Universidade Federal do Amazonas
Instituto de Computação
Programa de Pós-Graduação em Informática

GUILHERME MONTEIRO DA SILVA

Veiculação de Publicidade em Redes Sociais Utilizando Perfis de Usuários

Manaus
2014

Guilherme Monteiro da Silva

Veiculação de Publicidade em Redes Sociais Utilizando Perfis de Usuários

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas, como parte dos requisitos necessários para a obtenção do título de Mestre em Informática.

Orientador: Prof. Dr. Edleno Silva de Moura

Manaus

2014

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Silva, Guilherme Monteiro da
S586v Veiculação de publicidade em redes sociais utilizando perfis de usuários / Guilherme Monteiro da Silva. 2014
89 f.: il.; 31 cm.

Orientador: Marco Antônio Pinheiro de Cristo
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. redes sociais. 2. Wikipedia. 3. recomendação. 4. propaganda. 5. aprendizado de máquina. I. Cristo, Marco Antônio Pinheiro de II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"Veiculação de Publicidade em Redes Sociais Utilizando Perfis de Usuários"

GUILHERME MONTEIRO DA SILVA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Professores:

PROF. EDLENO SILVA DE MOURA – PRESIDENTE

PROF. DAVID BRAGA FERNANDES DE OLIVEIRA – MEMBRO

PROF. MARCO ANTÔNIO PINHEIRO DE CRISTO – MEMBRO

PROF. THIERSON COUTO ROSA – MEMBRO

Manaus, 31 de julho de 2014.

À minha mãe Iracema, a melhor mãe do mundo.

Ao meu pai Fernando, o qual me espelho pessoal e profissionalmente.

Agradecimentos

Em primeiro lugar aos meus pais e irmão que me concederam uma vida de educação, razão, respeito e apoio financeiro, emocional e moral.

À toda minha família, pelo apoio e incentivo em todas as horas.

Ao meu orientador Edleno Moura e meu "co-orientador" Klessius Berlt, pela oportunidade, conhecimento e sabedoria passadas, ajudando no meu crescimento profissional.

Aos meus amigos mestres Bruno Campos, Carlos Alessandro, Daniel Bittencourt, Davi Viana, Diego Froner, Felipe Hummel, Felipe Oliveira, Júlio Silva, Kaio Wagner, Maísa Vidal, Onilton Maciel, Petrina Kimura, Rafael Sousa, Rodrigo Braga, William Freitas, que além de grandes profissionais, são grandes amigos que levarei para o resto da minha vida.

Ao meu amigo Dr. Micael Granja, pela amizade e apoio.

À CAPES, pelo apoio financeiro.

À todos aqueles que tiveram contribuição direta ou indireta para o trabalho.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

As Redes Sociais estão entre os serviços mais utilizados na Web. Diariamente, milhões de usuários inserem informações pessoais em sites como Orkut e Facebook. Esse tipo de informação tem uma grande importância, pois o usuário está falando de si mesmo, representando um dado pessoal explícito. Nesta dissertação, é proposto um modelo de veiculação de publicidade em Redes Sociais, utilizando as informações contidas nos perfis de seus usuários. Para tal modelo, propusemos uma abordagem em dois passos: primeiro, uma abordagem de identificação de entidades utilizando os artigos da Wikipedia como fonte para filtrar e expandir a informação contida nos perfis; e então, utilizamos aprendizado de máquina para reformular o ranking das propagandas recomendadas.

Palavras-Chave: redes sociais, Wikipedia, recomendação, propaganda, aprendizado de máquina.

Abstract

Social Networks are among the most used services on the Web. Every day, millions of users insert personal information on websites such as Orkut and Facebook. Such information is of great importance, because the user is talking about himself, representing an explicit personal data. In this dissertation, we propose a model for advertising in social networks, using information contained in the profiles of its users. For this model, we proposed a two step approach: first, an approach for identifying entities using Wikipedia articles as source to filter and expand the information contained in the profiles; and then use machine learning to reshape the ranking of recommended advertisements.

KEYWORDS: social networks, Wikipedia, recommendation, advertisement, machine learning.

Capítulo 1

Introdução

Diariamente, nós somos apresentados a novos serviços que aumentam cada vez mais a interação entre usuários, principalmente nas redes sociais, blogs, etc. Esta interação e seus meios são rotulados de Web Social, a qual está crescendo ao passo que seu conteúdo é facilmente gerado pelos usuários. Este advento tem mudado o comportamento das pessoas em diversas formas. Por exemplo, imagine que deseje comprar um carro, com a Web Social você pode obter todo tipo de informação sobre o carro como fotos, videos ou até opiniões de usuários que possuem o carro de seu desejo. Esse tipo de informação está disponível em muitos sites, incluindo as redes sociais.

Sites como Facebook, Orkut e Twitter têm milhões de usuários, e representam um subconjunto muito importante da Web devido a sua popularidade. São locais onde há uma grande produção e, principalmente, consumo de conteúdo dos mais variados tópicos, recebendo muita atenção dos meios de comunicação. E devido a essa popularidade, muitos investimentos têm sido feitos nas redes sociais, criando um mercado bastante rentável na Web.

Nas redes sociais, encontra-se uma fonte rica de informação, principalmente em relação aos seus usuários. Estes inscrevem-se em comunidades (Orkut) ou explicita-

mente mostram seus gostos (Facebook), que representam nichos de interesses como bandas, jogos, atividades, etc. Normalmente, os usuários de redes sociais preenchem seus perfis com informações pessoais, o que é interessante, pois a descrição é feita pelas próprias palavras do usuário.

Esse fato possibilita a construção de aplicações que usam essas informações como fonte de conhecimento do usuário. Uma possível utilidade para essas informações é a propaganda baseada em contexto, onde a aplicação utiliza as informações dos usuários para conectar um usuário a um produto, de modo a reduzir o esforço na compra e maximizar a qualidade da experiência de compra do cliente, aumentando as chances de compra. Uma forma de viabilização é com o uso de técnicas de filtragem e recomendação baseada em conteúdo.

Nessa dissertação, focamos no problema de propaganda baseada em conteúdo, onde o conteúdo em questão são os dados de perfil dos usuários. Em outras palavras, nós utilizamos os dados contidos nos perfis dos usuários de redes sociais, e extraímos toda a informação possível, para realizar a recomendação de propagandas relevantes aos usuários. No entanto, devido a algumas características comuns em perfis de redes sociais, como uma quantidade de dados variáveis, ou seja, nem todos os usuários preenchem todos os campos de seu perfil. Além disso, por questões de privacidade, alguns destes campos não são acessíveis, pois o usuário seleciona quem pode ver ou não tal informação de seu perfil. Visto isso, decidimos utilizar outras abordagens de modo que perfis, seja com poucas ou muitas informações disponíveis, possam ser utilizados como fontes para recomendação de propagandas.

Primeiramente, a partir do dado coletado do perfil, utilizamos informações da Wikipedia como fonte para expandi-los. Deste modo, pudemos aumentar a chance de recomendarmos propagandas relevantes. E, em segundo, utilizamos técnicas de aprendizado de máquina para aperfeiçoar o ranking das propagandas recomendadas. A Wikipedia foi a base de dados escolhida devido a sua estrutura de organização de

artigos, por ser uma fonte de informação vasta e de boa qualidade [9], para os fins os quais serão usados neste trabalho. E, para fins de comparação, utilizamos o modelo vetorial como baseline para a recomendação e comparação com estas abordagens citadas acima.

Capítulo 2

Trabalhos relacionados

Embora a publicidade em redes sociais seja uma prática comum, como pode ser visto na parte superior da Figura 2.1, nenhum trabalho foi encontrado na literatura que proponha um método para selecionar anúncios com base no perfil do usuário. Isso ocorre, possivelmente, porque há várias questões de privacidade relacionadas com a disponibilidade dos dados do perfil do usuário. Este fato dificulta muito o amplo uso dessas informações dos usuários para a criação de quaisquer ferramentas que as utilize.

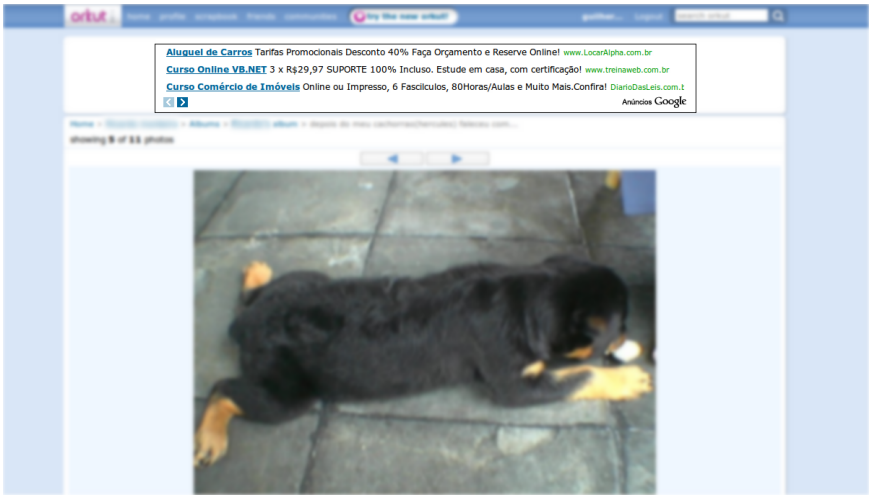

Figura 2.1: Publicidade no Orkut

Nesta dissertação, utilizou-se esses dados para estudar soluções para o problema de propaganda contextualizada. Em poucas palavras, favorecer-se dos dados do usuário para recomendar propagandas mais relevantes. Para isso, é importante realizar associações relevantes entre as propagandas e o usuário [15], visto que propagandas relevantes têm uma probabilidade maior de serem clicadas que as irrelevantes [7].

Para atingir tais objetivos, visando a melhoria na relevância das propagandas recomendadas aos usuários, existem na literatura algumas abordagens realizadas com sucesso, como o casamento de palavras-chave. Em propagandas, essas palavras-chave são palavras contidas no anúncio que melhor representam o produto ou o serviço anunciado. Realizar a seleção dessas palavras e como combiná-las a um contexto já configuram uma ampla área de pesquisa. Alguns trabalhos mostram como lidar com propriedades de propagandas para aumentar a sua relevância, utilizando casamento de palavras-chave, mostrando que características das palavras-chave selecionadas como natureza e tamanho têm impacto sobre a probabilidade de um anúncio ser clicado [14].

Na literatura atual, é comum a utilização de técnicas de aprendizado de máquina em complemento às técnicas citadas acima. Estas são utilizadas de modo a aprimorar a relevância de respostas de sistemas de busca, recomendação, entre outros. Por exemplo, estudos com o objetivo de melhorar o ranking de sistemas de recuperação de informação demonstraram bons resultados em [11] e [13]. No caso de ranking de propagandas, a parte de aprendizagem é feita através da seleção de características que mais determinam a relevância. Por exemplo, ao recomendarmos um conjunto de propagandas a um usuário, o mesmo as avaliará de acordo com sua relevância, então características como tamanho, similaridade e frequência de termos serão ponderadas de acordo com a relevância da propaganda. Esta ponderação será levada em conta para recomendação de novas propagandas.

Dentre as técnicas de aprendizado de máquina, uma que mostrou bons resultados na literatura foi o *Support Vector Machine* (SVM) [5]. O SVM é um método de classificação, mas que também pode ser utilizado como técnica de geração de modelo de ranking, como apresentado em [4]. Neste trabalho, o SVM foi utilizado para auxiliar na recomendação de propagandas mais relevantes para os usuários de redes sociais. Como o foco do trabalho desenvolvido nesta dissertação não é comparar formas de aprendizagem de máquina diferentes, e sim, propor um modelo para recomendação de produtos e propagandas a perfis de redes sociais, o escopo desta parte do trabalho foi limitado à utilização, apenas, desta técnica de geração de modelo de ranking com o SVM.

Indo além das informações contidas nos perfis, existem alguns estudos relacionados à publicidade em redes sociais, como o estudo de estratégias de preço para marketing viral em redes socais [1]. Tal método é mais focado no conceito de contágio social e como isso pode influenciar na venda de certos produtos. A computação de uma estratégia ótima, a qual maximiza o lucro esperado, é NP-Difícil. Isso adiciona uma complexidade que difere do objetivo desta dissertação.

Há também formas de se extrair informações de redes sociais inferidas de acordo com o comportamento de usuários em alguns sites, com o objetivo de seleção de uma boa audiência para propaganda de uma marca [16]. Estas informações vêm de dados sobre visitações e acessos ao site. Foram feitas medições de proximidade em rede, mostrando a afinidade de marcas para certas audiências.

Um estudo feito em [12], investigou os potenciais de veiculação de propagandas em redes sociais. Neste estudo, três problemas principais foram levantados sobre o tema: como veicular propagandas baseadas em relações e interações em redes sociais; como modelar um usuário de rede social de modo a representar seus interesses e necessidades; e como avaliar a efetividade desse sistema de propagandas. Para fins de comparação, o estudo fez um paralelo entre recomendação de notícias do Facebook

e quais destas lições poderiam ser utilizadas para recomendação de propagandas.

Um dos grandes problemas relacionados às informações vindas de redes sociais é sobre a privacidade destes dados. No estudo feito por [2], é elaborada uma abordagem para construção de perfis baseada em informações contidas no lado do cliente, como cookie e armazenamento local do navegador, visando aumentar a eficiência na veiculação de propagandas personalizadas. Mantendo os dados de usuário no lado do cliente é uma forma de suprir as necessidades do sistema de recomendação personalizada, mas sem sofrer muitas consequências dos critérios de privacidade e acesso aos dados do usuário. O trabalho realiza um estudo comparativo entre a construção de perfis nos lados de cliente e servidor.

Capítulo 3

Fundamentos

Neste capítulo, são explicados alguns conceitos necessários para o entendimento do modelo de veiculação de propagandas proposto.

3.1 Redes Sociais

O conceito de rede social é bastante difundido hoje na Web, apesar de possuir uma definição imprecisa. Isso ocorre, pois vários sites diferentes como o Twitter e o Orkut, por exemplo, são considerados redes sociais. O Twitter é uma plataforma de microblogs, onde os usuários interagem através de uma rede social. Os usuários escrevem textos curtos que são compartilhados com outros usuários conectados ao seu perfil. Já no Orkut, a interação do usuário é feita através de recados ou mensagens em comunidades.

Mesmo com algumas diferenças, essas redes sociais possuem características comuns como a interação entre usuários e um perfil descritivo do usuário. Guardadas as devidas proporções, um perfil de usuário contém informações pessoais que o descreve de alguma forma. No Orkut, por exemplo, o perfil do usuário é dividido em campos onde são inseridas suas informações, como pode ser visto na Figura 3.1.

Figura 3.1: Perfil do Orkut

Para esta dissertação, o conceito de Redes Sociais que se deve pensar é o de sites que usam perfis assim como o Orkut e o Facebook. Perfis estes que são definidos por um conjunto de campos preenchidos pelo dono do perfil. As relações entre perfis não foram explorada nesse trabalho, apenas o conteúdo do perfil do usuário da Rede Social.

