USO DE TAXONOMIAS
NA RECOMENDAÇÃO DE PRODUTOS

OSVALDO MATOS JÚNIOR

USO DE TAXONOMIAS NA RECOMENDAÇÃO DE PRODUTOS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

ORIENTADOR: NIVIO ZIVIANI

Belo Horizonte

de 2011

Matos-Junior, Osvaldo

M433u Uso de Taxonomias na Recomendação de Produtos / Osvaldo Matos Júnior. — Belo Horizonte, 2011.
xvi, 62 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais. Departamento de Ciência da Computação.

Orientador: Nivio Ziviani.

1. Computação - Teses. 2. Recuperação da informação. I. Orientador. II. Título.

CDU 519.6*73(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Uso de taxonomias na recomendação de produtos

OSVALDO CARNEIRO DE MATOS JÚNIOR

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. NIVIO ZIVIANI - Orientador
Departamento de Ciência da Computação - UFMG

PROF. ALBERTO HENRIQUE FRADE LAENDER
Departamento de Ciência da Computação - UFMG

PROF. WAGNER MEIRA JÚNIOR
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 28 de fevereiro de 2011.

Agradecimentos

Agradeço primeiramente, e dedico em especial, à minha mãe Noélia Carneiro da Silva, a qual não teve a satisfação de presenciar a colação de grau de seu filho e, por isso, fiz questão de sua presença na apresentação dessa defesa de mestrado. Também agradeço ao meu pai Osvaldo Matos, irmã Kilma e irmão Leandro pelo apoio, mesmo que distante, nos mais difíceis dias que passei em minha vida.

Ao grande amigo Fabiano Botelho e meu orientador Nivio Ziviani pela oportunidade e por acreditarem em meu profissionalismo e dedicação, sabendo que ao final triunfaria. Ao auxílio dos professores Altigran Silva e Marco Cristo pelas críticas importantes à conclusão deste trabalho.

Aos meus amigos da *Los Computeros*: Alan Castro, Wallace Favoretto e Rickson Guidoline; outros membros do LATIN: Anísio Mendes, Guilherme Menezes e Wladmir Cardoso; e amigos belo-horizontinos e manauaras que foram fundamentais à minha consolidação social.

Aos amigos jacuipenses, que fizeram muuuita falta nesse longo período longe. Para vocês foi apenas um amigo que se foi, mas para mim foram todos retirados de uma só vez.

Agradeço a Deus por não me fazer desistir e por ter fé, acreditar que tudo daria certo e o sonho se concretizaria.

"Os homens apressam-se mais a retribuir um dano do que um benefício, porque a gratidão é um peso e a vingança um prazer."
(Tácito)

Resumo

Sistemas de recomendação procuram compreender o interesse do usuário e gerar uma lista de recomendação com itens relacionados. Neste trabalho investigamos como obter vantagem da informação presente em taxonomias para melhorar a qualidade de sistemas de recomendação baseada em conteúdo. Adotamos o cenário onde o usuário está interessado em uma notícia publicada na Internet e um sistema de recomendação é usado para recomendar produtos, por exemplo, livros de uma livraria *online*. O sistema analisa o texto da notícia e, por meio de técnicas de Recuperação de Informação (RI), encontra livros semelhantes ao assunto da notícia. O uso de taxonomias abre a oportunidade de incorporar conhecimento de um domínio específico compilado por humanos. Esta dissertação apresenta um estudo de três estratégias para explorar o uso de taxonomias em sistemas de recomendação baseada em conteúdo, a saber: *descritores de categoria*, *características de classificação* e *filtro de categorias*. Embora algumas dessas estratégias tenham sido empregadas anteriormente para resolver outros problemas de RI, neste trabalho, essas estratégias foram aplicadas em um cenário diferente. Vários métodos de recomendação foram derivados a partir das três estratégias, explorando diversas configurações e premissas. Os experimentos foram realizados sobre uma coleção de 100 páginas de notícias (páginas alvo) do The New York Times, uma taxonomia e uma coleção de livros, ambas coletadas da Amazon.com. Os resultados experimentais mostram que as estratégias consideradas podem ser aplicadas com sucesso para melhorar os sistemas de recomendação baseada em conteúdo. Em particular, quando a página alvo é manualmente associada a uma categoria por humanos, os ganhos são próximos de 20% na precisão média. Por outro lado, se essa associação é automática, os ganhos ainda são representativos e próximos de 13% na precisão média.

Abstract

In this work we investigate how to take advantage of valuable information encoded in taxonomies to improve the quality of content-based recommender systems. The use of taxonomies opens the opportunity to incorporate domain-specific and common-sense knowledge compiled by humans. Our investigation is based on a case study over the book domain, in which the recommendation target is a news web page and the items to be recommended are books from an online bookstore. This is a representative real-case application that provides an adequate context to experiment with a number of distinct strategies. We present a comprehensive study of three strategies to exploit the use of taxonomies in content-based recommender systems. Although some of these strategies have been previously applied to other related IR problems, in this paper we present a fresh perspective for their application in a new scenario. For this, we implement several methods for content-based recommendation that apply these strategies individually and in combination, exploring diverse configurations and premises. We perform a comprehensive set of experiments with a collection of 100 news pages (i.e., target pages) from The New York Times, and a collection and a taxonomy of books crawled from *Amazon.com*. Experimental results indicate that our strategies can be successfully applied to improving traditional content-based recommender systems. In particular, when the target page is manually assigned to a category by a user, we obtain gains close to 20% in average precision. On the other hand, if such an assignment is automatic, the gains are still representative, reaching around 13% in average precision.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A enorme quantidade de informação disponível na Web torna difícil para as pessoas escolherem o que vale a pena consumir. Na verdade, existe certo consenso que as pessoas gastam mais tempo procurando por informação do que fazendo uso dela. Para lidar com essa sobrecarga de informação, muitas pesquisas têm sido realizadas na indústria e na academia para desenvolver sistemas capazes de (a) selecionar informação que seja mais relevante para o usuário e (b) tirar vantagem de alguma fonte de conhecimento disponível a fim de garantir a satisfação do usuário com a informação selecionada. Prover soluções para essas questões contribui para uma área de pesquisa com muitas aplicações práticas, tais como recomendar livros, músicas, notícias, filmes, parceiros para namoro, além de outros itens.

Os sistemas que tratam dessas questões são chamados sistemas de recomendação [Adomavicius & Tuzhilin, 2005]. Empresas de comércio eletrônico (por exemplo, Amazon.com[1]) utilizam sistemas de recomendação para encontrar e sugerir aos clientes novos produtos de interesse. Esses sistemas induzem os clientes a novas compras e, consequentemente, visam aumentar a receita dessas empresas. De acordo como a recomendação é feita, esses sistemas são classificados em [Adomavicius & Tuzhilin, 2005]: (i) sistemas de filtragem baseada em conteúdo (ii) sistemas de filtragem colaborativa, e (iii) sistemas híbridos. Nos sistemas baseados em conteúdo, o usuário recebe recomendações em função de itens similares a outros que ele gostou no passado, ou de interesse no momento. Nos sistemas colaborativos, o usuário é recomendado por itens que pessoas com preferência parecida também gostaram. Sistemas híbridos combinam as duas abordagens anteriores. Esses métodos são detalhados posteriormente na Seção 2.1.

Este trabalho estuda sistemas de recomendação com foco em filtragem baseada em conteúdo. Nesses sistemas, um item é representado por um conjunto de caracterís-

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

ticas usadas para descrever seu conteúdo. Em um sistema de recomendação de livros, por exemplo, as características correspondem a um conjunto de palavras, as quais, dependendo do cenário da aplicação, podem ser extraídas do conteúdo do próprio livro, de um sumário de seu conteúdo, da descrição de seu assunto, de seu título, etc. O alvo da recomendação é o usuário cuja necessidade de informação pode ser representada pelo perfil, por outro livro, por uma página web, etc. Então, é comum se referir a esses objetos como o "alvo" da recomendação. Os alvos são representados por características do mesmo tipo usado para representar os itens, isto é, palavras em nosso exemplo. Ao comparar os novos itens com o alvo é possível determinar quais itens são mais relevantes para o alvo.

Um possível problema com essa estratégia está no fato de que um mesmo conceito, ou relação semântica, pode ser descrito usando palavras diferentes. Por exemplo, a descrição do alvo pode incluir a palavra "cachorro", enquanto a descrição de um livro sobre cachorros não possui a palavra "cachorro", mas inclui um sinônimo, como "cão" ou uma palavra relacionada, como "canino". Assim, o casamento simples dos termos usados para descrever os itens pode não ser suficiente para detectar dois itens semelhantes.

Como o baixo casamento entre as características selecionadas para descrever o alvo e o item é algo comum, uma possível estratégia para melhorar a qualidade da recomendação é recorrer ao conhecimento obtido de fontes externas, tais como fontes que não foram usadas na representação do alvo e dos itens.

Este trabalho tira proveito da valiosa informação embutida em taxonomias para lidar com a dificuldade de detectar dois itens semelhantes, causada pelo baixo casamento de termos. O uso de taxonomias, ou outras bases de conhecimento construídas por humanos, abre a oportunidade de incorporar conhecimento de um domínio específico compilado por humanos. Essa valiosa fonte de informação não poderia ser obtida apenas a partir do alvo ou dos itens. Além disso, taxonomias de produtos são atualmente um recurso comum, mantido por empresas que operam sistemas de comércio eletrônico e que oferecem serviços de recomendação.

Neste trabalho a pesquisa é baseada em um estudo de caso sobre o domínio de livros, no qual o alvo da recomendação é uma página web de notícias e os itens a serem recomendados são livros de uma livraria *online*. Importante ressaltar que um sistema de recomendação de livros é um caso de aplicação real e que fornece um contexto adequado para experimentar estratégias distintas. Esse tipo de recomendação de produtos em tempo real está sendo explorada, por exemplo, pelo *Google Product Listing Ads*[2], que apresenta listas de recomendação de produtos associadas à consulta

[2]http://www.google.com/ads/innovations/productlistingads.html

do usuário na máquina de busca, sendo a receita gerada quando o usuário realiza uma compra (conhecido em inglês como CPA[3]).

A escolha de livros como itens a serem recomendados se deve à rica informação textual disponível, por exemplo, no título, editorial, revisão de usuários e autores. Já as páginas de notícias por terem uma estrutura bem definida e informação textual de qualidade que são úteis na tarefa de recomendação. Técnicas de recuperação de informação, de mineração de dados e de aprendizado de máquina são amplamente utilizadas para melhorar a qualidade da recomendação.

Para o estudo de caso tratado neste trabalho foi utilizada uma taxonomia representativa para o domínio de livros, com uma coleção contendo informação sobre 1.499.792 de livros disponíveis na Amazon.com. Na verdade, ter uma fonte tão rica de informação publicamente disponível para experimentação foi uma das razões pela qual o domínio de livros foi escolhido para a realização desse estudo de caso.

Dentro desse cenário, apresentamos um estudo abrangente de estratégias distintas para explorar o uso de taxonomias em sistemas de recomendação baseada em conteúdo. Mais especificamente, consideramos três estratégias: *descritores de categoria*, *características de classificação* e *filtro de categorias*.

Embora as duas primeiras estratégias já tenham sido previamente aplicadas a problemas como busca na Web [Carpineto et al., 2001; Carpineto & Romano, 1999], classificação textual [Gabrilovich & Markovitch, 2005] e associação de propagandas a páginas web [Anagnostopoulos et al., 2007], este trabalho apresenta uma nova perspectiva para sua utilização em um problema relacionado, porém distinto. Para isso, foram implementados vários métodos para recomendação baseada em conteúdo que aplicam as três estratégias individualmente e em conjunto, explorando diversas configurações e premissas. Em particular, consideramos o cenário no qual a página alvo é manual e antecipadamente associada por usuários ou editores a uma ou mais categorias, em contraste com outro cenário em que essa associação é feita por um classificador automático, sem qualquer intervenção humana.

Com os métodos implementados, foi realizada uma série de experimentos com uma coleção de páginas alvo da Web. Foram utilizadas uma coleção de 100 páginas de notícias do The New York Times, uma taxonomia e uma coleção de livros, ambas coletadas da Amazon.com. Os resultados experimentais indicam que, quando a categoria da página alvo é atribuída manualmente por um usuário, os ganhos obtidos são próximos a 20% de precisão média. Por outro lado, se a atribuição é automática, os ganhos ainda são representativos, e atingem cerca de 13% de precisão média.

[3] *Cost Per Action*, quando o anunciante paga por uma ação específica do usuário, por exemplo, uma compra.

1.1 Trabalhos Relacionados

Esta seção apresenta uma revisão da literatura dos principais trabalhos relacionados a recomendação de produtos. Os trabalhos anteriores que fazem uso de taxonomias para melhorar a busca na Web, a associação de propagandas a páginas web e sistemas de recomendação são considerados a seguir.

Com relação ao problema de busca na Web, melhorias são obtidas pela modificação da consulta, o que pode ser feito usando termos de toda a coleção (expansão global), um conjunto de documentos similar à consulta (expansão local) ou à categoria da consulta (expansão baseada na categoria). Este trabalho assemelha-se a expansão de consulta baseada na categoria. Por exemplo, Inquirus2 [Glover et al., 2001] usa perda de entropia para determinar os novos termos da consulta, enquanto *Keyword spices* [Oyama et al., 2001] e TAX-PQ [Pahlevi & Kitagawa, 2005] usam árvores de decisão para encontrar novos termos em documentos classificados de acordo com taxonomias hierárquicas e planas. O método discutido nesta dissertação difere dos trabalhos citados porque usa páginas web como consulta, as quais são bem maiores que as consultas realizadas nas máquinas de busca. Adicionalmente, os livros da coleção utilizada foram classificados manualmente, fato que provê uma informação confiável. Cabe observar que algumas medidas utilizadas foram usadas anteriormente em trabalhos de expansão local de consultas [Carpineto et al., 2001; Carpineto & Romano, 1999], os quais estão descritos na Seção 1.1.2.

Em relação à associação de propaganda a páginas web, um trabalho semelhante ao nosso foi proposto por Anagnostopoulos et al. [2007]. Os autores estudaram a associação de propaganda quando o verdadeiro conteúdo da página era visualizado apenas pelo usuário final, isto é, páginas geradas via JavaScript. Como não era possível adquirir todo o conteúdo da página acessando a URL, pequenos fragmentos com maior valor informacional (uma espécie de resumo) eram extraídos e enviados do cliente até o servidor de propagandas. Os autores empregaram uma extensa taxonomia hierárquica criada por humanos com aproximadamente 6.000 nós, que seria usada na geração de novas características para páginas web e propagandas. Essa taxonomia foi criada para fins comerciais pela Yahoo! US e inicialmente usada para classificar consultas de interesse comercial. Os autores usaram uma abordagem de geração de características, proposta por Gabrilovich & Markovitch [2005], para realizar a associação de propagandas usando o repositório de palavras e as novas características de classificação. Com essas novas evidências, eles concluíram que apenas 5% do texto da página original leva a perdas de apenas 1%-3% na relevância das propagandas. Assim como os autores, esta dissertação utiliza a estratégia de incluir novas características de classificação, agora

no contexto de recomendação de produtos.

Finalmente, em sistemas de recomendação, Ziegler et al. [2004a,b, 2005, 2008] exploram o conhecimento por trás dos conceitos de taxonomia para computar listas de recomendação de livros personalizadas em filtragem colaborativa. Como contribuição, os autores mudaram a representação vetorial dos usuários para tópicos (categorias), não mais itens (livros). Eles exploram a extensiva taxonomia de livros da Amazon.com, com mais de 13 mil nós e, por meio de análise *offline* e *online*, verificaram que sua proposta é superior aos métodos tradicionais quando a informação dos usuários é esparsa e avaliações implícitas prevalecem. Em seus experimentos, apesar de ter a precisão média reduzida, mostraram que os usuários ficaram mais satisfeitos com as novas recomendações. Em [Ziegler et al., 2005] foi apresentado com mais detalhes o impacto da diversificação dos tópicos, ajustados por parâmetros de *tunning* em sua abordagem. Esse trabalho mostrou ser possível melhorar a recomendação de livros por meio de taxonomia, que representa nosso objetivo de estudo, aplicado à recomendação baseada em conteúdo.

1.1.1 Sistemas de Recomendação

O primeiro sistema de recomendação citado na literatura foi o Tapestry [Goldberg et al., 1992]. Os autores usaram a expressão ***filtragem colaborativa*** para designar um sistema específico no qual a filtragem era auxiliada pela colaboração de outras pessoas. Alternativamente, Resnick & Varian [1997] preferiram usar o termo mais genérico ***sistemas de recomendação***, por duas razões principais: (i) recomendadores nem sempre explicitam colaboração com outras pessoas, pois sequer um conhece o outro e (ii) recomendadores também podem sugerir itens de interesse particular, incluindo aqueles que deveriam ser desconsiderados.

Os desenvolvedores do Fab System [Balabanović & Shoham, 1997] usaram uma abordagem híbrida (isto é, filtragem baseada em conteúdo em conjunto com a filtragem colaborativa) para recomendar páginas web aos usuários. A parcela baseada em conteúdo usava informação de páginas que o usuário havia visitado e avaliado no passado, enquanto que a parcela baseada na colaboração social usava informação de perfil dos "vizinhos mais próximos", ou seja, demais usuários com perfil similar. A recomendação por conteúdo usava técnicas de recuperação de informação, as quais eram baseadas na comparação entre o conteúdo das páginas e o perfil do usuário. Para isso, o sistema usava as palavras mais discriminativas do texto das páginas, ou seja, as palavras com maior valor computado usando um esquema de pesos. Neste trabalho também criamos uma representação da página usando suas palavras e um peso associado a elas.

