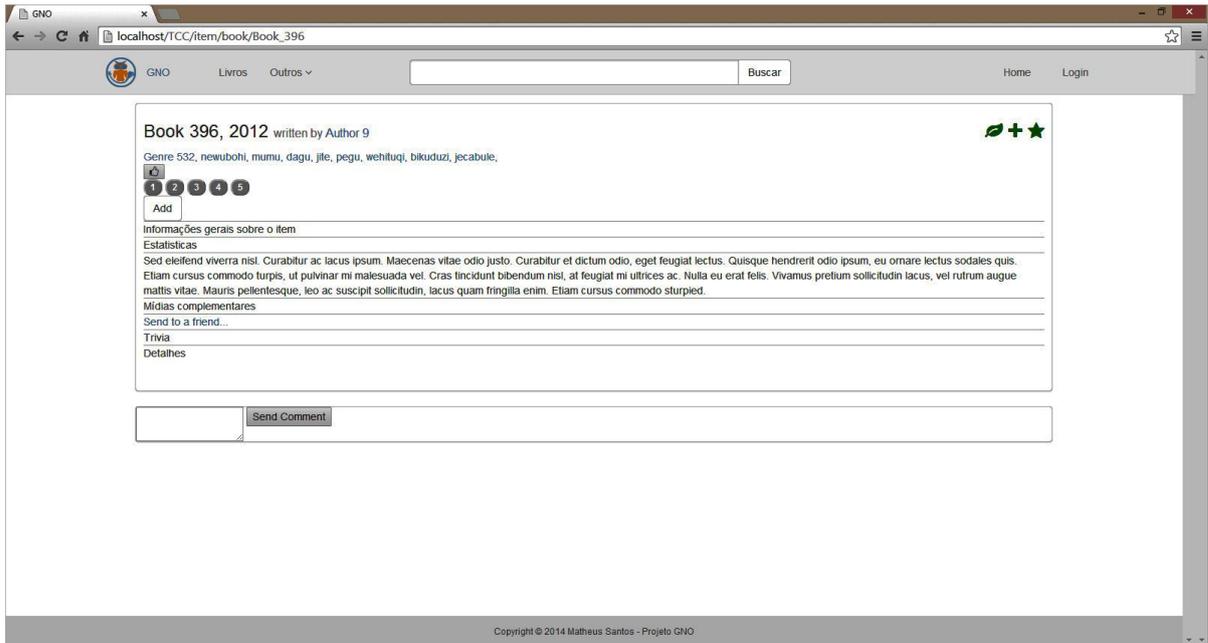


Figura 14 - Página de visualização de um item

Quando o usuário adiciona o item, esta ação cria uma marcação em seu perfil, com informações sobre o item adicionado. De forma análoga, representa o mesmo conceito do carrinho de compras, comum nos *e-commerces*. Nesta página (figura 15) o usuário pode efetuar ações sobre suas marcações diretamente; cada item da lista redireciona para sua página correspondente.

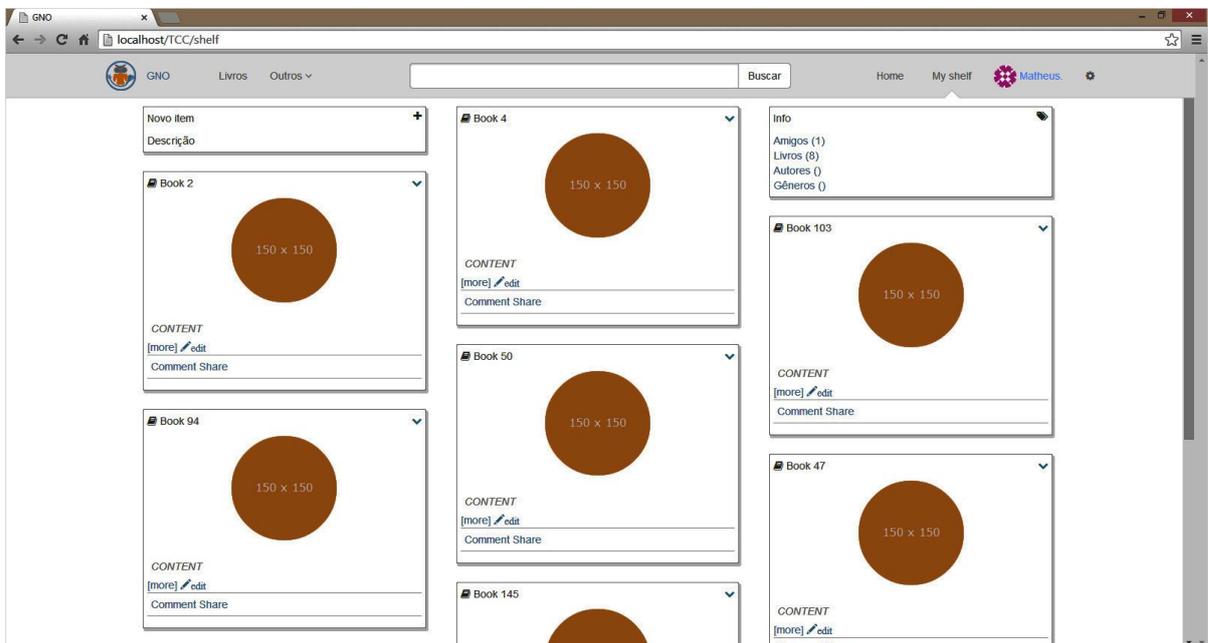


Figura 15 - Página pessoal do usuário.