3.1.1 Características no Perfil

Um perfil do Orkut tem um total de 81 campos de categorias diversas, que variam desde endereços de diversos serviços de e-mails até campos para texto livre sobre descrição pessoal, além de campos para definir gostos musicais, atividades, livros, etc. Alguns desses campos são preenchidos com texto livre, enquanto em outros seleciona-se o seu conteúdo dentre algumas alternativas. Estas são informações ricas, pois já estão classificadas em diferentes campos e foram descritas pelo próprio usuário.

Ao se pensar em propagandas para perfis, podemos ter uma intuição de que alguns campos fornecem melhores informações que outros para a tarefa de recomendação. Campos como músicas e livros, devem ser preenchidos com informações intimamente ligadas a produtos. Enquanto em campos como CEP e e-mail, uma associação com produtos pode não ser tão direta.

Desse ponto, é possível formar uma base de conhecimento sobre o usuário. Conhecimento o qual pode ser utilizado para realizar recomendações, por exemplo.

Uma observação interessante que podemos fazer é a de que, ao realizar recomendação baseada no conteúdo de perfis de redes sociais, a validação do método é realizada pelo próprio usuário, que é dono do perfil, o que traz confiabilidade.

3.2 Modelo Vetorial

Tendo propagandas e características dos perfis à disposição, é necessário uma forma de realizar uma filtragem para definir quais propagandas devem ser recomendadas para cada perfil. Uma forma de se realizar recomendação, é utilizando o modelo vetorial. Para os experimentos feitos neste trabalho, nós utilizamos o Apache Lucene[1], uma biblioteca de código aberto que fornece um sistema de recuperação de informação, incluindo indexação e busca.

3.2.1 Representações

No modelo vetorial [17], tem-se uma base de documentos textuais, sobre os quais desejamos realizar consultas. No caso do modelo apresentado por esta dissertação, os documentos são representados pelos anúncios e as consultas são os campos dos perfis das redes sociais. Como resultado, são recuperados documentos (propagandas) relevantes, em relação à consulta feita. E para implementar tal modelo, é necessário

[1] http://lucene.apache.org/

que algumas representações sejam feitas.

Primeiramente, os documentos e consultas são mapeados para vetores da seguinte forma:

$D_i = (d_{i1}, d_{i2}, d_{i3}, ..., d_{it})$

Onde D_i representa um documento ou consulta na forma de um vetor de t dimensões. Cada dimensão representa um termo do vocabulário da coleção de documentos, mais precisamente, uma dimensão d_{ij} representa o peso de um j-ésimo termo. Neste caso, a noção de termo pode significar apenas o radical das palavras ou expressões contendo mais de uma palavra, por exemplo. E para cada termo é atribuído um peso que, no caso do Lucene, é calculado o valor de TF-IDF, que representa o quão importante um termo é para um documento dentro de uma coleção. Para calcular o peso de d_{ij} fazemos:

$d_{ij} = (\text{TF-IDF})_{ij}$

$(\text{TF-IDF})_{ij} = TF_{ij} \times IDF_i$

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{3.1}$$

$$IDF_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \tag{3.2}$$

Na Equação 3.1, n_{ij} refere-se ao número de ocorrências do termo t_i no documento D_j e $\sum_k n_{kj}$ ao somatório do número de ocorrências de todos os termos no documento D_j. Na Equação 3.2, $|D|$ refere-se ao número total de documentos da coleção e $|\{d : t_i \in d\}|$ ao número de documentos que possuem t_i na coleção.

3.2.2 Similaridade

Com as consultas e documentos representados por vetores, possibilita-se a aplicação de operações sobre os mesmos. E para o modelo vetorial, uma operação muito utili-

zada é a de similaridade vetorial, que pode ser obtida pelo cálculo do cosseno entre os vetores. Por exemplo, na figura 3.2[2], podemos realizar a similaridade da consulta q com o documento d da seguinte forma:

$$sim(q,d) = \cos\theta = \frac{q \cdot d}{\|q\|\|d\|} \tag{3.3}$$

$$\|v\| = \sqrt{\sum_{i=1}^{n} v_i^2} \tag{3.4}$$

Na Equação 3.3, o numerador representa o produto interno dos vetores de consulta e documento, calculando-se o produto entre a norma dos mesmos vetores. A norma é definida na Equação 3.4.

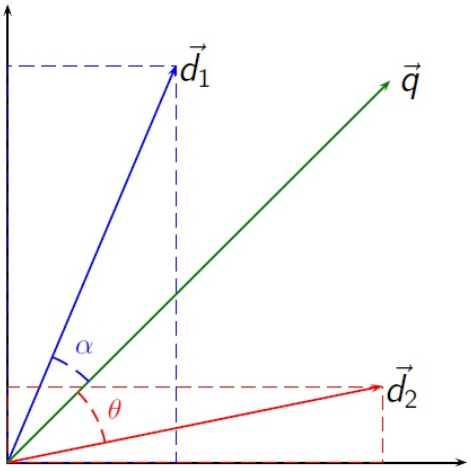

Figura 3.2: Espaço vetorial com documentos (d_1 e d_2) e consulta (q)

3.2.3 Indexação

Um processo importante no modelo vetorial é o de indexação. Neste passo, é criado um índice com todos os termos contidos na coleção de documentos, juntamente com alguns dados sobre estes termos como frequência e localização. Com

[2]http://en.wikipedia.org/wiki/Vector_space_model

essas informações, é possível realizar os cálculos de TF-IDF descritos acima.

O índice dos termos é organizado como uma estrutura de dados, como pode ser visto na Tabela 3.2.3:

Termo	(Documento, Frequência)
a	$(d_1, 3), (d_2, 9), (d_3, 4), (d_4, 6)$
amar	$(d_3, 9), (d_5, 5)$
bola	$(d_2, 8), (d_3, 4), (d_4, 1)$
casa	$(d_5, 7)$
zebra	$(d_1, 2)$

No exemplo acima, o termo *zebra* ocorre duas vezes no documento d_1, enquanto o termo *amar* ocorre nove vezes em d_3 e cinco em d_5, assim por diante. Desta forma, o cálculo e análise estatística sobre termos e documentos podem ser realizados de forma mais simplificada.

3.2.4 Processador de Consultas

Para concluir o processo de busca, é necessário um módulo que receba a consulta feita, analise o índice de termos e forneça um ranking com os documentos mais relevantes em relação à consulta feita. Esta é a função do processador de consultas.

No Lucene, o ranking de resposta pode ser feito através da ordenação dos *scores* dados aos documentos pelo processador de consultas. Isto é, a consulta é feita, então o processador de consultas compara a similaridade desta consulta com todos os documentos da base, baseando-se no índice. Assim, é possível calcular um valor, utilizando a Equação 3.5[3]:

[3]http://lucene.apache.org/java/3_0_3/api/core/org/apache/lucene/search/Similarity.html

$$score(q,d) = coord(q,d) \cdot queryNorm(q) \cdot \sum_{t \in q} (tf(t \in d) \cdot idf(t)^2 \cdot t.getBoost() \cdot norma(t,d)) \tag{3.5}$$

Onde $coord(q,d)$ é um score baseado em quantos termos da consulta q são encontrados no documento d. A função $queryNorm$ é um fator de normalização utilizado para tornar os scores comparáveis entre as consultas. O método $t.getBoost()$ é um aprimoramento no tempo de busca de um termo t na consulta q, especificado no texto da consulta. A função $norma(t,d)$ encapsula fatores de aprimoramento e tamanho (compressão) em tempo de indexação.

3.3 Wikipedia como Fonte de Entidades

Mesmo com a informação contida no perfil, é possível que essa informação não seja o suficiente para que sirva como base para recomendação. Nesta dissertação, optamos por extrair das redes sociais apenas os dados contidos no perfil do usuário. E em complemento, precisar-se-ia de uma fonte extra de informações para complementar as informações contidas no perfil.

Uma opção para se obter informações extras é a Wikipedia, uma grande fonte de informação sobre tópicos diversos. Estes assuntos estão divididos em conjuntos de artigos. Apesar de muitas vezes estarem incompletos ou imprecisos, a forma como estão organizados, categorias, links, etc, é um ponto de partida para obtermos assuntos relacionados. Como convenção, a partir de agora vamos nos referir aos artigos da Wikipedia como entidades.

Analisando alguns perfis de redes sociais, notamos certos fatos. Por exemplo, no campo “livros” do perfil, as pessoas devem preencher com informações relacionadas a livros como autores, títulos de livros, personagens, etc. Mas essa informação sozinha

pode não ser útil. Por outro lado, se pudermos associar esse tipo de informação com uma entidade da Wikipedia, podemos obter informações adicionais.

Digamos que o usuário diz em seu perfil que gosta de "Senhor dos Anéis" no campo de filmes. Então, analisando os links para outras entidades (Outlinks) que podem ser encontradas no artigo referente a "Senhor dos Anéis", temos uma relação como mostrada na Figura 3.3.

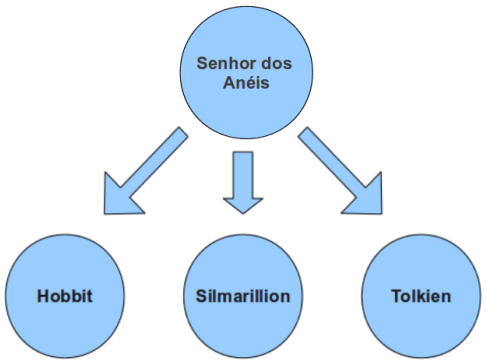

Figura 3.3: Outlinks do Artigo

Esse tipo de informação pode ser facilmente obtida. No entanto, as entidades relacionadas têm uma importância diferente entre si. Por exemplo, ainda na entidade de "Senhor dos Anéis", é razoável pensar em entidades relacionadas como "Hobbit", "Silmarillion" e "Tolkien", pois são relacionadas ao livro. Mas também é possível encontrar links para outras que mantêm uma relação mais fraca, como "África do Sul" (país de origem do autor) ou "BBC" (emissora de telecomunicação). E para solucionar esse problema, propusemos a seguinte fórmula:

A = Senhor dos Anéis

B = Hobbit

$n(A)$ = número de outlinks de A

$n(B)$ = número de outlinks de B

$n(A \cap B)$ = número de outlinks comuns entre A e B

$$Mutualidade(A,B) = \begin{cases} 1, & \text{se A possui pelo menos um outlink para B e vice-versa} \\ 0, & \text{senão} \end{cases} \tag{3.6}$$

$$Similaridade(A,B) = \frac{n(A \cap B)}{n(A) + n(B)} + Mutualidade(A,B) \tag{3.7}$$

Na Equação 3.7, insere-se um valor que chamamos de "mutualidade", o qual representa a reciprocidade entre duas entidades. Isto é, se duas entidades compartilham links entre si, mais similares são. Desta forma, pode-se selecionar uma entidade alvo e elencar as entidades obtidas pelos seus outlinks. Depois, compara-se estas entidades relacionadas uma a uma com a entidade alvo e ordenamos de acordo com seu grau de similaridade, definido na Equação 3.7.

3.4 Modelo de Ranking usando SVM

Classificadores baseados em aprendizagem de máquina, normalmente, possuem dados que serão utilizados para treino e outros que servirão para teste[10]. Cada instância utilizada para a fase de treino do classificador é representada por um conjunto de características, associadas a uma classe. Então, cabe ao classificador criar um modelo, a partir da base de treino, o qual é capaz de prever a classe de uma instância de teste, a partir do seu conjunto de características.

Um exemplo de classificador é o *Support Vector Machine* [5]. O SVM utiliza a ideia de hiperplanos para realizar a classificação de suas instâncias. Digamos que se deseja classificar uma instância, representada por um vetor de atributos, que pertence a um espaço vetorial de p dimensões. Para isso, possuimos uma base de treino D:

$D = \{(x_i, y_i) | x_i \in \Re^p, y_i \in \{-1, 1\}\}_{i=1}^n$

Onde y_i é a classe do ponto x_i no espaço de p dimensões. Assim, o SVM produz um hiperplano de $(p-1)$ dimensões que separa instâncias de classes diferentes na base de treino, como na figura 3.4[4]. Tal hiperplano é obtido com o auxílio dos vetores de suporte (pontos sobre a linha pontilhada na figura 3.4), os quais são definidos pela combinação linear entre vetores da mesma classe.

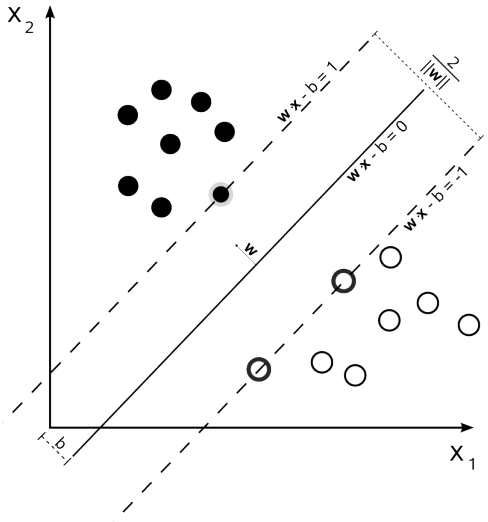

Figura 3.4: Hiperplano obtido com o SVM

Sendo w um vetor normal e perpendicular ao hiperplano, temos que para os pontos em $w \cdot x_i - b \geq 1$ são classificados como classe 1. Enquanto em $w \cdot x_i - b \leq -1$ são classe 2.

Com essa definição, diversas aplicações podem ser feitas, dentre as quais, a de construir um modelo de ranking a partir de uma base de treino. Isto é, utilizar aprendizado de máquina para aprimorar a qualidade (relevância) de resultados.

A técnica utilizada para os experimentos desta dissertação foi o SVMRank [4], que usa o algoritmo de classificação SVM para criar o modelo de ranking.

O algoritmo SVMRank é uma função de recuperação que, utilizando-se de aprendizagem de máquina, emprega métodos de ranking em pares para classificação dos

[4]http://en.wikipedia.org/wiki/Support_vector_machine

resultados, de forma adaptável, com base em sua relevância para uma consulta específica. O SVMRank usa uma função de mapeamento para descrever a correspondência entre uma consulta e as características de cada um dos resultados possíveis. Esta função de mapeamento projeta cada par de dados em um espaço de características. Tais características podem ser, por exemplo, uma lista de similaridades de cada campo no perfil do usuário e uma propaganda na base. Essas combinações dos campos do perfil e as propagandas, através das similaridades, são usadas como dados de treinamento para o SVMRank.

Geralmente, o SVMRank inclui três etapas no período de treinamento:

1. Mapeamento das semelhanças entre os campos dos perfis e as propagandas;
2. Cálculo das distâncias entre dois dos vetores obtidos no passo 1;
3. Formação de um problema de otimização que é semelhante a uma classificação SVM padrão e resolve esse problema com o algoritmo de SVM normal.