Em nosso caso, a representação da página define o interesse do usuário e não existe informação de outros usuários (parcela colaborativa).

Mooney & Roy [2000] utilizaram aprendizado de máquina na recomendação de livros da Amazon.com no contexto de filtragem colaborativa. Os autores propuseram um sistema chamado LIBRA (*Learning Intelligent Book Recommending Agent*), que utilizava um classificador de texto Bayesiano [Mitchell, 1997] para aprender como sugerir novos livros usando o texto dos livros (por exemplo, resumo e revisão de usuários) e as notas atribuídas pelos usuários. Finalmente, o modelo gerado era aplicado ao perfil do usuário e produzia uma lista de recomendação de livros com os melhores títulos do catálogo. Assim como os autores, nossos experimentos também foram realizados com uma amostra de livros da Amazon.com, só que a recomendação é baseada no conteúdo do item de interesse do usuário, ou seja, sem informação do perfil do usuário.

1.1.2 Expansão de Consultas

Experimentos realizados por Salton & Buckley [1988] e Harman [1992] usavam um método simples e eficiente para contornar o problema do vocabulário. Esse método, conhecido como *pseudorelevance feedback*, consistia na extração automática de termos de documentos no topo do *ranking*. A aplicação dessa técnica muitas vezes resultava na perda de precisão maior que o ganho correspondente na revocação. Em seguida, outros trabalhos reportaram melhoria dos resultados com expansão de consultas usando informação dos documentos do topo do *ranking* [Voorhees & Harman, 1998, 1999].

Carpineto et al. [2001] estudaram mais detalhadamente a expansão automática de consultas. Os autores comparam o *ranking* de termos usando medidas de teoria da informação contra outras técnicas usadas na expansão de consultas, como a fórmula de Rocchio [Rocchio et al., 1971]. Eles propuseram uma nova função de escore de termos, baseada na *entropia relativa* entre duas distribuições de probabilidade, medida também conhecida como *Kullback-Leibler divergence* (KLD) [Losee, 1990]. O peso computado por essa função foi usado na seleção de novos termos e combinado na fórmula Rocchio. Os experimentos foram realizados com as coleções TREC-7 e TREC-8, e o método proposto foi comparado com resultados sem expansão, Rocchio modificada [Srinivasan, 1996], Robertson Selection Value (RSV) [Robertson, 1990; Robertson et al., 1995], e outras funções de distribuição baseada no Chi-Squared [Doszkocs, 1978]. Os resultados experimentais mostram que os métodos baseados em teoria da informação produzem melhor eficácia entre diferentes coleções para quase todas as métricas avaliadas, e são mais eficientes quando usados dentro do arcabouço de Rocchio, não apenas para selecionar termos da expansão, mas também na atribuição de peso aos mesmos. Inspirado

nesses autores, nosso trabalho também utilizou medidas previamente usadas em expansão de consulta para selecionar novos termos e incluir à representação inicial da página.

Em aprendizado de máquina, pesquisadores passaram a combinar múltiplos classificadores visando melhorar a acurácia na classificação [Polikar, 2006; Rokach, 2010]. Individualmente, cada algoritmo de classificação possui um julgamento e podem discordar um do outro. Dessa forma, a combinação tenta acertar a predição pela maior quantidade de votos. Analogamente, Carpineto & Romano [1999] combinaram diferentes medidas de teoria da informação para eleger os melhores termos durante a expansão da consulta. Assim, os melhores termos seriam aqueles que apresentassem maior concordância no topo do *ranking* das medidas. A combinação de diferentes medidas na seleção de termos para expansão foi experimentada na estratégia de descritores de categoria, proposta neste trabalho (Seção 3.2).

Várias alternativas foram avaliadas por Ribeiro-Neto et al. [2005] na escolha de propaganda baseada em conteúdo. Enquanto usavam apenas os termos extraídos das páginas e dos elementos dos anúncios (por exemplo, título, descrição, palavras-chave), sua melhor estratégia obteve ganho de 60% na precisão média quando comparada ao modelo vetorial simples. No entanto, alguns anúncios relevantes não eram exibidos porque os termos extraídos da página não eram suficientes para recuperá-los. Esse problema semântico existente entre o vocabulário da página e anúncios foi chamado de *impedância do vocabulário* pelos autores. Então, uma outra estratégia usou Redes Bayesianas para identificar páginas associadas e daí extrair/incluir novos termos para resolver o problema da impedância. Essa última estratégia apresentou ganho de 50% na precisão média em relação à primeira. Nosso trabalho também utilizou técnicas de aprendizado de máquina para contornar o problema da impedância do vocabulário. Classificadores automáticos predizem as categorias da página e novos termos associados à essas categorias são incluídos à representação inicial da página.

1.2 Objetivos

O principal objetivo deste trabalho é desenvolver novos métodos que utilizem a informação de taxonomia para melhorar a recomendação de produtos. Nesta dissertação foi adotado como estudo de caso a recomendação de livros a páginas de notícias. Para atingir nosso propósito, adotamos os seguintes objetivos específicos:

1. Construção das coleções que serão fonte de estudo do trabalho: uma coleção de páginas alvo, composta por notícias extraídas da Web; e outra de produtos a

serem recomendados, constituída por livros adquiridos de uma loja *online*.

2. Implementar uma abordagem simplista para recomendar livros, tomando uma página de notícia como item de interesse do usuário. A abordagem utiliza técnicas de recuperação de informação baseadas no casamento dos termos dos itens envolvidos e foi utilizada como um *baseline*.

3. Propor estratégias de recomendação que utilizam informação de taxonomia. As estratégias foram elaboradas a partir de técnicas encontradas na literatura para resolução de outros problemas clássicos: associação de propagandas a páginas web, busca na Web e classificação textual.

4. Comparar a abordagem simplista com diferentes métodos de recomendação derivados das estratégias propostas.

1.3 Contribuições

Esta dissertação adotou como desafio recomendar produtos em tempo real na Web. O cenário é constituído pelo usuário que navega na Web, o qual recebe recomendações de itens relacionados à página atual. Uma associação de páginas e de produtos é feita utilizando informação de taxonomias para melhorar a recomendação. As principais contribuições deste trabalho são:

1. Proposição de diferentes estratégias no cenário de recomendação, algumas baseadas em técnicas conhecidas na literatura para resolução de outros problemas (Capítulo 3). As três estratégias propostas são:

   a) *Descritores de Categoria:* expansão da representação da página com termos que se destacam nas categorias. Isso inclui avaliar o impacto de métricas de teoria da informação empregadas na expansão de consultas.

   b) *Características de Classificação:* inclusão de conceitos extraídos de uma base de conhecimento (taxonomia). Esses conceitos, ou características de classificação, foram usados para reajustar a lista de recomendação de livros.

   c) *Filtro de Categorias:* restringir domínio dos resultados para o mesmo domínio da página alvo. Com isso, a lista final de recomendação seria composta apenas por livros da mesma área de conhecimento (categoria) indicada na notícia.

2. Elaboração de diferentes métodos de recomendação derivados das estratégias propostas.

3. Comparação dos métodos de recomendação. A comparação é feita por meio da análise por humanos, os quais avaliam a qualidade da recomendação.

4. Construção de uma coleção de produtos a serem recomendados, formada por título, descrição, autores e categorias de 1.499.792 livros coletados da Amazon.com (Seção 4.1.1).

5. Construção de uma taxonomia extraída da coleção de livros da Amazon.com, formada por 1.621 categorias (nós) inferidas a mão (Seção 4.1.1).

6. Construção de uma coleção de páginas alvo, formada por 100 notícias coletadas do The New York Times (Seção 4.1.2).

7. Avaliação com humanos da qualidade dos livros recomendados. Os avaliadores julgaram se o livro recomendado estava relacionado à página alvo. A saída desta avaliação foi usada na avaliação experimental dos métodos de recomendação estudados.

8. Publicação dos julgamentos sobre a relevância da recomendação, e que foram realizados por 15 pessoas. A publicação das coleções e dos julgamentos para a comunidade acadêmica permite que outros pesquisadores possam reproduzir os resultados deste trabalho e também comparar com novas abordagens experimentais.

1.4 Estrutura da Dissertação

O restante desta dissertação está dividido em quatro capítulos. No Capítulo 2, são discutidos os conceitos básicos necessários para uma melhor compreensão do trabalho. O Capítulo 3 apresenta três estratégias estudadas para o uso de taxonomia na recomendação baseada em conteúdo. A Seção 3.1 apresenta um tipo de abordagem simplista que usa um repositório de palavras, a Seção 3.2 estuda os descritores de categoria, a Seção 3.3 discute as características de classificação e a Seção 3.4 o filtro de categorias. O Capítulo 4 apresenta a avaliação experimental das estratégias discutidas no Capítulo 3. Os experimentos foram realizados usando técnicas de avaliação comuns em recuperação de informação. Duas coleções foram utilizadas nos experimentos: uma composta por páginas alvo de notícias e a outra por livros a serem recomendados. Neste mesmo

capítulo são apresentados os resultados dos experimentos e uma discussão sobre os mesmos. Finalmente, no Capítulo 5, apresentamos nossas conclusões finais e sugestões de trabalhos futuros para esta pesquisa.

Capítulo 2

Conceitos Básicos

Neste capítulo introduzimos os principais conceitos necessários para o melhor entendimento do restante do texto. Na Seção 2.1, descrevemos o funcionamento de sistemas de recomendação e principais implementações: filtragem colaborativa e filtragem baseada em conteúdo. Na Seção 2.2, descrevemos sobre um dos modelos de busca clássicos em recuperação de informação: o modelo de espaço vetorial. Na Seção 2.3, apresentamos medidas usadas em expansão de consultas para minimizar o problema do baixo casamento de termos. Finalmente, a Seção 2.4 apresenta as principais métricas usadas na avaliação de sistemas de recuperação de informação, e também empregadas na avaliação de sistemas de recomendação, que serão úteis para compreender a avaliação experimental do Capítulo 4.

2.1 Sistemas de Recomendação

Em comércio eletrônico, sistemas de recomendação são amplamente usados para atrair clientes a novas compras. Durante a navegação do usuário, esses sistemas funcionam internamente nos *sites* de compra para descobrir as relações entre clientes e produtos (ou apenas entre produtos). Finalmente, o sistema cria uma lista de produtos (recomendações) para ser ofertada ao usuário. Esses sistemas também são usados em outras áreas como, por exemplo, para recomendar filmes, músicas, notícias, livros e outros itens de interesse para o usuário. Na Tabela 2.1 são apresentados exemplos de *sites* que utilizam sistemas de recomendação e sua aplicação.

Podemos assumir que existem três tipos principais de abordagem em sistemas

[1] `http://www.amazon.com/`
[2] `http://www.last.fm/`
[3] `http://www.imdb.com/`

Tabela 2.1. Exemplos de *sites* que utilizam sistemas de recomendação

Site	*Descrição*
Amazon.com[1]	Uma das maiores lojas de comércio eletrônico do mundo. Quando o visitante acessa um produto, o sistema oferece produtos relacionados.
Last.fm[2]	Popular *site* da Internet para ouvir músicas online. À medida que o usuário ouve músicas, o sistema aprende seu gosto e sugere músicas de artistas semelhantes.
IMDb[3]	O *Internet Movie Database* é o maior repositório *online* sobre filmes. Os usuários inserem avaliação sobre filmes e o IMDb usa essa informação para gerar *ranking* dos melhores filmes.

de recomendação: filtragem colaborativa, filtragem baseada em conteúdo e filtragem híbrida. O termo *filtragem* foi aplicado aos sistemas de recomendação por filtrar para o usuário itens de interesse. A seguir, são apresentados mais detalhadamente os dois primeiros métodos de filtragem. O último método não foi detalhado porque corresponde a uma combinação desses dois.

2.1.1 Filtragem Colaborativa

Em filtragem colaborativa, a essência está na troca de experiências entre pessoas que tenham interesse em comum, e os itens são filtrados usando a informação proveniente de outros usuários [Reategui & Cazella, 2005]. Por exemplo, considere um sistema simples em que o cliente atribui uma pontuação a cada produto comprado. Nesse exemplo, o sistema de recomendação pode computar a média da pontuação dos produtos e, em seguida, sugerir aqueles que possuem maior nota.

Em uma abordagem mais complexa, os clientes são representados no espaço vetorial e cada dimensão corresponde à nota atribuída aos produtos do sistema. Por meio da similaridade vetorial (distância do cosseno), são recuperados clientes com gosto semelhante ao do usuário atual. O sistema usa a avaliação de seus semelhantes para predizer a avaliação do usuário atual e produtos com maior avaliação serão recomendados pelo sistema. Ainda em filtragem colaborativa, e ilustrado na Figura 2.1, uma matriz item-item pode ser construída com o relacionamento entre os itens: quem compra X também compra Y.

Entretanto, existem alguns problemas em utilizar informação de compra de usuários ou de avaliação de produtos. Alguns desses problemas referem-se à quantidade de informação, pois novos consumidores apresentam informação limitada e usuários antigos podem ter excesso de informação [Linden et al., 2003]. Além disso, para que este sistema funcione corretamente, é preciso veracidade nas opiniões fornecidas, senão

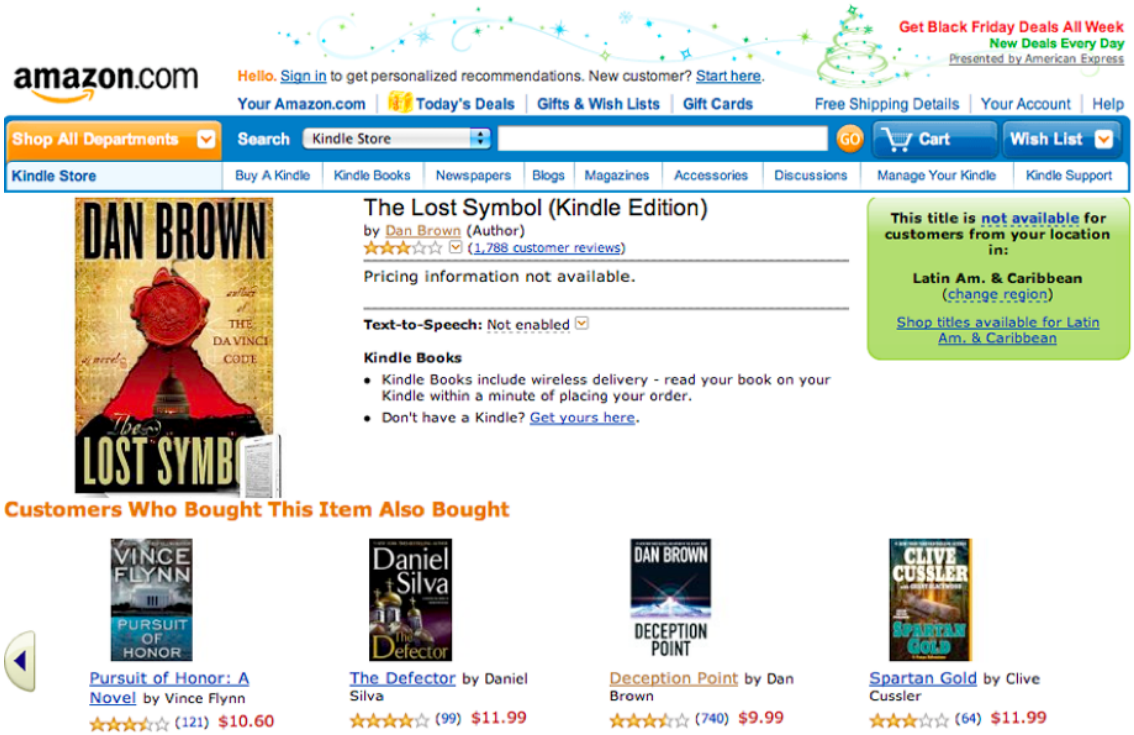

Figura 2.1. Filtragem colaborativa baseada em item-por-item na Amazon.com. Enquanto o livro *The Lost Symbol* de Dan Brown é visualizado, abaixo são recomendados livros adquiridos por clientes que também compraram este livro.

essas opiniões contribuirão negativamente na recomendação dos produtos.

2.1.2 Filtragem Baseada em Conteúdo

Na filtragem baseada em conteúdo, a recomendação é feita utilizando informação do próprio produto. Tomando CDs de músicas, por exemplo, é possível sugerir outros álbuns da mesma banda ou encontrar associação entre artistas do mesmo gênero. Na Figura 2.2 é apresentado um exemplo de filtragem baseada em conteúdo para artistas relacionados. Segundo Reategui & Cazella [2005], a filtragem colaborativa se diferencia da filtragem baseada em conteúdo exatamente por não exigir a compreensão ou reconhecimento do conteúdo dos itens.

A filtragem baseada em conteúdo pode ser feita utilizando técnicas de recuperação de informação. Nessa abordagem, título e descrição dos produtos são usados como consulta para recuperar produtos associados. O processador de consultas retorna uma lista de produtos ordenados por relevância e os produtos do topo da lista correspondem à lista de recomendação. Porém, esse tipo de abordagem pode produzir resultados de baixa qualidade [Linden et al., 2003] e algumas técnicas podem ser usadas para

Figura 2.2. Recomendação baseada em conteúdo no site Last.fm. Na página da cantora *Ivete Sangalo*, outros artistas do ritmo *Axé* são sugeridos.

melhorar a recomendação, como será apresentado neste trabalho.