O perfil do usuário, como mostra a figura 16, exibe informações sobre suas marcações em uma forma de listagem cronológica (*timeline*). Os resultados das interações geradas nesta página armazenam informações de interações, que mais tarde serão utilizadas pelo sistema de recomendação.

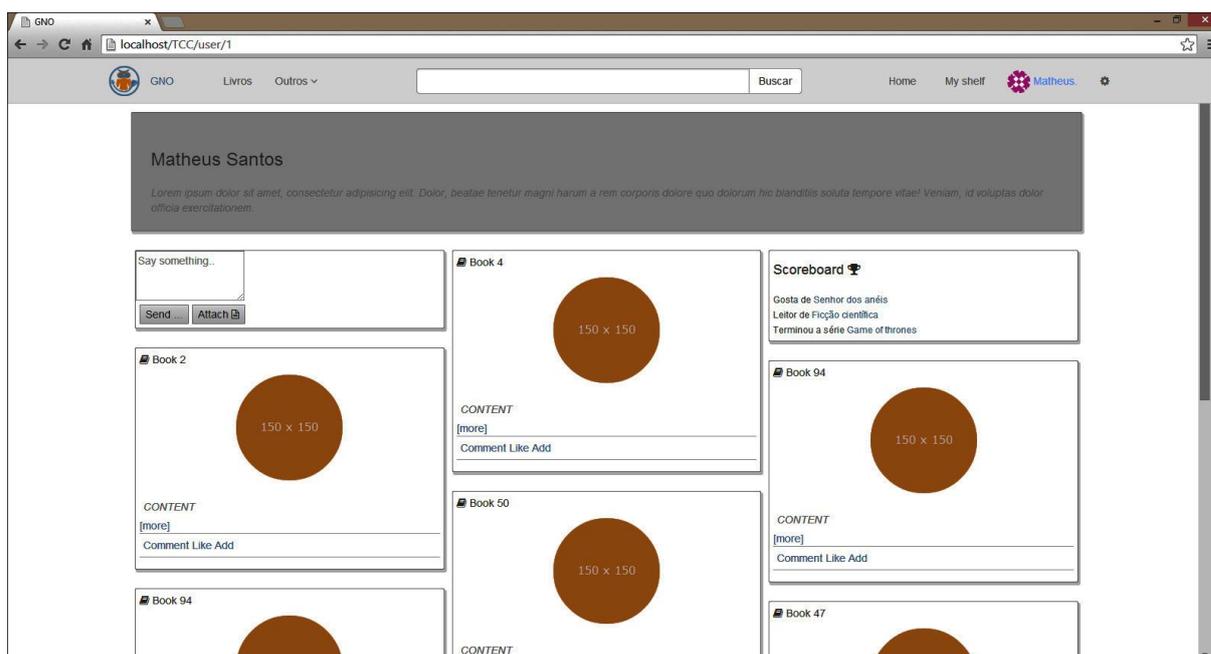


Figura 16 - Perfil do usuário.

A página inicial da aplicação apresenta os itens recomendados pelo sistema, e um campo de busca. Se o usuário não tiver efetuado seu *login*<sup>25</sup> no sistema, os itens serão exibidos aleatoriamente, visando mostrar itens diversos (gêneros e autores diferentes). Outra alternativa seria apresentar itens com boas avaliações, itens adicionados recentemente ou itens mais visitados (dentro de um período de tempo ou não). Embora a primeira opção tenha sido escolhida, as seguintes se mostram mais eficientes e podem ser consideradas numa futura etapa de refino do sistema.

“Day by day, the number of devices, platforms, and browsers that need to work with your site grows. Responsive web design represents a fundamental shift in how we’ll build websites for the decade to come”<sup>26</sup>, (JEFFREY, 2011).

Durante a pesquisa sobre plataformas e ferramentas a serem utilizadas para o desenvolvimento da aplicação, foi notada a necessidade de desenvolver uma aplicação que se adequasse a diversas resoluções de tela. Hoje, *websites* precisam estar preparados para apresentar suas informações sem que a estrutura do *layout*<sup>27</sup> atrapalhe. *Smartphones, tablets, notebooks*, computadores e televisões possuem resoluções de tela diferentes, e por isso é importante um *layout* que se encaixe em todos estes formatos. Para solucionar este problema, inicialmente os desenvolvedores optaram por criar estruturas diferentes para cada dispositivo (*mobile* e computadores). Porém, esta solução apresentava problemas de manutenibilidade, uma vez que um mesmo sistema teria que possuir múltiplos *layouts* e *templates*<sup>28</sup>. Com o avanço de tecnologias voltadas para a internet, e com o surgimento do conceito da Web 3.0,

<sup>25</sup> Refere-se ao ato de iniciar uma sessão em determinada aplicação.

<sup>26</sup> “Dia após dia, o número de dispositivos, plataformas e navegadores que precisam trabalhar com o seu site cresce. O design web responsivo representa uma mudança fundamental na forma como vamos construir sites para a próxima década”, (Tradução nossa).

<sup>27</sup> Layout é um termo utilizado para nomear os arranjos que representam a estrutura de uma determinada entidade.