Em [4], onde deseja-se reordenar o ranking dos resultados de uma busca, temos um modelo que pode ser adaptado e comparado com o proposto nesta dissertação, utilizando um processo inverso à busca: a recomendação. Devido aos resultados positivos em [4], foi decidido utilizar a mesma técnica para esta dissertação. Para isso, adotamos uma implementação disponível na web[5], usando os parâmetros padrão para a fase de treino. No método, assume-se que existe uma ordem entre os valores de rank. Estes podem ser, por exemplo, resultados de uma máquina de busca. Então, temos os seguintes rankings $r_1 > r_2 > r_3 > ... > r_k$ e cada instância a formar um ranking pode ser denotada como $x = (a_1, a_2, a_3, ..., a_n)$, onde a_i é o valor da característica i para a instância x. Nos experimentos realizados para esta dissertação, x representa o par entre perfil do usuário e propaganda recomendada e cada a_i é a

[5]http://svmlight.joachims.org/

similaridade da propaganda com um campo do perfil, a qual é calculada utilizando a Equação 3.5.

O SVMRank nada mais é que uma aplicação do SVM (classificador) para resolver problemas relacionados a ranking, um problema de otimização. No caso dos experimentos desta dissertação, o SVMRank é utilizado para ordenar de forma adaptativa propagandas/produtos de acordo com a similaridade (como na Equação 3.5) que estes possuem em relação a um campo de um perfil. Esta similaridade serve como função de cosseno, a qual combinada com a relevância ou não da propaganda, irá servir como base de treino para a criação do modelo usado pelo SVM. De forma prática, o SVMRank utiliza esse modelo para que se possa gerar um score diferente para cada propaganda recomendada a um perfil. É válido ressaltar, que o score auxilia na geração de um ranking mais aprimorado em relação à precisão.

Nos experimentos, cada propaganda recomendada para o usuário foi mapeada como um vetor de atributos, onde cada recurso representa um campo e seu valor é o vetor de similaridade entre a propaganda e o campo, como:

$Ad_1 > Ad_2, Ad_1 > Ad_3, Ad_4 > Ad_5, Ad_4 > Ad_6, Ad_7 > Ad_8, Ad_7 > Ad_9$

Tendo como modelo a base avaliada acima, o SVMrank faz uma regressão envolvendo os atributos e o alvo F de forma a minimizar o número de erros em relação às restrições de ranking observadas no treino. Ele procura a função de regressão que minimiza o número de ordenações incorretas. Ou seja, em lugar de minimizar algo relacionado a erros de classificação, minimiza-se algo relacionado com erros de rankeamento.

Capítulo 4

Propaganda em Redes Sociais

As redes sociais e seus milhões de usuários configuram um ambiente propício para a exploração de propagandas, não só pelo número de usuários presentes hoje nessas redes como também por haver nelas uma rica variedade de informação pessoal sobre cada usuário. No entanto, o uso de propaganda contextualizada em redes sociais não tem sido muito explorado até hoje. Nesta dissertação, é proposto um modelo que utiliza informações contidas em redes sociais para veiculação de propaganda a seus usuários, apresentando-se portanto como uma alternativa para a geração de receitas a estas redes.

No modelo estudado, utiliza-se apenas o conteúdo dos campos do perfil do usuário para a geração das recomendações de propagandas a serem mostradas a usuários em redes sociais. É importante observar que diversas outras alternativas poderiam ser estudadas. Por exemplo, uma outra forma que poderiamos ter adotado seria a utilização de filtragem colaborativa como em [6]. Contudo, nosso trabalho aqui restringiu-se ao estudo de formas de uso da informação de perfil na veiculação de propagandas.

Basicamente, é proposto um modelo que soluciona o problema de seleção de propagandas, com base no conteúdo de perfis, como definido na figura 4.1.

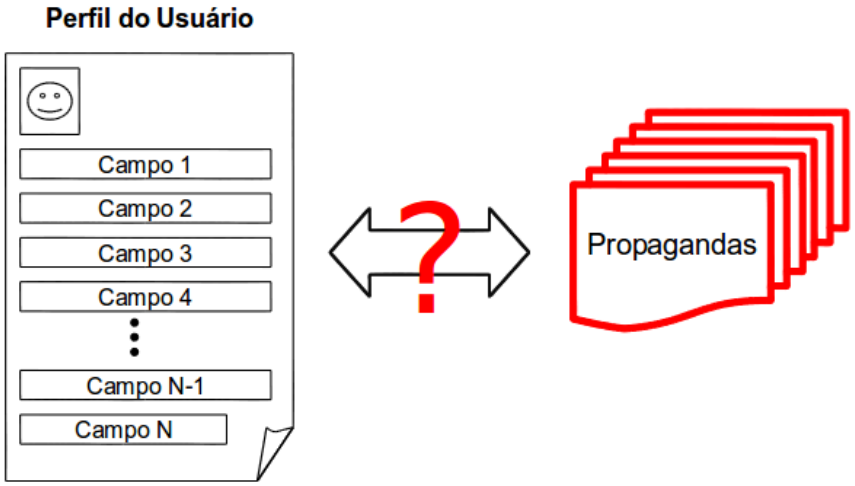

Figura 4.1: Problema

4.1 Visão Geral do Modelo

Antes da descrição detalhada, a figura 4.2 mostra uma visão geral do modelo. A ideia é construir uma base de treino utilizando as recomendações de propagandas feitas para uma base de perfis, a qual será utilizada como entrada para o SVMRank. Este cria uma função de ordenação e seleção que será utilizada para recomendar propagandas para novos perfis, como na figura 4.3.

Figura 4.2: Construção da Função de Ordenação e Seleção de Propagandas

Figura 4.3: Aplicação da Função de Ordenação de Seleção para Recomendação de Propagandas

4.2 Construção da Base de Treino

Os perfis são coletados da rede social e armazenados, formando uma base de dados de usuários. Estes servirão de consultas às propagandas indexadas. Para selecionar e ordenar as propagandas a serem mostradas aos usuários, utilizamos um método de recuperação de informação baseado em aprendizagem de máquina.

Esse tipo de método exige a criação de uma base de dados de treino para que o sistema possa então aprender a fazer a ordenação das propagandas a serem mostradas aos usuários. A base de treino deve ser composta de perfis de usuários e uma lista de propagandas avaliadas como relevantes ou não para serem mostradas junto com cada perfil. Para criar a base de treino utilizamos o modelo vetorial para recuperar um conjunto de propagandas a ser associado a cada perfil. Nesta fase é utilizada a expansão com dados da Wikipedia, somente para os perfis. As propagandas passam por um processo de avaliação de relevância, que é então utilizado como fonte para a formação do modelo do SVMRank, técnica de aprendizagem de máquina adotada na seleção de propagandas. Este processo é esquematizado de acordo com a figura 4.4.

4.2.1 Coleta

O primeiro passo para a criação da base de treino é a coleta de informação a respeito de perfis de usuários. Assim é possível entender que tipo e qual a qualidade do conteúdo presente nas redes sociais.

Durante a implementação do modelo, encontramos algumas dificuldades. Dentre as quais, a de coletar os perfis redes sociais. Primeiro, porque em redes sociais como o Orkut e Facebook, não se pode ter acesso aos dados sem uma conta. Segundo, mesmo com uma conta para essas redes, não se tem acesso a todas as informações contidas nos perfis dos usuários, a não ser que nos seja concedida esta permissão.

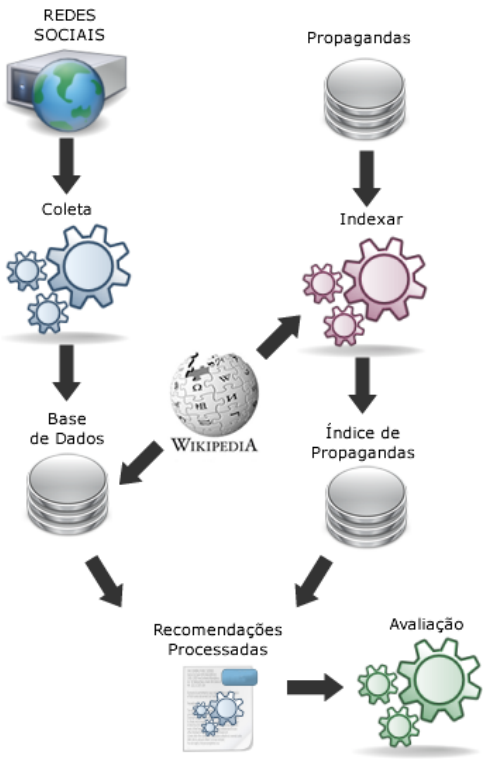

Figura 4.4: Construção da Base de Treino

Dentre as redes sociais mais conhecidas, decidimos utilizar perfis do Orkut, visto que é uma rede amplamente difundida entre usuários da Web brasileira, o que facilita coleta de dados e também a validação do modelo.

Os perfis foram coletados e armazenados num banco de dados local, formando uma base de dados de usuários. Uma análise sobre o conteúdo e qualidade destes perfis é fornecida no próximo capítulo desta dissertação.

4.2.2 Indexação

Para a validação do modelo, é necessário que se utilize uma base de propagandas. Essas serão recuperadas para que sirvam de recomendação aos perfis. Isto é, após a base de propaganda ser indexada como documentos no modelo vetorial. Dessa forma, é possível recuperá-las.

Ao observar os resultados de algumas recomendações, nota-se que ao indexar

termos individuais, perde-se algumas informações importantes. Por exemplo, o nome "São Paulo" refere-se à cidade ou o estado brasileiro, enquanto "São" e "Paulo", se separadas possuem significados distintos. Pode-se chamar esse conjunto de termos compostos que possuem um significado próprio de entidade, como "São Paulo". E uma forma para capturar essa informação, na indexação, é utilizando a Wikipedia.

4.2.2.1 Wikipedia como Fonte Complementar de Informação

Analisando essa relação entre dados do perfil e propagandas, nota-se que estas entidades, na forma citada acima, podem ser encontradas em ambos os lados, perfil e propagandas. Já que existe essa relação, isso pode ser utilizado para auxílio na tarefa de recomendação. Por exemplo, se o usuário preenche no campo de "livros" do seu perfil sobre "O Código Da Vinci", é evidente que refere-se ao livro de mesmo título. No entanto, se utilizarmos o modelo vetorial para recomendarmos propagandas baseadas nesse dado, é possível que sejam selecionadas propagandas relacionadas a termos distintos como "Código" ou "Da Vinci".

Para tratar esse problema, foi utilizada a Wikipedia como fonte de entidades para a representação dos dados, tanto do perfil quanto das propagandas. Usamos o dump da Wikipedia[1] com os títulos dos artigos, além de alterar o analisador de termos do Lucene para identificas as entidades. Assim, cada propaganda ou campo do perfil é representado não mais como um vetor de palavras, mas sim como um vetor de entidades como seus termos. Além da adição de entidades relacionadas, como explicado no capítulo anterior.

Para implementar o processo de identificação das entidades, utilizamos a ideia de N-gramas[3] com, no máximo, 5 termos. Retornando ao exemplo anterior, ao encontrar "O Código Da Vinci" no perfil ou numa propaganda, serão buscadas entidades com os 4 termos, mas caso não encontre nenhuma entidade, serão procuradas

[1] http://en.wikipedia.org/wiki/Wikipedia:Database_download

entidades com 3 termos “O Código Da”, assim por diante.

Caso uma entidade seja identificada no perfil, é realizado um cálculo de similaridade entre ela e as entidades apontadas pelos seus outlinks. Assim, obtém-se as entidades relacionadas, como definido no capítulo anterior. Estas, por sua vez, são adicionadas ao perfil. Desta forma, é possível complementar as informações contidas no perfil do usuário.

Para que se possa calcular a similaridade entre os campos do perfil e as propagandas, utilizando esta abordagem com a Wikipedia, é necessário que as propagandas também sejam representadas por essas entidades. No entanto, nós não adicionamos as entidades relacionadas na base de propagandas, pois acreditamos que poderia vir a inserir informação ruidosa na base, assim influenciando negativamente na qualidade das recomendações.

4.2.3 Recomendação

Com as características definidas, fez-se necessária uma forma de conectar os usuários às propagandas. No modelo proposto, utilizamos a similaridade vetorial [17], descrita no capítulo anterior, para realizar esta tarefa. Desta forma, cada campo do perfil representando uma característica, serviu como uma consulta para a biblioteca do Lucene, onde as propagandas estão indexadas. Os resultados das consultas são propagandas com algum grau de similaridade com as consultas feitas.Uma ilustração da geração dessas recomendações é vista na figura 4.5.

Cada campo gera um ranking diferente de propagandas, as quais são ordenadas de acordo com o *score* obtido, definido na Equação 3.5 no capítulo anterior. Assim, quanto maior o valor, mais similar a propaganda é do campo. Na figura 4.5, o vetor de características com valores de C_1 até C_N, representa o conjunto de similaridades do Campo 1 até o Campo N, com a propaganda recomendada.

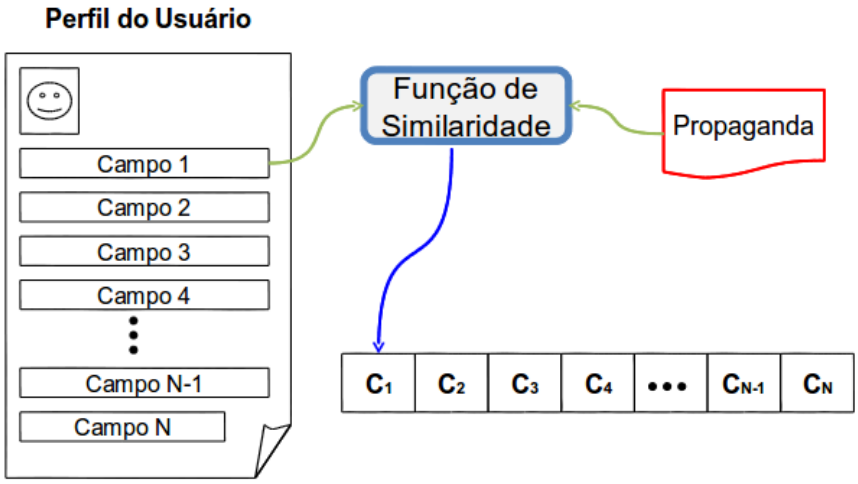

Figura 4.5: Similaridade entre campos e propagandas

4.2.4 Avaliação

As recomendações de propagandas foram geradas de acordo com o conteúdo dos perfis. Porém, necessita-se de uma avaliação dessas recomendações para assegurar a precisão do modelo proposto. Então os usuários, donos dos perfis, realizaram uma avaliação das propagandas veiculadas a ele. As propagandas foram apresentadas numa forma de lista sem ordem, ao lado do perfil do usuário. Cabe ao usuário julgar se as propagandas são relevantes ou não, isto é, se a propaganda despertou algum interesse de compra por parte do usuário. O conjunto de propagandas recomendadas foram rotuladas em “relevantes” e “não relevantes”, representados pelas marcações azuis e vermelhas na figura 4.6.