2.2 Modelo de Espaço Vetorial

Recuperação de Informação é a subárea de Ciência da Computação que estuda como organizar e recuperar dados eficientemente. Arquivo invertido (índice invertido) é a estrutura de dados mais usada em recuperação de informação para esse propósito [Witten et al., 1999] e um modelo de busca é usado para recuperar dados armazenados no índice invertido. Após especificada uma consulta, o modelo ordena os documentos recuperados de acordo com o grau de similaridade com a consulta [Baeza-Yates & Ribeiro-Neto, 2011].

No modelo espaço vetorial, documentos e consulta são representados como vetores de peso no espaço de termos. Portanto, o documento será representado pelo vetor $\vec{d_j} = \{w_{1,j}, w_{2,j}, ..., w_{t,j}\}$ e a consulta por $\vec{q} = \{w_{1,q}, w_{2,q}, ..., w_{t,q}\}$, onde w_i corresponde ao peso do i-ésimo termo. Então, a similaridade é computada pela distância do cosseno entre esses dois vetores, conforme a Equação 2.1.

$$\begin{aligned} cos\,\theta &= \frac{\vec{q} \cdot \vec{d_j}}{||\vec{q}|| \times ||\vec{d_j}||} \\ sim(q, d_j) &= \frac{\sum_{i=1} w_{i,q} \cdot w_{i,j}}{\sqrt{\sum_{i=1} w_{i,q}^2} \cdot \sqrt{\sum_{i=1} w_{i,j}^2}} \end{aligned} \tag{2.1}$$

2.3 Expansão de Consultas

Em uma máquina de busca convencional, o usuário entra com palavras-chave no campo de texto especificado e em seguida submete essas palavras como consulta. A consulta é então processada internamente pela máquina de busca, que devolve para o usuário um conjunto de documentos ordenados pela relevância à consulta.

Entretanto, nem sempre a entrada dos usuários possui informação suficiente para atender a consulta, por exemplo, a quantidade de palavras é pequena ou as palavras individualmente podem ter significados diferentes. Para resolver esse problema, as máquinas de busca utilizam técnicas de expansão de consultas que incluem novos termos à consulta original do usuário. Uma alternativa para a expansão de consulta, usada por volta de 1960, era o uso de *thesauri*, os quais correspondem a um conjunto de palavras relacionadas e sinônimos.

Nas máquinas de busca tradicionais, novas técnicas foram propostas para expansão de consulta de forma automática e semi-automática. Essas técnicas se baseiam na análise da coocorrência entre palavras na coleção, histórico de consultas, ou nos documentos do topo da lista do resultado. Para uma boa expansão, é preciso incluir novos termos que estejam relacionados ao contexto da consulta inicial, e expandir usando todas as palavras da consulta e não cada uma separadamente. Por exemplo, "sacola" talvez seja uma boa palavra para expandir o termo "bolsa" na consulta "bolsa de bebês", mas não é apropriada para a consulta "bolsa de ações".

Diferentes alternativas para expansão de consultas são sugeridas por Croft et al. [2009] e Carpineto et al. [2001]. A seguir, apresentamos algumas dessas medidas para relacionar dois temos a e b, algumas delas muito conhecidas na área de teoria da informação.

2.3.1 Informação Mútua

Mutual Information Measure (MIM) mede a amplitude em que as palavras ocorrem independentemente. Na Equação 2.2, $P(a)$ é a probabilidade da palavra a ocorrer no documento (ou janela de texto), $P(b)$ a probabilidade da palavra b, e $P(a,b)$ é a probabilidade de a e b ocorrerem. Se a ocorrência das palavras são independentes, então $P(a,b) = P(a)P(b)$ e a informação mútua será 0. Caso as duas palavras tendam a coocorrer, $P(a,b)$ será maior que $P(a)P(b)$ e a informação mútua será maior que 0.

$$\log \frac{P(a,b)}{P(a)P(b)} \tag{2.2}$$

Adotamos $P(a) = \frac{na}{N}$, $P(b) = \frac{nb}{N}$ e $P(a,b) = \frac{nab}{N}$, onde na é o número de documentos contendo a palavra a, nb é o número de documentos contendo a palavra b, nab é o número de documentos que possuem ambas as palavras, e N é a quantidade de documentos na coleção. Agora a Equação 2.2 pode ser re-escrita da seguinte forma:

$$\log \frac{P(a,b)}{P(a)P(b)} = \log \left(N \cdot \frac{nab}{na \cdot nb} \right) \equiv \frac{nab}{na \cdot nb} \tag{2.3}$$

É importante observar que MIM tende a beneficiar termos de baixa frequência. Por exemplo, seja a e b com frequência 10 e eles coocorrem metade das vezes ($nab = 5$), o valor de MIM seria 0,05. Agora, seja a e b com frequência 1.000 e coocorrem também metade das vezes ($nab = 500$), teríamos o valor de MIM em 0,0005. Esse problema é contornado pela métrica *Informação Mútua Esperada*, apresentada na sequência.

2.3.2 Informação Mútua Esperada

Expected Mutual Information Measure (EMIM) adiciona o peso de $P(a,b)$ à Equação 2.2 da Informação Mútua. Dessa forma, EMIM cobre todas as combinações de eventos em que as palavras coocorrem e não coocorrem, preferindo onde ambos os termos ocorrem. A equação resultante da EMIM é apresentada abaixo:

$$P(a,b) \cdot \log \frac{P(a,b)}{P(a)P(b)} = \frac{nab}{N} \cdot log \left(N \cdot \frac{nab}{na \cdot nb} \right) \equiv nab \cdot \log \left(N \cdot \frac{nab}{na \cdot nb} \right) \tag{2.4}$$

De volta ao exemplo anterior, e assumindo $N = 1.000.000$, EMIM seria 23,5 quando os termos tivessem baixa frequência, e 1.350 quando a e b tivessem alta frequência. Observe que agora o EMIM beneficia termos de alta frequência, e esta tendência pode ser um problema em algum momento.

2.3.3 Kullback-Leibler Divergence

Dada pela Equação 2.5, o *Kullback-Leibler divergence*, ou entropia relativa, estima a diferença entre duas probabilidades de massa $p(x)$ e $q(x)$ — a distância entre duas distribuições de probabilidade [Kullback & Leibler, 1951].

$$p(x) \cdot \log \frac{p(x)}{q(x)} \tag{2.5}$$

A Equação 2.5 assemelha-se à informação mútua esperada e, por meio de deduções matemáticas, podemos verificar que

$$P(a,b) \cdot \log \frac{P(a,b)}{P(a)P(b)}$$
$$P(a \cap b) \cdot \log \frac{P(a \cap b)}{P(a)P(b)}$$
$$P(b)P(a|b) \cdot \log \frac{P(b)P(a|b)}{P(a)P(b)}$$
$$P(b)P(a|b) \cdot \log \frac{P(a|b)}{P(a)}$$

e caso $P(b)$ seja constante ao longo do problema analisado e fazendo $p(x) = P(a|b)$ e $q(x) = P(a)$, verificamos que o *Kullback-Leibler divergence* produzirá a mesma saída que a informação mútua esperada.

2.3.4 Chi-Squared de Pearson (χ^2)

O Chi-Squared de Pearson corresponde a um teste estatístico χ^2 que computa o relacionamento entre frequência esperada, em toda população, e uma frequência observada [Pearson's, 1900]. O cálculo é feito pela Equação 2.6, em que E_i corresponde ao fenômeno esperado e O ao fenômeno observado.

$$\chi^2 = \frac{(O - E_i)^2}{E_i} \tag{2.6}$$

Em Croft et al. [2009], o chi-squared é usado para comparar o número de coocorrências de duas palavras a e b, fenômeno observado O, com o número esperado de coocorrências quando as duas palavras são independentes, fenômeno esperado E_i. A comparação é normalizada pelo valor esperado, que corresponde a $N \times \frac{na}{N} \times \frac{nb}{N}$ na Equação 2.7.

$$\frac{(nab - N \times \frac{na}{N} \times \frac{nb}{N})^2}{N \times \frac{na}{N} \times \frac{nb}{N}} \equiv \frac{(nab - \frac{na \cdot nb}{N})^2}{na \cdot nb} \tag{2.7}$$

Quando N é muito grande, a forma restrita (lado direito da Equação 2.7) produz o mesmo *ranking* de termos que a forma completa (lado esquerdo da Equação 2.7). Também é importante observar que chi-squared favorece termos de baixa frequência.

2.3.5 Coeficiente de Dice

O coeficiente de Dice (do inglês *Dice's coefficient*) é uma medida de similaridade usada em recuperação de informação como, por exemplo, na associação de termos [van Rijsbergen, 1979; Croft et al., 2009]. A medida vem sendo usada desde os primeiros estudos sobre similaridade de termos e na construção automática de *thesaurus*, por volta dos anos 1960. O coeficiente corresponde à simples proporção da ocorrência de termos que são coocorrentes, computado pela Equação 2.8:

$$\frac{2 \cdot n_{ab}}{n_a + n_b} \equiv \frac{n_{ab}}{n_a + n_b} \tag{2.8}$$

2.4 Métricas de Avaliação

Esta seção apresenta as principais métricas usadas na avaliação dos experimentos apresentados no Capítulo 4.

2.4.1 Precisão e Revocação

Revocação e *precisão* são as duas métricas mais comuns usadas na avaliação de máquinas de busca [Baeza-Yates & Ribeiro-Neto, 2011; Croft et al., 2009]. A *revocação* corresponde à proporção dos documentos relevantes que foram recuperados e a *precisão* a proporção dos documentos recuperados que são relevantes. Assumindo que a definição de relevância é binária, podemos definir A como o conjunto dos documentos relevantes e $\bar{A}$ como não relevantes, B como o cojunto dos documentos recuperados e $\bar{B}$ dos não recuperados. Assim, essas métricas serão computadas da seguinte forma:

$$Revocação = \frac{|A \cap B|}{|A|} \tag{2.9}$$

$$Precisão = \frac{|A \cap B|}{|B|} \tag{2.10}$$

onde $\cap$ significa a interseção entre dois conjuntos.

2.4.2 Precisão no Ponto e Precisão Média

O valor da precisão mostrado anteriormente computa a eficiência da máquina de busca no conjunto de documentos recuperados. Muitas vezes, é interessante observar o valor da precisão em uma posição específica do *ranking*, por exemplo, apenas para o primeiro documento no topo ou até o terceiro documento desse *ranking*. Nesse caso, a precisão no ponto mede a precisão para os k documentos mais relevantes, computada pela Equação 2.11:

$$p@k = \frac{1}{k}\sum_{i=1}^{k} rel(d_i), \tag{2.11}$$

onde $rel(d)$ é uma função binária que retorna 1 quando o documento d é relevante, ou 0 caso contrário.

Outra medida de precisão bastante utilizada na avaliação de *ranking* é a precisão média. O valor da precisão média é computado por meio da Equação 2.12, que corresponde à média da precisão no ponto dos documentos relevantes. É interessante observar que a precisão média valoriza os documentos no topo do *ranking*. Por exemplo, seja um *ranking* composto por dez documentos e apenas os cinco primeiros são relevantes, a precisão média será 1. Agora, se apenas os cinco últimos documentos são relevantes, $AP = 0{,}35$. Para ambos os casos a $p@10$ será igual a 0,5. Para um conjunto de consultas, a média dessa precisão é conhecida como MAP (*mean average precision*). Para computar a precisão média em qualquer posição do *ranking*, usamos a Equação 2.13, uma adaptação proposta por Ribeiro-Neto et al. [2005].

$$AP = \frac{1}{R}\left(\sum_{i=1}^{n} rel(d_i) \cdot p@i\right) \tag{2.12}$$

$$PAVG@k = \frac{1}{k}\sum_{i=1}^{k}\left(rel(a_i) \times \left(\frac{\sum_{j=1}^{i} rel(a_j)}{i}\right)\right) \tag{2.13}$$

2.4.3 Métricas para Julgamento Incompleto

Em experimentos de *ranking*, o julgamento de relevância é feito normalmente por humanos, avaliadores que analisam a resposta do sistema e assinalam os itens considerados relevantes. No entanto, durante a avaliação do resultados, nem sempre estão disponíveis o julgamento para todos os itens do *ranking*.

Algumas métricas, mais especificamente o MAP, geralmente são computadas usando 10 ou 100 itens, e alguns desses itens podem não ter sido avaliados. Quando existe julgamento incompleto, normalmente itens não avaliados são considerados irre-

levantes para o cálculo da precisão média. Entretanto, na literatura são encontradas algumas alternativas ao MAP, tais como *binary preference* [Buckley & Voorhees, 2004], *induced AP* e *inferred AP* propostas por Yilmaz & Aslam [2006].

Yilmaz & Aslam [2006] propuseram a precisão esperada, Equação 2.15, que é uma medida alternativa à precisão média quando existem julgamentos incompletos. Portanto, assim como o MAP, o *inferred AP* (*infAP*) corresponde à média da precisão esperada. Apesar da medida *bpref-10* (Equação 2.14) ser bastante usada em avaliação com julgamento incompleto, preferimos a outra porque estudos recentes mostram que o *infAP* se aproxima mais do MAP que o *bpref-10* [Yilmaz & Aslam, 2006; Carterette et al., 2010].

$$bpref\text{-}10 = \frac{1}{R}\sum_{r} 1 - \frac{|n \text{ ranked higher than } r|}{10+R}, \tag{2.14}$$

$$E[prec \text{ at rank } k] = \frac{1}{k} + \frac{k-1}{k} \cdot E[prec \text{ above } k], \tag{2.15}$$

$$E[prec \text{ above } k] = \frac{judged \text{ rel above } k}{judged \text{ rel above } k + judged \text{ nonrel above } k}$$

Capítulo 3

Uso de Taxonomias na Recomendação Baseada em Conteúdo

Neste capítulo discutimos as principais idéias estudadas para melhorar sistemas de recomendação baseada em conteúdo adicionando-se informação obtida de taxonomia. A partir do alvo da recomendação e de uma coleção de itens a serem recomendados, a recomendação baseada em conteúdo consiste em encontrar o subconjunto de itens semanticamente relacionado com o alvo. Neste contexto, a recomendação é feita pelo casamento semântico entre o alvo e itens a serem recomendados, representados por meio de um repositório de palavras. Essa abordagem simples, sem informação de taxonomia, será apresentada primeiramente. Logo em seguida serão apresentadas as estratégias que utilizam informação de taxonomia na tarefa de recomendação.

3.1 Repositório de Palavras

Repositório de palavras (*bag of words* – BOW) é uma representação bastante comum em tarefas que envolvem conteúdo textual (por exemplo, classificação textual, busca de documentos). Neste esquema, os itens são representados por vetores em que cada palavra do texto possui um peso associado, computado usualmente com o esquema *tf.idf*. Sistemas de recomendação baseada em conteúdo utilizam o modelo vetorial, apresentado na Seção 2.2, para computar um *ranking* de itens de acordo com a similaridade entre o alvo e os itens a serem recomendados.

A similaridade é normalmente computada por meio da medida do cosseno, apre-

sentada na Equação 2.1, ou uma de suas variações encontradas na literatura [Zobel & Moffat, 2006]. Neste trabalho, a similaridade entre a página alvo e o livro será computada pela Equação 3.1:

$$\begin{aligned} w_{p,t} &= \ln(1 + \frac{N}{f_t}) \\ w_{b,t} &= 1 + \ln f_{b,t} \\ W_b &= \sqrt{distinct\_terms} \\ sim(p,b) &= \frac{\sum w_{b,t} \times w_{p,t}}{W_b \times W_p} \end{aligned} \tag{3.1}$$

onde $f_{b,t}$ é a frequência do termo t na descrição do livro b, f_t é o número de livros em que t ocorre na coleção e N é o número total de livros na coleção. O $w_{p,t}$ é o peso do termo na página, $w_{b,t}$ é o peso do termo no livro e W_b e W_p são usados para normalização. O W_b é a raiz quadrada do número de termos distintos em b e o W_p pode ser ignorado, visto que é constante para todos itens da resposta.

A lista de recomendação, isto é, a lista com os itens que serão recomendados, corresponde aos k itens do topo do *ranking* descrito anteriormente. O tamanho de k geralmente é pequeno (por volta de dez itens) e varia de acordo com o cenário de aplicação.

A Figura 3.1 ilustra um exemplo de um sistema de recomendação onde o item alvo é uma página de notícia e os itens a serem recomendados são livros de uma loja *online*. Esse cenário corresponde ao estudo de caso explorado neste trabalho, onde a recomendação deve acontecer em tempo real e sem informação do perfil do usuário. Notícias são representadas pelo conteúdo textual das páginas e livros pela descrição textual de seus conteúdos.

Repositório de palavras é uma forma simples de representar o documento usando apenas informação presente no próprio texto. Entretanto, essa abordagem simplista pode gerar resultados de baixa qualidade e prejudicar a recomendação [Linden et al., 2003]. Isto acontece principalmente pelo baixo casamento entre os termos do alvo e da coleção, um fenômeno chamado de impedância do vocabulário por Ribeiro-Neto et al. [2005]. Por isso, é comum agregar novas características à representação dos itens para aumentar a qualidade da informação disponível.

Neste trabalho, aproveitamos a informação de qualidade presente em uma taxonomia para contornar o problema da impedância do vocabulário. O uso de taxonomias, e outras bases de conhecimento construídas por humanos, abre oportunidade de incorporar conhecimento de um domínio específico compilado por humanos. Essa valiosa

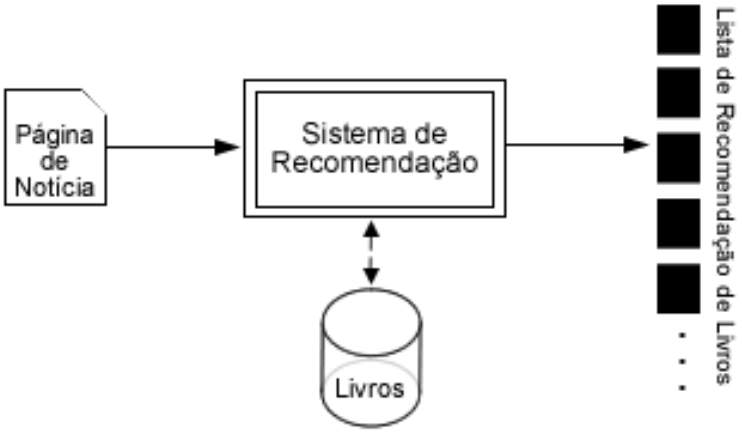

Figura 3.1. Arquitetura tradicional em recomendação baseada em conteúdo.

fonte de informação não poderia ser obtida a partir apenas do alvo ou dos itens.