<sup>28</sup> Ambiente definido como modelo. Utilizado como padrão, pode ser replicado ao longo da aplicação.

os *templates* se tornaram mais interativos, passando a ajustar o *website* de acordo com os dispositivos que o requisitam. Desta forma, a usabilidade por parte do usuário e manutenibilidade, por parte dos desenvolvedores, é melhorada. O *design* responsivo é uma abordagem de *web design* destinada a elaborar aplicações que ofereçam uma ótima experiência de visualização, fácil leitura e navegação com um mínimo de redimensionamento e visionamento, para uma ampla gama de dispositivos. O objetivo do *design* responsivo é ajustar a usabilidade para oferecer um maior conforto e uma melhor experiência ao usuário. Assim, o planejamento de uma interface interativa e *user-friendly* (amigável ao usuário) é uma etapa muito importante pois define como as informações serão exibidas e como usuário navegará entre as funcionalidades. A etapa de *user experience design* (UX Design) pode ser dividida em cinco passos, cada qual delegando uma competência: “[...] *information architecture, Interaction Design, Usability Engineering, Visual Design e Prototype Engineering*”<sup>29</sup> (PSOMAS, 2007). A figura 17 ilustra uma estrutura de interface responsiva em diversos tipos de dispositivos, e a figura 18 exhibe esta abordagem aplicada ao projeto:



Figura 17 - Uma aplicação responsiva adequa seu layout para diversos tipos de dispositivos.

---

<sup>29</sup> “[...] Arquitetura da informação, Design de Interação, Usabilidade, Engenharia Visual de Design e Engenharia de Protótipos”, (Tradução nossa).



Figura 18 - Visão do protótipo em resolução reduzida (um smartphone por exemplo).

## 4. ESTUDO DE CASO

Para contextualizar de forma mais clara o que foi explicado até agora, algumas aplicações que utilizam o conceito de recomendações de produtos serão apresentadas e resumidas, a fim de enfatizar a importância que estes tipos de solução têm na mineração de informações e suporte as buscas feitas pelos usuários.

Entre outros, os exemplos descritos aqui foram selecionados e suas funcionalidades foram estudadas para fins de aprendizado, exercendo também grande influência durante o desenvolvimento deste trabalho. Os conceitos estudados nestas aplicações serviram como base instrucional para o entendimento de funcionalidades e arquiteturas, pois utilizam princípios e técnicas que delineiam um sistema de recomendação.

Assim como já foi citado no capítulo Controle, um exemplo de aplicação é a recomendação de filmes que o Netflix utiliza baseado em avaliações dos usuários, a filtragem colaborativa.

### 4.1.1. AMAZON

O site de vendas de produtos Amazon talvez seja um dos exemplos mais populares sobre o sucesso de um sistema de recomendação. Especializado na venda de produtos de diversos setores, desde dispositivos eletrônicos até produtos infantis, a Amazon desenvolveu um método de recomendação denominado *item-to-item collaborative filtering*<sup>30</sup>. Basicamente, o algoritmo utilizado pela Amazon busca informações no carrinho de compras do usuário: o que ele já comprou, o que está comprando e os itens que ele visita.

Este sistema funciona nos moldes da filtragem colaborativa, porém ao invés de definir a similaridade entre usuários, o sistema foca na similaridade entre os itens encontrados no carrinho de compras do usuário. O que ele faz é analisar todos os itens que o usuário acessou, comparar com outros itens similares e criar uma lista de recomendações.

O diferencial desta solução está na escalabilidade: partindo do princípio que existem mais itens do que usuários dentro do sistema, a matriz que envolve os relacionamentos (similaridade) entre itens será menor do que o que envolve o relacionamento entre os usuários.

As vantagens desta abordagem estão na criação de listas de produtos similares baseado nos carrinhos de compras dos usuários. Isso quer dizer que a desempenho do sistema é otimizado, uma vez que estas listas permanecerão ativas, e como as características entre itens não tende a mudar drasticamente (como o gosto dos usuários), é também de fácil manutenção e acesso.

O método de cálculo de similaridade entre itens é o cálculo do Cosseno dos vetores.

Como o cálculo é baseado nos carrinhos de compras dos usuários, que possuem poucos itens em comparação com o total de produtos do sistema, sua execução é rápida, como explica Greg Linden:

---

<sup>30</sup> “Filtragem colaborativa de item-para-item”, (Tradução nossa).

“Given a similar-items table, the algorithm finds items similar to each of the user’s purchases and ratings, aggregates those items, and then recommends the most popular or correlated items. This computation is very quick, depending only on the number of items the user purchased or rated.”<sup>31</sup> (LINDEN, SMITH e YORK, 2003, p. 19).

Para concluir a análise deste sistema, uma amostra do algoritmo é apresentado à seguir, mostrando como as ações são executadas a partir da lista de produtos dos consumidores:

*For each item in product catalog, I1*  
*For each customer C who purchased I1*  
*For each item I2 purchased by customer C*  
*Record that a customer purchased I1 and I2*  
*For each item I2*  
*Compute the similarity between I1 and I2*  
 (LINDEN, SMITH e YORK, 2003, p. 19)

Como resultado, detectou-se um aumento na receita da Amazon. De acordo com Sucharita Mulpuru, analista da Forrester, os resultados obtidos pela Amazon com recomendações de produtos pode chegar a até 60% em relação a outros *websites* de comércio eletrônico. Como exemplo destes lucros, no primeiro trimestre de 2013, houve um aumento de 29% nos lucros da empresa, em relação ao trimestre passado, saltando de 9.9 bilhões de dólares para 12.98 bilhões (MANGALINDAN, 2012).

#### 4.1.2. GOODREADS

*Goodreads* é um site que oferece uma robusta base de dados sobre livros para os usuários (GOODREADS, 2007). Da mesma forma, também oferece uma base de revisões e comentários que auxiliam usuários a escolherem as próximas obras para ler.

