Sistema de Recomendação Baseado em Conteúdo Textual em Páginas Web

Haroldo Shigueaki Teruya[1], **Ronaldo Celso Messias Correia**[2]

[1] Faculdade de Ciências e Tecnologia
Universidade Estadual Paulista "Júlio de Mesquita Filho" (FCT/UNESP)
Departamento de Matemática Computação
Caixa Postal 468 – CEP 19060-900 – Presidente Prudente – SP – Brasil

haroldo.s.teruya@gmail.com, (ronaldo, joaosvaldo)@fct.unesp.br

Abstract. *Recommendation Systems has emerged to assist users against information overload in a environment where not all information is relevant or of interest to the user. Recommendation Systems came with the proposal of being an independent area in Computer Science, rather than filtering information, the focus is to recommend information based on preferences, tastes and interests surprising with new information and helping in the pursuit of the real need. In this paper we present a recommendation system, that is compose by an Middleware and an Framework, for Web environment based on textual content and the sequence of processes to arrive at a recommendation for the user, from the construction of the user's profile (user interest) to the recommendation based on the profile. This article also presents the strategies, technologies and areas related to Computer Science at each stage of the recommendation process.*

Resumo. *Sistemas de Recomendação surgiu para auxiliar os usuários em relação a sobrecarga de informação, em um meio onde nem toda informação é relevante ou de interesse para o usuário. Sistemas de Recomendação surgiu com a proposta de ser uma área independente na Ciência da Computação, mais do que filtrar informação, o foco é recomendar informação baseado nas preferências, gostos e interesses surpreendendo com novas informações e auxiliando na busca da real necessidade. Neste trabalho é apresentado um sistema de recomendação, composto por um Middleware e um Framework, para ambiente Web baseado em conteúdo textual e a sequencia de processos para chegar em uma recomendação para o usuário, desde a construção do perfil do usuário (interesse do usuário) até a recomendação baseado no perfil. Neste artigo ainda é apresentado as estratégias, tecnologias e áreas relacionadas com a Ciência da Computação em cada etapa do processo de recomendação.*

1. Introdução

A alta geração e disponibilidade de dados facilitada através da *Internet* possibilita encontrar uma diversidade muito grande de opções quando buscamos por uma informação específica. Explorar e analisar este vasto volume de informação na *Web* está se tornando cada vez mais complexo e demorado. Várias páginas *Web* podem nunca ser descobertos mesmo que o seu conteúdo seja mais relevante e útil para o usuário. Geralmente quando o usuário não possui experiência suficiente e deseja encontrar a informação que

procura o mais rápido possível, ele confia nas recomendações que são passadas de outras pessoas, as quais podem chegar de forma direta e confiável na fonte (”word-of-mouth”) [Shardanand e Maes 1995].

Os sistemas de recomendação surgiu com a proposta de auxiliar no aumento da capacidade e eficácia deste processo de recomendação já bastante conhecido na relação social explorando as preferências dos usuários do sistema [Resnick e Varian 1997]. Ao longo dos anos de estudos e pesquisas publicados sobre sistemas de recomendação, houve um grande interesse da indústria (*e-business*) sobre a área. Os sistemas de recomendação se tornaram indispensáveis para todo *Site*, *blogs*, de *notícias* entre outros, influenciando a maior parte do nosso dia-a-dia em tomadas de decisão. Trabalhos, conferências, livros, artigos e até prêmios de um milhão de dólares[1] são dedicados à essa área. Todos os usuários desses *Sites* confiam em algum grau nas recomendações ou são surpreendidos com o conteúdo oferecido. Esses *Sites* se apoiam nesses sistemas de recomendação para conquistar e aumentar o engajamento e confiança com os usuário.

No Sistema de Recomendação desenvolvido é utilizado a técnica de **Visualização de Informação**, o ***Tag Cloud*** (Seção 2.3), para ter uma ideia geral do conteúdo que o usuário está consumindo. Este artigo apresentará cada etapa (Seção 3) de acordo com o fluxo de processos (Figura 2) desde a iteração do usuário-sistema até a recomendação de tópicos. O sistema de recomendação consiste em um *Framework* na camada do *front-end* e um *Middleware* na camada do *back-end* integrado ao SGDB do *Site*. O *Framework* possui o papel de coletar, modelar e de enviar o ”perfil do usuário”para o *Middleware* que por sua vez criar as recomendações baseado no perfil do usuário.

O texto deste documento está organizado em cinco seções, contando em esta seção de Introdução. Na Seção 2 é realizada uma breve contextualização da área estudada para o desenvolvimento do presente trabalho. Na Seção 3 é descrito os processos e as abordagens utilizados para realizar uma recomendação desde a coleta dos dados do usuário. E finalmente na Seção 5 as considerações finais e a conclusão serão apresentadas, bem como sugestões de aperfeiçoamento e trabalhos futuros.

2. Fundamentação teórico

Sistemas de Recomendação possui várias áreas que o influencia de alguma forma, as **Ciências Cognitivas**, **Teoria da Aproximação**, **Recuperação de Informação**, **Teoria da Previsão**, **Ciência da Gestão** e até mesmo **Marketing ao Consumidor**. Ao longo dos anos de pesquisa, foi proposto **Sistemas de Recomendação** como uma área independente, devido ao processo de recomendação de conteúdo não classificados, e sim de acordo com as preferências do usuário [Resnick et al. 1994]. Inicialmente sistemas de recomendação foram chamados de sistemas de filtragem colaborativa, devido à utilização do primeiro sistema de recomendação em 1992, o ***Tapestry*** [Goldberg et al. 1992, Resnick et al. 1994, Medeiros 2013]. No entanto, foi Paul que deu início aos estudos na área, e mais tarde propôs um termo mais genérico em seu trabalho com Varian em 1997 [Resnick e Varian 1997]. O termo sistema de recomendação foi adotado pois não era necessariamente um produto da colaboração entre usuários. Mais do que filtrar informações, o foco era na recomendação de informações [Resnick e Varian 1997, Adomavicius e Tuzhilin 2005, Medeiros 2013]. Segundo Gediminas e Alexander

[1] http://www.netflixprize.com/

[Adomavicius e Tuzhilin 2005] o problema de recomendação pode ser simplificado ao problema da estimativa de classificação para as informações que ainda não foram visualizados pelo usuário.