Embora taxonomias tenham sido usadas anteriormente por outras pesquisas em tarefas de recuperação de informação [Anagnostopoulos et al., 2007; Gabrilovich & Markovitch, 2005], neste trabalho apresentamos um abrangente estudo de três diferentes estratégias para explorar taxonomias em sistemas de recomendação baseada em conteúdo. Como ilustrado na Figura 3.2, as estratégias utilizadas são *Descritores de Categoria*, *Características de Classificação* e *Filtro de Categorias*. Essas estratégias são descritas em detalhes nas seções seguintes.

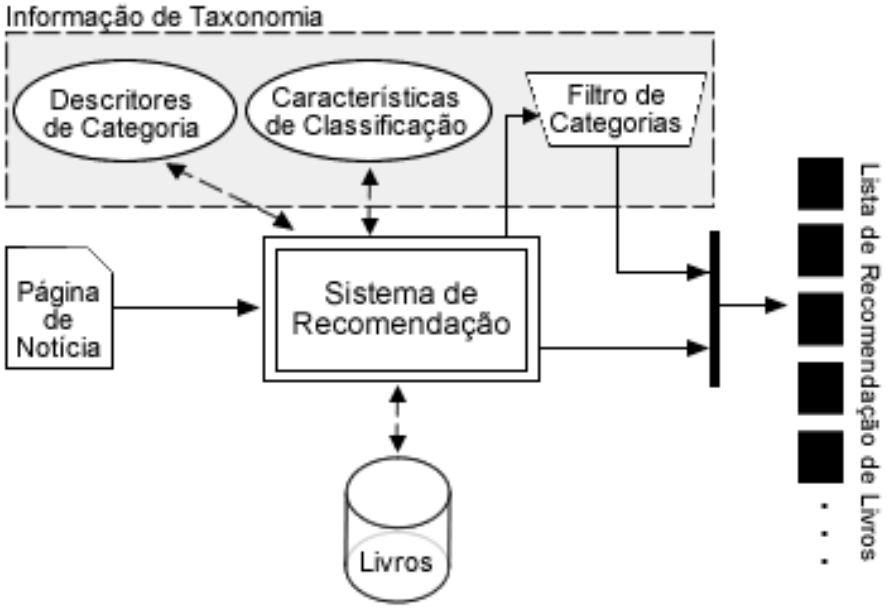

Figura 3.2. Arquitetura tradicional em recomendação baseada em conteúdo.

3.2 Descritores de Categoria

Uma estratégia comum que procura minimizar o problema de impedância do vocabulário é enriquecer a representação dos itens com novas fontes de informação. Por exemplo, no problema de seleção de propagandas relacionadas a páginas web estudado por Ribeiro-Neto et al. [2005], os autores sugeriram adicionar novos termos à representação inicial da página, os quais foram obtidos de páginas relacionadas. Uma estratégia parecida foi adotada por Carpineto et al. [2001] and Carpineto & Romano [1999] no problema de busca na Web, e os novos termos foram usados na expansão de consultas. Essa mesma idéia é experimentada no estudo desta dissertação, mas, ao invés de usar termos extraídos de páginas, é usada a informação obtida de uma taxonomia.

Alguns termos são mais discriminativos e melhores para representar determinado conceito, grupo ou categoria. Por exemplo, os termos *plantas*, *animais* e *mamíferos* caracterizam bem a categoria *Biologia*, entretanto, não são bons pra representar *Computadores*. Por isso, pretendemos selecionar um conjunto de termos representativos da categoria e usá-los para expandir a representação da página alvo. Os termos selecionados das categorias são chamados de *descritores de categoria*.

O processo para o uso de descritores de categoria consiste nas seguintes etapas: (1) associar a página alvo a uma ou mais categorias na taxonomia; (2) obter as descrições dos livros associados a essas categorias, e a partir delas selecionar os descritores, isto é, termos representativos da categoria; (3) adicionar esses descritores à representação da página alvo usando uma formula de combinação linear. Ao final desse processo, a representação da página alvo é enriquecida com termos que caracterizam as categorias às quais elas foram associadas. E como demonstrado pelos experimentos no Capítulo 4, essa representação enriquecida leva a resultados de recomendação mais precisos.

Para a etapa (1), as categorias que as páginas alvo estão associadas podem ser selecionadas manualmente ou usando um classificador automático. Cada uma dessas estratégias possuem suas vantagens e desvantagens. Por exemplo, a abordagem manual é geralmente mais precisa mas requer um esforço extra do usuário, o qual pode ser inconveniente se o número de categorias é muito alto. As duas abordagens foram avaliadas nos experimentos descritos no Capítulo 4. As etapas (2) e (3) requerem uma discussão mais detalhada, o que será feito a seguir.

3.2.1 Geração de Descritores de Categoria

Considere uma categoria C na taxonomia e seja $T(C)$ o conjunto de todos os termos que ocorrem nas descrições dos livros associados a essa categoria. Então, $D(C)$ corresponde

ao conjunto de termos descritores da categoria C, tal que $D(C) \subseteq T(C)$. Para isso, usamos uma função $m(t, C)$ que associa a importância do termo t à categoria C e construir o conjunto $D^K(C)$ dos top-K descritores, isto é, os K termos que possuem os valores mais alto de $m(t, C)$. Neste trabalho, foram testadas diferentes medidas que implementam essa função.

As medidas apresentadas a seguir usam a seguinte notação:

i) $P(t)$ é a probabilidade de um dado termo t na coleção de livros e $P(t|C)$ é a probabilidade do termo t dada a categoria C. Ambas as probabilidades são estimadas pela proporção de livros (descrição dos livros) que contêm o termo t sobre todos os livros ou apenas na categoria;

ii) B_t é o conjunto de todos os livros que contêm o termo t e B_C é o conjunto de todos os livros associados à categoria C.

Kullback-Leibler Divergence (KLD). Ou *entropia relativa*, é usada para estimar a diferença entre duas probabilidades de massa $p(x)$ e $q(x)$ — a distância entre duas distribuições de probabilidade [Carpineto et al., 2001; Kullback & Leibler, 1951], conforme a Equação 2.5. Em nosso caso, $p(x)$ representa as observações de $P(t|C)$ e $q(x)$ é $P(t)$. Assim, a equação usada para computar os descritores é:

$$P(t|C) \times \log \frac{P(t|C)}{P(t)}.$$

Chi-Squared de Pearson (CHI2). É um teste estatístico usado para computar o relacionamento entre uma frequência esperada na população geral e uma frequência observada [Carpineto et al., 2001; Croft et al., 2009; Pearson's, 1900]. Em nosso caso, a frequência esperada é a chance $P(t)$, e a frequência observada é a probabilidade de t condicionada à categoria C, ou seja, $P(t|C)$. Consequentemente, essa medida foi definida como:

$$\frac{(P(t|C) - P(t))^2}{P(t)}.$$

Coeficiente de Dice (DICE). Mede a similaridade entre dois conjuntos. Por exemplo, em [Croft et al., 2009; van Rijsbergen, 1979] é usada para medir a associação de termos. Em nosso caso, usamos essa medida para avaliar a similaridade entre B_t e B_C. Se esses dois conjuntos são similares, então t é considerado estreitamente relacionado a C. Nesse caso, o coeficiente é dado por:

$$\frac{2 \times |B_t \cap B_C|}{|B_t| + |B_C|}.$$

Document Frequency (DF). Também experimentamos uma medida simples, baseada na frequência do termo na categoria, que corresponde à interseção entre B_t e B_C, ou seja:

$$|B_t \cap B_C|.$$

Combinação de Medidas (ALL). Carpineto & Romano [1999] propuseram uma nova abordagem na expansão de consultas e, inspirados na combinação de classificadores, uma nova seleção de termos foi feita a partir da combinação de outras medidas. Os autores combinaram as medidas usando um esquema de votação e depois atribuíram um peso ao termo em função da sua posição no *ranking*. Motivados por essa abordagem, nós também propomos uma nova medida que explora a combinação. Nossa combinação corresponde à média aritmética dos escores das medidas participantes, sendo que essas medidas foram inicialmente normalizadas entre 0 (menor escore) e 1 (maior escore). No método combinado, ALL, foram combinadas KLD, CHI2 e DICE, as quais foram usadas anteriormente na expansão de consultas [Carpineto et al., 2001; Carpineto & Romano, 1999].

Um resumo das medidas, acompanhadas das fórmulas usadas, é apresentada na Tabela 3.1. Para ilustrar o uso dessas medidas, a Tabela 3.2 apresenta os 15 termos descritores mais importantes da categoria *Religião* para cada uma das medidas, isso considerando a taxonomia e descrição dos livros da amostra utilizada da Amazon (vide Seção 4.1). Em comparação a outras medidas, a medida DF seleciona termos que também são muito frequentes em toda coleção (por exemplo, *book*, *world*, *work*) e são pouco discriminativos.

Vale ressaltar que nem todas as medidas apresentadas no Capítulo 2 foram usadas como descritores. Por exemplo, a medida de *informação mútua* (MIM) produziu termos bastante específicos, considerados ruído à recomendação, e por isso foi desconsiderada na geração dos descritores. Quanto à *informação mútua esperada* (EMIM), produziu a mesma saída que a KLD por conta da probabilidade na categoria $P(C)$ ser constante – ver prova ao final da Seção 2.3.

Tabela 3.1. Medidas usadas para seleção dos descritores de categoria.

Medida	Fórmula
Kullback-Leibler divergence (KLD)	$P(t\|C) \cdot \log \frac{P(t\|C)}{P(t)}$
Chi-square Pearson (CHI2)	$\frac{(P(t\|C)-P(t))^2}{P(t)}$
Coeficiente de Dice (DICE)	$\frac{2\times\|B_t \cap B_C\|}{\|B_t\|+\|B_C\|}$
Document Frequency (DF)	$\|B_t \cap B_C\|$
Combinação de Medidas (ALL)	$\frac{KLD+CHI2+DICE}{3}$

Tabela 3.2. Exemplo para os top-15 descritores da categoria *Religião*.

DF	KLD	CHI2	DICE	ALL
book	god	god	god	god
life	christian	christian	spiritual	christian
god	spiritual	spiritual	christian	spiritual
new	bible	bible	life	bible
world	church	church	bible	church
spiritual	faith	jesus	church	faith
people	life	christ	faith	jesus
time	jesus	faith	book	christ
work	christ	biblical	religious	life
christian	religious	christians	jesus	biblical
author	biblical	scripture	people	religious
bible	religion	religious	christ	religion
history	christians	theology	world	christians
church	scripture	christianity	religion	scripture
study	theology	theological	new	theology

3.2.2 Recomendação com Descritores de Categoria

No processo de recomendação, os termos descritores de categoria são adicionados à representação vetorial da página alvo. No vetor de termos da página p, o peso do termo t na página foi computado usando o seguinte equação baseada no *tf.idf*:

$$w_{p,t} = f_{p,t} \times \ln(1 + \frac{N}{f_t}), \tag{3.2}$$

onde $f_{p,t}$ é a frequência do termo na página p e $\ln(1 + \frac{N}{f_t})$ corresponde à especificidade do termo da coleção[1]. A Tabela 3.3 apresenta o *ranking* dos termos de uma página exemplo, construído a partir do de $w_{p,t}$. Podemos observar que os termos próximos ao topo são mais relevantes ao assunto da página.

Tabela 3.3. Lista com 10 termos de maior (esquerda) e menor escore (direita) computados usando *tf.idf* normalizada pelo maior peso. Termos marcados com '*' são considerados relevantes para a página analisada, com o seguinte tema: *Papa Bento XVI fala sobre a crise de abuso sexual que afeta a Igreja Católica.*

Top 10		Bottom 10	
termo	peso	termo	peso
vatican*	1.00	called	0.02
pope*	0.88	city	0.02
abuse*	0.83	national	0.02
benedict*	0.58	second	0.02
church*	0.54	born	0.02
tuesday	0.39	history	0.02
sexual*	0.34	year	0.02
crisis*	0.33	later	0.02
remarks	0.31	years	0.02
nytimes	0.29	time	0.01

Por outro lado, a seleção dos descritores de categoria é feita usando as medidas listadas anteriormente. Adicionalmente, o valor associado por cada medida também será usado para reajustar o peso do termo. Para isso, recorremos a uma combinação linear similar à usada na fórmula de Rocchio [Rocchio et al., 1971].

Então, seja t um termo da representação BOW estendida da página alvo, o peso desse termo é computado como:

$$w'_{p,t} = (1 - \alpha) \times w_{p,t} + \alpha \times m(t, C), \tag{3.3}$$

onde α é uma constante positiva entre 0 e 1, $w_{p,t}$ é o peso de t na página p e $m(t, C)$ é o valor obtido por t na categoria C a qual a página está associada, usando uma das medidas descritas anteriormente. Para evitar discrepâncias, primeiramente os componentes da fórmula são normalizados para a mesma escala. Quando t é descritor e não está presente na página, $w_{p,t}$ será zero. E quando t está presente na página e não é descritor, $m(t, C)$ terá valor zero.

[1]No caso dos descritores de categoria, foi usada uma coleção externa composta por cerca de 3 milhões de artigos da Wikipedia (`http://www.wikipedia.org/`).

Chamamos atenção para quando a página está associada a várias categorias. O valor de $w'_{p,t}$ será calculado usando a média dos pesos $m(t, C)$ obtidos para todas as categorias. Este cenário foi experimentado quando um classificador automático é usado para predizer as categorias da página.

Na prática, a nova representação da página alvo p pode incluir um número muito grande de termos, por conta da adição dos descritores de categoria. Entretanto, a maioria desses termos não tem impacto algum na representação, visto que eles possuem um peso muito baixo. Portanto, essa representação foi limitada e foi mantido mantido apenas os N termos com mais alto escore, onde N é igual ao número de termos da representação inicial da página.

3.3 Características de Classificação

Uma abordagem alternativa para enriquecer a representação de itens envolvidos na recomendação é usar um outro *espaço de características* em adição ao usual espaço de termos da abordagem repositório de palavras. Por exemplo, na tarefa de classificação textual em Gabrilovich & Markovitch [2005], uma espaço de características adicional foi construído usando categorias de uma taxonomia, as quais são chamadas *características de classificação*.

A geração de características é desempenhada por um *gerador de características*, construído usando uma taxonomia como base de conhecimento. Por meio de técnicas de classificação, esse gerador aprende e infere conceitos, ou seja, as novas características de classificação. Na construção do gerador de características algumas regras devem ser obedecidas [Gabrilovich & Markovitch, 2005]. Por exemplo, a base de conhecimento escolhida deve conter uma coleção de conceitos organizada na forma de estrutura de árvore hierárquica. Essa base de conhecimento também deverá ter uma coleção de textos associada com cada conceito, que serão usados pelo gerador de características para aprender a definição e escopo de cada conceito e, enfim, poder atribuí-lo aos documentos de interesse.

Em Gabrilovich & Markovitch [2005], os autores utilizaram geração de características para melhorar a tarefa de classificação e adotaram uma base de conhecimento do Open Directory Project (ODP)[2], construída por humanos e com cerca de 400 mil conceitos (nós). Já Anagnostopoulos et al. [2007] usaram uma taxonomia construída para fins comerciais por Yahoo! US[3], com aproximadamente 6 mil nós, na associação de propagandas à páginas de Internet. Em nosso trabalho, experimentamos essa abor-

[2] `http://www.dmoz.org/`
[3] `http://www.yahoo.com/`

dagem e utilizamos a própria taxonomia de livros da Amazon.com (vide Seção 4.1), com 1.621 nós, na tarefa de recomendação de livros à páginas Web.

Para isso, a página p é representada por um vetor $\vec{p}$ no espaço de termos mais outro vetor $\vec{p_{cat}}$ no espaço de categorias, o qual é constituído por categorias preditas por um classificador. Portanto, assumimos que essas categorias são fortemente relacionadas à página. Assim como em Gabrilovich & Markovitch [2005]; Anagnostopoulos et al. [2007], o gerador de características utilizado também foi um classificador de texto baseado em centróides [Han & Karypis, 2000].

A seguir, na Seção 3.3.1 mostramos como os centróides são computados com o gerador de características. Depois, na Seção 3.3.2 apresentamos o procedimento para gerar as características de classificação para a página alvo p. Finalmente, na Seção 3.3.3 descrevemos como essas características são incorporadas no processo de recomendação.

3.3.1 Construção de um Gerador de Características

Antes da geração das características de classificação, um vetor de termos $\vec{c_j}$ tem de ser gerado para cada categoria C_j da taxonomia. Esse vetor corresponde ao centróide dos vetores de termos dos livros associados à categoria.

Para computar $\vec{c_j}$, um repositório de palavras é construído para cada livro b associado a C_j usando os termos presentes em seu título e descrição. *Stopwords*, números, palavras com caracteres numéricos e palavras muito raras (isto é, aquelas que não ocorreram em pelo menos 5 itens) foram removidas. *Stemming* foi aplicado aos termos remanescentes.