Possui um sistema de recomendações baseado em filtragem colaborativa, com foco em produtos (um sistema muito semelhante ao utilizado pela Amazon) e sua extensa base sobre livros proporcionam aos usuários alternativas sobre diversas obras, e conseqüentemente, possíveis escolhas (ISAACSON e SEBASTIAN, 2008). Funciona também de forma análoga ao NetFlix, porém com a premissa de oferecer recomendações de livros ao invés de filmes.

---

<sup>31</sup> “Dada uma tabela de itens semelhantes, o algoritmo encontra itens parecidos com cada um comprado ou avaliado pelo usuário, agregando esses itens e, em seguida, recomenda os itens mais populares ou correlatos. Este cálculo é muito rápido, dependendo apenas do número de itens que o usuário comprou ou classificou”, (Tradução nossa).

## 5. CONCLUSÕES

Como estratégia de interação inicial, para cada acesso do usuário à uma determinada página, o grafo de produtos dentro da aplicação será expandido em largura, ou seja, o algoritmo calculará a similaridade entre os itens mais próximos ao que está sendo acessado no momento. A propagação em largura será acionada toda a vez que o acesso de um novo usuário for computado.

Assim, baseado no número de acessos a um determinado item, o grafo a partir deste vértice será expandido e atualizado, e como o processamento está fragmentado ao longo da comunidade de usuários que irão interagir com o sistema, o esforço computacional também é fracionado ao longo dos acessos. A figura 19 ilustra estas conexões ao longo do tempo:

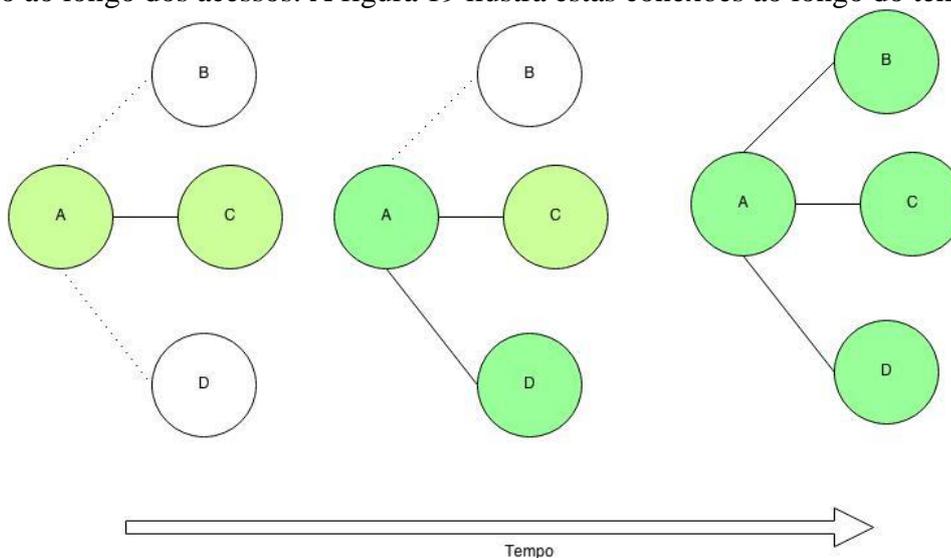


Figura 19 - Relacionamento entre itens sendo estabelecidos ao longo do tempo, pela interação dos usuários.

Embora a abordagem de dividir o cálculo do grafo ao longo dos acessos divida o processamento entre os usuários (*crowdsourcing*<sup>32</sup>), a aplicação se mostrou dependente dos usuários para expansão do grafo. Ainda, outro problema detectado durante os testes foi o fato de, inicialmente, somente os itens mais acessados tornavam-se relevantes, e em consequência disso, somente os itens que compartilhavam características semelhantes tornariam-se relevantes também, como exemplificado na figura 20:

<sup>32</sup> Prática de obter os serviços necessários, idéias ou conteúdo solicitando contribuições de um grande grupo de pessoas (WIKIPEDIA, 2006, Tradução nossa).

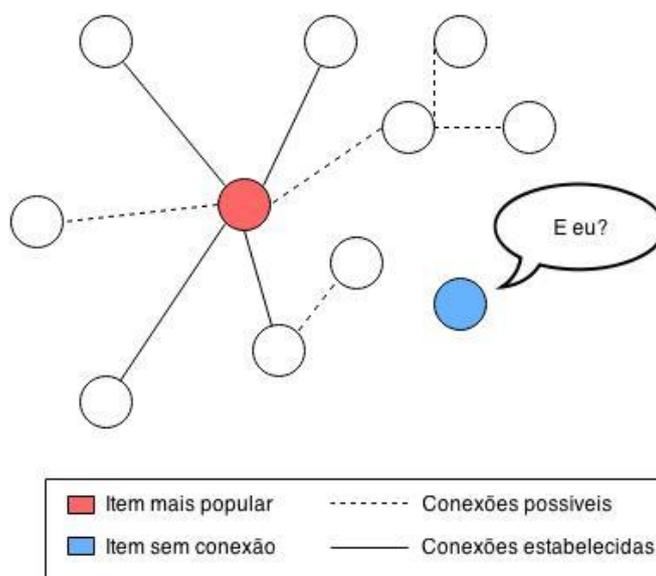


Figura 20 - Grafo exemplificando problema do item mais popular e os itens mais específicos (difíceis de encontrar).