Com a finalidade de um melhor entendimento, neste artigo é utilizado o termo "item" para designar o que o sistema recomenda ao usuário, podendo ser, filme, música, vídeo, roupa, *Hyperlinks* de outras páginas *Web* e até pessoas que fazem parte de um universo de informação. E *feedback* como a ação, interação, resposta e consumo de dados do usuário com o sistema.

Gediminas e Alexander [Adomavicius e Tuzhilin 2005] propôs a seguinte definição de que seja $\boldsymbol{U}$ um conjunto que represente todos os usuário, $\boldsymbol{I}$ um conjunto de todos os possíveis itens a serem recomendados, $\boldsymbol{\ddot{u}}$ a função de medição do quão relevante um determinado $\boldsymbol{i}$ (item) para um $\boldsymbol{u}$ (usuário) e U X I para um A, onde A é um conjunto ordenado. Resumidamente, $\boldsymbol{u} \in \boldsymbol{U}, \boldsymbol{i'u} = \boldsymbol{argmaxs} \in \boldsymbol{Iu}(\boldsymbol{\ddot{u}}, \boldsymbol{i})$.

Geralmente, os usuário avaliam um conjunto muito menor que $\boldsymbol{I}$. O problema pode ser estruturado como uma matriz esparsa, de acordo com a Figura 1:

Figura 1. Matriz esparsa *A*.

Usuários \ Itens	1	2	...	***i***	...	m
1	5		1		3	
1		2	4	2		1
...	1	5	3		3	2
u		2		1	4	1
...	1		1	2	1	
n	1	2			5	

Fonte própria.

A Figura 1 representa uma matriz de n usuários e m itens e cada célula *A(u,i)* corresponde à uma avaliação do usuário u ao item i. De acordo com a Figura 1, podemos notar com quais itens o usuário u se relaciona. O desafio é sempre superar a função $\boldsymbol{\ddot{u}}$ para todo o conjunto $\boldsymbol{U}$ X $\boldsymbol{I}$ para que o usuário u tenha novas experiências e conhecimento de todo o conteúdo em que ele está inserido [Adomavicius e Tuzhilin 2005].

2.1. Perfil do Usuário

Para que um sistema possa recomendar item(s) a um usuário, é preciso de alguma forma saber as preferências, gostos, interesses e se possível a real necessidade do usuário. Em Sistemas de Recomendação o perfil do usuário consiste em uma estrutura de dados que o

represente e precisa de funções específicas para tratar o usuário individualmente em um ambiente que existe vários usuários considerando estratégias de coleta para a construção do perfil.

2.1.1. Coletando dados do usuário

Quanto mais dados do usuário o sistema adquirir, a tendência é aumentar a qualidade do conteúdo que será recomendado, analogamente, o Sistema "aprenderá"gradativamente as preferências e a real necessidade do usuário. A área de **Aprendizagem de Máquina** pode auxiliar com o refinamento do perfil do usuário com base nos dados que são coletados ao longo da interação usuário-sistema. As técnicas de Aprendizado de Máquina são baseados nos princípios de aprendizado indutiva da Área de **Inteligência Artificial**. O aprendizado indutivo é implementado por meio de algoritmos que processam um conjunto de dados e extraem um modelo capaz de representar. É de extrema importância a eficiência desta etapa pois a recomendação de conteúdo depende da qualidade dos dados coletados e as necessidades do usuário podem mudar. Quanto mais específico os dados do usuário, mas específico serão as recomendações e a acurácia da recomendação mesmo com a mudança das necessidades do usuário. É apresentado em seguida as estratégias de coleta de dados utilizadas de forma explícita e implícita do usuário baseado nas interação com o sistema.

As seguintes estratégias de coletas são utilizadas:

Monitoração de atividade: Essa abordagem é ideal para os usuários que não estão dispostos a desprender tempo em questionários ou cadastros, desse modo podemos coletar dados demográficos, interesses de tópicos visitados, *Hyperlinks* e páginas *Web*. *Hyperlinks* podem ser adicionados, modificados, removidos, reorganizados ou comentados. O sistema pode apresentar, esconder ou enfatizar fragmentos de uma página, assegurando que seu conteúdo inclua a informação apropriada, em um nível adequado de dificuldade ou detalhe [Cazella et al. 2010, Reategui et al. 2006].

Cadastro: O Sistema pode oferecer uma interface para o usuário, que disponibiliza ao usuário uma área de cadastro com informações pessoais. Essas informações ficam armazenados em uma base de dados [Reategui et al. 2006].

Cookie: Essa abordagem é citado no artigo de Eliseo, Sílvio e Fernando [Reategui et al. 2006] como identificação no cliente, referência ao navegador de *Internet*. Essa identificação utiliza normalmente *cookies*, um mecanismo pelo qual um *Web Site* consegue identificar o acesso do computador no sistema. Útil para armazenar e recuperar o perfil do usuário, dessa forma, não é preciso recomeçar a construção do perfil do usuário em um futuro acesso ao *Site*.

Exibindo Exemplos: Utilização de um conjunto de exemplos de interação de usuário com o sistema. Esse conjunto é utilizado para inferir a construção do perfil do usuário. Por exemplo, o sistema solicita ao usuário avaliar um conjunto de informações exemplos entre relevante e irrelevante. Esse método é adequado para o sistema identificar de maneira fácil as preferências do usuário, identificar conteúdo que gostaria e que não gostaria que fossem recomendado [Reategui et al. 2006].

2.1.2. Construção do perfil do usuário

Após a etapa de coleta dos dados do usuário, é preciso construir o perfil do usuário, existem várias abordagens que podem ser aplicadas, todos de alguma forma possuem alguma base nos **Modelos de Recuperação de Informação** (**??**). O trabalho de Eliseo, Sílvio e Fernando [Reategui et al. 2006], para auxiliar na representação do perfil do usuário:

Histórico: É uma lista de conteúdo já visitados ou avaliados pelo usuário de alguma forma. Alguns sistemas utilizam duas listas, uma para conteúdo dito interessantes e outra não interessante. É muito utilizado em *Sites* na área de *e-business* e de busca, onde é armazenado o histórico de *Hyperlinks* acessados. Com o histórico de *Hyperlinks* acessados, o sistema recomendador possui a capacidade de ter conhecimento do conteúdo acessado. Atualmente as páginas *Web* já possuem o seu conteúdo estruturado ou classificado, não sendo necessário a análise do conteúdo [Silva 2016b].