O vetor de termos $\vec{b}$ é então gerado para o livro b usando os *stems*, de acordo com modelo espaço vetorial. Para isso, o vocabulário da coleção corresponde ao espaço de termos.

Seja B_j o conjunto de vetores que representam os livros associados à categoria C_j, o centróide $\vec{c_j}$ que representa C_j é definido por:

$$\vec{c_j} = \frac{1}{|B_j|} \sum_{\vec{b} \in B_j} \vec{b} \tag{3.4}$$

No modelo espaço vetorial, a similaridade entre dois documentos é comumente medida usando a função cosseno [Salton, 1989] e, analogamente, a classificação será baseada no cosseno do ângulo entre o livro e o centróide:

$$C_{max} = \arg\max_{C_j \in C} \frac{\vec{c_j}}{\|\vec{c_j}\|} \cdot \frac{\vec{b}}{\|\vec{b}\|} = \arg\max_{C_j \in C} \frac{\sum w_{c,t} \cdot w_{b,t}}{\sqrt{\sum w_{c,t}^2}\sqrt{\sum w_{b,t}^2}} \tag{3.5}$$

onde $w_{c,t}$ e $w_{b,t}$ correspondem ao peso dos termos em $\vec{c_j}$ e $\vec{b}$, respectivamente, computados com um esquema de peso baseado no padrão "ltc" do *tf.idf* [Salton & Buckley, 1988],

Seguindo Gabrilovich & Markovitch [2005], o vetor $\vec{c_j}$ é construído usando apenas os 1.000 termos mais frequentes da categoria. Outro método de seleção também poderia ter sido usado, no entanto, termos mais frequentes (DF) apresentou melhores resultados nos experimentos de classificação, mostrados na Tabela 3.4. Por questão de desempenho, foi utilizada uma amostragem estratificada de 1% do conjunto original de instâncias, o qual foi dividido em 70% para treino e 30% para teste. Os nós com poucas instâncias foram removidos e o número máximo de instâncias por nó foi limitado, a fim de minimizar o desbalanceamento da base.

Tabela 3.4. Avaliação da seleção de atributos na amostra de livros.

	Recall	*Precisão*	*Acurácia*
Todos atributos	0.33	0.47	0.25
Document Frequency	0.28	0.40	0.21
Information Gain	0.24	0.34	0.17
Chi-Squared	0.23	0.33	0.16

3.3.2 Geração de Características de Classificação

O procedimento para geração de características de classificação para a página p é descrita pelo Algoritmo 1 e detalhado a seguir.

Após uma etapa de inicialização, na linha 3, o algoritmo extrai segmentos da página alvo p. O objetivo aqui é ser capaz de selecionar categorias relacionadas à toda página por meio das categorias relacionadas a cada segmento da página. Assim como em Gabrilovich & Markovitch [2005], modos distintos de segmentação foram avaliados, entretanto, dois deles tiveram os melhores resultados: (i) fazendo todo conteúdo da página como um único segmento e (ii) com a abordagem *multi-resolution*, isto é, janela de palavras (bigramas), parágrafos, e conteúdo completo documento.

O algoritmo então itera pelos segmentos (laço 4–14). Para cada segmento s, um vetor de termos $\vec{s}$ é gerado usando o esquema de peso $tf.idf$ (linha 5), então é possível

Algoritmo 1 Geração das Características de Classificação.

Entrada: página alvo p
Saída: vetor de categorias $\vec{p}_{cat}$ para p

1: **Para todo** categorias C_j na taxonomia **faça**
2: $\quad \vec{p}_{cat}[C_j] \leftarrow 0$
3: $S \leftarrow segmentos(p)$
4: **Para todo** segmentos $s \in S$ **faça**
5: $\quad$ Gerar vetor de termos $\vec{s}$ para s
6: $\quad$ **Para todo** categorias C_j na taxonomia **faça**
7: $\quad\quad$ **Seja** $\vec{c_j}$ o vetor do centróide gerado para C_j
8: $\quad\quad sim(s, C_j) \leftarrow \cos(\vec{s}, \vec{c_j})$
9: $\quad \mathcal{C}^K \leftarrow$ as K categorias mais similares a s
10: $\quad$ **Para todo** $C_j \in \mathcal{C}^K$ **faça**
11: $\quad\quad \vec{p}_{cat}[C_j] \leftarrow \vec{p}_{cat}[C_j] + sim(s, C_j)$
12: $\quad$ **Seja** $\mathcal{A}$ o conjunto dos ancestrais das categorias em $\mathcal{C}^K$
13: $\quad$ **Para todo** $C_\ell \in \mathcal{A}$ **faça**
14: $\quad\quad \vec{p}_{cat}[C_\ell] \leftarrow \vec{p}_{cat}[C_\ell] + sim(s, C_\ell) \cdot \epsilon$

determinar a similaridade entre o segmento e cada categoria, calculando-se o cosseno do ângulo entre os vetores $\vec{s}$ e $\vec{c_j}$ (linha 8).

Nas linhas 10–11 o vetor de categorias $\vec{p_{cat}}$ é construído usando as K categorias mais relevantes para cada segmento. Vale observar que o peso de cada categoria depende do número de segmentos que está relacionado a ela e do valor da similaridade entre o segmento e a categoria. Adicionalmente, nas linhas 13–14, os ancestrais dessas top-K categorias na taxonomia são também incluídos como característica de classificação, com o peso amortizado pela constante ϵ. O objetivo é capturar conceitos de mais alto nível (por exemplo, *esportes* adicionalmente a *futebol*), evitando, no entanto, sua dominância no vetor. Similar ao que foi feito em Gabrilovich & Markovitch [2005]; Anagnostopoulos et al. [2007], em nossos experimentos usamos $K = 5$ e $\epsilon = 0{,}5$.

3.3.3 Recomendação com Características de Classificação

As características de classificação geradas para a página alvo p usando o Algoritmo 1 agora podem ser usadas no processo de recomendação. Para isso, o vetor de termos $\vec{p}$ no espaço de termos é usado juntamente com o vetor $\vec{p_{cat}}$ no espaço de categorias na associação da página alvo p com o livro b, como segue:

$$sim_{CLF}(p, b) = \alpha \cdot cos(\vec{p}, \vec{b}) + \beta \cdot cos(\vec{p_{cat}}, \vec{b_{cat}}), \tag{3.6}$$

onde $\vec{b}$ e $\vec{b_{cat}}$ são, respectivamente, o vetor de termos e o vetor de categorias para b, e α e β são constantes, as quais tiveram valor fixado em 1 em nossos experimentos.

O vetor $\vec{b}$ é gerado de acordo com o modelo espaço vetorial usando repositório de palavras na representação do livro b. No caso de $\vec{b_{cat}}$, recordamos que todos os livros foram manual e antecipadamente associados a um conjunto de categorias da taxonomia. Aproveitamos dessa informação de alta qualidade e criamos o vetor $\vec{b_{cat}}$, associando 1 às dimensões correspondentes a cada uma dessas categorias.

3.4 Filtro de Categorias

As duas estratégias descritas anteriormente são usadas para melhorar a associação entre o alvo da recomendação e os itens a serem recomendados. Nossa terceira estratégia é baseada no *filtro de categorias*, que trabalha na saída dessa associação e acontece independente de qualquer estratégia de associação.

O filtro de categorias realiza uma filtragem na saída do casamento de termos, representados por repositório de palavras. Os termos extraídos da página podem introduzir ruído (termos de baixa qualidade) ou esses termos também podem estar relacionados a outros assuntos, diferente daquele abordado na página. Ambas as situações contribuem para recuperar livros não interessantes. Por exemplo, a palavra "suína" poderá recuperar livros sobre o animal, quando o assunto da página é a doença "gripe suína". Para resolver esse problema, restringimos o domínio para sugerir apenas livros pertencentes à mesma área de conhecimento da página. Assim, apenas livros sobre saúde com a palavra "suína" seriam sugeridos no exemplo anterior.

Neste contexto, utilizamos a própria taxonomia de livros da Amazon, apresentada na Seção 4.1, como base de conhecimento e a restrição de domínio é realizada usando suas categorias. Neste trabalho, essa estratégia de filtro de categoria funciona conforme é descrito a seguir. Inicialmente, a página alvo é mapeada em uma ou mais categorias da taxonomia dos livros. Em seguida, a lista de recomendação de livros adquirida usando algumas das estratégias de casamento de termos descritas anteriormente. Essa lista então é filtrada e apenas livros pertencentes àquelas categorias (pelo menos uma) associadas à página alvo são mantidos.

Podemos fazer um comparativo com a estratégia que aumenta a representação da página com descritores de categoria, o que efetivamente favorece a revocação, visto que o número de livros relacionados tende a crescer. Isto, entretanto, pode direcionar a resultados fora do assunto da página. O filtro de categoria, por sua vez, realiza uma poda nos resultados fora do assunto e evita este desvio, o que favorece a precisão.

Para que essa estratégia funcione sem problemas, é crucial a associação apropriada das categorias às páginas. Nos experimentos deste trabalho, essa associação foi avaliada tanto no cenário manual quanto automático, detalhada na Seção 4.1.

Capítulo 4

Experimentos

Este capítulo descreve a metodologia usada na avaliação experimental das estratégias descritas no Capítulo 3. A Seção 4.1 apresenta as coleções utilizadas na avaliação: a primeira, livros da Amazon.com; a segunda, notícias do The New York Times. A Seção 4.2 descreve a metodologia experimental com detalhes sobre a avaliação da recomendação dos livros, realizada por humanos. Na Seção 4.3 são apresentados os vários métodos para recomendação baseada em conteúdo que aplicam aquelas estratégias individualmente ou de forma combinada. Por fim, na Seção 4.4 são apresentados os resultados e discussão dos experimentos.

4.1 Coleções

Esta seção descreve as coleções usadas no desenvolvimento deste trabalho: uma amostra dos livros da Amazon.com e páginas de notícias do The New York Times[1].

4.1.1 Livros da Amazon.com

Uma amostra de livros da Amazon.com foi utilizada para gerar listas de recomendação. A construção da amostra foi dividida em duas etapas: *download* de páginas web e consulta a *Web Services*. Na primeira etapa, um simples *crawler* foi usado para coletar algumas centenas de páginas HTML de livros do *site* da Amazon.com[2] e um *parser* específico foi usado para extrair os termos do título, autores e descrição dos livros, bem como a revisão dos usuários. Na segunda etapa, esses termos foram submetidos como

[1] As coleções e os julgamentos da avaliação experimental (posteriormente descritas) estão disponíveis em `http://www.latin.dcc.ufmg.br/collections/sigir2011/` para efeito de reprodutibilidade.

[2] http://www.amazon.com/

consulta para o AWS[3] e os livros extraídos dos documentos XML retornados foram usados para compor a amostra.

Milhares de livros foram coletados, os quais muitos eram duplicados. O processo de remoção de duplicatas usou o identificador de produtos da Amazon (ASIN) e o ISBN dos livros para distinguir um dos outros. Depois desse processo, cerca de 6 milhões de livros (em 34 categorias) permaneceram na coleção. Em seguida, foram removidos livros sem editorial, sem categoria, sem ISBN e também livros em diferentes tipos de mídia –por exemplo, livros no formato de áudio (audiobooks). Outros produtos diferentes de livros (por exemplo, software, DVD) retornados pela API também foram removidos. Essa filtragem nos deu uma lista de 1.499.792 livros e 28 categorias no nível 1 da taxonomia. A amostra final era desbalanceada, o que pode ser observado na Figura 4.1, considerando apenas o nível 1 da taxonomia.

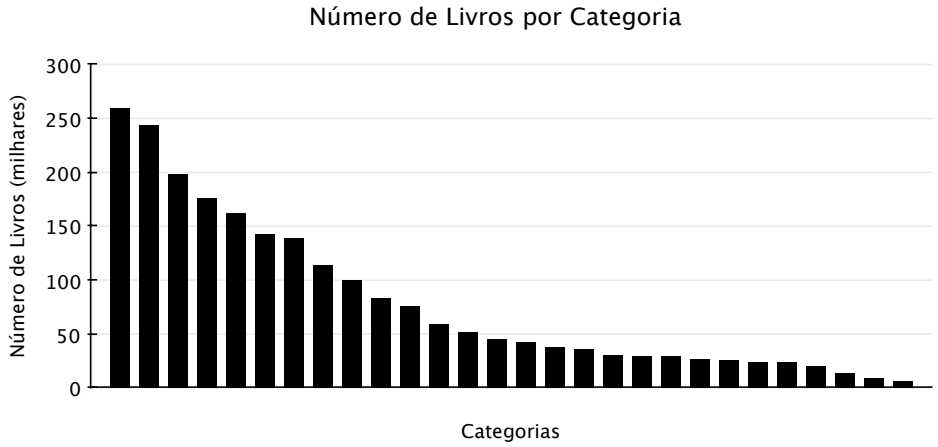

Figura 4.1. Número de livros por categoria.

A taxonomia de livros da Amazon pode ser representada como uma estrutura de árvore (Figura 4.2), onde *'Books'* é a raiz da árvore e os nós filhos são as categorias. Usando a informação de taxonomia retornada pelo AWS, a árvore inicial da taxonomia possuía 11.299 nós. Após remover *Customs Stores*, *Feature Stores* e outros nós de navegação, a árvore da base final de livros possuía 1.621 nós e profundidade máxima igual a sete. Como os livros podem aparecer em diversas categorias, a quantidade de instâncias de livros na árvore foi maior que o número real da amostra final de livros (classificação multi-rótulo). Essa informação e o sumário dos dados dessa amostra da Amazon é apresentado na Tabela 4.1. Na Figura 4.3, é possível observar que o terceiro

[3]A documentação do Amazon Web Service (AWS) está disponível em `http://docs.amazonwebservices.com/AWSECommerceService/2007-04-04/DG/`

nível possui mais categorias e, pela Figura 4.4, que a maioria dos nós possuem poucos filhos.

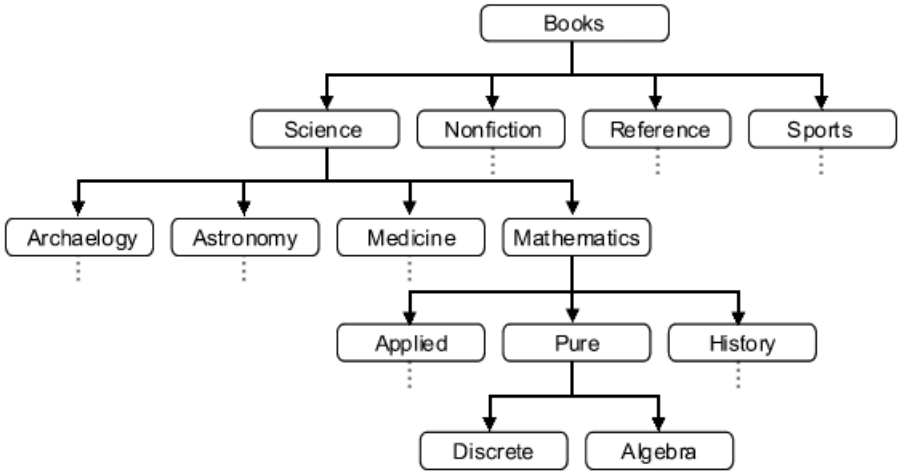

Figura 4.2. Fragmento da taxonomia de livros da Amazon [Ziegler et al., 2005]

Tabela 4.1. Sumarização da amostra de livros coletada da Amazon.com

Atributo	Valor
Total de livros coletados	6.372.612
Tamanho da amostra final de livros	1.499.792
Altura da árvore	8
Número de nós no nível 1	28
Número de nós em todos os níveis	1.621
Número de instâncias de livros	2.198.850

4.1.2 Notícias do The New York Times

Esta segunda coleção corresponde ao alvo de recomendação. Os métodos propostos podem ser usados com qualquer tipo de página web, pois afinal requerem apenas o conteúdo textual das páginas, o qual inclui os textos presentes em menu de navegação, rodapé e outros elementos da página. Entretanto, páginas de notícias foram escolhidas por conter maior informação textual e também já foram usadas anteriormente em outros estudos [Ribeiro-Neto et al., 2005]. Então, as páginas são usadas como alvo para a recomendação de livros, e a recomendação deve acontecer em tempo real, sem qualquer informação sobre o perfil do usuário que acessa a página.

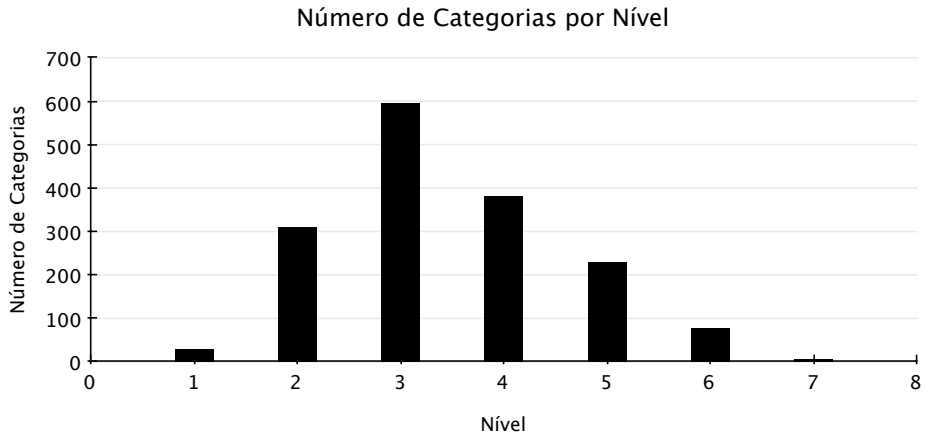

Figura 4.3. Número de categorias por nível: mediana = 3 e nível máximo = 7.