A solução para este problema poderia ser o uso de uma base de dados inicial, cujas similaridades entre todos os itens já estivessem pré-definidas e calculadas, alterando o foco do usuário na expansão do grafo para atualização. O papel da composição de novos vértices no sistema seria então necessário integrar os novos itens inseridos na base dados, enquanto que os itens já calculados seriam somente atualizados pelos usuários, sendo esta forma de atualização executada por meio de contribuições como inserção de novas características, ou revisões e avaliações sobre o item.

Para complementar a solução, um segundo sistema de recomendação poderia atuar sobre a base dados. Agindo sobre as avaliações dos itens, um sistema de recomendação baseado em filtragem colaborativa auxiliaria na produção de recomendações sobre itens avaliados por outros usuários, atribuindo um formato de recomendação ativo sobre revisões e notas. Em suma, a implementação de um sistema híbrido de recomendações auxiliaria em pontos críticos da base dados, tendo em vista que os pontos fortes de um poderiam suplementar os pontos fracos do outro.

Como exemplo disso, pode-se citar uma situação onde um item acabou de ser inserido no sistema. Como não possui nenhuma avaliação dos usuários, um sistema de recomendação baseado em filtragem colaborativa poderia encontrar problemas para descrever este item. Assim, com um sistema complementar baseado em conteúdo, e tendo as características todas já descritas, o sistema seria capaz de calcular a similaridade entre diversos outros itens para então gerar uma predição.

Em outra situação, tem-se uma base de dados cujas similaridades foram pré-definidas e calculadas. Ao longo do tempo, as características dos itens inseridos não mudarão, e a base de dados pode correr o risco de se tornar obsoleta. Com um sistema de filtragem colaborativa, os usuários podem auxiliar o sistema a atualizar a base de dados, oferecendo uma nova alternativa de recomendação por meio das revisões e notas sobre o item. Também, caso um item inserido seja totalmente novo e não compartilhe características com nenhum outro da

base, ainda sim poderá ser avaliado e classificado pelos usuários que utilizam o sistema, criando assim novos meios de recomendação.

Para a construção do grafo, temos de selecionar, para cada item, todos os outros itens que compartilham pelo menos uma característica. O Método 1 determina as palavras-chave de um item e possui uma complexidade  $O(N)$ , onde  $N$  é o número de palavras-chave na base de dados. Tendo selecionado as palavras-chave, o Método 2, que fará a seleção dos itens que possuem palavras-chave em comum, é de complexidade  $O(N*M)$ , onde  $N$  representa o número de itens presentes na base de dados e  $M$  representa o número de características do item em questão (central).

```

Para cada par <item, palavra-chave>, faça
  Se item = ITEM então
    Lista.adicionar(<palavra-chave>)
  FimSe;

```

```

FimPara
Retorna Lista

```

Método 1 - Algoritmo Seleção de palavras-chave.

```

Para cada par <item, palavra-chave>, faça
  Se Item.hashTable(<item>) não existe, então
    Para cada palavra-chave de Lista, faça
      Se <palavra-chave> = palavra-chave, então
        Item.adiciona(<item>)
      FimPara;
    FimSe;
  FimPara;
FimSe;

```

```

FimPara;

```

Método 2 - Algoritmo Seleção de palavras-chave.

Após a determinação de todos os itens similares, o Método 3 fará o cálculo da similaridade, buscando as palavras-chave em comum e então executando o cálculo por cosseno de vetores. A complexidade deste método é da ordem de  $O(N*M)$ , onde  $N$  = número de itens que compartilham pelo menos uma característica, e  $M$  = características do item central. Além disso, o cálculo do cosseno envolve um esforço computacional extra devido à potenciação e radiação de valores. Tendo em vista esta deficiência, a estratégia adota foi de selecionar uma amostra fixa dos itens similares.

```

Função COSSENO(item, ITEM):
    V1 = palavrasChave(item); // <palavras-chave>
    V2 = Lista; // (Palavras-chave de método 1)
    Retorna (V1 . V2 / |V1| * |V2|)
FimFunção

```

```

Para cada item de Item, faça
    Sim[Item] = COSSENO(Item, ITEM);
FimPara;

```

Método 3 - Cálculo da similaridade.

A forma de recomendação de itens para o usuário, como mostrado no Método 4, se dá selecionando as características de todos os itens cujo usuário teve alguma interação, como um comentário, uma avaliação ou mesmo uma adição à sua lista. Assim, tem-se inicialmente um método de complexidade  $O(N*M)$ , sendo N diferentes tipos de interação usuário/item e M o número de itens dentro da base de dados:

```

Para cada interação de Interações, faça
    Para cada par <item, palavra-chave> dentro de interação, faça
        Se Lista.hashTable(item) não existe, então
            Se item = ITEM então
                Lista.adicionar(<palavra-chave>)
            FimSe;
        FimSe;
    FimPara;
FimPara;

```

Método 4 - Selecionando características do usuário.

Ao caracterizar o usuário desta forma, pode-se aplicar o mesmo método visto anteriormente (método 3) para determinar quais os itens possuem maior semelhança baseado nos atributos descritos e então o sistema poderá realizar recomendações.

Novamente, é necessário enfatizar o uso de sistemas complementares no suporte à recomendação de itens. Outro ponto importante é a recuperação de informações na internet. Em um sistema de recomendação baseado em conteúdo, nota-se a utilidade que agentes inteligentes têm na descrição dos itens da base de dados. Com o uso de agentes inteligentes vasculhando a internet em busca de informações sobre itens, é possível descrevê-los de forma mais concisa, definindo palavras-chave e relevância de informações mais fortes para o sistema de recomendação.

