MÉTODOS DE SELEÇÃO DE PALAVRAS-CHAVE PARA PUBLICIDADE CONTEXTUAL NA WEB

MARCELA SÁVIA PICANÇO PESSOA

MÉTODOS DE SELEÇÃO DE PALAVRAS-CHAVE PARA PUBLICIDADE CONTEXTUAL NA WEB

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

Orientador: Marco Antônio Pinheiro de Cristo

Manaus

Maio de 2010

Pessoa, Marcela Sávia Picanço
D1234p Métodos de Seleção de palavras-chave para Publicidade Contextual na Web / Marcela Sávia Picanço Pessoa. — Manaus, 2010
xvi, 47 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal do Amazonas
Orientador: Marco Antônio Pinheiro de Cristo

1. Seleção de palavras-chave. 2. Classificação. 3. Extração de Informação. 4. Publicidade Baseada em Conteúdo. I. Título.

CDU 519.6*82.10

[Folha de Aprovação]

Quando a secretaria do Curso fornecer esta folha,
ela deve ser digitalizada e armazenada no disco em formato gráfico.

Se você estiver usando o `pdflatex`,
armazene o arquivo preferencialmente em formato PNG
(o formato JPEG é pior neste caso).

Se você estiver usando o `latex` (não o `pdflatex`),
terá que converter o arquivo gráfico para o formato EPS.

Em seguida, acrescente a opção `approval={`*nome do arquivo*`}`
ao comando `\ppgccufmg`.

À minha maravilhosa família, especialmente à minha mãe Dinair Pessoa.

Agradecimentos

Meus agradecimentos serão em ordem cronológica dos acontecimentos, em nada refletem a importância das pessoas:

- Agredeço acima de tudo a Deus por, além de me ter permitido nascer em uma família extraordinária me fez encontrar pessoas maravilhosas pelo percurso da minha vida.
- À minha mãe Dinair Pessoa por ter vencido todos os obstáculos da vida de forma que tenha me proporcionado condições, sobriedade e perseverança para chegar onde eu cheguei. A ela toda minha gratidão e reconhecimento.
- Às minhas irmãs Sebastiana, Francinete e Diélia pelo companheirismo e confiança depositados em mim.
- Não posso esquecer do meu irmão Paulo César que mesmo sendo tímido sempre esteve ao meu lado me incentivando e, muitas vezes, abrindo mão da própria vida para facilitar a minha. Além de ter permitido que eu soubesse o que é ser mãe me emprestando um filho, o Paulo André.
- Gostaria de agradecer o professor Glauco, que só esteve em Manaus em um curto espaço de tempo, mas o suficente para me dar uma carta de recomendação para ingressar no tão sonhado mestrado.
- Já o professor Marco Cristo... este não tem nem como agradecer. Pois nada que eu escreva vai alcançar a milenésima parte de minha gratidão. Não só pela exímia orientação que me dispensou mas também pelo amigo que ganhei. Tenho perfeita consciência que se hoje estou concluindo este mestrado foi porque ele me recomendou generosamente para ingressar neste programa de pós-graduação, aliás, ele me adotou. Agradeço pelo apoio e força que me deu nos momentos que batia o desespero, sempre sereno com aquela resposta padrão e convincente "é assim mesmo".

- Não tem como deixar de registrar o apoio do Mauricio e Nakamura em todas as etapas deste processo. Na fase de seleção dando sempre as melhores dicas de como fazer. Durante o mestrado dando apoio e até mesmo ensinando as disciplinas mais difíceis como PAA e FTC. É fato que se eu aprendi autômatos devo às aulas que o Nakamura generosamente me deu.

- Agradeço à Fucapi, especialmente ao Antônio Luiz e Niomar por sempre confiarem em mim.

- Esse mestrado foi muito importante, especialmente para me fazer superar limites e reconquistar autoconfiança. Neste contexto sou muito grata ao professor Edleno pois quando achava que nada poderia piorar ele sempre aparecia para provar o contrário e, o melhor de tudo, eu superei.

- Se nada mais valesse, já teria valido a pena por ter conhecido o Eric Landin e o Éfren Lopes. Estes, foram amigos de estudo de PAA e que quero tê-los como amigos por toda a minha vida. Saudosos são os momentos que estivemos juntos. As desconcentrações, as vezes que Eric nos chamou a atenção para que não deixássemos o nervosismo atrapalhar nosso aprendizado, os medos de não conseguir, as inseguranças... todos os sentimentos superados pela amizade e força de vontade... poxa, como eu agradeço a Deus por tê-los encontrado.

- E o Kaio? com seu conhecimento em Linux. Tanto tentou que me convenceu a usá-lo. Meu eterno professor de Linux. Ainda bem que temos o Gtalk para nos manter próximos...

- Sem falar no Christophe, Édna e Sionise. Bons momentos de estudo: FTC, PAA... nervosismo, brincadeiras, estresse... momentos luluzinhas.

- Sou muito grata ao Raoni que esteve presente em todas as minhas angústias. Menino bom, humilde, sempre disposto a ajudar. Um coração generosíssimo.

- Ao grupo de publicidade, em especial ao Klessius cuja ajuda foi essencial para ter conseguido realizar os experimentos. Mostrando-se humilde quando respondia minhas dúvidas que pareciam triviais em relação ao seu conhecimento e experiência. Como foi maravilhoso ter encontrado pessoas como o Klessius.

- Quero agradecer ao Mauro Rojas pela experiência compartilhada e, não posso esquecer, pelos códigos em Perl que me disponibilizou.