4.3 Função de Ordenação e Seleção

No modelo proposto nesta dissertação, foram utilizadas as avaliações como uma base de conhecimento para a aplicação do SVMRank. Esta técnica recebe como parâmetros várias instâncias que representam as propagandas recomendadas a cada perfil. Assim temos:

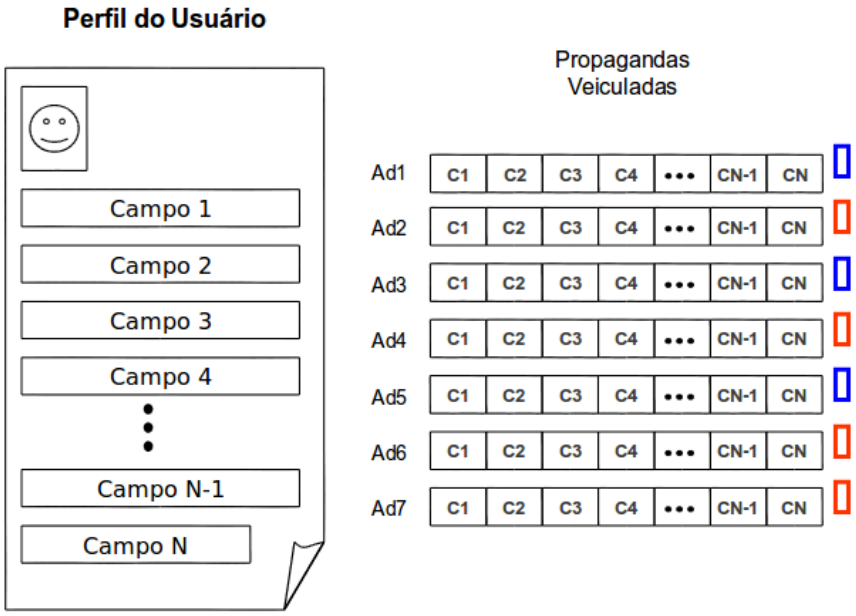

Figura 4.6: Avaliação das propagandas recomendadas

- Perfil 1:

$$Ad_1 : (R_1, \{F_{11}F_{12}F_{13}...F_{1t}\})$$

$$Ad_2 : (R_2, \{F_{21}F_{22}F_{23}...F_{2t}\})$$

$$Ad_3 : (R_3, \{F_{31}F_{32}F_{33}...F_{3t}\})$$

- Perfil 2:

$$Ad_4 : (R_4, \{F_{41}F_{42}F_{43}...F_{1t}\})$$

$$Ad_5 : (R_5, \{F_{51}F_{52}F_{53}...F_{2t}\})$$

$$Ad_6 : (R_6, \{F_{61}F_{62}F_{63}...F_{3t}\})$$

- Perfil 3:

$$Ad_7 : (R_7, \{F_{71}F_{72}F_{73}...F_{1t}\})$$

$$Ad_8 : (R_8, \{F_{81}F_{82}F_{83}...F_{2t}\})$$

$$Ad_9 : (R_9, \{F_{91}F_{92}F_{93}...F_{3t}\})$$

Para cada propaganda, temos um valor R_i que representa se a i-ésima propaganda foi julgada relevante ou não. Além do valor F_{ij} que representa a similaridade da i-ésima propaganda com o j-ésimo campo do perfil.

Assim, é gerado um conjunto de pares, entre as propagandas de um perfil. Desta forma, pode-se aplicar a mesma ideia do SVM para o treino do classificador. Então, aplicando uma nova propaganda com seu vetor de atributos ao modelo criado, é possível prever a qual classe pertence, relevante ou não.

Capítulo 5

Experimentos e Resultados

Neste capítulo, serão detalhados os experimentos feitos, assim como os resultados obtidos dos mesmos, de modo a validar o modelo de veiculação de propaganda, proposto nesta dissertação.

Nos experimentos, utilizou-se uma base de perfis, duas bases de propagandas e produtos, além de um dump da Wikipedia disponível na Web, como são descritos na seção seguinte.

5.1 Ambiente de Experimentação

Nesta seção são descritas ferramentas e bases utilizadas para a realização dos experimentos.

5.1.1 Perfis do Orkut

O acesso aos perfis do Orkut são restritos aos seus usuários. Portanto, para obtermos o conteúdo desses perfis, criamos uma conta e os 50 perfis de usuários voluntários foram adicionados como contatos do perfil. Deste modo, temos acesso a todas as informações necessárias para os experimentos que realizamos.

O uso de apenas 50 perfis de usuários para a realização dos experimentos deve-se ao fato da coleta de perfis ser uma tarefa complicada por vários motivos. Dentre os principais, temos o problema da privacidade em relação aos dados do usuário. Então, para cada perfil coletado, foi solicitado ao usuário dono do perfil uma autorização prévia para tal aquisição dos dados. Ainda dentro do assunto de privacidade, devido à política do Orkut, não somos autorizados a executarmos qualquer tipo de programa que exerça a função de crawler dentro da rede social. Tal fato dificulta a aquisição de dados de muitos e variados perfis. Logo, esta foi a quantidade máxima de perfis que conseguimos coletar em um tempo hábil para a realização dos experimentos desta dissertação.

De modo a simplificar o acesso às informações dos perfis, optamos por coletar as páginas referentes aos perfis e armazená-las num banco de dados. Para a coleta, utilizamos o pacote GNU Wget[1] juntamente com os *cookies* do usuário, para autenticação no Orkut. Para armazenar os dados, desenvolvemos um script que extrai apenas o conteúdos dos campos do perfil, eliminando os códigos HTML e Javascript contidos na página.

5.1.1.1 Qualidade do conteúdo dos perfis

Nas redes sociais, é difícil encontrar perfis com os campos totalmente preenchidos com informações. Muitas vezes, o usuário preenche com informações que não condizem com a verdade, ou apenas informações que não têm utilidade para a tarefa de recomendação.

Utilizamos uma base com 50 perfis que possuem uma taxa de preenchimentos dos campos de acordo com a Tabela 5.1

Ao falarmos de conteúdo de perfis de redes sociais, podemos pensar na hipótese de que existe muita informação dispensável para a tarefa de recomendação. Com

[1] http://www.gnu.org/software/wget/

Campo	Porcentagem de preenchimento
Livros	100%
Paixões	97%
Filmes	97%
Cozinhas	93%
Esportes	87%
Atividades	83%
Faculdade/Universidade	83%
Cidade Natal	77%
Ocupação	73%
Religião	73%

Tabela 5.1: Tabela de taxa de preenchimento dos campos da base de perfis

base nisso, elaboramos uma avaliação dos termos contidos nos perfis.

Foram listadas todas as palavras do vocabulário dos perfis, assim como as entidades que pudemos encontrar. Então, foi perguntado ao avaliador se aquele termo ou entidade representa algo relacionado a propagandas ou produtos. Obtivemos os seguintes resultados por campo:

Campo	Termo Individual	Entidade da Wikipedia
Músicas	16%	79%
Cargo	33%	78%
Filmes	13%	76%
Religião	30%	75%
Programas de TV	74%	74%
Esportes	37%	73%
Curso	48%	71%
Cozinhas	23%	61%
Livros	12%	58%
Paixões	14%	39%

Tabela 5.2: Tabela de porcentagem de termos e entidades julgados positivos para propaganda

Pode-se notar, com o resultado desta avaliação, que o uso das entidades da

Wikipedia no conteúdo dos perfis agrega significado aos termos que os compõem. Um fator que contribui para tal é a possibilidade das entidades serem formadas por termos compostos, os quais, se separados, podem ter outro significado.

5.1.2 Base da Wikipedia

Para os experimentos que envolvem as entidades da Wikipedia, utilizou-se uma versão do dump na língua portuguesa, disponível na Web[2]. Esta versão está atualizada até o mês de Julho de 2010 e contém os artigos da Wikipedia na língua portuguesa no formato de XML. O armazenamento e acesso à base foi feito através da biblioteca Tokyo Cabinet[3], que implementa funções para o gerenciamento de um banco de dados.

5.1.3 Bases de Propagandas e Produtos

Foram obtidas duas bases de propagandas para realizarmos a validação do modelo junto aos perfis. Uma base conta com 93.972 propagandas e a outra com pouco mais de 347.674 produtos, obtida junto a empresa Neemu[4], representando um subconjunto da base de produtos pertencentes à empresa.

Deve-se destacar as semelhanças e diferenças de ambas as bases. Por exemplo, abaixo pode-se ver um exemplo de propaganda:

[2]http://download.wikimedia.org/ptwiki/20100701/ptwiki-20100701-pages-articles.xml.bz2
[3]http://fallabs.com/tokyocabinet/
[4]http://www.neemu.com/

Título	superman
Descrição	superman no precomania compare precos entre centenas de lojas. veja o preco total com impostos e custo de frete. economize seu dinheiro ao comprar na loja com o melhor preco. compra comparativa superman precomania populares opinioes comerciantes busca produtos computadores fotografia eletronicos software video games filmes musica livros brinquedos papelaria joias roupas health beauty casa jardim babies kids flowers gourmet busca ocorrencias superman distribuido ordem popularidade canais complete superman collection starring christopher reeve margot kidder gene hackman 1978 1987 action adventure rating escreva critica 34 99 11 sellers complete superman collection diamond anniversary edition director max fleischer dave fleischer 1941 1943 childrens rating escreva critica 60 10 sellers superman director richard lester starring christopher reeve margot kidder 1980 science fiction fantasy rating opinioes 10 95 13 sellers ver 36 resultados filmes pessoas encontradas superman superman pajamas boys pijamas opiniao opiniao 40 66 loja size pedal car superman outdoors 70180 opiniao opiniao 813 74 loja superman fleece bath robe pijamas opiniao opiniao 56 31 loja ver resultados babies kids superhero robe superman pijamas camisolas opiniao opiniao 93 87 loja superman blue juvenile shirt camiseta tops opiniao opiniao 37 40 loja superman blue shirt camiseta tops opiniao opiniao 56 18 loja ver 14 resultados roupas incredible hulk superman autor roger stern opiniao escreva critica 99 loja batman superman world finest autor karl kessel batman wonder woman...

Agora um exemplo de um produto:

Título	blu-ray batman begins - importado
Descrição	liam neeson; cd, dvds e blu-rays / blu-ray

De uma forma geral, esses exemplos representam um formato das instâncias contidas em ambas as bases. Como pode-se notar, comparações entre as duas bases de resultados de experimentos devem ser feitas com cautela, pois tratam-se de tipos de dados diferentes, com formas diferentes de serem apresentados.

5.1.4 Métricas de Avaliação

Para a avaliação dos resultados, foram utilizadas 3 métricas comuns em sistemas de recuperação de informação: precisão; revocação; e medida-f.

5.1.4.1 Precisão

A precisão representa a porcentagem de propagandas recuperadas que foram consideradas relevantes para um determinado perfil. É representada pela Equação 5.1.

$$Precisão = \frac{|\{propagandas\ relevantes\} \cap \{propagandas\ recomendadas\}|}{|\{propagandas\ recomendadas\}|} \quad (5.1)$$

Para os experimentos, utilizou-se uma precisão a 5 ou P@5, isto é, a porcentagem de propagandas relevantes entre as 5 primeiras. Esta precisão é calculada por campo do perfil.

5.1.4.2 Revocação

A revocação representa a porcentagem de propagandas recomendadas para um perfil que foram recomendadas com sucesso. É representada pela Equação 5.2.

$$Revocação = \frac{|\{propagandas\ relevantes\} \cap \{propagandas\ recomendadas\}|}{|\{propagandas\ relevantes\}|} \quad (5.2)$$

No experimento, foi calculado uma revocação relativa ao perfil. Para isso, o número de propagandas relevantes e que foram recomendadas a um certo perfil, foi dividido pelo número de propagandas relevantes para o mesmo perfil. Este valor é limitado a 5, caso haja mais de 5 propagandas relevantes recomendadas ao perfil.

5.1.4.3 Medida-F

A medida-f representa uma média harmônica entre a precisão e a revocação. É representada pela Equação 5.3.

$$\text{Medida-F} = \frac{2 \cdot (precisão \cdot revocação)}{(precisão + revocação)} \quad (5.3)$$

5.2 Experimentos

Para cada base de propagandas e produtos, foram realizados experimentos independentes, apesar da utilização dos mesmos perfis como base de dados.

5.2.1 Propagandas

Para a base de propagandas, aplicamos o modelo proposto no capítulo anterior. Primeiramente, utilizamos a biblioteca Lucene para indexar a base de propagandas.

Neste passo, dois índices invertidos diferentes foram criados: um usando apenas termos; e outro com as entidades da Wikipedia incluídas no índice.

Depois, submetemos os campos dos perfis como consultas ao Lucene. Foram utilizadas duas abordagens visando os dois diferentes índices. Na primeira, o conteúdo do campo do perfil foi submetido como uma consulta comum ao modelo vetorial. Na segunda, identificamos as entidades e expandimos o conteúdo dos campos do perfil, inserindo estas entidades relacionadas ao campo. Entidades que representam artigos de desambiguação na Wikipedia não foram expandidos, pois há dúvida sobre qual entidade está sendo referida no perfil.

Como proposto no modelo, a seleção das entidades relacionadas foi feita com base na Equação 3.7, comparando a entidade alvo com as entidades encontradas através dos seus outlinks. No entanto, a inserção de todas as entidades relacionadas pode vir a inserir um excesso de informação no campo, podendo prejudicar a qualidade das recomendações. Os valores calculados com a Equação 3.7 foram normalizados pelo maior valor de similaridade dentre as entidades relacionadas. E somente as entidades relacionadas que obtiveram um valor superior a 0.8 foram inseridas no campo do perfil.

Após, os campos dos perfis, já com a expansão das entidades identificadas, foram submetidos como consultas para o Lucene, de modo que gerasse as recomendações de propagandas. Estas, por sua vez, foram avaliadas pelos usuários donos dos perfis como relevantes ou não.

Construiu-se a base de treino para a aplicação da técnica de aprendizado de máquina SVMRank, a qual gera um ranking aprimorado para as propagandas recomendadas. Para a execução do SVMRank, como proposto no capítulo anterior, nós utilizamos validação cruzada com 10 folds, pois foi a mesma validação utilizada em [4] com sucesso. Isto é, a base com as recomendações para os 50 perfis avaliados foi dividida em 10 partes, com recomendações referentes a 5 perfis cada. Então, o

SVMRank foi aplicado 10 vezes sobre esta base de treino, contendo 9 partes para a fase de treino e 1 parte para a fase de testes. Cada vez que o SVMRank é aplicado, troca-se a parte de teste, de modo que no final das 10 execuções, todas as partes tenham servido como teste uma única vez e como treino nas outras 9 vezes.

Com a aplicação do modelo concluída, obtivemos os resultados com precisão a 5, obtendo os seguintes resultados para cada combinação de experimentos, como se pode ver nas abordagens abaixo:

5.2.1.1 Sem Entidades da Wikipedia na indexação e sem expansão dos campos dos perfis

Campo	Medida-F	Precisão	Revocação
SVMRank	**28,40%**	**28,40%**	**28,40%**
Filmes	27,38%	29,30%	25,70%
Músicas	27,07%	27,23%	26,90%
Todos os Campos	24,00%	24,00%	24,00%
Livros	21,37%	24,10%	19,20%
Programas de TV	18,84%	20,54%	17,40%
Aniversário	18,80%	18,80%	18,80%
Esportes	13,01%	13,68%	12,40%
Humor	12,81%	15,47%	10,93%
Título Pessoal	11,74%	17,65%	9,80%
Paixões	11,02%	11,89%	10,27%

Tabela 5.3: Valores de Precisão, Revocação e Medida-F obtidos com os experimentos para a base de Propagandas, sem Entidades da Wikipedia na indexação e sem expansão dos campos dos perfis

Na Tabela 5.3, podemos ver que algumas presunções iniciais puderam ser confirmadas, pois campos como “Filmes”, “Músicas” e “Livros”, obtiveram melhores resultados que os demais. Podemos também notar que a junção de todos os campos também apresentou bons resultados, sendo assim, uma boa respresentação do perfil. Assim, o SVMRank obteve 28,40% de Medida-F, Precisão e Revocação, apresentando um resultado superior ao melhor campo (“Filmes”). A comparação dos resultados obtidos pelo SVMRank estão na Tabela 5.9, encontrada adiante.