Vetorial: Baseado nos Modelos de Recuperação de Informação 2.4.3, é o modelo mais utilizado para modelar o perfil do usuário com base em conteúdo textual (não estruturado).

Matriz de Avaliação Usuário-Item: É o mais utilizado em sistemas com muita interação entre os usuários em um mesmo ambiente, como as redes sociais. Alguns sistemas de filtragem colaborativa mantêm uma matriz de avaliações de usuário-item, como um perfil do usuário, pois nesta matriz é possível verificar o que é mais ou menos interessante para o usuário. Cada célula da matriz contém uma avaliação representando a avaliação do usuário para um determinado item, caso exista um vazio, significa que o usuário não avaliou o referido item.

2.2. Abordagens de Sistema de Recomendação

Ao longo dos anos de pesquisas e estudos, algoritmos, estratégias e heurísticas foram surgindo e se aprimorando. Vários pesquisadores [Resnick et al. 1994, Shardanand e Maes 1995, Balabanovic e Shoham 1995, Goldberg et al. 1992, Resnick et al. 1994, Shardanand e Maes 1995, Balabanovic e Shoham 1995, Reategui et al. 2006, Cazella et al. 2010, Lops et al. 2011, Medeiros 2013] utilizam várias classificações de Sistemas de Recomendação. Cada abordagem possui a sua própria estratégia para realizar a recomendação, com as suas vantagens e limitações.

2.2.1. Abordagem Baseado em Conteúdo

O processo de selecionar itens semelhantes do usuário com os itens do sistema, é um problema de Recuperação de Informação (2.4), onde os itens preferidos é utilizado para a atividade de consulta no universo de informação do sistema [Reategui et al. 2006, Silva 2016a, Baeza-Yates et al. 1999, Medeiros 2013].

Essa abordagem se baseia na premissa de que os usuários gostariam de receber recomendações de itens semelhantes aos itens preferidos do passado [Balabanovic e Shoham 1995, Medeiros 2013]. Os sistemas que aditam esse paradigma não necessitam da colaboração de avaliação de itens e da comparação dos perfis de outros usuários do sistemas. Os recomendadores baseados em conteúdo lida somente com as

avaliações de itens do próprio usuário e a construção do perfil do usuário deve ser muito mais elaborado em comparação com as outras abordagens, em compensação, o sistema se torna muito mais personalizado e possui a capacidade de recomendar novos itens ao usuário que ainda não foram avaliados, incluindo itens que nunca foram avaliados por nenhum usuário do sistema.

No entanto, essa abordagem possui algumas limitações. Nenhum sistema de recomendação pode fornecer sugestões adequadas se o perfil do usuário analisado não contém informação suficiente para discriminar itens entre relevantes e não relevantes.

2.2.2. Abordagem Baseado em Filtragem Colaborativa

Essa abordagem foi desenvolvida para aplicar em alguns pontos que estavam em aberto na abordagem baseado em conteúdo e a principal diferença é que não exige a análise do conteúdo dos itens [Reategui et al. 2006]. Essa abordagem utiliza da premissa de que se um usuário do sistema gostou de um item, então outros usuários com perfil semelhante também irão gostar [Silva 2016b].

2.2.3. Abordagem Híbrida

A abordagem híbrida consiste no uso das duas abordagens, a abordagem baseado em conteúdo e abordagem baseado em filtragem colaborativa.

Há um aspecto complementar para as duas abordagens principais de recomendação, pois justamente onde uma técnica limita-se a outra avança transformando este fato em uma oportunidade. Este tipo de sistema é denominado híbrido e pode, por exemplo, surpreender o usuário apresentando itens de interesse de outro usuário com gosto similar.

2.2.4. Abordagem Baseado em Contexto

Com a finalidade de aperfeiçoar as abordagens anteriores, vários pesquisadores desenvolveram novos algoritmos ou adicionaram novos elementos ao recomendado, além de dados sobre os itens ou informações, é considerado outras informações, como localização geográfica, orientação religiosa ou política dos usuário para identificar as preferências do usuário a curto e longo prazo [das Neves et al. 2013, Adomavicius e Tuzhilin 2005, Silva 2016b].

Para que os sistemas de recomendação se desenvolvam na nova geração, deve aprimorar a construção do modelo de perfil do usuário, onde este perfil deve ser considerado como experiências reais. Considerando o percurso de um sistema de computador para atingir a mínima compreensão desta otimizada interação de como os humanos procedem nos seus processos de recomendação na ”vida real”, pesquisadores da *W3C Emoticon Incubator Group* [W3C 2009] têm desenvolvido uma *Linguagem de Marcação* para representar emoções [Cazella et al. 2010].

2.3. *Tag Cloud*

Tag Cloud é uma técnica da área de **Visualização de Informação** que tem por objetivo, por meio da computação gráfica, dar sentido a estes dados [Andreotti e Eler 2016].

A partir da visualização é possível ter uma ideia geral do assunto tratado, sem que seja necessário ler todo o conteúdo e para selecionar as informações relevantes e tornar a exploração mais amigável. Esta técnica consiste na criação de nuvem de palavras fazendo distinção entre elas pelo tamanho da fonte ou cor, destacando-as na visualização levando em consideração a quantidade de vezes que aparecem no texto, ou seja, criação de uma *tag cloud*, nuvem de palavras, que pertence ḿ conjunto de palavras, texto ou documentos, possuem base na área de **Recuperação de Informação** (Seção 2.4), mais especificamente, no **Modelo Vetorial** (Seção 2.4.3) [Andreotti e Eler 2016].

2.4. Recuperação de Informação

Para recomendar uma informação, deve-se buscar de alguma forma os dados em um conjunto maior de documentos, muitas vezes o conteúdo informativo recuperadas são irrelevantes ou com uma quantidade muito elevado de resultados e podem apresentar uma grande variância ou dispersão de informação.