- Uma grande dificuldade em RI é encontrar pessoas dispostas a doar um pouco de seu tempo para as estressantes avaliações. Para esta tarefa eu pude contar com muitos amigos. Em especial quero agradecer o Guilherme Toda, Jonatas Nascimento, Miguel, Raoni e Yanna Assunção por terem participado de todas as minhas avaliações. Mas, muitas outras pessoas participaram e não tem como deixar de citá-las: Adria, Ana Carolina Valentin, André Pedralho, Andrew Pacifico, Antônio Júnior, Bruno Ferreira, Carlos Alessandro Venezuela, Cláudio Santos, Chistophe Xavier, Cristiane Pacheco, Cynara Benarrós, Davi Fernandes, Daniel Elias, Daiana Oliveira, Douglas Araújo, Efren Lopes, Eli Cortez, Emerson Nogueira, Estephanie Jéssica, Eduardo João, Eder Pessoa, Eder Franco, Edna Magalhães, Everlin Marques, Felipe Hummel, Flávia Gomes, Gabriel de Souza, Geisa Dayana, Guilherme Monteiro, Gisele Amaral, Heitor Ferreira, Igor Cippolone, Igor Lopes, Icaro Oliveira, Izabel Barros, Jander Nascimento, Jerry Bastos, Jéssica Orlane, João Suarez, Jucimar, Julio Silva, Karla Okada, Kelen Acquati, Ketlen Teles, Leonardo Santos, Liliam Simão, Luiz Estrella, Lucas Arruda, Marxlening, Marcus Hilson, Márcio Palheta, Maria do Carmo, Nilson Silva, Neuma Pereira, Natascha Raposo, Osman Seixas, Onilton Maciel, Priscila Britto, Rebeca Silva, Rivaldo Assis, Rodrigo Avelino, Rodrigo Rezende, Raoni Ferreira, Sancha, Sionize, Sidney Júnior, Tomaz Noleto, Thales Alexandre e Zeina Oliveira.

- Quero destacar, ainda a participação de meus sobrinhos José Haroldo Júnior, Roberto Pessoa, Sávia Pessoa, Tayanne Bentes e Tayenne Bentes e da minha irmã Diélia Pessoa... não valeu só a avaliação e sim a companhia em algumas madrugada regadas a pizza e guaraná.

- Tem ajuda que não podemos mensurar mais são de infinita importância. Neste contexto gostaria de registrar o apoio do Elimarcos Arouca não só pela criteriosa, indispensável e onipresente revisão desta dissertação como também pelas agradabilíssimas conversas no percurso de volta para casa, especialmente nestes dias que antecedem a defesa da dissertação. Que companhia maravilhosa.

- Por fim, gostaria de agradecer ao Jerry por ter sido meu companheiro, meu amigo e meu porto seguro. Estando ao meu lado em todos os momentos de angústia e alegria. Não se importando com minha ausência, ou pelo menos não externando isso. Como foi importante cada "boa noite", cada abraço acalentador, cada tentativa de entender as coisas para me ajudar, cada olhar de confiança, cada comportamento de pai buscando me fazer concentrar. Cada momento junto foi e é mágico.

"Há três fatores para conquistar qualquer coisa: primeiro trabalho duro; segundo persistência; terceiro bom senso."
(Thomas Edison)

Resumo

Atualmente, o formato de publicidade dominante na Internet é baseado em mecanismos de leilão. Em tais formatos, os anunciantes fazem lances de palavras-chave que acreditam retornar anúncios interessantes para serem mostrados aos usuários. Desta forma, para selecionar os melhores anúncios levando em consideração o conteúdo textual das páginas é necessário saber quais palavras-chave as descrevem melhor ao mesmo tempo que possuem o melhor apelo para publicidade. Vários trabalhos encontrados na literatura propõem estratégias de aprendizagem de máquina para determinar estas palavras-chave em diferentes contextos, tais como mensagens de correio eletrônico e páginas web. Estas estratégias normalmente têm o objetivo de selecionar palavras-chave que os seres humanos consideram boas. Assim, o aprendizado é baseado em listas de pelavras-chave fornecidas por avaliadores humanos. Este tipo de estratégia chamamos de seleção derivada de lista manual ou, mais sucintamente, "seleção manual". Neste trabalho propomos uma nova estratégia de aprendizagem de máquina, onde a seleção é baseada no impacto esperado da palavra na qualidade final do sistema de seleção de anúncio. Chamamos nossa estratégia de seleção baseada em precisão de algoritmos de casamento de propagandas, ou mais sucintamente, "seleção por precisão". Nossa intuição é que as palavras-chave selecionadas pelos usuários como as melhores para a publicidade podem não ser as mais apropriadas para algorítmos de casamento de anúncios. Nos experimentos realizados com uma coleção de anúncios e considerando características propostas em um trabalho anterior, observamos que a seleção por precisão produziu melhores resultados que a seleção manual, com ganhos variando de 48,8% a 82,8% na precisão final apresentada por um sistema de seleção de anúncios. Paralelo ao novo método para selecionar palavras-chave, também estudamos a utilização de evidências extraídas da coleção de anúncios não estudadas anteriormente para este problema. Embora a adição de tais características não tenha melhorado o resultado do método de seleção manual, elas melhoraram os resultados obtidos pela seleção por precisão em, aproximadamente, 15%. Como consequência, quando comparamos os resultados da nossa abordagem, a seleção por precisão, com os resultados da seleção

humana, temos um ganho final de 79% a 110% após adicionarmos as novas evidências.

Palavras-chave: Seleção de palavras-chave, Classificação, Extração de Informação, Publicidade Baseada em Conteúdo.