Ainda em estágio inicial, a aplicação apresentada aqui possui um caráter de protótipo. Ao longo do desenvolvimento do projeto, foram constatadas diversas possibilidades de expansão para esta solução, e o principal intuito deste trabalho acabou se tornando o de ser um pontapé inicial para futuras novas aplicações.

Este trabalho abrangeu diversos aspectos sobre o desenvolvimento de uma solução. Assim como a pesquisa sobre os métodos de recomendação de produtos, outros pontos também foram visitados, como soluções estruturais (banco de dados, linguagem utilizada, ferramentas)

e de design (conceitos de disposição de elementos dentro de um *website*, experiência do usuário, *design* responsivo).

O projeto ainda teve, desde seu início, a proposta de visitar diversas áreas do desenvolvimento de uma solução, e abordar algumas de suas mais variadas técnicas e ferramentas, com a finalidade de absorver de forma ampla os diversos pontos dos processos que envolvem o desenvolvimento de uma aplicação, mais especificamente, um *website* moderno de caráter comercial.

Por fim, sob a necessidade de refinar e incrementar pontos do projeto, diversas vertentes foram abertas para pesquisa, dentre elas, se pode citar: desenvolvimento para dispositivos móveis; pesquisas sobre métodos híbridos de recomendação; inteligência artificial, mais especificamente o uso de agentes inteligentes na extração de informações; conceitos de experiência do usuário (*user experience*<sup>33</sup>) para melhorar a navegação do usuário e busca por produtos por meio do melhoramento e adequação da interface do sistema; ferramentas e linguagens para a implementação do projeto.

“The scientific man does not aim at an immediate result. He does not expect that his advanced ideas will be readily taken up. His work is like that of the planter - for the future. His duty is to lay the foundation for those who are to come, and point the way”, (TESLA, 1934).

Em suma, o conhecimento necessário para a construção de um sistema de recomendação pode ser aplicado em diversos setores, com diversas finalidades. A vantagem do uso desta abordagem é diminuir o esforço em buscas feitas pelos usuários e oferecer serviços personalizados e diretos. O grande desafio hoje na internet é a construção de uma base de dados que catalogue todas as informações da *web*, facilitando sua busca e alcance (W3C, 2001). De forma reduzida, a utilização de serviços que são capazes de vasculhar a base de dados de uma aplicação (no caso, uma lista de produtos) pode alavancar vendas e acessos, ao oferecer produtos (itens) que manifestem o interesse dos usuários. Contudo, ainda há questões delicadas a se tratar quanto à profundidade permitida destes agentes nas buscas de informações, observando o respeito à privacidade de informações dos usuários.

Quanto aos resultados do protótipo, se pode concluir que um sistema de recomendação baseado em conteúdo possui uma desvantagem clara quanto à classificação dos itens. Ao contrário da filtragem colaborativa, que se baseia na opinião dos usuários do sistema, a recomendação baseada em conteúdo precisa inicialmente caracterizar cada item, ou seja, definir atributos e pesos correspondente, para então poder oferecer o conteúdo de forma dirigida. Na filtragem colaborativa, por exemplo, a similaridade entre os produtos é definida de forma natural, se baseando nas ações de cada usuário para definir o quão próximos são dois itens distintos. Porém, a desvantagem no uso deste processo está nos itens mais “frescos”, ou seja, itens que foram adicionados recentemente e não possuem nenhuma avaliação. De forma análoga, este sistema também tem de lidar com usuários recém-chegados ao sistema, que ainda não efetuaram nenhuma avaliação ou executaram poucas ações no *website*.

Para minimizar estes problemas, uma solução proposta é a de usar mais de um método para a análise e recomendação, além da aplicação de sistemas de suporte à extração de informações e

---

<sup>33</sup> Experiência do usuário se refere à percepções de uma pessoa e as respostas que resultam do uso ou uso antecipado de um produto, sistema ou serviço (ISO, 2010, Tradução nossa).

análise de ações, para auxiliar na recomendação dos produtos. De fato, os métodos híbridos apresentam excelentes resultados e vêm se tornando cada vez mais populares em diversos tipos de aplicações, principalmente aquelas de caráter comercial.

Ainda sim, estes tipos de sistemas requerem constante atualização e revisão de seus métodos, uma vez que lidam com as incertezas de um item recomendado ser ou não selecionado pelo usuário. Embora tenha, em sua essência, a finalidade de predizer qual será o próximo item que determinado usuário irá se interessar, baseado no seu perfil e histórico de ações, não é certo que realmente o seja.

Portanto, por representar uma ferramenta de suporte que oferece estimativas para direcionar de forma mais otimizada os resultados de busca e por ter esta premissa de predileção de itens, este campo de recomendação (propaganda direcionada) é vasto e repleto de pontos a se pesquisar e aprofundar.

Além da revisão destes métodos, como manipular as informações e como calcular a similaridade entre os itens, outro ponto considerado importante foi o de ter uma visão geral sobre todo o projeto. Mais do que se preocupar com o cerne do sistema, entender como as informações geradas serão exibidas e oferecer a melhor interface para isso são questões que devem ser consideradas durante a concepção de uma ideia.