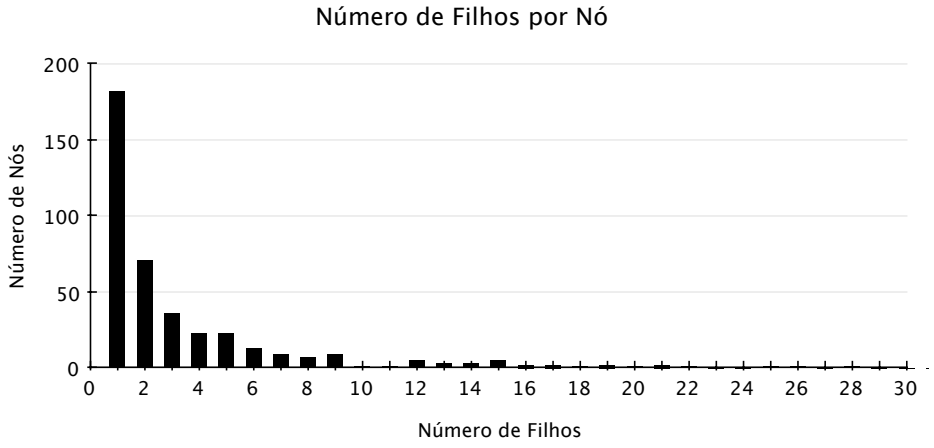

Figura 4.4. Número de nós filhos por nó.

Uma coleção de páginas-alvo foi construída com notícias extraídas do The New York Times[4] (NYT), publicadas entre novembro e dezembro de 2009. Essa coleção é composta por 100 notícias, igualmente (ou parcialmente) distribuídas em nove tópicos diferentes: *Arts, Business, Technology, Education, Health, Politics, Science, Sports and Travel.*

Como mencionado anteriormente, nossas estratégias requerem uma ou mais categorias associadas com a página alvo, onde consideramos dois cenários de associação: um manual e outro automático. Intuitivamente, aqueles tópicos poderiam ser mapeados por alguma pessoa em uma categoria correspondente na taxonomia de livros da Amazon, como pode ser visto na Tabela 4.2. Entretanto, isso levou a um mapeamento

[4] `http://www.nytimes.com/`

genérico e superficial. A partir do contexto da notícia e buscando enriquecer o nível de detalhes, cada uma das 100 notícias foi mapeada no nó mais específico de toda hierarquia da taxonomia. Assim, uma notícia sobre a *"Copa do Mundo FIFA"* não seria mapeada apenas para a categoria *Sports*, e sim em *Sports>Soccer*. O mapeamento completo das notícias do NYT é apresentado no Anexo. Para o segundo cenário, poderia ter sido usado qualquer classificador automático de texto, e nós adotamos o mesmo classificador baseado em centróides descrito na Seção 3.3 para associar uma ou mais categorias nessa mesma taxonomia.

Tabela 4.2. Mapeamento dos tópicos das páginas de notícias do The New York Times para categorias nível 1 da taxonomia de livros da Amazon.com.

Tópico no NYT	Categoria na Amazon
Arts	Arts & Photography
Business	Business & Investing
Technology	Computers & Internet
Education	Nonfiction
Health	Health, Mind & Body
Politics	Nonfiction
Science	Science
Sports	Sports
Travel	Travel

4.2 Configuração Experimental

Nos experimentos, foi adotado um esquema de *pooling* largamente empregado na avaliação de sistemas de recuperação de informação, por exemplo, na avaliação da TREC [Hawking et al., 1998]. Fazendo uma analogia com a TREC, as páginas de notícias correspondem aos tópicos e os métodos de recomendação produzem listas ordenadas de recomendação que correspondem às *runs*. Os 100 tópicos do NYT foram considerados entrada para os diferentes métodos de recomendação, os quais geram vários pares página-livro. Como a tarefa de recomendação geralmente envolve poucos itens, para cada *run* coletamos os k livros mais relevantes ao tópico, onde $k = 5$. Então, os resultados do topo do *ranking* de livros foram selecionados e enviados para julgamento de avaliadores. Após remover pares duplicados, restaram 7.380 pares distintos, o que garante quase 95% de confiança [Carterette et al., 2006].

Nossa avaliação contou com 15 avaliadores para avaliar os pares página-livro. Os avaliadores utilizaram um sistema para avaliação de livros, desenvolvido exclusivamente

para este trabalho. Conforme a tela ilustrada na Figura 4.5, no topo do sistema há o endereço para a página original da notícia e, logo abaixo, a lista de recomendação de livros (desordenada). A lista de recomendação possui o título, a descrição dos livros e no título há um apontador para o anúncio na loja *online* da Amazon. Após entrar no sistema, o avaliador recebe as instruções e, daquela lista de recomendação de livros, assinala os livros como relevantes ou não, de acordo com a pergunta:

"Se você estivesse acessando esta página na Internet, quais livros você consideraria uma boa recomendação?"

Figura 4.5. Sistema de avaliação de livros.

Os experimentos foram avaliados com as métricas apresentadas na Seção 2.4. Para computar a precisão no ponto foram considerados apenas os livros que possuíam julgamento, ou seja, os cinco livros do topo do *ranking*. Já para o MAP e *infAP*, os dez livros do topo foram usados. Isto porque recomendação geralmente envolve poucos itens e, neste caso, precisão é mais importante que revocação.

Todos os resultados experimentais foram verificados usando o teste t de Student [Jain, 1991]. Para os resultados apresentados nas seções seguintes, o símbolo † indica um nível de confiança 95% e o símbolo ‡ indica um nível de confiança 99%.

4.3 Métodos de Recomendação

A Tabela 4.3 apresenta a lista de métodos de recomendação implementados baseados em nossas três estratégias: DESC (descritores de categoria), CLF (características de classificação) and CTF (filtro de categorias). O método BOW corresponde à abordagem simplista que usa pesos do *tf.idf*, como descrito na Seção 3.1. Ele desempenha a função de *baseline* em nossa avaliação.

Para a estratégia DESC, cinco medidas nos deram os seguintes métodos: *Document Frequency* (DESC-DF), *Kullback-Leibler divergence* (DESC-KLD), *Chi-Squared de Pearson* (DESC-CHI2), *Coeficiente de Dice* (DESC-DICE) e o método que combina as medidas (DESC-ALL). Na estratégia CLF, o método CLF-EC usa todo o conteúdo da página para a geração das características de classificação e o método CLF-SE inclui segmentos da página, de acordo com a abordagem *multi-resolution* [Gabrilovich & Markovitch, 2005] (veja Seção 3.3). Na estratégia CTF, o método CTF-A usa categorias associadas à página usando um classificador automático, enquanto o método CTF-M usa categorias manualmente associadas à página alvo.

Além dos métodos derivados de cada estratégia, a combinação de diferentes estratégias também foi experimentada, o que originou novos métodos híbridos.

Tabela 4.3. Métodos de recomendação

Abreviação	Método
BOW	Repositório de Palavras
DESC	Descritores de Categoria
DESC-CHI2	DESC com *Chi-Squared de Pearson*
DESC-KLD	DESC com *Kullback-Leiber divergence*
DESC-DICE	DESC com *Coeficiente de Dice*
DESC-DF	DESC com *Document Frequency*
DESC-ALL	DESC com combinação de medidas
CLF	Características de Classificação
CLF-EC	CLF usando toda a página
CLF-SE	CLF usando a página segmentada
CTF	Filtro de Categoria
CTF-A	CTF Automático
CTF-M	CTF Manual

4.4 Resultados

4.4.1 DESC - Descritores de Categoria

Influência dos Descritores de Categoria

Na Equação 3.3, a influência dos descritores de categoria nos métodos da estratégia DESC é ajustada pelo parâmetro α. A Figura 4.6 mostra o impacto no *infAP* pela variação de α entre 0 (apenas termos da página são usados) e 1 (apenas termos descritores de categoria são usados). Devido ao grande volume de experimentos, essa figura apresenta apenas quatro valores para α e um ajuste mais elaborado poderá ser realizado em trabalhos futuros. Neste experimento, foi considerado apenas o cenário em que a categoria é associada manualmente à página alvo.

O gráfico mostra que todos os métodos baseados em descritores de categoria tiveram seus melhores resultados quando $\alpha = 0{,}25$. Isso significa que descritores de categoria podem realmente melhorar a qualidade da recomendação, mas também é importante preservar a influência dos termos da página, os quais no fim das contas preservam seu assunto. Por essa razão adotamos $\alpha = 0{,}25$ nos próximos experimentos com descritores de categoria. Por outro lado, quando $\alpha = 1$, os termos mais gerais de DESC-DF prejudicam a qualidade da recomendação, mas DESC-ALL e DESC-KLD mantiveram o $infAP$ acima de 0.5.

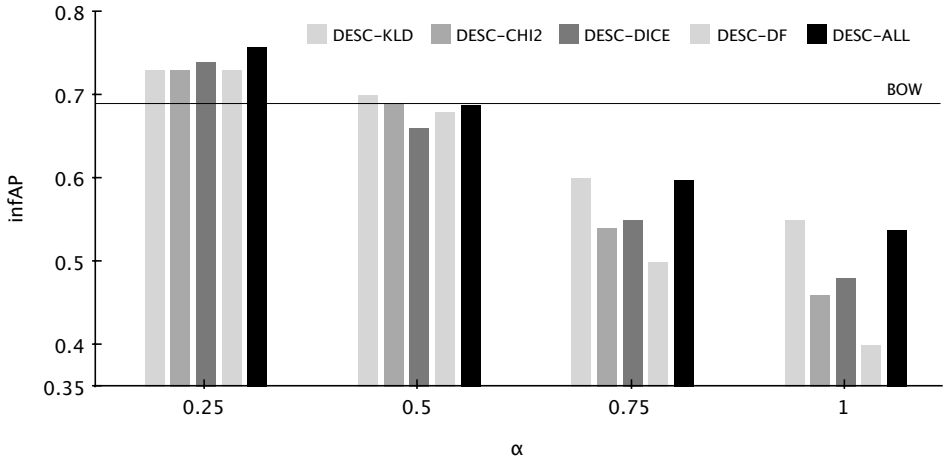

Figura 4.6. Comportamento dos métodos com descritores de categoria variando-se o valor de α. Com $\alpha = 0{,}25$, métodos com expansão superam o método sem expansão (BOW).

Comparando Medidas para Seleção de Descritores

Na Tabela 4.4, o método DESC-ALL, aquele que combina diferentes medidas, superou o método BOW com 99% de confiança (símbolo ‡) em todos valores de precisão e mostrou ser superior a todos os os métodos que utilizam apenas uma medida. O ganho do método DESC-ALL sobre o *baseline* BOW é de aproximadamente 20% em p@1. A mesma tabela apresenta os resultados para o método DESC-ALL considerando o cenário em que as categorias são atribuídas automaticamente à página. Esses métodos são chamados DESC-ALL-nA, onde n é o número de categorias associadas à página.

Vale observar que todos os métodos que usam associação manual de categoria foram superiores a todos os métodos que usam associação automática de categoria. Entre os métodos que usam associação automática, DESC-ALL-5A obteve os melhores resultados. Uma possível explicação é que: (1) usando mais de uma categoria permite que diversos conceitos sejam adicionados, os quais contribuem para uma rica representação; (2) eventuais erros de classificação, os quais classificadores automáticos estão propensos, podem ser prejudiciais na seleção dos descritores de categoria – no caso do DESC-ALL-1A, se a única categoria está incorreta, apenas descritores não relacionados podem ser adicionados, enquanto que no caso de DESC-ALL-10A, muitos desses descritores podem ser adicionados.

Tabela 4.4. Valores de precisão para a estratégia DESC.

	p@1	p@3	p@5	MAP	infAP
DESC-ALL	$0.79^{\ddagger}$	$0.72^{\ddagger}$	$0.72^{\ddagger}$	$0.80^{\ddagger}$	$0.76^{\ddagger}$
DESC-DICE	$0.78^{\dagger}$	$0.71^{\ddagger}$	$0.72^{\ddagger}$	$0.79^{\dagger}$	$0.75^{\ddagger}$
DESC-CHI2	$0.75^{\dagger}$	$0.70^{\ddagger}$	$0.70^{\ddagger}$	$0.78^{\ddagger}$	$0.74^{\ddagger}$
DESC-DF	0.73	$0.72^{\ddagger}$	$0.72^{\ddagger}$	$0.77^{\ddagger}$	$0.73^{\dagger}$
DESC-KLD	0.74	$0.69^{\ddagger}$	$0.70^{\ddagger}$	$0.77^{\dagger}$	$0.73^{\dagger}$
DESC-ALL-5A	0.70	$0.71^{\ddagger}$	$0.70^{\ddagger}$	0.76	0.72
DESC-ALL-10A	0.72	$0.70^{\ddagger}$	0.67	0.74	0.70
DESC-ALL-1A	0.69	$0.67^{\ddagger}$	$0.68^{\dagger}$	0.74	0.70
BOW	0.66	0.61	0.62	0.70	0.67

4.4.2 CLF - Características de Classificação

A Figura 4.7 mostra os resultados para os métodos CLF em comparação com o *baseline* BOW. Ambos os métodos melhoram os resultados da recomendação, mas para os primeiros 70% níveis de revocação o método CLF-EC (que usa todo o conteúdo da página) é melhor que o CLF-SE (que usa segmentos).

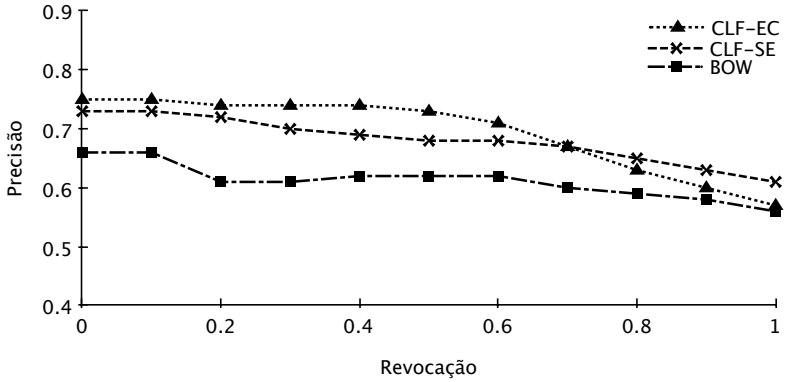

Figura 4.7. Comparação do *baseline* BOW com os métodos CLF-EC e CLF-SE. Ambos os métodos melhoram os resultados da recomendação, mas o método CLF-EC que usa todo o conteúdo da página apresentou melhores resultados que o método CLF-SE que usa a página segmentada.

Os valores de precisão apresentados na Tabela 4.5 confirmam que os métodos baseados na geração de características são mais eficientes que o método *baseline* BOW. Por exemplo, considerando a métrica p@3, os métodos CLF-EC e CLF-SE superam o *baseline* em aproximadamente 21% e 15%, respectivamente. Como pode ser observado nessa tabela, o método CLF-EC obteve o melhor desempenho em todas as métricas.

Tabela 4.5. Valores de precisão para a estratégia CLF.

	p@1	p@3	p@5	MAP	infAP
CLF-EC	0.75	$0.74^{\ddagger}$	$0.71^{\ddagger}$	$0.78^{\ddagger}$	$0.74^{\ddagger}$
CLF-SE	0.73	$0.70^{\ddagger}$	$0.68^{\ddagger}$	$0.76^{\ddagger}$	$0.72^{\ddagger}$
BOW	0.66	0.61	0.62	0.70	0.67

Ressaltamos que, para os métodos da estratégia CLF, o cenário de associação manual de categoria não é aplicado, visto que a estratégia CLF é baseada no uso de um classificador automático.

4.4.3 CTF - Filtro de Categorias

A Figura 4.8 apresenta o gráfico de precisão e revocação para os métodos baseados na estratégia de filtro de categorias. Três métodos usam um classificador automático para associar categorias à página alvo, denominados CTF-1A, CTF-5A e CTF-10A, os quais usam uma, cinco e dez categorias, respectivamente; e um método, CTF-M, usa associação manual de uma categoria à página alvo.

O gráfico mostra que CTF-5A e CTF-10A superam o método *baseline* BOW considerando todos os níveis de revocação e que o método CTF-M superou todos os métodos.

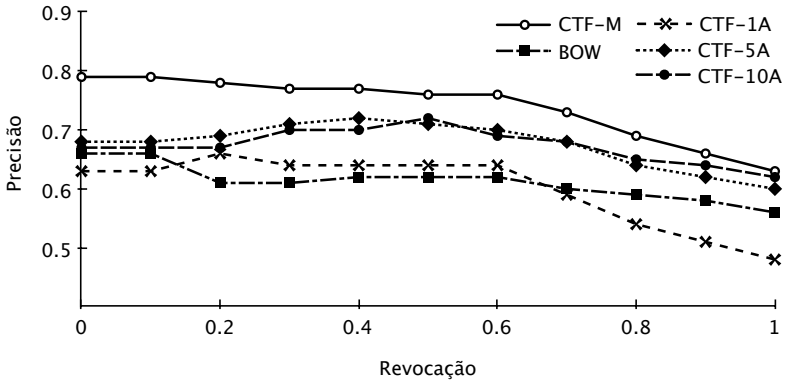

Figura 4.8. Impacto do filtro automático de categoria para 1, 5 e 10 categorias indicadas pelo classificador. Com apenas a primeira categoria os resultados foram inferiores ao *baseline* (BOW), e com 5 foi suficiente para superar o *baseline*. Como esperado, a classificação manual apresentou melhores resultados que a classificação automática.