Abstract

Nowadays, dominant adversiting formats in the Internet are based on auction mechanisms where advertisers bid on keywords they believe are interesting for the users. Thus, to select the best ads for a textual content, it is necessary to know which keywords better describe it and present the most effective marketing appeal. Several previous research found in literature have proposed machine learning strategies to determine these keywords in different contexts, such as emails and web pages. Such machine learning approaches usually have the goal of selecting keywords considered as good by humans, which we name here as *human-driven keyword* selection approaches. We propose in this work a new machine learning strategy where the selection is driven by the expected impact of the keyword in the final quality of the ad placement system, which we name here as *score-driven keyword* selection. Our intuition is that the keywords selected by the users as the best for advertising may not be the most appropriate for the ad matching algorithm. In experiments we performed with an ad collection, considering features proposed in a previous work, we found that the new *score-driven* approach led to significant gains over the previous human-driven approach, with gains ranging from 48,8% to 82,8% in the precision of results presented by an ad selection system. Beside the new paradigm to select keywords, we also study the use of new evidence extracted from the ad inventory in the task of selecting keywords. Even though the addition of such features did not improve the result of keyword selection when using the *human-driven* approach, they were able to improve the results obtained by the *score-driven* approach in roughly 15%. When considering these new features, the *score-driven* approach achieved a final improvement ranging from 79% to 110%, when compared to a recently proposed *human-driven* approach.

Keywords: Selection Keyword, Classification, Information Extraction, Content-based adversiting.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A Internet tornou-se um meio muito eficaz de publicidade visto que é possível a utilização de novos métodos e suporte à comunicação interativa, direcionada e persuasiva. Na verdade, todo o mercado mudou impulsionado pelos novos comportamentos dos consumidores uma vez que, em sua busca por novidades, informações e entretenimento, estão migrando rapidamente de mídias tradicionais, como televisão e jornal, para as alternativas digitais, todas convergindo para a Internet. Além disso, para os anunciantes, a Internet representa a possibilidade de exposição global a baixo custo, com grande potencial de medição direta dos resultados e interação com os consumidores. Essas características, juntamente com uma audiência crescente, nos ajudam a entender a adoção generalizada de publicidade na web.

Atualmente, os formatos de publicidade dominantes na web são os chamados métodos de pagamento por desempenho. Em tais estratégias, o anunciante recebe uma posição no *ranking* em uma lista patrocinada de acordo com o valor que ele está disposto a pagar e a previsão de relevância do anúncio para o usuário final. Neste caso o anunciante só paga quando recebe um clique. O primeiro método de pagamento por desempenho foi a publicidade de busca (ou busca patrocinada) [Broder et al., 2008a], onde são apresentados *links* patrocinados juntamente com a lista de resultados da pesquisa nas máquinas de busca. Neste caso, os anunciantes associam palavras-chave aos anúncios e esperam que estas sejam as mesmas palavras que os usuários irão utilizar para realizar a consulta na máquina de busca. O sucesso dessa estratégia levou empresas como Google, Yahoo e Microsoft a oferecer seus serviços de publicidade em outros contextos, como mensagens de correio eletrônico e páginas web. Em todos estes casos, os sistemas tentam encontrar palavras-chave que possam ajudar os sistemas de seleção de anúncios a trazer bons anúncios de acordo com o contexto encontrado em, por exemplo, mensagens de correio eletrônico, páginas web e vídeos.

Uma estratégia simples para determinar a importância de uma candidata a palavra-chave é utilizar informações estatísticas clássicas de recuperação de informação tais como frequência do termo e frequência inversa dos termo [Baeza-Yates & Ribeiro-Neto, 1999; Salton et al., 1974]. Tais estatísticas, entretando, não necessariamente consideram aspectos específicos do problema, como por exemplo, o apelo comercial do termo. Um determinado termo pode ter mais anúncios na base de propagandas associados a ele. Quando consideramos somente características da base de página desprezamos estas informações que acreditamos ser importantes para determinar a importância de um termo para retornar boas propagandas.

Com o intuito de encontrar estatísticas e indicadores mais apropriados, vários trabalhos abordam o problema de determinar quais os termos de um conteúdo textual seriam boas palavras-chave para um anúncio. Em particular, um trabalho muito influente [Yih et al., 2006], que estamos usando como base para o nosso estudo, adotou uma estratégia de aprendizagem de máquina em que os termos candidatos são descritos por várias características. O objetivo do método é aprender os padrões específicos que distinguem claramente as palavras-chave do anúncio.

Nestas abordagens, o que se aprende são as palavras-chave consideradas boas por humanos. Assim, o aprendizado é baseado em listas de palavras-chave fornecidas por seres humanos. Chamamos este tipo de estratégia de seleção derivada de lista manual ou, mais sucintamente, "seleção manual". Neste trabalho, estamos propondo uma nova estratégia, onde a seleção é feita levando em consideração o impacto esperado da palavra na qualidade final do sistema de casamento de anúncios. Assim, chamamos esta estratégia de seleção derivada baseada em precisão de algoritmos de casamento de propagandas ou, mais precisamente, "seleção por precisão".

Mais especificamente, podemos tirar vantagem em relação ao método anterior alterando a estratégia utilizada para compor a coleção de treino. Em vez de pedir aos usuários para dizerem quais são as boas palavras encontradas nas páginas de treino, nós ordenamos os anúncios que casam com cada palavra-chave candidata encontrada nas páginas de treino e pedimos aos usuários para avaliarem se os anúncios são relevantes para a página. Nossa intuição é que nem sempre avaliadores humanos são capazes de escolher as palavras-chave mais efetivas para um sistema de seleção de anúncios. Uma das principais razões para isso é o fato de que o avaliador humano não conhece a base de propagandas e, portanto, muitas das oportunidades de negócios que poderiam ser exploradas nela, como por exemplo se a base possui mais propaganda relacionada a um assunto específico ou mesmo, se as propagandas estão diretamente relacionadas a assuntos que não são o principal tema da página mas que tem boas propagandas na base de anúncios relevantes a tal tema. Como mostraremos, além desta estratégia

fornecer resultados competitivos ela pode se aproveitar de estimativas de relevância obtidas automaticamente, tais como a informação da taxa de clique, o que contribui para minimizar o custo de treino.