Em complemento a esses resultados, dois outros experimentos foram realizados com o objetivo de melhor analisar o trabalho de seleção de campos para a recomendação. O primeiro é o que pode ser visto no gráfico 5.1, onde apresenta os valores de Medida-F para o SVMRank ao se remover os campos da Tabela 5.3, de forma crescente em relação a Medida-F. O segundo experimento, apresentado no gráfico 5.2, foi realizado de forma análoga ao primeiro, mas removendo os campos de forma decrescente em relação a Medida-F. Os mesmos experimentos foram feitos para outras variações adiante.

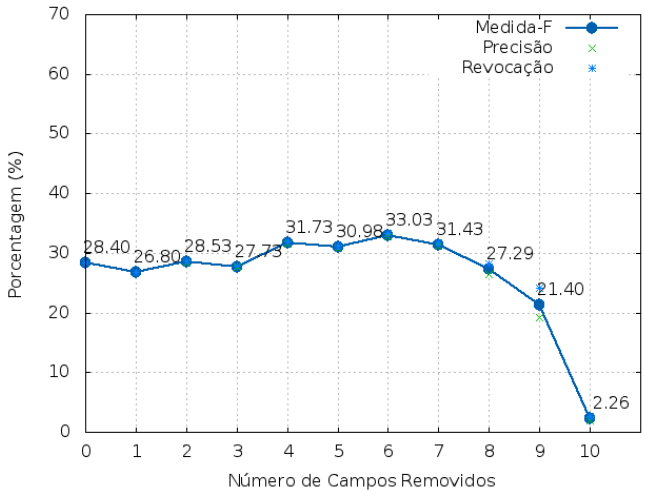

Figura 5.1: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma crescente em relação a Medida-F, utilizando a base de Propagandas, sem Entidades da Wikipedia na indexação e sem expansão dos campos dos perfis

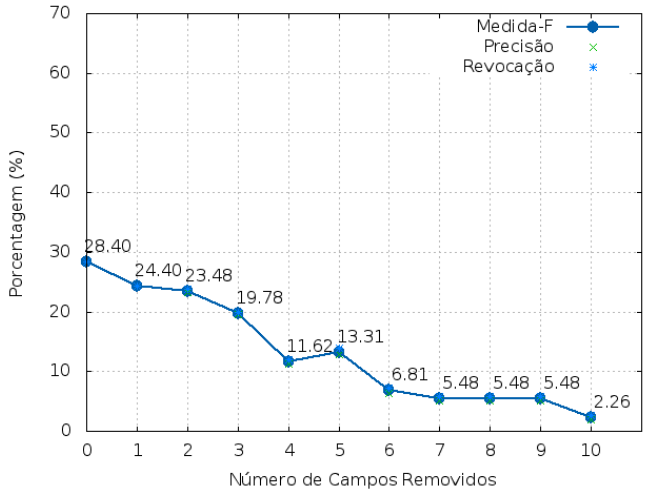

Figura 5.2: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma decrescente em relação a Medida-F, utilizando a base de Propagandas, sem Entidades da Wikipedia na indexação e sem expansão dos campos dos perfis

No gráfico 5.1, pode-se notar que os valores de Medida-F variam pouco conforme as remoções dos campos, principalmente entre os campos menos relevantes, apenas apresentando uma degradação grande ao se remover todos os 10 campos mais relevantes. Essa variação pequena, para mais ou para menos, poderia ser esperada, pois como se pode ver na Tabela 5.3, as precisões não apresentam grandes diferenças entre si. Já no gráfico 5.2 pode-se ter uma visão melhor de como os campos mais relevantes influenciam no resultado do SVMRank. Os melhores campos causam impactos importantes no SVMRank se removidos do método. Ao passo que cerca de 40% dos campos não devem ser utilizados, pois introduzem ruído. Isso ocorre devido a fatores como a diminuição de de amostras positivas e amostras em geral, os quais estão ligados aos campos mais relevantes. Ao passo que o SVMRank não consegue aprender devidamentes sem esses campos mais relevantes.

5.2.1.2 Sem Entidades da Wikipedia na indexação e com expansão dos campos dos perfis

Campo	Medida-F	Precisão	Revocação
SVMRank	**45,60%**	**45,60%**	**45,60%**
Músicas	35,73%	36,59%	34,90%
Todos os Campos	33,60%	33,60%	33,60%
Aniversário	27,48%	29,01%	26,10%
Filmes	26,79%	28,10%	25,60%
Programas de TV	24,33%	26,49%	22,50%
Livros	24,07%	27,18%	21,60%
CCQNPVS [a]	22,99%	29,60%	18,80%
Melhor Característica	22,15%	32,50%	16,80%
Esportes	21,74%	25,76%	18,80%
Título Pessoal	20,76%	35,94%	14,60%

[a] Cinco Coisas Que Não Posso Viver Sem

Tabela 5.4: Valores de Precisão, Revocação e Medida-F obtidos com os experimentos para a base de Propagandas, sem Entidades da Wikipedia na indexação e com expansão dos campos dos perfis

Para o caso apresentado na Tabela 5.4, o SVMRank obteve 45,60% de Medida-F, Precisão e Revocação, apresentando um resultado superior ao melhor campo ("Músicas"). O "Todos os Campos" segue logo após o campo de Música, o que indica que a expansão dos campos dos perfis não conseguiu ser efetiva individualmente para os campos restantes, de uma forma geral. Este fato é um indicativo de que, para essa abordagem, a expansão dos perfis adicionou informação ruidosa, a qual não

auxiliou positivamente na precisão.

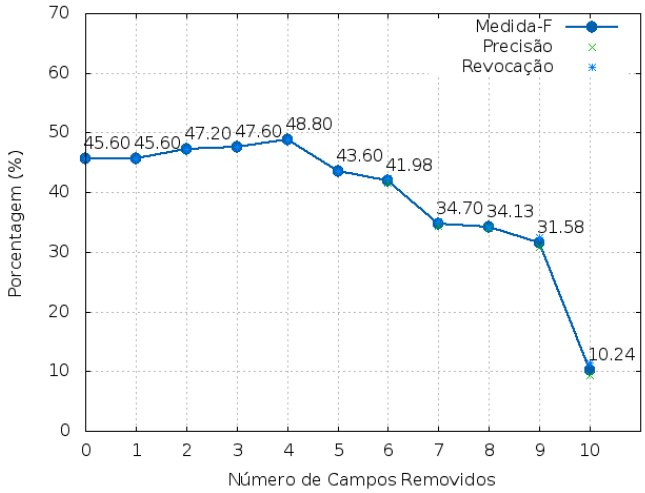

Figura 5.3: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma crescente em relação a Medida-F, utilizando a base de Propagandas, sem Entidades da Wikipedia na indexação e com expansão dos campos dos perfis

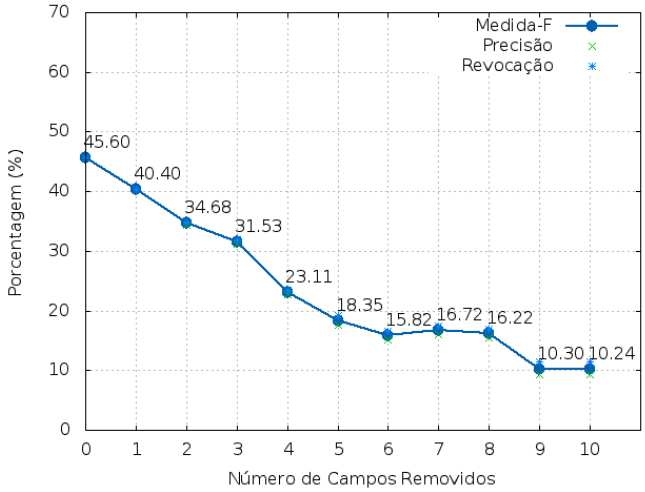

Figura 5.4: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma decrescente em relação a Medida-F, utilizando a base de Propagandas, sem Entidades da Wikipedia na indexação e com expansão dos campos dos perfis

Pode-se observar que no gráfico 5.3 houve um comportamento parecido com o

gráfico anterior, isto é, há uma pequena variação ao se remover os campos menos relevantes, até antes do campo de "Livros". A partir desse ponto, a qualidade do treinamento cai bastante, indicando que campos como "Livros", "Filmes" e "Músicas" influenciam bastante no resultado do SVMRank, como foi suposto anteriormente na dissertação. O gráfico 5.4 também apresenta um comportamento similar ao anterior, o qual apresenta uma degradação significativa ao se remover os campos mais significativos.

5.2.1.3 Com Entidades da Wikipedia (filtro) na indexação e sem expansão dos campos dos perfis

Campo	Medida-F	Precisão	Revocação
Todos os Campos	35, 63%	35, 20%	36, 07%
Músicas	35, 51%	37, 48%	33, 73%
SVMRank	**34,69%**	**34,40%**	**35,00%**
Filmes	20, 49%	23, 11%	18, 40%
Programas de TV	18, 52%	20, 63%	16, 80%
Livros	17, 33%	20, 40%	15, 07%
País	16, 37%	17, 11%	15, 70%
Paixões	15, 84%	18, 24%	14, 00%
Esportes	15, 22%	17, 60%	13, 40%
Visão Política	12, 42%	34, 00%	7, 60%
Título Pessoal	12, 03%	20, 00%	8, 60%

Tabela 5.5: Valores de Precisão, Revocação e Medida-F obtidos com os experimentos para a base de Propagandas, com Entidades da Wikipedia (filtro) na indexação e sem expansão dos campos dos perfis

Para o caso apresentado na Tabela 5.5, o SVMRank obteve 34,69%, 34,40% e 35,00% de Medida-F, Precisão e Revocação, respectivamente, apresentando um resultado levemente inferior ao melhor campo (“Todos os Campos”).

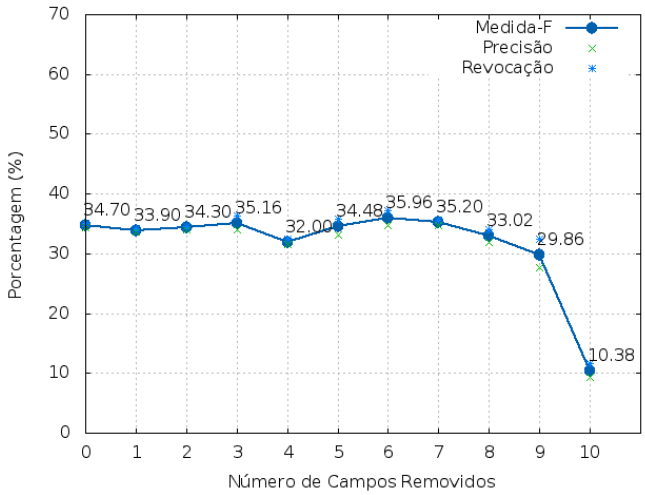

Figura 5.5: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma crescente em relação a Medida-F, utilizando a base de Propagandas, com Entidades da Wikipedia (filtro) na indexação e sem expansão dos campos dos perfis

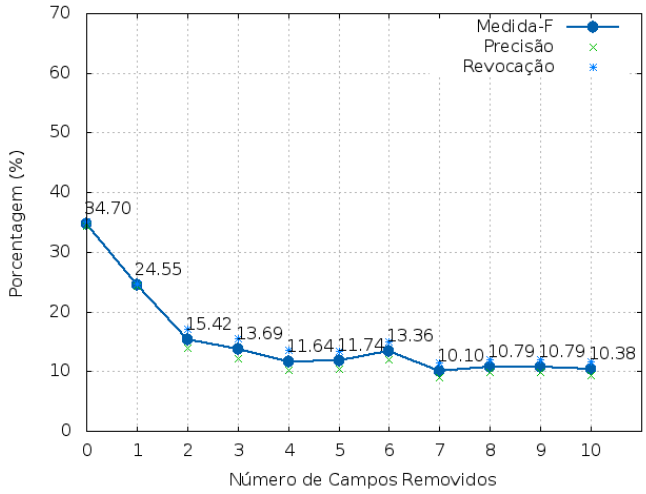

Figura 5.6: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma decrescente em relação a Medida-F, utilizando a base de Propagandas, com Entidades da Wikipedia (filtro) na indexação e sem expansão dos campos dos perfis

Em ambos os gráficos, 5.5 e 5.6, apenas a concatencação de todos os campos apresentou um resultado relevante para o SVMRank. Ao se remover esse campo, houve uma degradação na qualidade, ao passo que o mesmo não ocorre no mesmo grau em outros campos. Com a utilização do filtro da Wikipedia, a quantidade de termos na propagandas diminuiu naturalmente, o que causa esse impacto.

5.2.1.4 Com Entidades da Wikipedia (filtro) na indexação e com expansão dos campos dos perfis

Campo	Medida-F	Precisão	Revocação
Músicas	42.14%	44.89%	39.70%
SVMRank	**36.45%**	**36.40%**	**36.50%**
Todos os Campos	36.40%	36.40%	36.40%
Filmes	30.34%	35.47%	26.50%
Programas de TV	23.40%	29.03%	19.60%
Línguas que Falo	21.52%	28.75%	17.20%
Visão Política	21.03%	85.00%	12.00%
CCQNPVS [a]	20.66%	29.13%	16.00%
Paixões	18.24%	21.21%	16.00%
País	15.55%	16.89%	14.40%
Sobre Mim	15.33%	16.94%	14.00%

[a] Cinco Coisas Que Não Posso Viver Sem

Tabela 5.6: Valores de Precisão, Revocação e Medida-F obtidos com os experimentos para a base de Propagandas, com Entidades da Wikipedia (filtro) na indexação e com expansão dos campos dos perfis

Para o caso apresentado na Tabela 5.6, o SVMRank obteve 36.45%, 36.40% e 36.50% de Medida-F, Precisão e Revocação, respectivamente, apresentando um resultado levemente superior ao melhor campo (“Todos os Campos”).