O impacto positivo da intervenção humana também é evidente na Tabela 4.6, onde o método CTF-M superou o método *baseline* BOW em aproximadamente 26% na métrica p@3. Por outro lado, o melhor método que usa associação automática de categorias, CTF-5A, apresentou um ganho de aproximadamente 8% sobre o *baseline*.

Tabela 4.6. Valores de precisão para a estratégia CTF.

	p@1	p@3	p@5	MAP	infAP
CTF-M	$0.79^{\ddagger}$	$0.77^{\ddagger}$	$0.76^{\ddagger}$	$0.80^{\ddagger}$	$0.76^{\ddagger}$
CTF-5A	0.68	$0.71^{\ddagger}$	$0.71^{\ddagger}$	$0.76^{\dagger}$	$0.72^{\dagger}$
CTF-10A	0.67	$0.70^{\ddagger}$	$0.72^{\ddagger}$	0.75	0.71
CTF-1A	0.63	0.64	0.64	0.69	0.64
BOW	0.66	0.61	0.62	0.70	0.67

O gráfico na Figura 4.8 e os valores da Tabela 4.6, mostram que, para o cenário automático, usar uma categoria no filtro (CTF-1A) não produz bons resultados e que cinco categorias são suficientes para obter melhorias significativas. As razões para este comportamento são as mesmas discutidas previamente na Seção 4.4.1.

4.4.4 Comparação dos Métodos

A Tabela 4.7 apresenta a comparação dos melhores métodos para cada estratégia, considerando associação manual e automática das categorias. As linhas na tabela que correspondem aos métodos HYBRID serão brevemente discutidos na próxima seção.

Na tabela, vale primeiro observar que os métodos manuais são sempre melhores que os métodos automáticos. Nota-se que os dois métodos manuais alcançaram aproximadamente os mesmos resultados. Entre os métodos automáticos, CLF-EC supera DESC-ALL-5A e CTF-5A. Vale a pena notar que os métodos automáticos DESC-ALL-5A e CTF-5A, ambos usando 5 categorias, foram os melhores entre os métodos dentro de suas respectivas estratégias. A razão para isso foi examinada na Seção 4.4.1.

Finalmente, destacamos que CLF-EC parece ser uma opção muito interessante para se considerar, caso a associação manual não seja viável. Na verdade, é possível observar que, enquanto a diferença no ganho em p@1 favorece CTF-M em mais de 6%, a diferença em termos de $pavg$@5 é perto de 1%.

Tabela 4.7. Comparação dos melhores métodos de cada estratégia de recomendação.

	p@1	ganho*	p@3	ganho*	p@5	ganho*	pavg@3	ganho*	pavg@5	ganho*
HYBRID-M	$0.81^{\ddagger}$	22.73	$0.78^{\ddagger}$	27.87	$0.76^{\ddagger}$	22.58	$0.87^{\ddagger}$	19.18	$0.86^{\ddagger}$	19.44
CTF-M	$0.79^{\ddagger}$	19.70	$0.77^{\ddagger}$	26.23	$0.76^{\ddagger}$	22.58	$0.84^{\ddagger}$	15.07	$0.83^{\ddagger}$	15.28
DESC-ALL	$0.79^{\ddagger}$	19.70	$0.72^{\ddagger}$	18.03	$0.72^{\ddagger}$	16.13	$0.84^{\ddagger}$	15.07	$0.83^{\ddagger}$	15.28
CLF-EC	0.75	13.64	$0.74^{\ddagger}$	21.31	$0.71^{\ddagger}$	14.52	$0.82^{\dagger}$	12.33	$0.82^{\ddagger}$	13.89
DESC-ALL-5A	0.70	6.06	0.71	16.39	0.70	12.90	0.80	9.59	$0.79^{\dagger}$	9.72
CTF-5A	0.68	3.03	$0.71^{\ddagger}$	16.39	$0.71^{\ddagger}$	14.52	0.77	5.48	0.78	8.33
HYBRID-A	0.66	0.00	$0.72^{\ddagger}$	18.03	$0.70^{\dagger}$	12.90	0.78	6.85	0.77	6.94
BOW	0.66	-	0.61	-	0.62	-	0.73	-	0.72	-

* Ganhos em porcentagem (%).

4.4.5 Combinação dos Métodos

Até agora, os métodos estudados usaram apenas uma estratégia individualmente. Com o objetivo de melhorar a qualidade dos resultados, podemos também criar métodos híbridos combinando mais de uma estratégia.

Por exemplo, podemos expandir a representação da página com descritores de categoria (DESC) e usar filtragem por categoria (CTF) sobre os livros relacionados. A razão para essa combinação vem do fato de que descritores de categoria contribuem para melhorar a revocação, enquanto a filtragem por categoria ajuda a melhorar a precisão.

Baseado nessa idéia, implementamos métodos híbridos (HYBRID) que utilizam tanto os descritores quanto o filtro de categorias. No cenário de associação manual de

categoria, o HYBRID-M combina DESC-ALL e CTF-M. Já no cenário de associação automática, o método HYBRID-A combina DESC-ALL-5A e CTF-5A. Como mostrado na Tabela 4.7, HYBRID-M é o melhor de todos os métodos, superando também os métodos que combina.

Esse método melhora *pavg*@3 e *pavg*@5 em mais de 3% em comparação com o melhor método baseado em uma estratégia, CTF-M e DESC-ALL. Como resultado, os ganhos obtidos com esse método sobre o *baseline* (BOW) são 27% melhor na precisão média que os ganhos obtidos com o melhor método baseado em uma estratégia. Com respeito ao HYBRID-A, verificamos na Tabela 4.7 que a combinação de estratégias não foi tão benéfica como foi o caso do HYBRID-M.

4.4.6 Impacto da Taxonomia

As métricas usadas até agora sumarizam os resultados obtidos de várias páginas alvo em um único valor. Por exemplo, o *infAP* corresponde à média da precisão esperada (ver Seção 2.4.3) de todas as páginas usadas como alvo da recomendação. Por isso, para explorar uma perspectiva diferente dos resultados, apresentamos na Figura 4.9 o impacto de usar taxonomia para cada página alvo da coleção de notícias, isto é, os pontos no gráfico. Para isso, o melhor método, HYBRID-M, foi comparado com o método *baseline* (BOW) usando os valores de precisão esperada das páginas. Para efeito de análise, os pontos abaixo de 0,5 são interpretados como recomendação de baixa qualidade e os pontos acima de 0,5 recomendação de boa qualidade. O gráfico é dividido em quatro quadrantes, Q1, Q2, Q3 e Q4, nos quais pode-se observar:

Q1: melhorias em muitos casos em que a precisão era muito baixa ou até mesmo nula;
Q2: valores da precisão esperada já eram elevados (acima de 0,5) e, ainda assim, HYBRID-M apresentou melhorias;
Q3: não houve melhorias e nem deficiências;
Q4: pouquíssimos casos em que HYBRID-M prejudicou uma boa recomendação.

Adicionalmente, é importante notar em Q1 e Q2 que o método enriquecido com taxonomia manteve grande parte das páginas com precisão acima de 0,5.

Quando a precisão esperada foi zero no método BOW (Q1), isto acontece porque um mesmo termo contribui para recuperar livros em diferentes áreas. A palavra *giants* (gigante) na notícia sobre o time Giants da NFL, por exemplo, contribuiu para recuperar livros sobre gigantes da caverna, gigantes dos mares e outros. Nesse caso, o filtro de categorias mantém apenas os livros da categoria esporte e elimina os livros de categorias diferentes.

Por outro lado, a precisão esperada passou a ser zero no método HYBRID-M (Q4) quando (i) ocorreu erro de classificação ou (ii) a categoria associada foi muito geral. Ambos os casos contribuem para inclusão de descritores não relacionados ao tópico da página ou remover os livros de categorias coerentes. Por exemplo, uma notícia sobre o esporte boxe foi categorizada em *Individual Sports* e descritores como *horses*, *cycling* e *running* foram selecionados e prejudicaram a recomendação. Entretanto, foi constatado que, apesar da amostra da Amazon conter livros sobre boxe, a taxonomia não possuía a subcategoria *Boxing* filiada a *Individual Sports*. Recentemente a Amazon adicionou esse novo nó na hierarquia a atualização da taxonomia resolveria esse caso.

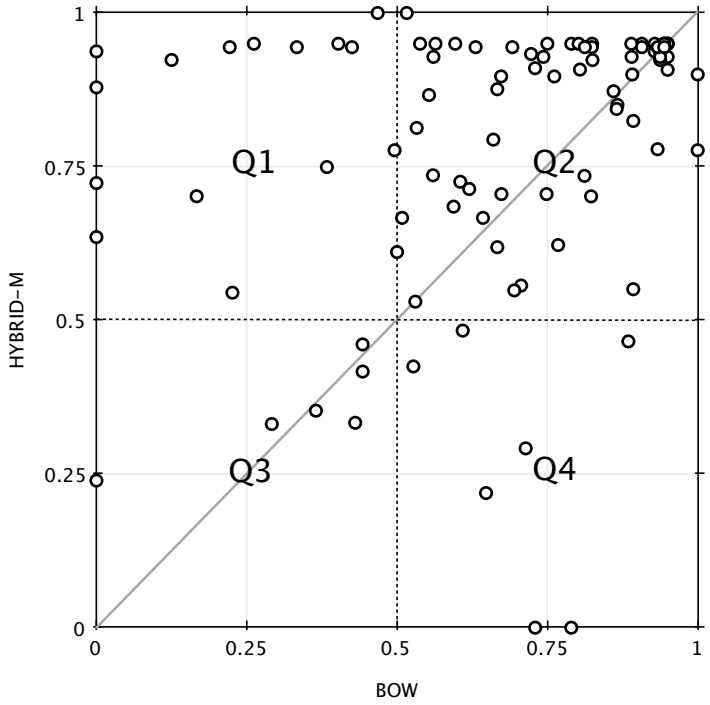

Figura 4.9. Um método baseado em taxonomia (HYBRID-M) *versus* um método baseado em conteúdo puro (BOW). Os pontos representam o valor da precisão esperada para cada página alvo.

Capítulo 5

Conclusões e Trabalhos Futuros

5.1 Conclusões

Este trabalho mostrou que o uso do conhecimento humano, embutido em uma taxonomia de domínio específico, pode melhorar significativamente a qualidade de sistemas de recomendação baseada em conteúdo. Três estratégias que aproveitam da informação embutida em taxonomias foram usadas em um arcabouço para sistemas de recomendação baseada em conteúdo. Duas dessas estratégias (o uso de descritores de categoria e características de classificação) foram usadas anteriormente em diferentes contextos. A terceira estratégia, a qual aplica uma filtragem por categoria na lista de recomendação, foi proposta neste trabalho. A primeira estratégia (descritores de categoria) favorece a revocação, enquanto a terceira estratégia (filtro de categorias) favorece a precisão. Portanto, faz total sentido combiná-las.

Foram implementados vários métodos de recomendação que aplicam essas estratégias individualmente e em combinação. Os métodos foram avaliados em um estudo de caso onde os itens a serem recomendados eram livros e o alvo da recomendação eram páginas de notícias. Foram também considerados dois cenários: (i) a página de notícias é manualmente associada à uma única categoria da taxonomia de livros e (ii) a notícia é automaticamente associada a uma ou múltiplas categorias usando um classificador baseado em centróides.

Os métodos que potencializam a importância da informação presente na classificação manual das páginas alvo foram melhores que os demais métodos. O ganho para $p@k$ varia de 16,13 a 27,87% e o ganho para $pavg@k$ varia de 15,07 a 19,44%, ambos acima do *baseline* repositório de palavras. Os ganhos foram obtidos com nível de confiança de 99%. Embora em algumas situações talvez seja custoso aplicar classificação manual, há outras situações onde isso não é. Por exemplo, um site de notícias que

também vende livros *online*, os autores das notícias poderiam associar uma categoria na taxonomia de livros enquanto criam as páginas de notícia, assim como eles fazem para notícias relacionadas.

Por outro lado, os métodos que fazem classificação automática são totalmente independentes dos usuários e, por isso, escalam melhor que aqueles que requerem classificação manual. Entretanto, classificação automática introduz ruído e esse é o motivo porque eles não obtiveram a mesma eficácia que a classificação manual. A boa notícia é que eles também melhoram consideravelmente a qualidade da lista de recomendação. O ganho para $p@k$ alcançou 21,31% e o ganho para $pavg@k$ alcançou 13,89%.

5.2 Trabalhos Futuros

Uma proposta para trabalhos futuros seria avaliar a recomendação dos livros em uma coleção mais geral da Web. Para o caso de uma aplicação real na Web, por questão de escala apenas o cenário de associação automática de categoria pode ser aplicado às páginas web e, então, aplicar as estratégias de recomendação que usam informação de taxonomia. Alternativamente, as estratégias também poderiam ser avaliadas para uma coleção mais específica, como páginas e artigos de medicina. Nesse caso, seria necessário uma nova taxonomia da área para executar os procedimentos inerentes a cada estratégia, por exemplo, geração de descritores. E, baseado nos resultados obtidos neste trabalho, espera-se obter uma melhoria considerável na recomendação dos artigos.

Existem novas abordagens que utilizam taxonomia que podem ser exploradas para melhorar a recomendação. Por exemplo, descritores de categoria podem ser usados para enriquecer a representação dos livros, não apenas das páginas.

Na estratégia que utiliza características de classificação, uma nova dimensão foi criada para combinar a informação de taxonomia com a página não enriquecida. Analogamente, podemos incluir outras dimensões. Seguindo Gabrilovich & Markovitch [2007], por exemplo, podemos gerar outras características com as entidades encontradas nos artigos da Wikipedia. Alternativamente, outras bases de conhecimento com informação de taxonomia podem ser usadas. Assim como Gabrilovich & Markovitch [2005], pode-se optar pela taxonomia do Open Directory, uma extensa taxonomia de páginas compilada por humanos. Em ambos os casos, as novas características seriam geradas tanto para as páginas alvo quanto para os livros a serem recomendados.

Em nosso estudo de caso, o cenário não envolvia informação sobre o perfil do usuário. Entretanto, esse problema pode ser estendido e incluir outros dados do usuário, tais como gosto, histórico de navegação e informação de amigos em uma rede social.

Essas informações representariam uma nova dimensão de características que, da mesma forma, seriam adicionadas por meio de uma combinação linear.

Os valores das constantes usadas neste trabalho foram adquiridas de outros trabalhos da literatura ou são aproximações dos resultados experimentais. Um estudo mais aprofundado pode ser feito utilizando aprendizado de máquina. Por exemplo, programação genética seria usada para obter os valores ótimos (ou aproximações) dessas constantes. Então, após ajustadas as constantes, espera-se uma recomendação de maior qualidade.

Anexo

Associação manual de categorias às páginas de notícias do The New York Times. De acordo com o conteúdo da notícia, o humano escolhia a categoria mais específica da hierarquia da taxonomia de livros da Amazon.com.

ID	URL
1	http://www.nytimes.com/2009/12/15/business/economy/15bank.html Categoria: Business_and_Investing>Finance
2	http://www.nytimes.com/2010/06/05/business/global/05honda.html Category: Business_and_Investing>Economics
3	http://www.nytimes.com/2009/12/15/business/global/15dubai.html Categoria: Business_and_Investing>Economics
4	http://www.nytimes.com/2010/06/04/business/economy/04fed.html Categoria: Business_and_Investing>Small_Business_and_Entrepreneurship
5	http://www.nytimes.com/2010/06/08/business/08markets.html Categoria: Business_and_Investing>International
6	http://www.nytimes.com/2010/06/02/business/global/02rates.html Categoria: Business_and_Investing>Economics
7	http://www.nytimes.com/2010/06/02/business/02credit.html Categoria: Business_and_Investing>Finance
8	http://www.nytimes.com/2009/12/12/business/12warrant.html Categoria: Business_and_Investing>Finance
9	http://boss.blogs.nytimes.com/2010/06/04/targeting-new-business/ Categoria: Business_and_Investing>Small_Business_and_Entrepreneurship
10	http://www.nytimes.com/2009/12/15/business/15auto.html Categoria: Business_and_Investing>Management_and_Leadership
11	http://www.nytimes.com/2009/12/14/technology/internet/14virus.html Categoria: Computers_and_Internet>Networking
12	http://www.nytimes.com/external/venturebeat/2010/06/07/ 07venturebeat-yahoo-hollows-itself-out-further-with-facebo-44972.html Categoria: Computers_and_Internet>Business_and_Culture
13	http://www.nytimes.com/2009/12/17/technology/personaltech/17dell.html Categoria: Computers_and_Internet>Hardware
14	http://www.nytimes.com/external/readwriteweb/2010/06/03/ 03readwriteweb-microsoft-connects-windows-live-essentials-62410.html Categoria: Computers_and_Internet>Software
15	http://www.nytimes.com/external/venturebeat/2010/06/04/ 04venturebeat-google-tests-out-twitter-ads-both-companies-83865.html Categoria: Computers_and_Internet>Business_and_Culture