Nós também estamos propondo e estudando novas evidências extraídas da coleção de anúncios. Mostramos, ainda, que as estratégias propostas levam a ganhos significativos sobre abordagens anteriores. Usando apenas as características propostas por Yih et al. [2006], conseguimos melhorias variando de 48,8% a 82,8% na precisão, em relação ao sistema que tomamos como base quando usamos as palavras chaves retornadas em um sistema de seleção de anúncios. Esta melhoria é ainda maior quando consideramos as características extras que propusemos. Nossa abordagem apresenta um ganho em precisão de 79% a 110% sobre nossa base de comparação que usa a opinião de humanos para selecionar as palavras-chave.

1.1 Objetivos

1.1.1 Geral

Demonstrar que tomar como base para treino a precisão no casamento de propagandas, bem como utilizar características da coleção de anúncios, pode levar a uma seleção mais efetiva de palavras-chave por métodos automáticos, baseados em Aprendizagem de Máquina, para publicidade contextual.

1.1.2 Específicos

- Montar coleção de palavras-chave, avaliadas por usuários;
- Implementar extratores para os principais indícios usados previamente na literatura;
- Propor e estudar novos indícios, em particular, aqueles derivados da coleção de propagandas;
- Comparar o desempenho de um método de aprendizagem de máquina baseado em listas de palavras-chave fornecidas diretamente por avaliadores humanos, com o método baseado em palavras-chave que resulta em boa seleção de propagandas;

- Comparar os métodos considerando a presença ou não de indícios derivados da coleção de anúncios.

1.2 Estrutura do Texto

Esta dissertação está dividida em 5 capítulos, assim organizados:

Neste primeiro capítulo introduzimos o nosso trabalho, incluindo justificativa, motivação e os objetivos.

No capítulo seguinte descrevemos o conhecimento necessário para o entendimento do trabalho, mais precisamente sistemas de publicidade na web, conceitos básicos relacionados com sistema de recuperação de informação, métodos de classificação baseados em aprendizagem de máquina, métricas e técnicas de avaliação. Finalmente apresentamos os trabalhos relacionados.

O terceiro capítulo descreve a metodologia utilizada incluindo nossa estratégia de seleção de palavras-chave e as características que representam as palavras-chave, considerando tanto as características presentes na coleção de páginas quanto as presentes na coleção de anúncios.

O quarto capítulo apresenta nossos experimentos, incluindo a descrição das bases de dados utilizadas, nossa metodologia de avaliação e os resultados obtidos para as abordagens de seleção manual e por precisão.

E, por fim, o quinto capítulo apresenta as conclusões de nossa pesquisa e possibilidades de trabalhos futuros.

Capítulo 2

Referencial Teórico

Neste capítulo apresentamos os conceitos necessários para a compreensão dos métodos estudados e propostos neste trabalho. Em particular, vamos introduzir algumas definições básicas, descrever as técnicas e métricas empregadas para a avaliação e, finalmente, apresentar os trabalhos relacionados ao nosso, encontrados na literatura.

2.1 Conceitos Básicos

Nesta seção apresentamos os conceitos básicos associados a este trabalho. Mais especificamente vamos descrever sistemas de publicidade para web, sistemas de recuperação de informação e o método de classificação que usamos no trabalho.

2.1.1 Sistemas de Publicidade na Web

A publicidade na web envolve uma série de diferentes formatos. Nos primeiros anos, em meados da década de 90, ela foi fortemente influenciada pelas estratégias comuns em mídias tradicionais, onde o anunciante pagava pelo número de vezes que o anúncio era mostrado na página. Assim, formatos dominantes eram aqueles baseados na exposição de marcas e produtos em *sites* segmentados ou de grande audiência e a veiculação direta de ofertas por meio de correio eletrônico.

Com o tempo, novos formatos surgiram que exploravam melhor o ambiente de veiculação de informação, a interatividade, as possibilidades de medição direta de desempenho e a possibilidade de direcionamento preciso, baseado nos interesses dos usuários. Dentre os mais importantes, citamos os formatos de publicidade de busca e publicidade contextual [Broder et al., 2008b; Jin et al., 2007], descritos em detalhes nas próximas seções.

2.1.1.1 Publicidade de Busca

Neste tipo de publicidade, o usuário faz uma consulta em uma máquina de busca e recebe como resposta tanto páginas relevantes para a consulta como propagandas relacionadas a ela. Ou seja, é mostrada uma lista de anúncios na página retornada pela máquina de busca. Esta lista é chamada de lista paga. Os anúncios exibidos estão relacionados às palavras usadas pelo usuário para fazer a consulta. A Figura 2.1 mostra um exemplo de publicidade de busca, onde os anúncios são mostrados tanto no topo como ao lado direito da página.

Figura 2.1. Exemplo de publicidade de busca onde é feita uma busca utilizando a frase "clinicas de estética no Brasil". As propagandas relacionadas com a busca são mostradas tanto no topo da página quanto ao lado direito, marcadas como *links* patrocinados.

Normalmente os anúncios contêm um título, uma breve descrição e um endereço para uma página em que o usuário pode iniciar uma transação (página de pouso). Além disso, cada anúncio está associado a um conjunto de palavras-chave que incluem uma ou mais palavras e são utilizadas pelos anunciantes para descrever um eventual interesse do usuário. Por exemplo, o primeiro anúncio presente nos *links* patrocinados, à direita da página, na Figura 2.1 tem como título "Especialização Estética" com a descrição "Pós Graduação lato sensu - FAMESP. Ligue: (11) 5074-1010 Metrô Saúde/SP" e um *link* para a página "www.famesp.com.br". Neste exemplo, possíveis palavras chave associadas ao anúncio são "estética", "clínicas de estética" ou "profissionais de estética".