Para que o sistema identifique que uma informação seja relevante ou irrelevante é um tema central da Ciência da Informação. A relevância possui um caráter muito subjetivo na área de Recuperação de Informação, uma vez que está relacionado com as necessidades do usuário. Já que uma informação pode ser relevante para um usuário e irrelevante para outro. Os primeiros Sistemas de Recomendação tiveram seu desenvolvimento baseado em algoritmos da área de Sistemas de Recuperação de Informação. No artigo de Renato [Souza 2006], ele afirma que Sistemas de Recuperação de Informação fazem interface entre uma coleção de informação com um ou mais usuários, ou seja, Sistemas de Recuperação de Informação fazem a ponte entre os criadores de informações e os usuários dessas informações. Então a tarefa de um Sistema de Recomendação é de tomar para si a responsabilidade de recomendar esses conteúdos para os usuários consumidores.

Como descrito na Seção 2.2.1, o problema de comparar o perfil do usuário com o conteúdo é um problema da Área de Recuperação de Informação. É tratado as preferências dos usuários como uma consulta em um conjunto de itens, no entanto, a relação entre um item e o seu conteúdo informativo é muito vago, principalmente com conteúdo textual [Foskett 1972].

Várias estratégias possuem uma base na área de inteligência artificial, alguns utilizam o poder de processamento dos computadores, outros se baseiam em modelos probabilísticos. Esses modelos possuem em comum a utilização do perfil do usuário como base na realização da consulta [Baeza-Yates et al. 1999, Souza 2006].

2.4.1. Modelos Clássicos

Nos modelos clássicos cada documento é descrito por um conjunto de palavras-chave representativas, também chamadas de termos de indexação, que busca representar o assunto do documento e sumarizar seu conteúdo de forma significativa.

2.4.2. Booleano

Inicialmente os Sistemas de Recuperação de Informação tiveram a sua base no modelo booleano, onde os usuários combinavam as operações ”AND”, ”OR”e ”NOT”para estabelecer relações de ocorrência com as palavras-chave. No entanto, esses documentos são analisados de forma que, a decisão será binária, entre interessantes ou não interessantes. Além de que não é criada nenhuma espécie de ordenação dos resultados que atendam às condições de consulta. Existem alguns modelos alternativos ao *booleano*, apresentados a seguir:

Lógica Difusa ou Nebulosa (*fuzzy*): Esse modelo atribui um grau de relevância para cada documento, com isso, é possível criar um conjunto de documentos ordenados.

Booleano Estendido: Esse modelo atribui pesos aos termos, com isso, existe uma superação nas decisões binárias e se aproxima do modelo vetorial.

2.4.3. Vetorial

No trabalho de Renato [Souza 2006], é descrito um conjunto de palavras como ”*bag-of-words*”(sacos de palavras) que são representados estruturalmente em um vetor possivelmente com várias dimensões. Cada dimensão representa o total de termos. Esse modelo atribui um peso para cada termo tanto aos termos da expressão de busca como aos termos de indexação que representam o conjunto. Um documento é representado por um conjunto de termos de indexação, cada qual associado a um valor numérico entre 0 e 1, que representa a relevância do respectivo termo na representação do conteúdo informacional do documento. Uma expressão de busca é também representada por um conjunto de termos e seus respectivos pesos, que representam a importância do termo na expressão de busca [Silva 2016b, Kuramoto 1996].

”bag-of-words:”: Esse modelo leva em conta a frequência de cada termo em um texto ou documento, definido como *tokenização*. Em seguida deve ser removido os *stopwords*, que são os termos pouco relevantes de acordo com o contexto [Alencar 2013]. Posteriormente, também conhecido como modelo vetorial, o conteúdo é representado por um vetor de frequências dos termos que ele ocorrem.

Indexação Semântica Latente: Esse modelo propõe estabelecer o casamento conceitual entre o documento e a expressão dos termos utilizado na busca pelo usuário.

Redes Neurais: De acordo com os autores Silva e Souza [**?**, Souza 2006], redes neurais artificiais utilizam de padrões para relacionar as expressões de busca dos usuários com os documentos de um conjunto maior, de modo que cada expressão de busca libera um sinal que ativa os termos do sistema e que se propaga aos documentos relacionados. Por meio desse processo, o usuário pode se deparar com documentos com termos que não foram utilizados na busca, mas que demonstraram ter relação com a expressão pesquisada.

2.4.4. Probabilístico

Nesse modelo, supõe-se que exista um conjunto ideal de documentos que satisfaz a cada uma das consultas ao sistema, e que este conjunto pode ser recuperado. Através

de tentativa inicial com um conjunto de documentos (para a qual se podem utilizar técnicas de outros modelos, como o vetorial) e do *feedback* do usuário em sucessivas interações, busca-se aproximar cada vez mais deste conjunto ideal, por meio de análise dos documentos considerados pertinentes pelo usuário. O valor desse modelo está em considerar a interação contínua com o usuário como um caminho para refinar o resultado continuamente [Souza 2006].

Redes de Inferência: Nesse modelo, associam-se variáveis aleatórias ao evento do atendimento de uma busca específica por um documento específico. Essas variáveis podem ser alteradas de acordo com os eventos futuros, de forma a estabelecer relacionamentos baseados nos eventos observados [Souza 2006].

Redes de Crença: Nesse modelo, similares às redes de inferência, documentos são modelados como subconjuntos de um espaço de conceitos. A cada documento, associa-se a probabilidade de que o mesmo cubra os conceitos presentes no espaço de conceitos. Cada busca é mapeada no espaço de conceitos, que por sua vez, está conectado ao espaço de documentos [Souza 2006].

2.4.5. Classificação de Conteúdo Textual

O processo de avaliar o conteúdo consiste em predizer a qual classe de interesse pertence baseado no perfil do usuário construído. Os classificadores, além de definir a classe, interessantes e não interessantes, também podem definir a relevância do conteúdo em grau. O **Modelo Vetorial** descrito na Seção 2.4.3 é o mais utilizado.

Alguns dos principais classificadores de textos que podem ser aplicados em Sistemas de Recomendação são Naïve Bayes e Algoritmo de Rocchio [Aas e Eikvil 1999, Baharudin et al. 2010, Silva 2016a].

2.5. Tarefa de Recomendação

No trabalho de Rafael [Silva 2016b] é apresentado os seguintes cenários de itens e para avaliá-los é necessário compreender quais objetivos e tarefas eles realizam:

Anotação de contexto : Essa categoria são de sistemas que apoiam o usuário na tomada de decisão baseado no contexto. Sugerir outros destinatários em uma mensagem de correio eletrônico, baseando-se na presença de outros destinatários, ou mesmo no assunto, é um exemplo de anotações em contextos utilizados em programas de e-mail.