## REFERÊNCIAS BIBLIOGRÁFICAS

AGARWAL, D.; CHEN, B.-C. Machine Learning for Large Scale Recommender Systems. **University of Wisconsin Madison Yahoo! Research**, 2011. Disponível em: <<http://pages.cs.wisc.edu/~beechung/icml11-tutorial/>>. Acesso em: 19 jun. 2014.

AMATRIAIN, ; JAIMES, ; OLIVER, ; PUJOL, J. M. **Data Mining Methods for Recommender Systems**. [S.l.]: [s.n.], 2011. Disponível em: <[http://www.springer.com/cda/content/document/cda\\_downloaddocument/9780387858197-c1.pdf?SGWID=0-0-45-1007442-p173841681](http://www.springer.com/cda/content/document/cda_downloaddocument/9780387858197-c1.pdf?SGWID=0-0-45-1007442-p173841681)>. Acesso em: Abril 2014.

ANDERSON, C. **The Long Tail Why The Future Of Business Is Selling Less Of More**. [S.l.]: [s.n.], 2014. Disponível em: <<http://www.emarketingmexico.com/images/stories/ebooks/chris-anderson-the-long-tail-why-the-future-of-business-is-selling-less-of-more.pdf>>.

BERNERS-LEE, T.; FISCHETTI, M. **Weaving the Web**. San Francisco: HarperCollins, 1999. ISBN 9780062515872. Acesso em: 10 out. 2013.

BREAS, M. E. A. Constrained Clustering Algorithms: Practical Issues and Applications. **Universidad da Coruña Departamento de Computación**, 2013. Disponível em: <<http://www.dc.fi.udc.es/~edu/pubs/meares-phd.pdf>>. Acesso em: Março 2014.

BURKE, R. Hybrid Web Recommender Systems. **University of Warwick Department of Computer Science**, 2010. Disponível em: <<http://www.dcs.warwick.ac.uk/~acristea/courses/CS411/2010/Book%20-%20The%20Adaptive%20Web/HybridWebRecommenderSystems.pdf>>. Acesso em: Março 2014.

DEZA, E.; DEZA, M. **Encyclopedia of Distances**. [S.l.]: Springer, 1994. 94 p. Disponível em: <[http://en.wikipedia.org/wiki/Euclidean\\_distance](http://en.wikipedia.org/wiki/Euclidean_distance)>.

GOODREADS. Goodreads. **Goodreads**, 2007. Disponível em: <<https://www.goodreads.com>>. Acesso em: Janeiro 2014.

ISAACSON, ; SEBASTIAN,. Book Recommendations on GoodReads.com. **Stanford University**, 2008. Disponível em: <<http://cs229.stanford.edu/proj2008/IsaacsonSebastian-GoodReadsRecommendations.pdf>>. Acesso em: Janeiro 2014.

ISO. Ergonomics of human system interaction - Part 210: Human-centered design for interactive systems. **International Organization for Standardization**, 2010. ISSN 9241-210:2010. Disponível em: <[http://www.iso.org/iso/catalogue\\_detail.htm?csnumber=52075](http://www.iso.org/iso/catalogue_detail.htm?csnumber=52075)>. Acesso em: Maio 2014.

JEFFREY, V. What The Heck Is Responsive Web Design? **John Polacek**, 2011. Disponível em: <<http://johnpolacek.github.io/scrolldeck.js/decks/responsive/>>. Acesso em: Março 2014.

- JIE, Z.; GUIBING, G.; YORKE-SMITH, N. A Novel Bayesian Similarity Measure for Recommender Systems. **Nanyang Technological University**, 2013. Disponível em: <<http://www.ntu.edu.sg/home/zhangj/paper/ijcai13-guibing.pdf>>. Acesso em: Março 2014.
- LAUFER, M. A difusão do conhecimento. **Scielo**. Disponível em: <[http://www.scielo.org/ve/scielo.php?script=sci\\_arttext&pid=S0378-18442008001100004&lng=en&nrm=iso](http://www.scielo.org/ve/scielo.php?script=sci_arttext&pid=S0378-18442008001100004&lng=en&nrm=iso)>. Acesso em: Novembro 2013.
- LEVANDOWSKY, M.; WINTER, D. Distance between sets. **Nature**, v. 234, p. 34-35, May 1971. ISSN 10.1038/234034a0.
- LINDEN, G.; SMITH, B.; YORK, J. Amazon.com Recommendations Item-to-Item Collaborative Filtering. **Computer Science University of Maryland**, 2003. Disponível em: <<http://www.cs.umd.edu/~samir/498/Amazon-Recommendations.pdf>>. Acesso em: Janeiro 2014.
- LOPS, ; DE GEMMIS, ; SEMERA,. Content-based Recommender Systems: State of the Art and Trends. **Donald Bren School of Information and Computer Sciences**, 2012. Disponível em: <<http://www.ics.uci.edu/~welling/teaching/CS77Bwinter12/handbook/ContentBasedRS.pdf>>. Acesso em: Março 2014.
- MANGALINDAN,. Amazon's recommendation secret. **CNN**, 2012. Disponível em: <<http://tech.fortune.cnn.com/2012/07/30/amazon-5/>>. Acesso em: Abril 2014.
- MANNING, D.; RAGHAVAN, ; SCHÜTZE,. An Introduction to information retrieval. **The Stanford Natural Language Processing Group**, 2009. Disponível em: <<http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>>. Acesso em: Março 2014.
- MINKA,. Estimating a Dirichlet distribution. **Microsoft Research**, 2000. ISSN 16. Disponível em: <<http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirichlet.pdf>>. Acesso em: Março 2014.
- MYSQL. Why MySQL. **MySQL**, 2014. Disponível em: <<http://www.mysql.com/why-mysql/>>. Acesso em: Abril 2014.
- NETFLIX. Netflix. **Netflix**, 1997. Disponível em: <<https://www.netflix.com>>. Acesso em: Abril 2014.
- OCHI, ; DIAS, C. R.; STÊNIO, S. S. F. Clusterização em Mineração de Dados. **Universidade Federal Fluminense Instituto de computação**, 2004. Disponível em: <<http://www2.ic.uff.br/~satoru/conteudo/artigos/ERI-Minicurso-SATORU.pdf>>. Acesso em: Março 2014.
- OXFORD DICTIONARIES. Cluster meaning. **Oxford Dictionaries**. Disponível em: <[http://www.oxforddictionaries.com/us/definition/american\\_english/cluster](http://www.oxforddictionaries.com/us/definition/american_english/cluster)>. Acesso em: Abril 2014.