Como uma mesma palavra pode ser escolhida por vários anunciantes, sistemas de publicidade de busca avaliam o valor monetário da palavra de acordo com a sua procura.

Para tanto, eles organizam um leilão de palavras onde o valor que o anunciante está disposto a pagar é determinante para definir a posição da sua propaganda, relacionada à palavra-chave, na lista paga. Broder et al. [2008b] descreve busca patrocinada como uma cadeia de eventos envolvendo três atores: anunciante, máquina de busca e usuário. O anunciante é quem define uma campanha, ou seja, o conjunto de propagandas com uma determinada temática. A máquina de busca disponibiliza um espaço para colocar as propagandas nas páginas de resultados da busca e seleciona as propagandas mais relevantes para o usuário. O usuário visita as páginas e interage com as propagandas.

O sucesso da publicidade de busca e a necessidade de diversificar os meios de veiculação estimulam operadores destes sistemas a explorar outros contextos, tais como páginas web, mensagens de correio eletrônico e vídeo. Assim, surgiu a publicidade contextual, descrita em maiores detalhes na próxima seção.

2.1.1.2 Publicidade Contextual

O sistema de publicidade contextual é similar à busca patrocinada no que diz respeito ao anunciante pois este escolhe as palavras que julga interessantes e paga por elas. Mas, ao invés do anúncio estar relacionado à busca realizada, ele está relacionado ao conteúdo da página visitada pelo usuário. A Figura 2.2 mostra um exemplo de publicidade contextual.

Os anúncios utilizados na publicidade contextual são exatamente os mesmos usados na publicidade de busca. Duas diferenças importantes da publicidade de busca para a contextual são (1) a natureza mais passiva do usuário no caso da publicidade contextual. O usuário não precisa escolher palavras que descrevam seu interesse. A única ação do usuário é selecionar a página com o assunto que gostaria de ler e o sistema de publicidade escolherá os anúncios de acordo com o conteúdo da página que está sendo exibido, e (2) a necessidade de se processar muito mais termos em uma página quando comparada a uma consulta, visto que as consultas normalmente são compostas de poucas palavras. Já as páginas, na maioria das vezes, possuem muitas palavras. Esta segunda diferença, em particular, tem duas implicações importantes. Primeiro, páginas de natureza mais dinâmica podem necessitar que sistemas de busca avaliem constantemente o seu conteúdo, o que pode se tornar proibitivo, inviabilizando a publicidade contextual. O mesmo não ocorre na publicidade de busca, onde o pequeno tamanho da consulta viabiliza o seu processamento *on-line* eficiente o bastante para não prejudicar a experiência do usuário. Segundo, o maior conteúdo proporciona um casamento de mais termos, muitos deles casando apenas parcialmente (este é o caso de frases). Isto dificulta a análise de quão importante cada termo escolhido foi para a

campanha, uma informação preciosa para o anunciante. Ambos os casos apontam para a necessidade de se avaliar quais termos são mais úteis para a publicidade. De posse desta informação, termos poderiam ser filtrados, minimizando o primeiro problema. A mesma informação poderia auxiliar os anunciantes no processo de escolha de termos e análise do desempenho obtido.

Neste trabalho, focamos este problema específico de determinar quão provável é que um dado termo seja útil para a publicidade contextual. Para que isso seja possível, são necessárias várias tarefas, que incluem pré-processamento de texto, casamento, ordenação e classificação. Na seção seguinte são descritos em mais detalhes todos estes conceitos relacionados com sistemas de recuperação de informação textual.

Figura 2.2. Exemplo de publicidade baseada em conteúdo. O texto da página está relacionado com frutas e limpeza dos dentes. Nela são mostrados anúncios relacionados a saúde bucal e saúde em geral.

2.1.2 Sistemas de Recuperação de Informação Textual

Sistemas de recuperação de informação textual correspondem a uma série de sistemas de processamento de texto cru, não-estruturado. Entre as tarefas mais comuns de tais sistemas, podemos citar a busca eficiente por padrões de informação, a extração de padrões e o agrupamento/classificação de texto.

Um sistema de publicidade contextual pode ser visto como um sistema de busca textual onde, dado o conteúdo de uma página (a consulta), se pretende encontrar as propagandas mais relevantes para aquela página (os documentos). Em geral, tais sistemas envolvem várias tarefas de apoio como pré-processamento do texto, casamento, ordenação e classificação. Tais tarefas são descritas nas próximas seções.

2.1.2.1 Pré-processamento de documentos

Extrair os textos da página é a primeira etapa no processo de construção da base a ser usada por um sistema de recuperação da informação seja para busca, classificação ou publicidade contextual. Esta fase envolve a seleção de termos que melhor expressam o conteúdo das páginas. Ou seja, toda a informação que não refletir alguma idéia importante poderá ser desconsiderada. Desta forma, a seleção de termos reduz a quantidade de termos e, consequentemente, a dimensão dos vetores que representam os documentos. Esta diminuição possibilita que uma menor quantidade de memória seja utilizada e minimiza o tempo de processamento. Além disso, diminui a probabilidade de *overfitting*, fenômeno observado quando o classificador se adapta a detalhes muito específicos da base de treino, perdendo a capacidade de generalizar, levando com isso, a uma diminuição na taxa de acertos.

De modo geral, a etapa de pré-processamento tem por finalidade melhorar a qualidade dos dados já disponíveis e orgaizá-los de forma que possam ser submetidos a algum algoritmo de indexação ou mineração de dados.

Duas técnicas comuns de pré-processamento de texto são o *stemming* e a filtragem de *stopwords*. O *stemming* consiste em representar palavras através do seu radical, após remover sufixos e prefixos. Já *stopwords* são as palavras funcionais, ou seja, palavras consideradas não relevantes para a análise de um texto. Em grande parte dos textos são palavras auxiliares ou conectivas, não fornecendo nenhuma informação discriminativa na expressão do conteúdo do texto. Palavras como pronomes, artigos, preposições e conjunções podem ser consideradas *stopwords*.