Encontre bons itens : O recomendador fornece ao usuário uma lista de itens ordenados baseado em seu interesse e limitada a uma quantidade K de itens.

Encontre todos os bons itens : Semelhante ao itens anterior, este cenário apresenta todos os itens recuperados na ordem de relevância.

Recomendação em sequência : Alguns sistemas precisam prever uma sequência de itens que melhore a experiência do usuário. Por exemplo, para um pesquisador que deseja iniciar uma pesquisa em uma área, o sistema poderia sugerir uma sequência de artigos que deseja iniciar uma pesquisa em uma área.

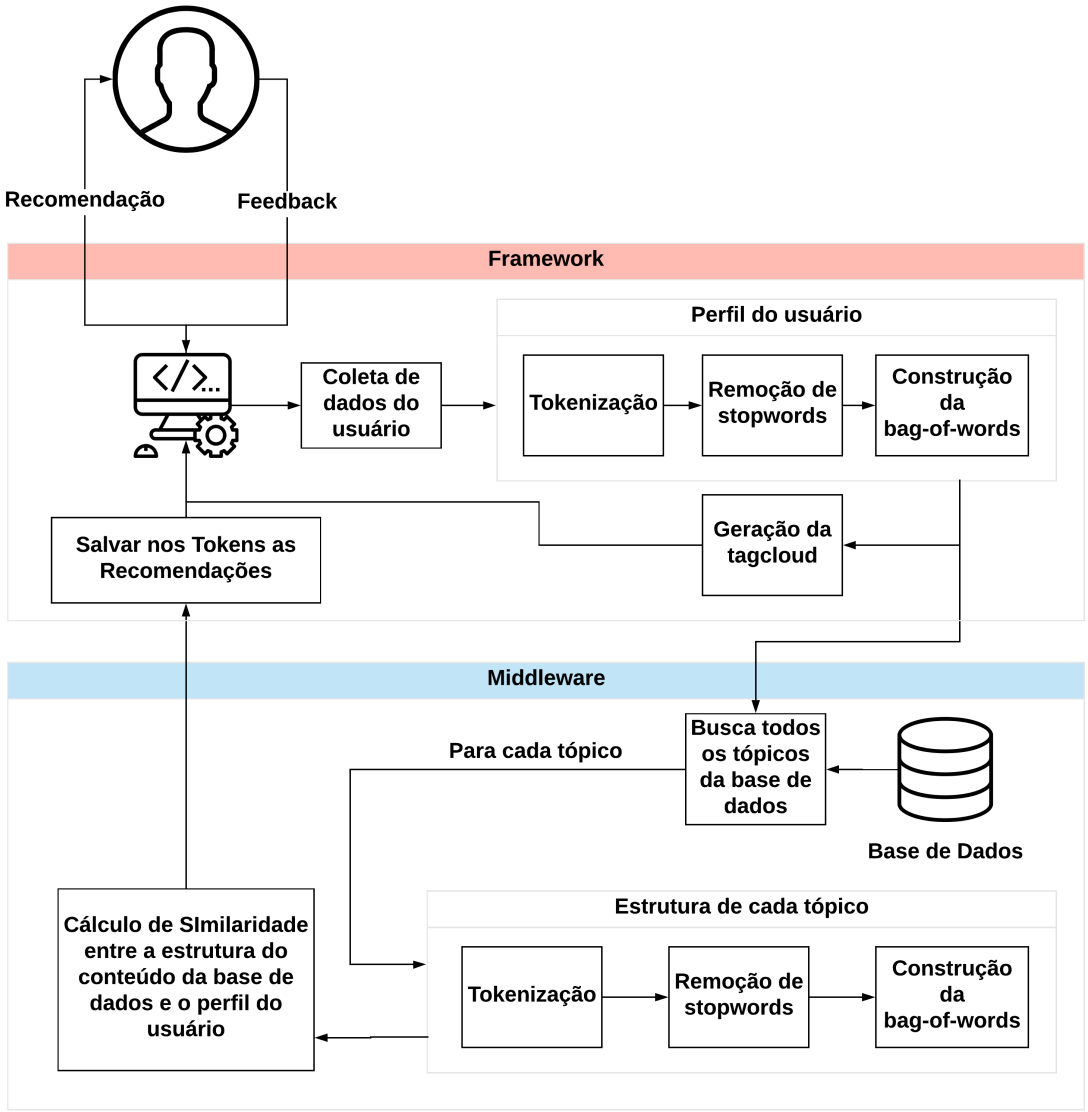
Somente navegando : Muitas vezes, para uma certa ocasião, pode ser interessante ao usuário receber recomendações sem que haja um motivo em particular, neste caso, a

qualidade da apresentação da recomendação é mais importante que a qualidade da mesma.

Encontre a recomendação confiável : Usuários podem realizar ataques aos sistemas de recomendação aos sistemas de recomendação alterando suas preferências simplesmente para identificar seu grau de confiança. Neste caso o usuário não está interessado nas recomendações, mas sim em testar o quão robusto é o sistema.

3. Abordagem proposta

Figura 2. Fluxo de processos do Sistema de Recomendação Baseado em Conteúdo Textual em Páginas Web.



Fonte própria.

A Figura 2 ilustra as etapas do Sistema de Recomendação proposto de forma sucinta e objetiva, partindo da interação do usuário com o sistema até a recomendação.

O sistema de recomendação proposto utiliza as seguintes tecnologias:

- A linguagem de marcação HTML5 visando a garantia de portabilidade para os navegação mais atuais;
- A folha de estilos (*Cascading Style Sheet*) para customizar o estilo dos documentos *Web*;
- O pré-processador *SASS* (*Syntactically Awesome Style Sheets*). Uma extensão do *CSS* que permite implementar utilizando variáveis, regras entre outros.
- A linguagem alto nível interpretada de programação JavaScript para lidar com o comportamento da página, comunicação com o servidor, as interações usuário-interface e os vários processos a serem realizados;
- Biblioteca *open source* Bootstrap para desenvolvimento *Web* com *HTML5*, *CSS* e *JavaScript*.
- Biblioteca *open source jQCloud* [2] para construir em puro HTML e CSS os *tagcloud* (Seção 2.3) baseados na *bag-of-words*.
- *Framework* de alto nível utilizando a linguagem Python *Django*. Está presente no *front-end* com a sua linguagem de *template* e no *back-end*;
- Mesmo que implícito, foi utilizado o Sistema de Gerenciamento de Base de Dados Relacional *SQLite*;

A seguir são descritas com mais detalhes cada uma das etapas do processo de recomendação proposto, dividido em duas frentes, *Framework* e *Middleware*.