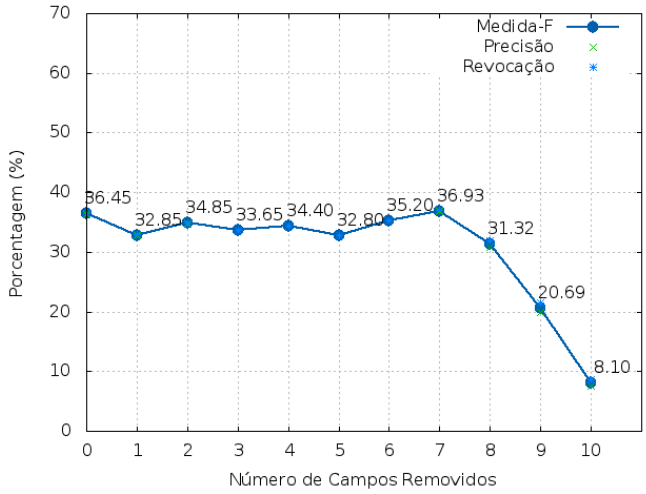

Figura 5.7: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma crescente em relação a Medida-F, utilizando a base de Propagandas, com Entidades da Wikipedia (filtro) na indexação e com expansão dos campos dos perfis

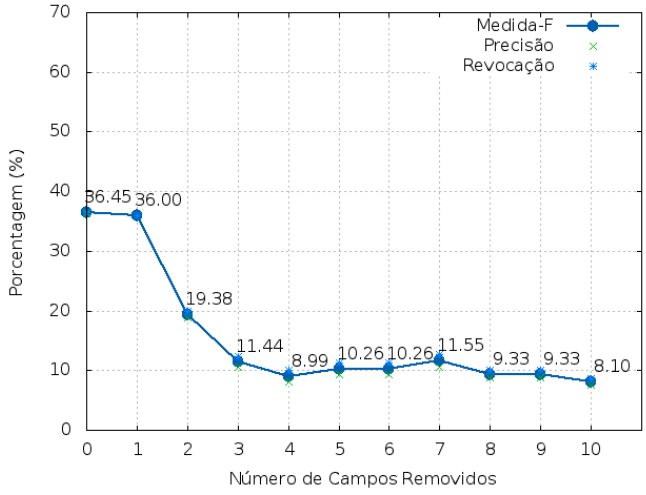

Figura 5.8: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma decrescente em relação a Medida-F, utilizando a base de Propagandas, com Entidades da Wikipedia (filtro) na indexação e com expansão dos campos dos perfis

Utilizando o filtro da Wikipedia, mas desta vez com expansão dos campos, ajudou a melhorar um pouco os resultados, em comparação com a mesma abordagem e sem

expansão. Nos gráficos 5.7 e 5.8, é possível ver que após a remoção dos dois campos mais relevantes, a qualidade dos resultados é prejudicada.

5.2.1.5 Com Entidades da Wikipedia na indexação e sem expansão dos campos dos perfis

Campo	Medida-F	Precisão	Revocação
SVMRank	**37.60%**	**37.20%**	**38.00%**
Todos os Campos	35.64%	35.20%	36.10%
Músicas	32.53%	33.48%	31.63%
Filmes	24.34%	26.10%	22.80%
Livros	23.60%	28.56%	20.10%
Programas de TV	21.29%	24.26%	18.97%
Paixões	15.81%	18.89%	13.60%
País	15.04%	16.83%	13.60%
Sobre Mim	13.23%	13.57%	12.90%
O Que Me Atrai	12.50%	15.65%	10.40%
Título Pessoal	12.36%	21.33%	8.70%

Tabela 5.7: Valores de Precisão, Revocação e Medida-F obtidos com os experimentos para a base de Propagandas, com Entidades da Wikipedia na indexação e sem expansão dos campos dos perfis

Para o caso apresentado na Tabela 5.7, o SVMRank obteve 37.60%, 37.20% e 38.00% de Medida-F, Precisão e Revocação, respectivamente, apresentando um resultado superior ao melhor campo (“Todos os Campos”).

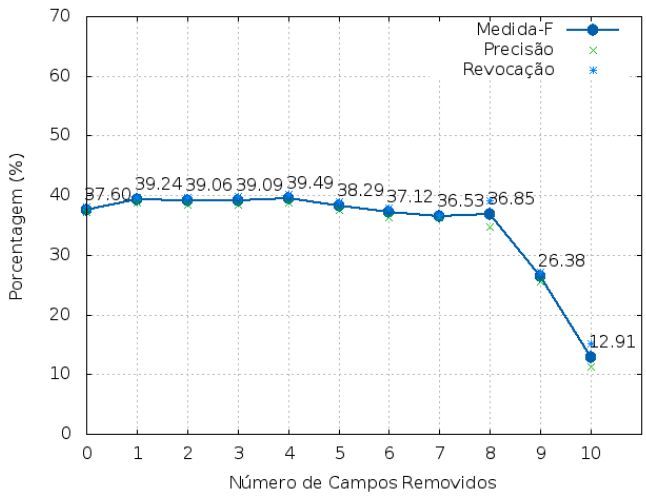

Figura 5.9: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma crescente em relação a Medida-F, utilizando a base de Propagandas, com Entidades da Wikipedia na indexação e sem expansão dos campos dos perfis

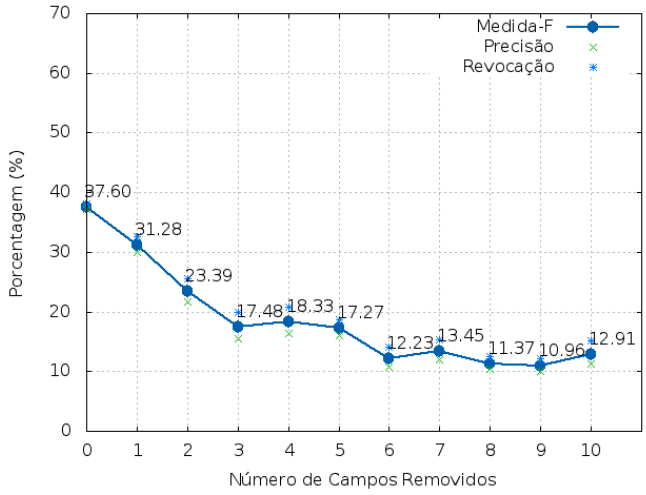

Figura 5.10: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma decrescente em relação a Medida-F, utilizando a base de Propagandas, com Entidades da Wikipedia na indexação e sem expansão dos campos dos perfis

Nesta abordagem, percebe-se um comportamento similar aos anteriores, isto é, quando dois campos destacam-se em seus valores de precisões, ao serem removidos,

prejudicam bastante o SVMRank. Isso é visível nos gráficos 5.9 e 5.10 ao se remover os campos “Músicas” e “Todos os Campos”.

5.2.1.6 Com Entidades da Wikipedia na indexação e com expansão dos campos dos perfis

Campo	Medida-F	Precisão	Revocação
Músicas	44.74%	47.11%	42.60%
SVMRank	**42.90%**	**42.80%**	**43.00%**
Todos os Campos	36.80%	36.80%	36.80%
Filmes	31.15%	34.79%	28.20%
Programas de TV	28.90%	36.09%	24.10%
O Que Me Atrai	22.15%	32.50%	16.80%
CCQNPVS [a]	18.38%	28.33%	13.60%
Sobre Mim	16.91%	20.49%	14.40%
Línguas Que Falo	16.20%	22.05%	12.80%
Livros	16.04%	19.56%	13.60%
Paixões	15.22%	18.24%	13.07%

[a] Cinco Coisas Que Não Posso Viver Sem

Tabela 5.8: Valores de Precisão, Revocação e Medida-F obtidos com os experimentos para a base de Propagandas, com Entidades da Wikipedia na indexação e com expansão dos campos dos perfis

Para o caso apresentado na Tabela 5.8, o SVMRank obteve 42.90%, 42.80% e 43.00% de Medida-F, Precisão e Revocação, respectivamente, apresentando um resultado levemente inferior ao melhor campo (“Músicas”).

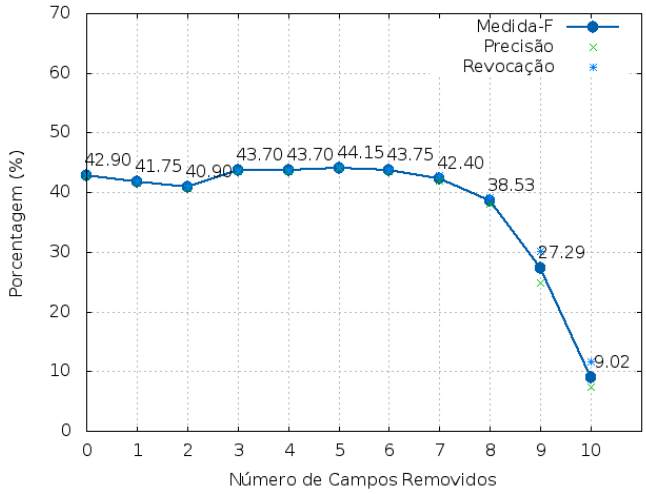

Figura 5.11: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma crescente em relação a Medida-F, utilizando a base de Propagandas, com Entidades da Wikipedia na indexação e com expansão dos campos dos perfis

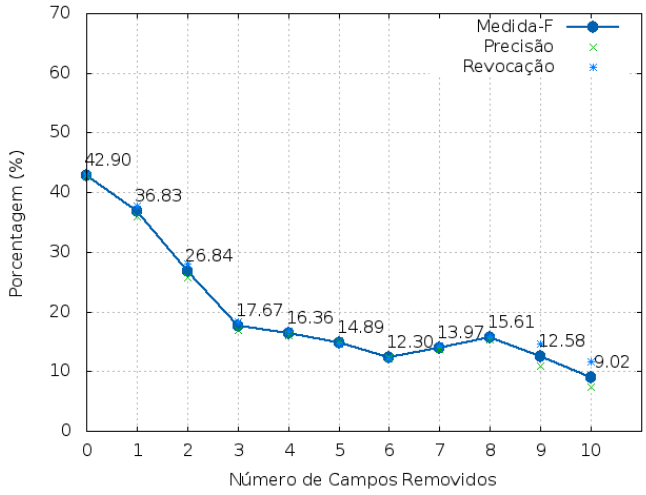

Figura 5.12: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma decrescente em relação a Medida-F, utilizando a base de Propagandas, com Entidades da Wikipedia na indexação e com expansão dos campos dos perfis

A expansão dos campos utilizando a Wikipedia favoreceu certos campos para essa abordagem, como o “Músicas” e “Filmes”. Melhorando também o resultado

do SVMRank em relação à abordagem sem a expansão. Com isso, pode-se ver nos gráficos 5.11 e 5.12 que ao se remover "Músicas", "Todos os Campos" e "Filmes", o SVMRank atinge valores inferiores que o normal.

5.2.1.7 Comparação das abordagens para a base de Propagandas

SVMRank	Medida-F	Precisão	Revocação
Sem Wikipedia, Com Expansão	45.60%	45.60%	45.60%
Sem Filtro, Com Wikipedia, Com Expansão	42.90%	42.80%	43.00%
Sem Filtro, Com Wikipedia, Sem Expansão	37.60%	37.20%	38.00%
Com Filtro, Com Wikipedia, Com Expansão	36.45%	36.40%	36.50%
Com Filtro, Com Wikipedia, Sem Expansão	34.69%	34.40%	35.00%
Sem Wikipedia, Sem Expansão	28.40%	28.40%	28.40%

Tabela 5.9: Valores de Precisão, Revocação e Medida-F obtidos com o SVMRank para a base de Propaganda, utilizando todas as variações dos métodos aplicados

Como pode ser visto na Tabela 5.9, o SVMRank foi aplicado a cada metodologia proposta. No melhor caso, temos o método em que a base da Wikipedia não foi utilizada no processo de indexação, mas os dados contidos nos campos foram expandidos com as entidades da Wikipedia.

O SVMRank utiliza as features (similaridade dos campos e propagandas) como base para criar um novo score que resulta num ranking melhor de propagandas a serem recomendadas. As medidas de precisão, revocação e medida-f são mensuradas de acordo com as 5 propagandas mais similares ao texto do campo do perfil. Já no SVMRank, essa medida é feita de acordo com as 5 propagandas de maior score (gerado pelo algoritmo de SVMRank) dentre todas recomendadas àquele perfil. Assim, como para cada perfil obteve-se 5 ou mais propagandas relevantes, as

fórmulas de precisão e revocação acabam por ficarem iguais. Isto é, o número de propagandas relevantes entre as 5 de maior score, dividido por 5, o máximo de propagandas relevantes possíveis para o cálculo. Assim, os resultados obtidos para precisão e revocação são iguais e, por consequência da Equação 5.3, quando a precisão é igual a revocação, a medida-f também resulta no mesmo valor de ambos. O mesmo resultado é esperado para a base de produto, como pode ser confirmado posteriormente.

5.2.2 Produtos

O mesmo procedimento descrito na seção anterior foi realizado para a base de produtos, obtendo os resultados também com precisão à 5, como se pode ver nas abordagens abaixo:

5.2.2.1 Sem Entidades da Wikipedia na indexação e sem expansão dos campos dos perfis

Campo	Medida-F	Precisão	Revocação
SVMRank	**63.20%**	**63.20%**	**63.20%**
Filmes	53.33%	57.67%	49.60%
Livros	49.72%	55.25%	45.20%
Músicas	47.43%	48.51%	46.40%
Curso	45.92%	57.12%	38.40%
Programas de TV	37.68%	41.08%	34.80%
Setor (Trabalho)	34.75%	44.76%	28.40%
Esportes	34.40%	37.19%	32.00%
Todos os Campos	28.40%	28.40%	28.40%
Título	27.87%	42.22%	20.80%
Habilidades Profissionais	26.60%	48.00%	18.40%

Tabela 5.10: Valores de Precisão, Revocação e Medida-F obtidos com os experimentos para a base de Produtos, sem Entidades da Wikipedia na indexação e sem expansão dos campos dos perfis

Para o caso apresentado na Tabela 5.10, o SVMRank obteve 63.20%, 63.20% e 63.20% de Medida-F, Precisão e Revocação, respectivamente, apresentando um resultado superior ao melhor campo (“Filmes”). A comparação dos resultados obtidos pelo SVMRank estão na Tabela 5.16, encontrada adiante.

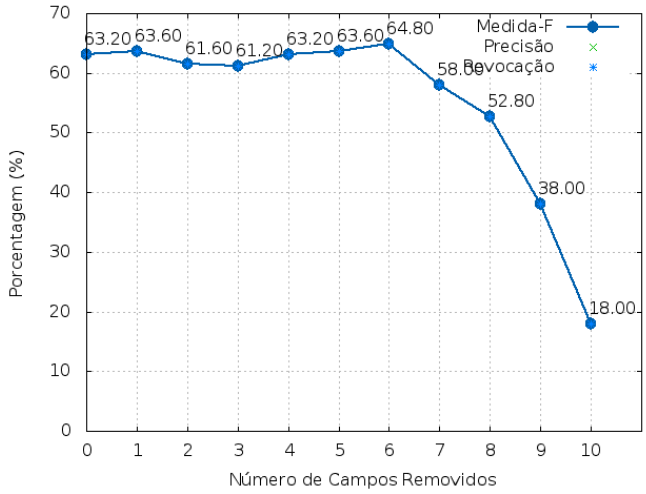

Figura 5.13: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma crescente em relação a Medida-F, utilizando a base de Produtos, sem Entidades da Wikipedia na indexação e sem expansão dos campos dos perfis

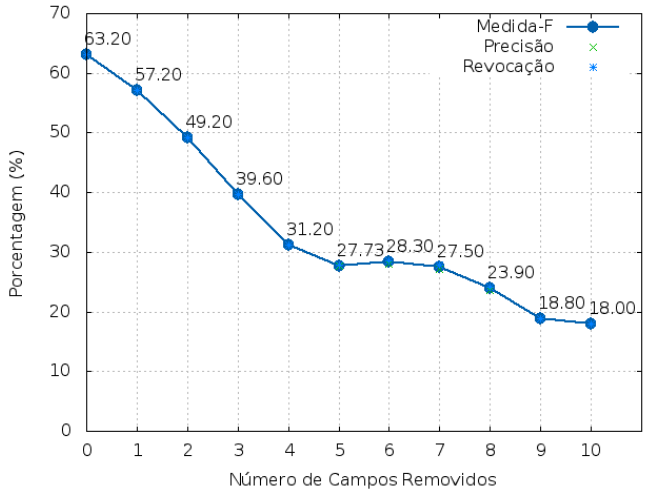

Figura 5.14: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma decrescente em relação a Medida-F, utilizando a base de Produtos, sem Entidades da Wikipedia na indexação e sem expansão dos campos dos perfis

A base de produtos mostra um comportamento similar ao apresentado na base de propagandas. Até mesmo campos como “Filmes”, “Livros” e “Músicas” apresentam

resultados relevantes e que influenciam mais que outros na geração do SVMRank, como se pode ver nos gráficos 5.13 e 5.14.