Neste trabalho, seguindo a linha de outros trabalhos na literatura [Yih et al., 2006], não aplicamos *stemming* nem filtramos *stopwords*.

O diagrama da Figura 2.3 ilustra as fase do pré-processamento que está sendo utilizado neste trabalho.

A primeira etapa do nosso pré-processamento de texto é a tokenização. Nela, o texto, que é representado por uma sequência de caracteres, é inicialmente agrupado segundo fronteiras delimitadas por caracteres primitivos como espaço, vírgula, ponto, entre outros. Cada um desses grupos de caracteres é cha-

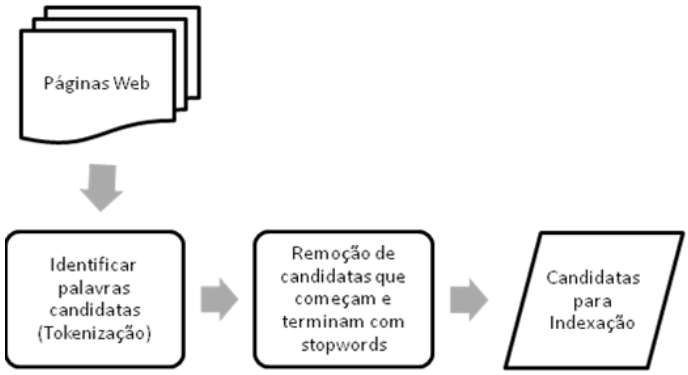

Figura 2.3. Etapas que compõem o pré-processamento das páginas Web.

mado de token. Tanto os grupos de caracteres, como os delimitadores se tornam tokens na nova sequência. O único caracter descartado é o espaço em branco. Por exemplo, no fragmento de texto "A uva Malbec é muito melhor na Argentina do que na França. Em seu país" resulta nos seguintes tokens [A][uva][Malbec][é][muito][melhor][na][Argentina][do][que][na][França][.][Em][seu][país].

O resultado desse processo é uma sequência de palavras intercaladas por espaço e algumas vezes por símbolos delimitadores. Os tokens que são constituídos por caracteres delimitadores são descartados restando somente os tokens constituídos por palavras do texto.

Como observado no exemplo, os tokens são constituídos por uma única palavra. Neste trabalho estamos considerando candidata qualquer palavra ou frase de até cinco palavras que façam parte de uma página. Por esta razão, expandimos nossos tokens para serem constituídos por até cinco palavras.

Os acentos das palavras foram retirados porém mantivemos a formatação de letras maiúsculas e minúsculas já que esta informação é importante para extrair uma de nossas evidências.

Por último, apesar de não termos feito remoção de stopwords, desconsideramos todas as candidatas que começam ou terminam com stopwords.

Após a etapa de pré-processamento, as páginas são representadas pelos termos que a compõem. Tais termos necessitam ser indexados para garantir sua recuperação eficiente. A tarefa de indexação será descrita na próxima seção.

2.1.2.2 Indexação

Após pré-processadas, as páginas são representadas por palavras. Para que uma consulta seja realizada é necessário percorrer toda a coleção de páginas, analisando documento a documento, o que demanda tempo e esforço computacional. Para minimizar

estes problemas são utilizadas técnicas de indexação. Ou seja, são criadas estruturas de dados (índices), capazes de permitir que uma consulta seja realizada sem ser necessário analisar toda a coleção de dados [Manning et al., 2008].

Para indexar o conteúdo de nossa coleção de páginas utilizamos a ferramenta Lucene[1]. O Lucene é uma Biblioteca de busca de texto desenvolvida pela Apache, com código fonte aberto e bem aproveitado por sistemas que precisam fazer buscas em textos completos.

Uma vez que documentos são recuperados através de seus índices, eles precisam ser ordenados de acordo com o grau de relevância para a pesquisa. Esta operação é chamada de *Ranking*.

Para ordenar dois documentos, é necessário representá-los de forma que possam ser comparados. Para tanto, em geral, um modelo de representação é usado, através do qual uma função de comparação é derivada. Um modelo de *ranking* bastante utilizado é o baseado em espaço vetorial. Este modelo é apresentado em mais detalhes na seção a seguir.

2.1.2.3 Modelo de Espaço Vetorial

O modelo vetorial representa documentos e consultas como vetores de termos. Termos são ocorrências únicas nos documentos e possuem um valor associado que indica o seu grau de importância no documento. Assim, para todo termo w_i no documento d_j é associado um peso w_{ij}. Um documento d_j é representado como um vetor de pesos $d_j = \{w_{1j}, w_{2j}, ..., w_{ij}\}$, onde i é a quantidade de termos distintos no conjunto de documentos. Os pesos quantificam a relevância de cada termo para as consultas e para os documentos no espaço vetorial.

Neste trabalho utilizamos a estratégia adotada por Baeza-Yates & Ribeiro-Neto [1999] onde anúncios e palavras-chave candidatas são representadas como vetores em um espaço composto por termos indexados. Assim, um anúncio a_j é representado como um vetor de termos t com pesos $a_j = \{w_{1j}, w_{2j}, ..., w_{tj}\}$. Cada peso w_{ij} reflete a importância do termo k_i em um anúncio a_j e é calculado como $w_{ij} = tf_{ij} \times log\frac{N}{ni}$, onde tf_{ij} é o número de anúncios onde k_i ocorre, e N é o número total de anúncios na coleção. Note que tf_{ij} é conhecido como *fator TF* e $log\frac{N}{ni}$ é conhecido como *fator IDF*. Para calcular a similaridade entre uma palavra-chave candidata c e um anúncio a_j, usamos o cosseno do ângulo entre c e a_j, como segue:

[1] `http://lucene.apache.org/java/2_0_0/index.html`

$$sim(a_j, c) = \frac{\sum_{i=1}^{t} w_{ij} \times w_{ic}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} w_{ic}^2}} \quad (2.1)$$

Apesar do modelo de espaço vetorial ser um bom método de *ranking*, existem abordagens mais sofisticadas que podem ser adaptadas, como por exemplo, realizar uma combinação dos *rankings* de todos os anúncios retornados para cada página e reordená-los de acordo com a similaridade obtida utilizando como chave secundária o identificador do anúncio.