3.1. O *Framework*

O *Framework* atua na camada do *front-end* da aplicação *Web* e a sua finalidade é de coletar, estruturar e enviar para o *Middleware* os dados do usuário. Assincronamente, o *Framework* é responsável por receber as recomendações baseado nos dados enviados do *Middleware* e exibir na interface do usuário.

3.1.1. Coleta de dados do Usuário: Monitoração de Atividades

Esta etapa coleta o conteúdo que usuário está consumindo em tempo real com base em uma biblioteca [3] desenvolvida em *JavaScript*, o *screentime.js*. A coleta dos dados é realizado implicitamente, ou seja, o usuário não é notificado enquanto navega pelo *Site*. Esta etapa visa a usabilidade do usuário e utiliza a técnica de monitoração [Reategui et al. 2006, Cazella et al. 2010]. A coleta de dados é realizado periodicamente.

3.1.2. Construção do perfil do usuário: *"bag-of-words"*

Nesta etapa ocorre a classificação dos dados coletado na etapa anterior, já que o texto puro é um conjunto de dados não estruturado, é utilizado as técnicas de Recuperação de Informação 2.4 para estruturar em um vetor, mais especificamente, o Modelo de Recuperação de Informação Vetorial 2.4.3. A *"bag-of-words"* será construído a partir do texto coletado e os termos mais relevantes serão aqueles com maior frequência ou quantidade no vetor. Essa estrutura de dados representará o perfil do usuário e será enviado para o *Middleware* para construir as recomendações.

[2] http://mistic100.github.io/jQCloud/
[3] http://screentime.parsnip.io/

3.1.3. Geração da *tag cloud*

Com a *"bag-of-words"* construída, nesta etapa, é gerado e disponibilizado na interface do usuário a "nuvem de palavras"utilizando a biblioteca *jQCloud*, desta forma, pode ser visualizado a estrutura de dados que representa o perfil do usuário.

3.1.4. Tarefa de Recomendação: Encontrando Bons Itens

Nesta etapa, o *Framework* recebe de forma assíncrona utilizando o *AJAX* o vetor de recomendação construído pelo serviço *Django*, mais especificamente pelo *Middleware*. Utilizando a técnica apresentado no trabalho de Rafael [Silva 2016b], será apresentado os três *Bons Itens Encontrado* descrito na Seção 2.5.

3.2. O *Middleware*

O *Middleware* atua na camada do *back-end* da aplicação *Web* e a sua finalidade, assim como a definição de *middleware*, é de receber uma entrada de dados, processar e retornar para o *Framework* uma saída de dados sem influenciar as outras funcionalidades do *Site*. Para um melhor entendimento da manipulação do conteúdo da base de dados, é utilizado o termo "item"para representar um tópico, na qual para a aplicação *Web*, tópico é considerado como um conjunto de dados pré-definido. O *Middleware* recebe uma entrada, o perfil do usuário, busca todos os itens da base de dados, estrutura em *"bag-of-words"* cada item, realiza o cálculo de similaridade dos itens com o perfil do usuário para gerar as recomendações. O *Middleware* foi desenvolvido em *Django* e todos os processos nesta camada foi utilizado a linguagem de programação *Python*.

3.2.1. Busca todos os tópicos

Nesta etapa ocorre o primeiro contato com o *Framework* recebendo a *"bag-of-words"* do perfil do usuário e realizando *queries* dos itens da base de dados.

3.2.2. Estruturação dos tópicos

Nesta etapa é realizado o processo de estruturação do conteúdo de cada item. O conteúdo informativo de cada item é textual, portanto são dados não estruturados, é utilizado o **Modelo Vetorial** descrito na Seção 2.4.3 e como resultado, cada item será estruturado em um vetor de termos, onde cada termo possui um peso de acordo com a quantidade deste termo no conteúdo informativo do tópico, ou seja, a "*bag-of-words*".

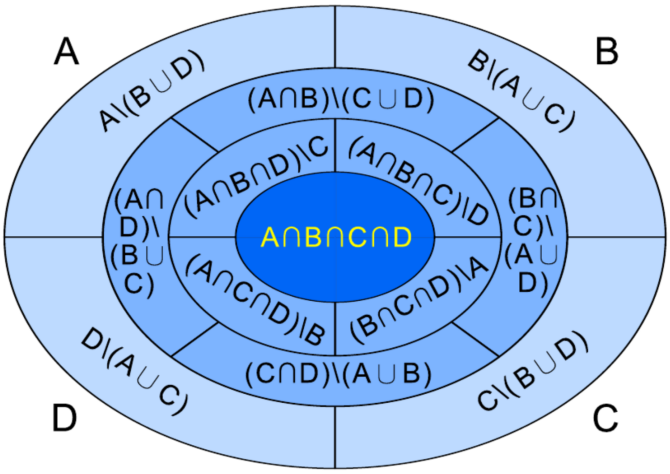
3.2.3. Cálculo de Similaridade

Nesta etapa, é definido o modelo vetorial de cada tópico como um conjunto para um melhor entendimento do processo de similaridade entre itens. Por meio de uma fusão sistemática, pontos em comum e as diferenças entre os conjuntos são identificados, além de evitar redundância.

Este processo de cálculo de similaridade é baseado na técnica *ConcentriCloud* descrito no trabalho de André e Danilo [Andreotti e Eler 2016], quanto mais termos em comum maior será o grau de similaridade.

Por exemplo, na Figura 3, para quatro itens A, B, C e D. Cada item é representado por um conjunto de palavras levando em consideração seus pesos.

Figura 3. Esquema da composição de *ConcentriCloud (letras A à D representam os Modelos Vetorias dos tópicos).*



Fonte: Adaptado do exemplo de André e Danilo [Andreotti e Eler 2016].

Quanto mais próximo do centro o termo daquele conjunto está, mais semelhante ele será do conteúdo base, ou seja, o grau de similaridade é maior [Andreotti e Eler 2016]. Tomando como base o grau de similaridade, é construído o vetor de recomendação e enviado para o *Framework*.