PARR, B. Google: There Are 129,864,880 Books in the Entire World. **Mashable**, 2010. Disponível em: <<http://mashable.com/2010/08/05/number-of-books-in-the-world/>>. Acesso em: 20 Abril 2014.

POPE, T.; KRASNER, E. A Description of the Model-View-Controller User. **Universitetet I Kobenhavn**, 1988. Disponível em: <[http://www.itu.dk/courses/VOP/E2005/VOP2005E/8\\_mvc\\_krasner\\_and\\_pope.pdf](http://www.itu.dk/courses/VOP/E2005/VOP2005E/8_mvc_krasner_and_pope.pdf)>. Acesso em: Fev 2014.

PSOMAS,. The Five Competencies of User Experience Design. **Uxmatters**, 2007. Disponível em: <<http://www.uxmatters.com/mt/archives/2007/11/the-five-competencies-of-user-experience-design.php>>. Acesso em: Março 2014.

RESNICK, P.; VARIAN, R. H. Recommender systems. **University of Texas at Austin**, 1997. Disponível em: <[https://www.ischool.utexas.edu/~i385d/readings/Resnick\\_Recommender\\_97.pdf](https://www.ischool.utexas.edu/~i385d/readings/Resnick_Recommender_97.pdf)>. Acesso em: Janeiro 2014.

RUSSO, R. Bayesian and Neural Networks for Motion Picture Recommendation. **Boston College**, 2006. Disponível em: <[https://www.bc.edu/content/dam/files/schools/cas\\_sites/cs/local/bach/2006/06RobertRusso.pdf](https://www.bc.edu/content/dam/files/schools/cas_sites/cs/local/bach/2006/06RobertRusso.pdf)>. Acesso em: 19 jun. 2014.

SCHIVINSKI, B.; DABROWSKI, D. The effect of social media communication on consumer perceptions of brands. **Journal of Marketing Communications**, p. pp. 1-26, 20 Janeiro 2014. Disponível em: <<http://www.tandfonline.com/doi/full/10.1080/13527266.2013.871323#.U2u19PldWZx>>.

SHIRKY, C. Power Laws, Weblogs, and Inequality. **Clay Shirky's Writings About the Internet**, 2003. Disponível em: <[http://www.shirky.com/writings/powerlaw\\_weblog.html](http://www.shirky.com/writings/powerlaw_weblog.html)>. Acesso em: Abril 2014.

TANIMOTO,. **An Elementary Mathematical theory of Classification and Prediction**. IBM. [S.l.], p. 10. 1957.

TESLA, N. Radio Power Will Revolutionize the World. **Modern Mechanics and Inventions**, Julho 1934. 6. Disponível em: <<http://blog.modernmechanix.com/2008/07/14/radio-power-will-revolutionize-the-world/>>. Acesso em: Abril 2014.

TOSCHER, ; JAHNER,. The BigChaos Solution to the Net. **Commendo**, 2009. Disponível em: <[http://www.commendo.at/UserFiles/commendo/File/GrandPrize2009\\_BigChaos.pdf](http://www.commendo.at/UserFiles/commendo/File/GrandPrize2009_BigChaos.pdf)>. Acesso em: Março 2014.

UNGAR, H.; FOSTER,. Clustering Methods for Collaborative Filtering. **ASSOCIATION FOR THE ADVANCEMENT OF ARTIFICIAL INTELLIGENCE**, 1998. Disponível em: <<http://www.aaai.org/Papers/Workshops/1998/WS-98-08/WS98-08-029.pdf>>. Acesso em: Abril 2014.

W3C. The Semantic Web Made Easy. **W3C**, 2001. Disponível em:  
<<http://www.w3.org/RDF/Metalog/docs/sw-easy>>. Acesso em: Janeiro 2014.

WIKIPEDIA. Semantic Web. **Wikipedia the free Encyclopedia**, 2001. Disponível em:  
<[http://en.wikipedia.org/wiki/Semantic\\_Web](http://en.wikipedia.org/wiki/Semantic_Web)>. Acesso em: Novembro 2013.

WIKIPEDIA. Pearson product-moment correlation coefficient. **Wikipedia**, 2003. Disponível em: <[http://en.wikipedia.org/wiki/Pearson\\_product-moment\\_correlation\\_coefficient](http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient)>. Acesso em: Mar 2014.

WIKIPEDIA. Crowdsourcing. **Wikipedia The Free Encyclopedia**, 2006. Disponível em:  
<<http://en.wikipedia.org/wiki/Crowdsourcing>>. Acesso em: Abril 2014.