No entanto, optamos pelo algoritmo simples derivado do modelo vetorial, conhecido como AAK (*ads and keywords*) [Ribeiro-Neto et al., 2005], por já ter sido utilizado na literatura para ordenar anúncios. Este algoritmo consiste no casamento entre o conteúdo da página e o conteúdo do anúncio, incluindo título, descrição e palavra-chave do anúncio. Neste algoritmo, são considerados relevantes somente os pares página-anúncio cujo conteúdo textual da página possua todas as palavras encontradas no campo palavra-chave do anúncio.

Além disso, nosso foco principal é validar o método de seleção de palavras-chave proposto sem nos preocupar com o método de ordenação envolvido. O *ranking* final utilizando todas as nossas evidências foi implementado usando o sistema de busca Lucene.

2.2 Classificação

O processo de seleção de palavras-chave pode ser visto como um problema de classificação. Dado um termo candidato à palavra-chave, o problema consiste em determinar se este termo deve ser classificado ou não como palavra-chave.

A tarefa de classificação consiste na predição de uma categoria ou classe discreta, usando uma função de classificação que mapeia dados para esta classe. A classificação pode ser supervisionada ou não-supervisionada. Na classificação supervisionada, amostras pré-classificadas são usadas para criar um modelo das classes possíveis. As amostras pré-classificadas constituem a coleção de treino e são ditas rotuladas.

Matematicamente, dado um conjunto $Y = \{y_1, y_2, ..., y_n\}$ de itens e um conjunto $X = \{x_1, x_2, ..., x_m\}$ de classes, supõem-se a existência de uma função $f : y \rightarrow X$ que mapeia um item $y_i \in Y$ para uma classe $x_j \in X$.

No aprendizado supervisionado o objetivo é, a partir de um subconjunto de itens $y' \subset Y$, tal que o valor $f(y')$ é conhecido, determinar a função $f' : y' \rightarrow X$, tal que f' é uma boa aproximação de f. Neste caso, cada item y' é representado por um conjunto

de características $Z = \{Z_1, Z_2, ..., Z_q\}$, tal que $y' = \{z_1, z_2, ..., z_q\}$. Assim, o algoritmo de aprendizado tenta determinar o valor de $f'(y')$ a partir da combinação de valores z_i. Há muitos métodos propostos na literatura para encontrar f'. Vários deles consideram que f' é uma função linear. Assim, eles tratam o problema de classificação como um problema geométrico de separação linear. Métodos que usam esta abordagem são, em geral, chamados métodos lineares e possuem custo de aprendizado e teste relativamente baixos. Um destes métodos é a Regressão Logística [Witten & Frank, 2005] que permite realizar uma modelagem preditiva, usando um modelo de dados binário que calcula a probabilidade referente ao sucesso de um determinado evento que será descrita a seguir.

2.2.1 Classificação por meio de Regressão logística

Como em todos os modelos de regressão, a finalidade da regressão logística é um modelo probabilístico que descreve a relação existente entre uma variável resposta e uma ou mais variáveis explicativas independentes, apresentando a resposta de maneira suscinta, geralmente como um número ou uma série de números.

O que difere esse modelo de regressão do modelo linear é o fato de a variável resposta que se deseja encontrar ser binária (ou dicotômica), enquanto que na linear essa variável é contínua.

No modelo utilizado por Hosmer & Lemeshow [2000], dada uma variável resposta Y binária, pode-se tratá-la como uma variável aleatória capaz de assumir apenas os valores 0 e 1. Em um modelo de regressão logística simples, com apenas uma variável independente X, define-se $\pi(x)$ como sendo a probabilidade de a variável resposta ser igual a 1 dado que $X = x$. Como Y só pode assumir os valores 0 e 1, essa probabilidade $\pi(x)$ é igual a $E(Y|X = x)$, podendo ser expressa conforme a equação (2.2):

$$E(Y_i) = \pi(x) = \frac{e^{\beta_0+\beta_1}}{1 + e^{\beta_0+\beta_1}} \tag{2.2}$$

A definição de $\pi(x)$ mostrada em (2.2) deriva de uma importante função de transformação de variáveis denominada transformação *logit* da probabilidade $\pi(x)$. A principal função dessa transformação é permitir que a função de regressão logística $g(x)$ tenha características lineares e, portanto, permita que valores entre $-\infty$ e $+\infty$ sejam assumidos pela variável dependente. A equação (2.3) mostra como é feita a transformação *logit*, onde a razão $\frac{\pi(x)}{1-\pi(x)}$ é comumente chamada de *Odds* e a função resposta $\beta_0 + \beta_1 x$ é denominada como a função resposta *logit*.

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x \tag{2.3}$$

Onde: -$\infty \prec g(x) \prec +\infty$ para todo x tal que -$\infty \prec x \prec +\infty$

Quando existe uma variável resposta binária podemos dizer que π é a probabilidade de ocorrência de determinado evento e $1-\pi$ é a probabilidade de não ocorrência. Nesse caso podemos assumir que a variável resposta Y é uma variável de Bernoulli, com parâmetro $E(Y) = \pi$ e, portanto, $Y_i = E(Y_i) + \varepsilon_i$. Onde ε é o erro associado à estimativa da variável resposta Y_i, expressando o desvio das observações em relação a média.