4. Validação e Resultados

Para a validação da abordagem proposta descrita na Seção 3, apresentamos a aplicação *Web* que utilizamos para integrar o Sistema de Recomendação desenvolvido. O *Atlas Ambiental Escolar de Presidente Prudente*[4] é o resultado de um trabalho coletivo envolvendo profissionais de diferentes áreas do conhecimento da *Faculdade de Ciência e Tecnologia (Unesp - Campus de Presidente Prudente, São Paulo, Brasil)*[5] juntamente em parceria com a *Fundação de Amparo à Pesquisa do Estado de São Paulo*[6] e a *Prefeitura de Presidente Prudente - SP*[7].

A proposta do Atlas é de mostrar à comunidade prudentina as formas diferentes de compreensão do uso e ocupação do solo urbano e rural. Este processo histórico de produção do espaço geográfico urbano e rural transformou as paisagens do município gerando alterações socioambientais, tais como: impactos hidrológicos, morfológicos, climáticos, biogeográficos, socioeconômicos e culturais [Nunes 2017].

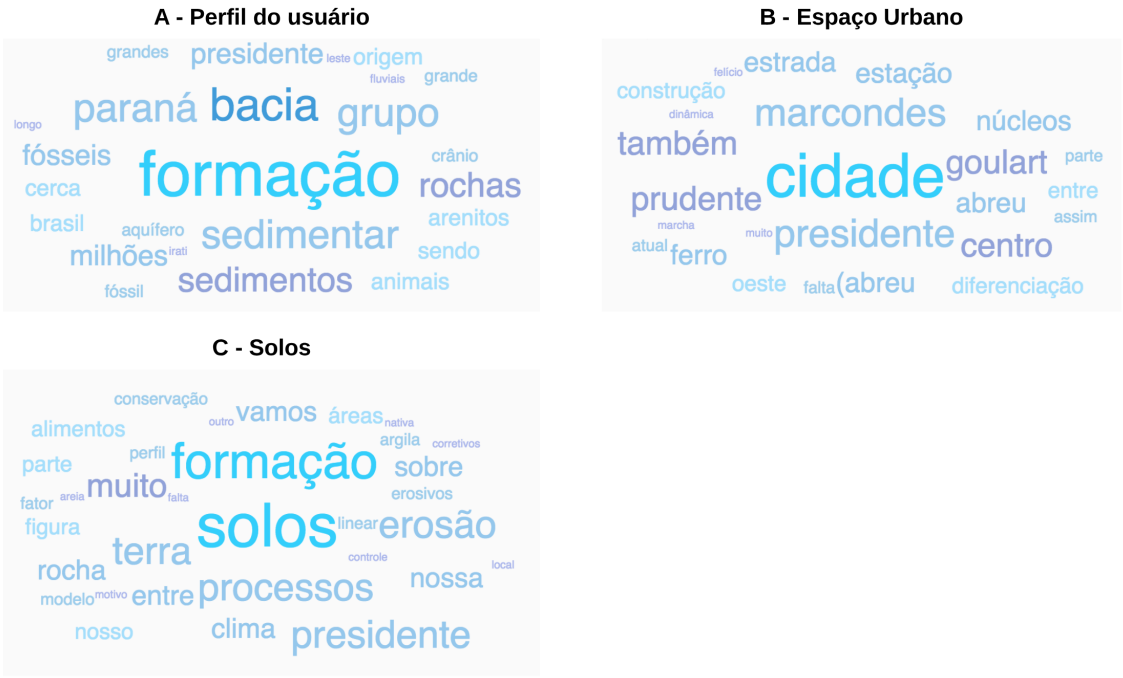
[4]http://fct.unesp.br/atlasambiental
[5]http://www.fct.unesp.br/
[6]http://www.fapesp.br/
[7]http://www.presidenteprudente.sp.gov.br/

Os itens que o Sistema de Recomendação manipula durante o processo de recomendação no *Atlas* são os seguintes tópicos: *Evolução Geológica*, *Relevo*, *Solos*, *Clima*, *Hidrografia*, *Cobertura Vegetal*, *Histórico do município*, *Espaço Urbano*, *Espaço Rural*, *Geografia da População*, *Indicadores Socioespaciais*, *Saneamento Básico*, *Áreas Verdes Urbanas*, *Fragilidade, Vulnerabilidade e Riscos* e *Derivações Ambientais*.

Na Figura 4 temos ilustrado as *Tagclouds* dos conjuntos A, B e C, na qual o conjunto A representa o perfil do usuário e os demais as recomendações. Logo na Figura 5 podemos visualizar o Modelo Vetorial de cada *Tagcloud* e os pesos de cada termo.

Temos como resultado da recomendação baseado no perfil os tópicos *Espaço Urbano* e *Solos*, ou seja, o Sistema de Recomendação considera os tópicos semelhantes ao Perfil do Usuário.

Figura 4. *Tagcloud* A representa o perfil do usuário. *Tagcloud* B e C representa as recomendações



Fonte própria.

Figura 5. *"Bag-of-words"* A representa o perfil do usuário. *"Bag-of-words"* B, C e D representa as recomendações

A - Perfil do usuário

"formação", peso: 27
"bacia", peso: 17
"sedimentar", peso: 15
"grupo", peso: 14
"paraná", peso: 14
"rochas", peso: 13
"presidente", peso: 9
...

B - Espaço urbano

"cidade", peso: 37
"presidente", peso: 22
"marcondes", peso: 19
"prudente", peso: 18
"goulart", peso: 18
"também", peso: 16
...