ID	URL
16	http://www.nytimes.com/2010/06/03/technology/personaltech/03basics.html Categoria: Computers_and_Internet>Networking
17	http://www.nytimes.com/2010/06/01/technology/01loopt.html Categoria: Computers_and_Internet>Business_and_Culture
18	http://gadgetwise.blogs.nytimes.com/2010/03/12/sony-adds-motion-control-to-the-ps3/ Categoria: Entertainment>Puzzles_and_Games
19	http://www.nytimes.com/external/venturebeat/2010/06/04/ 04venturebeat-hackers-find-holes-in-sprints-new-4g-phone-48094.html Categoria: Computers_and_Internet>Networking
20	http://bits.blogs.nytimes.com/2009/12/10/microsoft-is-losing-fight-for-consumers-analyst-says/ Categoria: Computers_and_Internet>Programming>Software_Design>_Testing_and_Engineering
21	http://www.nytimes.com/2010/06/01/science/01obiguana.html Categoria: Science>Biological_Sciences>Animals
22	http://www.nytimes.com/2009/12/15/science/15obdino.html Categoria: Science>Biological_Sciences>Animals>Dinosaurs
23	http://www.nytimes.com/2010/06/02/us/02coral.html Categoria: Science>Biological_Sciences>Biology>Marine_Biology
24	http://www.nytimes.com/2009/12/13/science/earth/13savannah.html Categoria: Science>Physics>Nuclear_Physics
25	http://www.nytimes.com/2010/04/22/business/energy-environment/22NUKE.html Categoria: Science>Earth_Sciences
26	http://www.nytimes.com/2009/12/12/science/earth/12quake.html Categoria: Science>Earth_Sciences>Geology
27	http://www.nytimes.com/2009/12/11/science/earth/11basel.html Categoria: Science>Earth_Sciences>Geology
28	http://www.nytimes.com/2010/06/05/us/05pelican.html Categoria: Science>Biological_Sciences>Animals
29	http://www.nytimes.com/2010/02/05/science/05dino.html Categoria: Science>Biological_Sciences>Animals>Dinosaurs
30	http://www.nytimes.com/2009/12/10/science/10collide.html Categoria: Science>Physics>Nuclear_Physics
31	http://www.nytimes.com/2009/11/17/health/research/17risk.html Categoria: Health>_Mind_and_Body>Disorders_and_Diseases
32	http://well.blogs.nytimes.com/2010/06/02/the-voices-of-alzheimers/ Categoria: Health>_Mind_and_Body>Disorders_and_Diseases
33	http://www.nytimes.com/2009/11/17/health/research/17prog.html Categoria: Health>_Mind_and_Body>Personal_Health>Women's_Health
34	http://well.blogs.nytimes.com/2009/11/12/dan-barber-knows-vegetables/ Categoria: Cooking>_Food_and_Wine>Vegetables_and_Vegetarian
35	http://well.blogs.nytimes.com/2009/11/12/the-alternative-medicine-cabinet-marigolds-to-soothe-skin/ Categoria: Health>_Mind_and_Body>Beauty_and_Fashion
36	http://health.nytimes.com/ref/health/healthguide/esn-herpes-ess.html Categoria: Health>_Mind_and_Body>Disorders_and_Diseases
37	http://well.blogs.nytimes.com/2009/11/11/going-vegetarian-for-thanksgiving/ Categoria: Cooking>_Food_and_Wine>Vegetables_and_Vegetarian
38	http://health.nytimes.com/ref/health/healthguide/esn-lupus-ess.html Categoria: Health>_Mind_and_Body>Disorders_and_Diseases
39	http://www.nytimes.com/2010/06/01/health/research/01obesity.html Categoria: Health>_Mind_and_Body>Nutrition
40	http://www.nytimes.com/2009/11/12/health/nutrition/12best.html Categoria: Health>_Mind_and_Body>Personal_Health
41	http://www.nytimes.com/2009/11/13/sports/13pacquiao.html Categoria: Sports>Individual_Sports
42	http://www.nytimes.com/2009/11/12/sports/football/12jets.html Categoria: Sports>Football_(American)
43	http://www.nytimes.com/2009/11/13/sports/basketball/13dejuan.html Categoria: Sports>Basketball
44	http://www.nytimes.com/2010/06/05/sports/tennis/05tennis.html Categoria: Sports>Racket_Sports>Tennis
45	http://www.nytimes.com/2009/12/14/sports/golf/14woods.html Categoria: Sports>Golf

ID	URL
46	http://www.nytimes.com/2009/12/14/sports/football/14giants.html Categoria: Sports>Football_(American)
47	http://www.nytimes.com/2009/12/14/sports/football/14jets.html Categoria: Sports>Football_(American)
48	http://www.nytimes.com/2009/12/14/sports/basketball/14triangle.html Categoria: Sports>Basketball
49	http://www.nytimes.com/2009/12/14/sports/olympics/14skate.html Categoria: Sports>Winter_Sports
50	http://www.nytimes.com/2009/12/14/sports/soccer/14soccer.html Categoria: Sports>Soccer
51	http://www.nytimes.com/2010/02/08/arts/music/08chip.html Categoria: Entertainment>Music
52	http://www.nytimes.com/2009/12/14/movies/14box.html Categoria: Entertainment>Movies
53	http://www.nytimes.com/2010/06/05/arts/music/05june.html Categoria: Entertainment>Music
54	http://www.nytimes.com/2009/12/13/nyregion/13artsli.html Categoria: Arts_and_Photography>Performing_Arts>Dance
55	http://www.nytimes.com/2009/12/13/arts/dance/13feet.html Categoria: Arts_and_Photography>Performing_Arts>Dance
56	http://www.nytimes.com/2009/12/14/arts/dance/14brown.html Categoria: Arts_and_Photography>Performing_Arts>Dance
57	http://artsbeat.blogs.nytimes.com/2010/06/02/founders-of-gallery-owned-by-christies-to-step-down/ Categoria: Arts_and_Photography>Museums_and_Collections
58	http://www.nytimes.com/2010/06/04/arts/design/04klein.html Categoria: Arts_and_Photography>Painting
59	http://www.nytimes.com/2010/06/06/arts/music/06taylor.html Categoria: Entertainment>Music
60	http://www.nytimes.com/2009/08/09/arts/music/09pare.html Categoria: Entertainment>Music
61	http://travel.nytimes.com/2009/11/13/travel/escapes/13cycle.html Categoria: Travel>Specialty_Travel>Adventure
62	http://travel.nytimes.com/2006/12/17/travel/17hours.html Categoria: Travel>Europe>Italy
63	http://travel.nytimes.com/2009/11/15/travel/15nextstop.html Categoria: Travel>Caribbean
64	http://frugaltraveler.blogs.nytimes.com/2009/11/11/qa-with-jeanne-dee-the-nomadic-family-traveler/ Categoria: Travel
65	http://travel.nytimes.com/2010/06/06/travel/06heads.html Categoria: Travel>Europe>Italy
66	http://travel.nytimes.com/2005/11/13/travel/13going.html Categoria: Travel>United_States>States>Utah
67	http://www.nytimes.com/1996/11/10/magazine/to-machu-picchu-the-hard-way.html Categoria: Travel>Latin_America>South_America
68	http://travel.nytimes.com/2009/07/19/travel/19surface.html Categoria: Travel>Europe>Spain
69	http://travel.nytimes.com/2009/12/06/travel/06explorer.html Categoria: Travel>Caribbean
70	http://travel.nytimes.com/2009/06/14/travel/14surface.html Categoria: Travel>Europe>Spain

ID	URL
71	http://www.nytimes.com/2010/06/07/us/politics/07halter.html Categoria: Nonfiction>Politics
72	http://www.nytimes.com/2010/06/07/us/politics/07townhall.html Categoria: Nonfiction>Politics
73	http://www.nytimes.com/2010/06/04/us/politics/04pentagon.html Categoria: Nonfiction>Politics
74	http://www.nytimes.com/2009/12/14/us/politics/14mccain.html Categoria: Nonfiction>Politics
75	http://www.nytimes.com/2009/12/14/us/politics/14obama.html Categoria: Nonfiction>Politics
76	http://www.nytimes.com/2010/06/04/us/politics/04obama.html Categoria: Nonfiction>Politics
77	http://www.nytimes.com/2010/06/04/us/politics/04romanoff.html Categoria: Nonfiction>Politics
78	http://www.nytimes.com/2009/12/14/us/politics/14spendweb.html Categoria: Nonfiction>Politics
79	http://www.nytimes.com/2009/12/15/world/asia/15mullen.html Categoria: Nonfiction>Politics
80	http://www.nytimes.com/2007/12/20/us/politics/20clinton.html Categoria: Nonfiction>Politics
81	http://www.nytimes.com/2010/06/05/nyregion/05queensborough.html Categoria: Nonfiction>Education>College_and_University
82	http://www.nytimes.com/2009/11/12/education/12community.html Categoria: Nonfiction>Education>College_and_University
83	http://www.nytimes.com/2010/06/01/nyregion/01gifted.html Categoria: Nonfiction>Education
84	http://www.nytimes.com/2010/05/29/nyregion/29trinity.html Categoria: Nonfiction>Education>College_and_University
85	http://www.nytimes.com/2010/05/21/us/21students.html Categoria: Nonfiction>Education
86	http://www.nytimes.com/2010/05/19/nyregion/19duncan.html Categoria: Nonfiction>Education
87	http://www.nytimes.com/2009/11/11/us/11foodfight.html Categoria: Nonfiction>Education
88	http://www.nytimes.com/2010/06/01/nyregion/01schools.html Categoria: Nonfiction>Education
89	http://www.nytimes.com/2009/11/07/education/07mit.html Categoria: Nonfiction>Education
90	http://www.nytimes.com/2010/04/18/education/edlife/18philosophy-t.html Categoria: Nonfiction>Education
91	http://www.nytimes.com/2010/05/12/world/europe/12pope.html Categoria: Religion_and_Spirituality>Christianity>Catholicism
92	http://goal.blogs.nytimes.com/2010/05/09/chelsea-emphatically-wins-premier-league/ Categoria: Sports>Soccer
93	http://goal.blogs.nytimes.com/2010/05/10/us-world-cup-roster-to-be-announced-tuesday/ Categoria: Sports>Soccer
94	http://www.nytimes.com/2010/05/11/health/11case.html Categoria: Health>_Mind_and_Body>Disorders_and_Diseases
95	http://www.nytimes.com/2010/05/13/technology/personaltech/13smart.html Categoria: Computers_and_Internet>Hardware
96	http://www.nytimes.com/2010/05/11/arts/design/11restore.html Categoria: Arts_and_Photography>Painting
97	http://travel.nytimes.com/2009/10/11/travel/hotprague.html Categoria: Travel>Europe>Czech_Republic
98	http://www.nytimes.com/2010/05/12/business/global/12yuan.html Categoria: Business_and_Investing>International
99	http://www.nytimes.com/2010/05/10/education/10teacher.html Categoria: Nonfiction>Education
100	http://www.nytimes.com/2010/05/11/science/space/11neemo.html Categoria: Science>Astronomy

Referências Bibliográficas

Adomavicius, G. & 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.

Anagnostopoulos, A.; Broder, A. Z.; Gabrilovich, E.; Josifovski, V. & Riedel, L. (2007). Just-in-time Contextual Advertising. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, pp. 331–340.

Baeza-Yates, R. & Ribeiro-Neto, B. (2011). *Modern Information Retrieval: The Concepts and Technology behind Search*. Addison-Wesley Publishing Company, USA, 2 edição.

Balabanović, M. & Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM*, 40(3):72.

Broder, A.; Fontoura, M.; Josifovski, V. & Riedel, L. (2007). A Semantic Approach to Contextual Advertising. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 559–566.

Buckley, C. & Voorhees, E. (2004). Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 25–32.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Carpineto, C.; de Mori, R.; Romano, G. & Bigi, B. (2001). An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems*, 19(1):1–27.

Carpineto, C. & Romano, G. (1999). Towards more effective techniques for automatic query expansion. *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries*, pp. 126–141.

Carterette, B.; Allan, J. & Sitaraman, R. (2006). Minimal test collections for retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 268–275.

Carterette, B.; Kanoulas, E. & Yilmaz, E. (2010). Low cost evaluation in information retrieval. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 903.

Croft, B.; Metzler, D. & Strohman, T. (2009). *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA.

Doszkocs, T. (1978). AID, an associative interactive dictionary for online searching. *Online Review*, 2(2):163–173.

Gabrilovich, E. & Markovitch, S. (2005). Feature generation for text categorization using world knowledge. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, volume 19, pp. 1048–1053.

Gabrilovich, E. & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pp. 6–12.

Glover, E.; Flake, G.; Lawrence, S.; Kruger, A.; Pennock, D.; Birmingham, W. & Giles, C. (2001). Improving category specific web search by learning query modifications. In *Proceedings of the 2001 Symposium on Applications and the Internet*.

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

Han, E. & Karypis, G. (2000). Centroid-Based Document Classification: Analysis and Experimental Results. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 424–431.

Harman, D. (1992). *Relevance feedback and other query modification techniques*, pp. 241–263. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Hawking, D.; Craswell, N. & Thistlewaite, P. (1998). Overview of TREC-7 very large collection track. *NIST Special Publications*, pp. 93–106.

Jain, R. K. (1991). *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. Wiley/Interscience, New York, NY, USA.

Joachims, T. (1997). Text categorization with support vector machines: Learning with many relevant features. Technical Report 23, Universität Dortmund, LS VIII-Report.

Kautz, H.; Selman, B. & Shah, M. (1997). Referral web: combining social networks and collaborative filtering. *Communications of the ACM*, 40(3):63–65.

Konstan, J. A.; Miller, B. N.; Maltz, D.; Herlocker, J. L.; Gordon, L. R. & Riedl, J. (1997). Grouplens: applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3):77–87.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lacerda, A.; Cristo, M.; Gonçalves, M.; Fan, W.; Ziviani, N. & Ribeiro-Neto, B. (2006). Learning to Advertise. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 549–556.

Linden, G.; Smith, B. & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, pp. 76–80.

Losee, R. M. J. (1990). *The Science of Information: Measurement and Applications*. Academic Press Professional, Inc., San Diego, USA.

Mitchell, T. (1997). Machine learning. WCB. *Mac Graw Hill*.

Mooney, R. J. & Roy, L. (2000). Content-based book recommending using learning for text categorization. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, pp. 195–204.

Oyama, S.; Kokubo, T.; Ishida, T.; Yamada, T. & Kitamura, Y. (2001). Keyword spices: a new method for building domain-specific web search engines. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pp. 1457--1463.

Pahlevi, S. & Kitagawa, H. (2005). Conveying taxonomy context for topic-focused web search: Research articles. *Journal of the American Society for Information Science and Technology*, 56(2):173–188.

Pearson's, K. (1900). *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling.* Philos. Magazine.

Perugini, S.; Gonçalves, M. A. & Fox, E. A. (2004). Recommender systems research: A connection-centric survey. *Journal of Intelligent Information Systems*, 23(2):107–143.

Polikar, R. (2006). Ensemble based systems in decision making. *Circuits and Systems Magazine, IEEE*, 6(3):21–45.

Reategui, E. & Cazella, S. (2005). Sistemas de recomendaçao. In *XXV Congresso da Sociedade Brasileira de Computação*, pp. 306–348.

Resnick, P.; Iacovou, N.; Suchak, M.; Bergstorm, P. & Riedl, J. (1994). Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, pp. 175–186, Chapel Hill, North Carolina.

Resnick, P. & Varian, H. R. (1997). Recommender Systems. *Communications of the ACM*, 40(3):56–58.

Ribeiro-Neto, B.; Cristo, M.; Golgher, P. & Silva de Moura, E. (2005). Impedance Coupling in Content-targeted Advertising. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 496–503.

Robertson, S.; Walker, S. & Jones, S. (1995). Okapi at TREC-3.

Robertson, S. E. (1990). On term selection for query expansion. *Journal of Documentation*, 46(4):359–364.

Rocchio, J. et al. (1971). Relevance feedback in information retrieval. *The SMART retrieval system: experiments in automatic document processing*, pp. 313–323.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1):1–39.

Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval* 1. *Information Processing & Management*, 24(5):513–523.

Schafer, J.; Konstan, J. & Riedl, J. (2001). E-commerce Recommendation Applications. *Data Mining and Knowledge Discovery*, 5(1):115–153.

Schafer, J. B.; Frankowski, D.; Herlocker, J. & Sen, S. (2007). Collaborative Filtering Recommender Systems. *The Adaptive Web: Methods and Strategies of Web Personalization*, pp. 291–324.

Srinivasan, P. (1996). Query expansion and MEDLINE. *Information Processing and Management*, 32(4):431–443.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth.

Voorhees, E. & Harman, D. (1998). Overview of the Sixth Text REtrieval Conference (TREC-6). In *Information technology: the Sixth Text REtrieval Conference*, p. 1. US Dept. of Commerce, Technology Administration, National Institute of Standards and Technology.

Voorhees, E. & Harman, D. (1999). Overview of the seventh text retrieval conference (TREC-7). In *Proceedings of the Seventh Text REtrieval Conference*, pp. 1–23.

Weideman, M. & Haig-Smith, T. (2002). An investigation into search engines as a form of targeted advert delivery. In *Proceedings of the 2002 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on Enablement through Technology*, p. 258. South African Institute for Computer Scientists and Information Technologists.

Witten, I. H.; Moffat, A. & Bell, T. C. (1999). *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann, San Francisco, CA, 2. edição.

Yilmaz, E. & Aslam, J. (2006). Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp. 102–111.

Ziegler, C.-N.; Lausen, G. & Konstan, J. A. (2008). On exploiting classification taxonomies in recommender systems. *AI Communications*, 21(2-3):97–125.

Ziegler, C.-N.; Lausen, G. & Lars, S.-T. (2004a). Taxonomy-driven computation of product recommendations. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, pp. 406–415.

Ziegler, C.-N.; McNee, S. M.; Konstan, J. A. & Lausen, G. (2005). Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, p. 32.

Ziegler, C.-N.; Schmidt-Thieme, L. & Lausen, G. (2004b). Exploiting semantic product descriptions for recommender systems. In *Proceedings of the 2nd ACM SIGIR Semantic Web and Information Retrieval Workshop*.

Zobel, J. & Moffat, A. (2006). Inverted files for text search engines. *ACM Computing Surveys*, 38(2):6.