5.2.2.2 Sem Entidades da Wikipedia na indexação e com expansão dos campos dos perfis

Campo	Medida-F	Precisão	Revocação
SVMRank	**59.60%**	**59.60%**	**59.60%**
Músicas	49.14%	50.78%	47.60%
Livros	45.91%	51.84%	41.20%
Curso	39.12%	54.87%	30.40%
Filmes	34.44%	38.42%	31.20%
Esportes	33.06%	45.37%	26.00%
Todos os Campos	29.20%	29.20%	29.20%
Programas de TV	27.44%	31.34%	24.40%
Sobre Mim	23.94%	25.71%	22.40%
Atividades	23.85%	31.46%	19.20%
Profissão	23.12%	44.62%	15.60%

Tabela 5.11: Valores de Precisão, Revocação e Medida-F obtidos com os experimentos para a base de Produtos, sem Entidades da Wikipedia na indexação e com expansão dos campos dos perfis

Para o caso apresentado na Tabela 5.11, o SVMRank obteve 59.60%, 59.60% e 59.60% de Medida-F, Precisão e Revocação, respectivamente, apresentando um resultado superior ao melhor campo (“Músicas”).

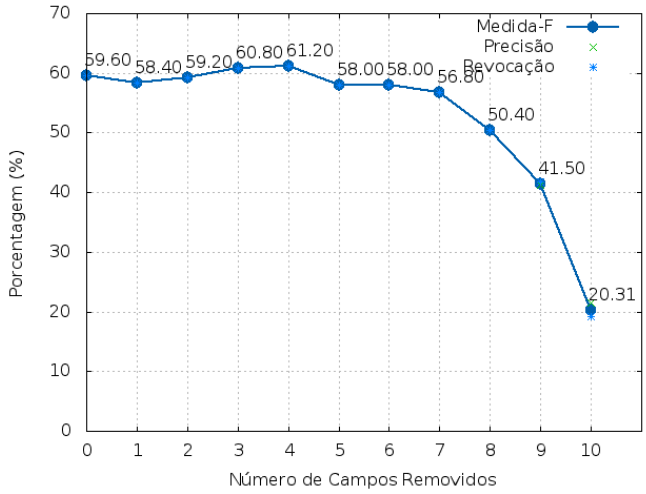

Figura 5.15: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma crescente em relação a Medida-F, utilizando a base de Produtos, sem Entidades da Wikipedia na indexação e com expansão dos campos dos perfis

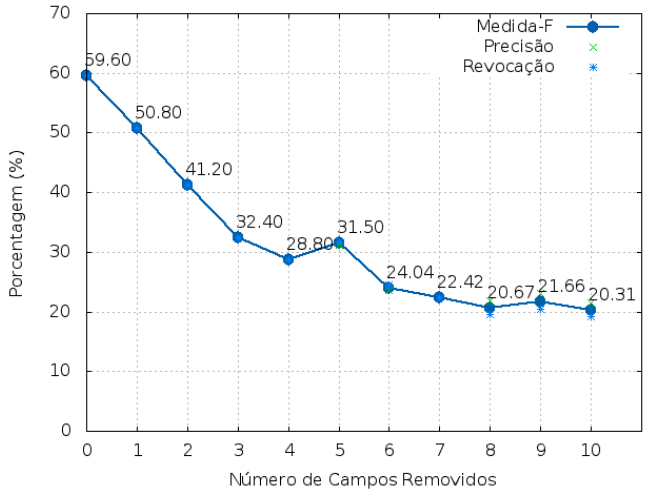

Figura 5.16: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma decrescente em relação a Medida-F, utilizando a base de Produtos, sem Entidades da Wikipedia na indexação e com expansão dos campos dos perfis

Com a expansão dos campos, houve uma piora na precisão da maioria dos campos em relação à abordagem sem expansão. Isso ocorre devido a como os produtos são

descritos, os quais faltam não abundam em termos que possam ajudar no trabalho de expansão. Com isso, o SVMRank apresenta um valor menor. Mesmo assim, ao retirar os campos mais relevantes, o comportamento é similar à abordagem anterior, como se pode ver em 5.15 e 5.16.

5.2.2.3 Com Entidades da Wikipedia (filtro) na indexação e sem expansão dos campos dos perfis

Campo	Medida-F	Precisão	Revocação
Músicas	45.86%	51.56%	41.30%
SVMRank	**40.00%**	**40.00%**	**40.00%**
Curso	33.88%	60.00%	23.60%
Paixões	32.44%	41.08%	26.80%
Todos os Campos	30.40%	30.40%	30.40%
Setor (Trabalho)	27.55%	36.84%	22.00%
Livros	27.46%	33.63%	23.20%
Programas de TV	24.73%	28.92%	21.60%
Filmes	24.14%	29.35%	20.50%
Esportes	23.50%	44.24%	16.00%
Profissão	22.21%	41.25%	15.20%

Tabela 5.12: Valores de Precisão, Revocação e Medida-F obtidos com os experimentos para a base de Produtos, com Entidades da Wikipedia (filtro) na indexação e sem expansão dos campos dos perfis

Para o caso apresentado na Tabela 5.12, o SVMRank obteve 40.00%, 40.00% e 40.00% de Medida-F, Precisão e Revocação, respectivamente, apresentando um resultado inferior ao melhor campo (“Músicas”).

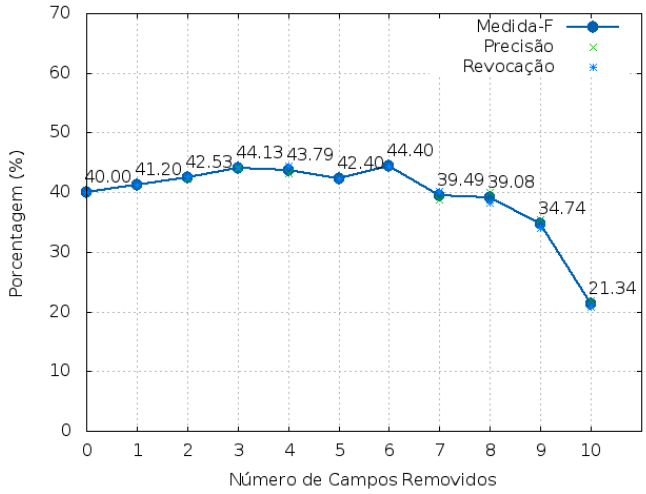

Figura 5.17: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma crescente em relação a Medida-F, utilizando a base de Produtos, com Entidades da Wikipedia (filtro) na indexação e sem expansão dos campos dos perfis

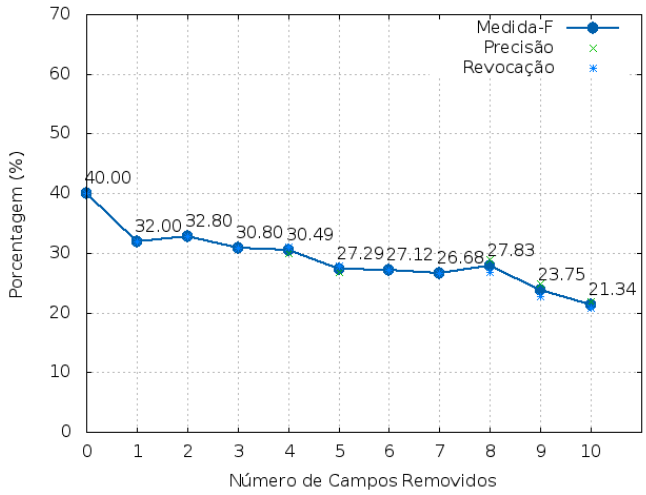

Figura 5.18: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma decrescente em relação a Medida-F, utilizando a base de Produtos, com Entidades da Wikipedia (filtro) na indexação e sem expansão dos campos dos perfis

Assim como na base de propagandas, a abordagem de Entidades da Wikipedia com filtro na indexação e sem expansão dos campos é restritiva, no sentido de

limitar, por exemplo, a quantidade de propagandas para o treinamento. Com isso, o SVMRank tem um treinamento insuficiente, acabando por não conseguir ser mais eficiente que o campo de "Músicas". Tal fato também pode ser observado nos gráficos 5.17 e 5.18, onde o valor de precisão do SVMRank degrada bastante ao remover o campo de "Músicas", mas tem uma queda bem menos acentuada na remoção dos outros campos.

5.2.2.4 Com Entidades da Wikipedia (filtro) na indexação e com expansão dos campos dos perfis

Campo	Medida-F	Precisão	Revocação
SVMRank	**44.00%**	**44.00%**	**44.00%**
Músicas	43.92%	52.03%	38.00%
Curso	37.09%	62.33%	26.40%
Todos os Campos	34.80%	34.80%	34.80%
Paixões	31.03%	43.87%	24.00%
Esportes	29.42%	50.25%	20.80%
CCQNPVS [a]	27.52%	44.12%	20.00%
Livros	27.50%	36.67%	22.00%
Setor (Trabalho)	24.63%	37.25%	18.40%
Programas de TV	22.11%	25.36%	19.60%
Profissão	21.47%	80.00%	12.40%

[a]Cinco Coisas Que Não Posso Viver Sem

Tabela 5.13: Valores de Precisão, Revocação e Medida-F obtidos com os experimentos para a base de Produtos, com Entidades da Wikipedia (filtro) na indexação e com expansão dos campos dos perfis

Para o caso apresentado na Tabela 5.13, o SVMRank obteve 44.00%, 44.00% e 44.00% de Medida-F, Precisão e Revocação, respectivamente, apresentando um resultado levemente superior ao melhor campo (“Músicas”).

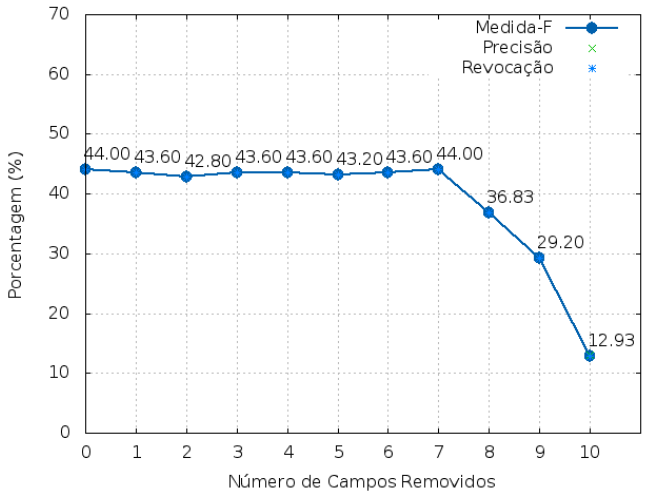

Figura 5.19: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma crescente em relação a Medida-F, utilizando a base de Produtos, com Entidades da Wikipedia (filtro) na indexação e com expansão dos campos dos perfis

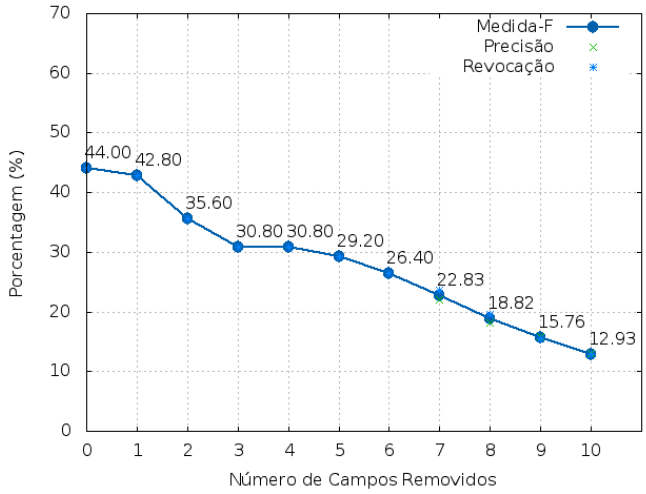

Figura 5.20: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma decrescente em relação a Medida-F, utilizando a base de Produtos, com Entidades da Wikipedia (filtro) na indexação e com expansão dos campos dos perfis

Nesta abordagem, que é menos restritiva devido a expansão dos campos, o SVM-Rank consegue se adaptar e treinar melhor sua base, assim obtendo melhores resultados também. Nos gráficos 5.19 e 5.20, é possível ver um comportamento mais parecido com os as de abordagens anteriores, onde a remoção de campos mais relevantes causa mais impacto no valor de precisão do SVMRank.

5.2.2.5 Com Entidades da Wikipedia na indexação e sem expansão dos campos dos perfis

Campo	Medida-F	Precisão	Revocação
Músicas	47.97%	51.63%	44.80%
SVMRank	**40.80%**	**40.80%**	**40.80%**
Curso	36.39%	60.62%	26.00%
Todos os Campos	34.00%	34.00%	34.00%
Setor (Trabalho)	30.57%	42.11%	24.00%
Esportes	30.12%	40.42%	24.00%
Livros	29.87%	35.09%	26.00%
Paixões	27.30%	34.95%	22.40%
Programas de TV	25.19%	29.46%	22.00%
Profissão	24.61%	46.00%	16.80%
Filmes	23.86%	26.97%	21.40%

Tabela 5.14: Valores de Precisão, Revocação e Medida-F obtidos com os experimentos para a base de Produtos, com Entidades da Wikipedia na indexação e sem expansão dos campos dos perfis

Para o caso apresentado na Tabela 5.14, o SVMRank obteve 40.80%, 40.80% e 40.80% de Medida-F, Precisão e Revocação, respectivamente, apresentando um resultado inferior ao melhor campo (“Músicas”).