C - Solos

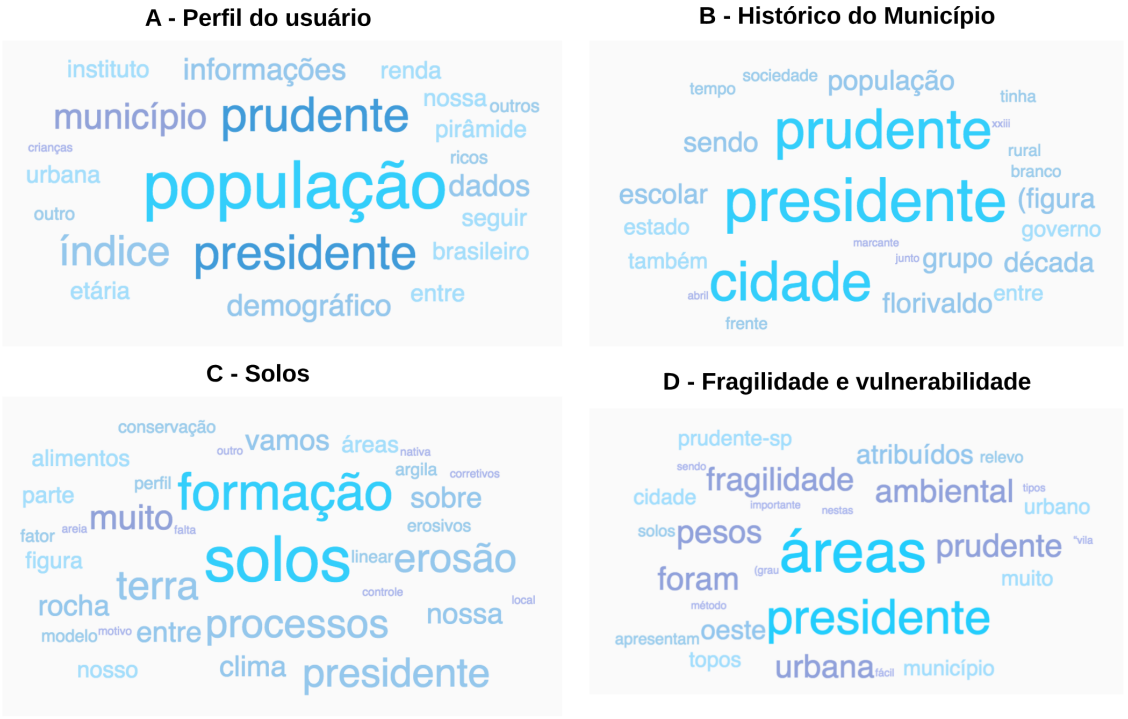
"solos", peso: 25
"formação", peso: 20
"processos", peso: 15
"terra", peso: 14
"erosão", peso: 14
"presidente", peso: 13
...

Fonte própria.

Na Figura 6

Na Figura 4 temos ilustrado as *Tagclouds* dos conjuntos A, B, C e D, na qual o conjunto A representa o perfil do usuário e os demais as recomendações.

Figura 6. *Tagcloud* A representa o perfil do usuário. *Tagcloud* B, C e D representa as recomendações



Fonte própria.

5. Considerações finais

Neste trabalho foi estudado e aplicado as técnicas de Sistema de Recomendação de tópicos com base em conteúdo textual em Páginas Web. Para a avaliação foi integrado o *Framework* e o *Middleware* na aplicação *Web Atlas Ambiental Escolar de Presidente Prudente*, visando facilitar a exploração de tópicos.

A estratégia de calcular a similaridade entres os tópicos ainda persiste a seleção de tópicos com poucos termos em comum em alguns casos. Ainda existe a visualização de alguns termos irrelevantes, no entanto, ainda destaca as mais relevantes. Também existe alguns termos tanto no plural quanto no singular e que poderiam ser unidas afim de representar de forma única na visualização.

Cabe ressaltar as recomendações podem ser desinteressantes para o usuário e afim de auxiliar na exploração de conteúdo, a aplicação poderia apresentar uma interface que possibilitasse a inserção de palavras chaves.

Futuramente, a aplicação poderá contar com a recomendação de conteúdo, *Hyperlinks* e imagens fora do domínio do *Site*, buscando conteúdo por meio de outras plataformas, como no Google. É possível ainda aperfeiçoar a construção do Perfil do Usuário com a utilização de técnicas de Aprendizado de Máquina, consequentemente a qualidade do conteúdo a ser recomendado.

Referências

(2009). Emotion incubator group. `http//WWW.org/2005/Incubator/emotion`.

Aas, K. e Eikvil, L. (1999). Text categorisation: A survey.

Adomavicius, G. e Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6):734–749.

Alencar, A. B. (2013). *Visualização da evolução temporal de coleções de artigos científicos*. PhD thesis, Universidade de São Paulo.

Andreotti, A. L. D. e Eler, D. M. (2016). Análise visual da evolução de coleções de documentos utilizando tag cloud.

Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

Baharudin, B., Lee, L. H., e Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1):4–20.

Balabanovic, M. e Shoham, Y. (1995). Learning information retrieval agents: Experiments with automated web browsing. pages 13–18.

Cazella, S. C., Nunes, M., e Reategui, E. B. (2010). A ciência da opinião: Estado da arte em sistemas de recomendação. In *XXX Congresso da Sociedade Brasileira de Computação—Jornada de Atualização em Informática (JAI)*, pages 161–216.

das Neves, É., Oliveira, U., Vargas, P. K., e Mangan, U. (2013). Predição de prescrição médica eletrônica: análise de um sistema especialista.

Foskett, A. C. (1972). Subject approach to information.

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

Kuramoto, H. (1996). Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais.

Lops, P., de Gemmis, M., e Semeraro, G. (2011). *Content-based Recommender Systems: State of the Art and Trends*, pages 73–105. Springer US, Boston, MA.

Medeiros, I. R. G. (2013). Estudo sobre sistemas de recomendação colaborativos.

Nunes, J. O. R. N. (2017). Atlas ambiental escolar de presidente prudente. `http://fct.unesp.br/atlasambiental`.

Reategui, E. B., Cazella, S. C., e Osório, F. S. (2006). Personalização de páginas web através dos sistemas de recomendação. *Tópico em Sistemas Interativos e Colaborativos. São Carlos*.

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

Resnick, P. e Varian, H. R. (1997). Recommender systems. *Commun. ACM*, 40(3):56–58.

Shardanand, U. e Maes, P. (1995). Social information filtering: Algorithms for automating “word of mouth”. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '95, pages 210–217, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

Silva, N. (2016a). Recommendation system textual content based: Avaliation and comparation. *Institute of Math*.

Silva, R. G. N. (2016b). Sistema de recomendação baseado em conteúdo textual: avaliação e comparação.

Souza, R. R. (2006). Sistemas de recuperaÃ§Ãde informaÃ§Ãe mecanismos de busca na web: panorama atual e tendÃ. *Perspectivas em CiÃda InformaÃ§Ã*, 11:161 – 173.