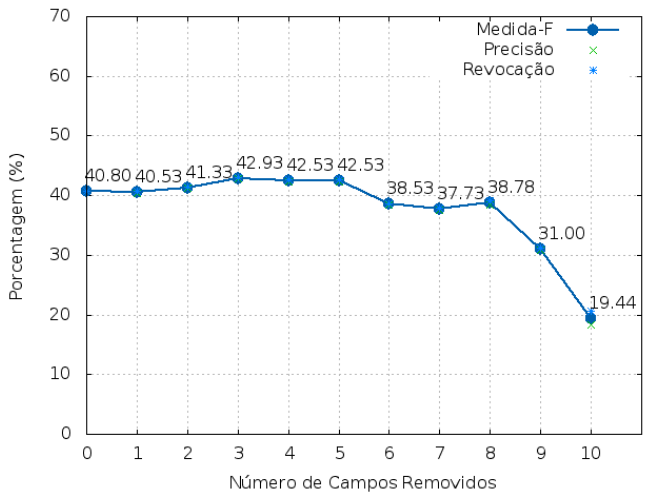

Figura 5.21: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma crescente em relação a Medida-F, utilizando a base de Produtos, com Entidades da Wikipedia na indexação e sem expansão dos campos dos perfis

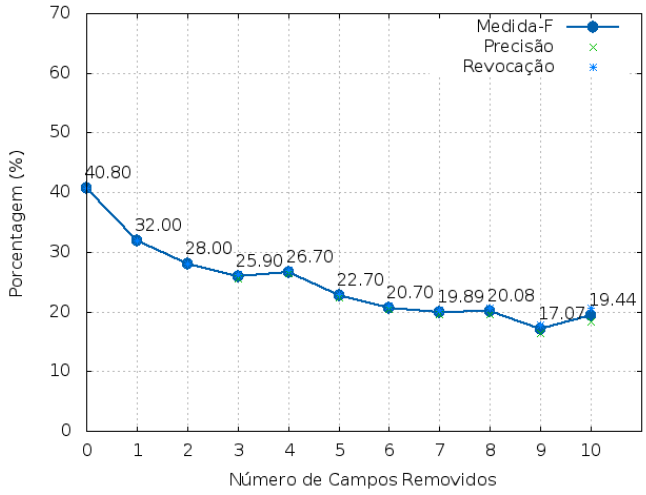

Figura 5.22: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma decrescente em relação a Medida-F, utilizando a base de Produtos, com Entidades da Wikipedia na indexação e sem expansão dos campos dos perfis

Os gráficos 5.21 e 5.22, mostram como nas abordagens anteriores que os campos de mais relevâncias influenciam mais no treinamento do SVMRank e, consequen-

temente, na qualidade final das recomendações. Mesmo assim, a base de produtos não ofereceu informações o suficiente para o treinamento do SVMRank, ocasionando num pior resultado se comparados a campos como “Músicas”.

5.2.2.6 Com Entidades da Wikipedia na indexação e com expansão dos campos dos perfis

Campo	Medida-F	Precisão	Revocação
Músicas	43.41%	51.34%	37.60%
SVMRank	**39.20%**	**39.20%**	**39.20%**
Curso	37.25%	65.67%	26.00%
Todos os Campos	31.20%	31.20%	31.20%
Esportes	28.63%	45.91%	20.80%
Livros	27.14%	33.54%	22.80%
CCQNPVS [a]	26.85%	42.59%	19.60%
Programas de TV	26.61%	34.63%	21.60%
Paixões	25.60%	34.38%	20.40%
Setor (Trabalho)	23.82%	36.86%	17.60%
Profissão	21.47%	57.50%	13.20%

[a] Cinco Coisas Que Não Posso Viver Sem

Tabela 5.15: Valores de Precisão, Revocação e Medida-F obtidos com os experimentos para a base de Produtos, com Entidades da Wikipedia na indexação e com expansão dos campos dos perfis

Para o caso apresentado na Tabela 5.15, o SVMRank obteve 39.20%, 39.20% e 39.20% de Medida-F, Precisão e Revocação, respectivamente, apresentando um resultado inferior ao melhor campo (“Músicas”).

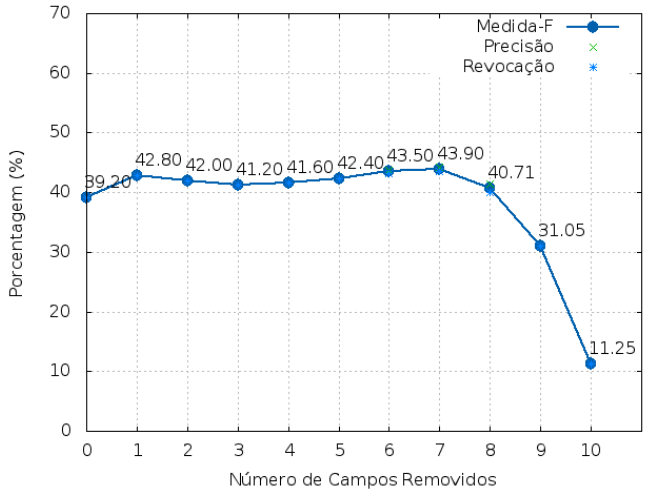

Figura 5.23: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma crescente em relação a Medida-F, utilizando a base de Produtos, com Entidades da Wikipedia na indexação e com expansão dos campos dos perfis

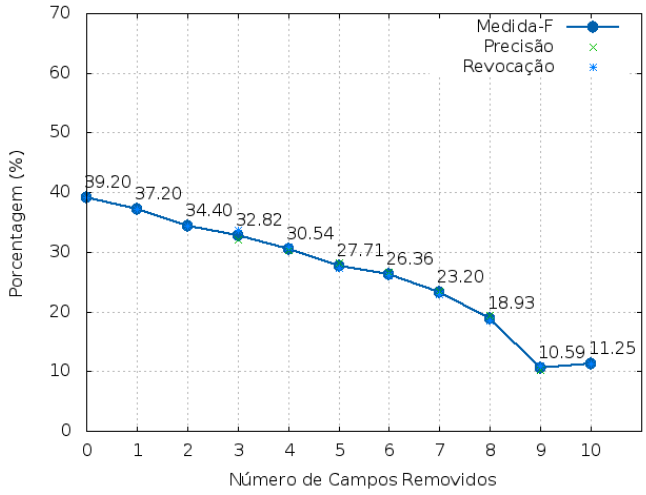

Figura 5.24: Gráfico dos valores de Medida-F resultantes da remoção de campos de forma decrescente em relação a Medida-F, utilizando a base de Produtos, com Entidades da Wikipedia na indexação e com expansão dos campos dos perfis

Nesta abordagem, a expansão dos campos não influenciou positivamente na maioria dos casos, o que ocasionou num resultado pior que a abordagem anterior,

em termo de Medida-F. No entanto, nota-se que a razão disso no caso do campo "Músicas", por exemplo, deve-se a queda na revocação. Em outras palavras, a expansão ajudou a buscar propagandas mais diversas, em detrimento de propagandas que seriam mais precisas, configurando um caso de inserção de informação ruidosa, por parte da expansão. Com isso, ao se remover os campos mais relevantes, como visto em 5.23 e 5.24, a degeneração das precisões do SVMRank são mais tênues.

5.2.2.7 Comparação das abordagens para a base de Produtos

SVMRank	Medida-F	Precisão	Revocação
Sem Wikipedia, Sem Expansão	63.20%	63.20%	63.20%
Sem Wikipedia, Com Expansão	59.60%	59.60%	59.60%
Com Filtro, Com Wikipedia, Com Expansão	44.00%	44.00%	44.00%
Com Filtro, Com Wikipedia, Sem Expansão	40.00%	40.00%	40.00%
Sem Filtro, Com Wikipedia, Sem Expansão	40.80%	40.80%	40.80%
Sem Filtro, Com Wikipedia, Com Expansão	39.20%	39.20%	39.20%

Tabela 5.16: Valores de Precisão, Revocação e Medida-F obtidos com o SVMRank para a base de Produtos, utilizando todas as variações dos métodos aplicados

Como pode ser visto na Tabela 5.16, o SVMRank foi aplicado a cada metodologia proposta. No melhor caso, temos o método onde não foram utilizadas a base da Wikipedia no processo de indexação nem as entidades da mesma.

Capítulo 6

Conclusões e Trabalhos Futuros

Como já mencionado anteriormente, não foram encontrados modelos para recomendação de propagandas como o proposto nessa dissertação. Logo, de forma a comparar os resultados, deve-se assumir um baseline como um sistema de recomendação simples, baseado apenas no modelo vetorial, sem qualquer adição de técnicas ou métodos auxiliares. Sendo assim, com o resultado dos experimentos realizados, pode-se dizer que o modelo proposto, com o SVMRank, obteve resultados significativamente melhores que o baseline, num caso geral. O SVMRank mostrou-se melhor para abordagens que possuem mais informações textuais.

As presunções iniciais feitas sobre quais campos seriam apropriados para recomendação foram satisfeitas com os experimentos. Pode-se ver que certos campos como "Músicas", "Livros" e "Filmes", por exemplo, são sempre constantes entre os melhores resultados de precisão. Ao passo que também pode-se concluir que a junção de todos os campos mostrou-se uma boa forma de se representar o perfil do usuário, visto que aparece entre as melhores precisões em todos os casos, sendo o melhor em alguns.

Uma observação interessante a se fazer é que houve resultados diferentes ao comparar as mesmas abordagens para bases diferentes (produtos e propagandas). Isso

ocorre pois o tipo de informação e dados são diferentes nas bases. Onde a base de produtos possui mais informações precisamente relacionadas aos produtos, facilitando o trabalho de recomendação, obtendo melhores resultados por conseguinte. Enquanto, na base de propagandas, muitos termos que não são necessariamente ligados ao produto, mas sim formas de marketing. Em outras palavras, na base de produtos, cada item possui uma descrição deste, onde na base de propagandas, cada item possui um texto de promoção ao produto. Tal diferença influencia diretamente na recomendação, pois termos diferentes resultam em recomendações diferentes. Por exemplo, temos o seguinte produto: "MacBook Pro MC024BZ Intel Core i5 2.53 ghz 4 gb 500 gb led 17 - Apple; Apple; eletroeletrônicos / informática / notebook". Comparemos com a seguinte propaganda: "Venda de computadores; venda de computadores e impressoras; computadores e impressoras. Sem duvida o melhor preço e a melhor qualidade. Confira nossas promoções! Today informática 0xx11 5521 6763 loja informática desenvolvimento today informática paulo sp brasil fax 55 11 5521 6763". Podemos notar que termos como "MacBook", "i5" e "Apple", por exemplo, são termos mais específicos e ligados a um produto. Enquanto os termos da propaganda são bem vagos em comparação.

A partir do corrente trabalho desenvolvido nessa dissertação, pode-se tentar diversas outras abordagem de modo a se encontrar técnicas mais eficientes. Assim como em outros trabalhos que envolvem o uso de base de dados para experimentos, uma abordagem futura seria realizar testes com outras bases, de tamanhos e variedades de produtos diferentes dos quais já foram cobertos com os experimentos desta dissertação.

Outra possível abordagem que pode ser tomada em estudos futuros seria sobre o uso de outras fontes externas de informações. Nesse ponto, outras fontes como

IMDB[1] ou FreeBase[2] poderiam ser utilizadas em complemento ou em substituição à Wikipedia. Estudos de casos específicos poderiam ser feitos, visto que bases como o IMDB pertencem à um nicho específico.

Esta ideia de utilizar informação adicional ao perfil do usuário, com a ajuda de entidades, é algo a se explorar separadamente. Durante os experimentos, pode-se ver resultados muito interessantes, obtendo relações com termos que não existiriam no perfil sem a expansão de informação feita. Por exemplo, certos usuários possuiam o termo "Friends" (Seriado de TV), a qual possui uma forte relação com a entidade "Matt LeBlank" (ator do Seriado). O mesmo ocorreu com: "Big Bang Theory" e "Kayley Couco"; "De Volta pro Futuro" e "Michael J. Fox"; "Matrix" e "Keanu Reeves"; "Harry Potter" e "Daniel Radcliffe"; entre outros.

Assim como no Orkut, existem outras redes sociais como o Facebook[3] ou Hi5[4], as quais possuem campos com informações também, tais como livros, músicas e filmes. Os métodos e experimentos feitos nessa dissertação também podem ser realizados nesses tipos de redes sociais, e podem ser explorados em trabalhos futuros.

Neste trabalho foi proposta a utilização de técnicas de aprendizado de máquina para aperfeiçoar as recomendações, no entanto não foi feito um estudo mais profundo sobre quais técnicas de aprendizado de máquina se adequariam para o caso apresentado. Este cenário poderia ser estudado de forma mais específica em um trabalho futuro.

[1] https://www.imdb.com/interfaces/
[2] https://developers.google.com/freebase/
[3] https://www.facebook.com/
[4] https://www.hi5.com/

Referências Bibliográficas

[1] Arthur, D., Motwani, R., Sharma, A., and Xu, Y. Pricing strategies for viral marketing on social networks. In *International Workshop on Internet and Network Economics* (2009), Springer, pp. 101–112.

[2] Bilenko, M., and Richardson, M. Predictive client-side profiles for personalized advertising. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (2011), ACM, pp. 413–421.

[3] Brown, P., Desouza, P., Mercer, R., Pietra, V., and Lai, J. Class-based n-gram models of natural language. *Computational linguistics 18*, 4 (1992), 467–479.

[4] Ciaramita, M., Murdock, V., and Plachouras, V. Online learning from click data for sponsored search. In *Proceeding of the 17th international conference on World Wide Web* (2008), ACM, pp. 227–236.

[5] Cortes, C. Support vector machine. *Learning 20*, 3 (1995), 273–297.

[6] Domingos, P. Mining social networks for viral marketing. *IEEE Intelligent Systems 20*, 1 (2005), 80–82.

[7] Feng, J., Bhargava, H., and Pennock, D. Implementing sponsored search in web search engines: Computational evaluation of alternative mechanisms. *INFORMS Journal on Computing 19*, 1 (2007), 137.

[8] Furasté, P. A. ***Normas Técnicas para o trabalho científico**: elaboração e formatação.* 14.ed., Porto Alegre:Dáctilo-Plus, 2006.

[9] Giles, J. Internet encyclopaedias go head to head. *438* (dec 2005), 900–901.

[10] Hsu, C., Chang, C., Lin, C., et al. A practical guide to support vector classification, 2003.

[11] Joachims, T. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (2002), ACM, pp. 133–142.

[12] Karimzadehgan, M., Agrawal, M., and Zhai, C. Towards advertising on social networks. *Information Retrieval and Advertising (IRA-2009) 28* (2009).

[13] Lin, H., and Li, L. Large-margin thresholded ensembles for ordinal regression: Theory and practice. In *Algorithmic Learning Theory* (2006), Springer, pp. 319–333.

[14] OneUpWeb. ***How keyword length affects conversion rates**:*. jan. 2005, <http://www.oneupweb.com/landing/keywordstudy_landing.htm>, Acesso em 18 Dez. 2010.

[15] Parsons, J., Gallagher, K., and Foster, K. Messages in the medium: An experimental investigation of Web Advertising effectiveness and attitudes toward Web content. In *System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on* (2002), IEEE, p. 10.

[16] Provost, F., Dalessandro, B., Hook, R., Zhang, X., and Murray, A. Audience selection for on-line brand advertising: privacy-friendly social network targeting. In *Proceedings of the 15th ACM SIGKDD international*

conference on Knowledge discovery and data mining (2009), ACM, pp. 707–716.

[17] Salton, G., Wong, A., and Yang, C. A vector space model for automatic indexing. *Communications of the ACM 18*, 11 (1975), 613–620.