Em regressão linear esse erro apresenta a distribuição normal de probabilidade com variância constante. Já na regressão logística, justamente pelo fato de todas as variáveis seguirem uma distribuição de Bernoulli, esse erro apresenta uma distribuição de média zero e variância igual a $\pi(x)\,[1-\pi(x)]$, podendo assumir apenas dois valores:

- $\varepsilon = 1 - \pi(x)$ com probabilidade $\pi(x)$, se $Y = 1$;
- $\varepsilon = -\pi(x)$ com probabilidade $1 - \pi(x)$, se $Y = 0$;

Para determinar o modelo de regressão logística é necessário estimar os valores dos coeficientes β_0 e β_1 através do método da máxima verossimilhança [Hastie et al., 2001].

O método da verossimilhança tem como objetivo retornar valores para os parâmetros desconhecidos β_0 e β_1 de forma a maximizar a probabilidade de se obter a sequência observada de dados.

Inicialmente a aplicação do método consiste na definição e modelagem da função de verossimilhança, que expressa a probabilidade de obtenção da sequência observada como função dos parâmetros β_0 e β_1. Como a variável resposta Y assume somente os valores 0 ou 1, a expressão (2.2) para $\pi(x)$ apresentada anteriormente fornece a probabilidade condicional de Y ser igual a 1 dado x, ou seja, $P(Y = 1|x)$. Da mesma forma, temos que $1 - \pi(x)$ representa a probabilidade condicional de Y ser igual a 0 dado x, $P(Y = 0|x)$.

O modelo de regressão logística simples se difere da regressão logística composta somente por contar com mais de uma variável preditora. Com isso, podemos reescrever a transformação *logit* como demonstrado na equação (2.4)

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x + ... + \beta_{p-1}x_{p-1} \quad (2.4)$$

A única diferença será em $\beta' x = \beta_0 + \beta_1 x + ... + \beta_{p-1}x_{p-1}$. Desta forma, podemos reescrever a fórmula (2.2) como (2.5) para regressão logística múltipla.

$$\pi(x) = \frac{e^{\beta' x}}{1 + e^{\beta' x}} \quad (2.5)$$

A seção seguinte descreve como usamos regressão logística aplicando nossos dados.

2.2.2 Regressão Logística aplicada aos Métodos de Seleção de Palavra-chave

Como visto, a regressão logística é, de fato, um método de regressão e não de classificação. Ou seja, este método procura encontrar um valor real associado ao par documento-classe (por exemplo, a probabilidade da classe dado o documento) e não um valor discreto (a classe, dado o documento), como métodos tradicionais de classificação [Jin et al., 2007]. Neste método, uma função logística é usada para calcular a probabilidade de x ser a classe de y', conforme a Equação (2.6), onde $w = \beta_0 + \beta_1.z_1 + \beta_2.z_2 + ... + \beta_q.z_q$ para $y' = \{z_1, z_2, ..., z_q\}$. Os valores $\{\beta_0 + \beta_1 + \beta_2 + ... + \beta_n\}$ são coeficientes da regressão que indicam quão importante são as características $\{Z_1, Z_2, ..., Z_q\}$ para determinar a probabilidade de x_i ser a classe de y'.

$$P(x|y') = \frac{e^w}{1 + e^w} \quad (2.6)$$

Como estamos interessados na classe de y' e não na probabilidade $P(x|y')$, é necessário transformar a probabilidade $P(x|y')$ em uma classe. Assim, para determinar que classe x_j, em particular, deve ser atribuída para y_i, o seguinte processo é realizado. Primeiro, para cada classe x_j, todos os itens $y'_i \in Y'$ rotulados como x_j (ou seja, $f(y'_i) = (x_j)$) são pré-rotulados com 1. Os demais itens são rotulados com 0.

Então, dado um novo item candidato $y"_i \in Y'$, é calculado o valor da expressão de regressão logística tanto para os itens rotulados com 1 quanto para aqueles rotulados com 0. O valor para itens rotulados com 1 representa a probabilidade de x_j ser a classe de y'. Assim, dados os valores para todas as classes x_j, é escolhida a classe com maior valor.

Nós decidimos usar regressão logística como tarefa de classificação pois esta foi escolhida por Yih et al. [2006], trabalho que está sendo utilizado como linha de base

de nossos experimentos. Além disso, a exemplo do que fez Yih et al. [2006], testamos vários outros métodos, como Naive Bayes, árvore de decisão e SVM linear. Nossos experimentos só reforçaram o que Yih et al. [2006] já haviam concluído: que a regressão logística apresenta precisão satisfatória para esta tarefa com custo computacional baixo o suficiente para viabilizar a implementação de um sistema real.

2.3 Métricas e Técnicas de Avaliação

Para analisar os resultados do nosso modelo, foram adotadas métricas de avaliação de desempenho da área de recuperação da informação baseadas na noção de relevância. Esta é uma noção subjetiva relacionada com a idéia de que um tópico é importante para o documento selecionado. Assim, para determinar a relevância de um tópico para um documento, é necessário perguntar a opinião de seres humanos.

De posse desta opinião, é possível calcular a precisão de um método automático por determinar o quanto ele concorda com os avaliadores humanos. Mais precisamente, para um dado conjunto de itens recuperados, a precisão é definida como a proporção entre o número de itens relevantes recuperados e o número total de itens recuperados. Especificamente no nosso caso, a exemplo de experimentos anteriores de seleção de anúncios, a precisão é definida como o percentual médio de anúncios relevantes nos N primeiros anúncios retornados considerando um conjunto de n páginas de teste, como indicado na Equação (2.7).

$$P@N = \sum_{j=1}^{n} \frac{\#relevantes}{N} \tag{2.7}$$

Para garantir que os métodos foram treinados em conjuntos de dados não viciados foi utilizada a técnica de validação cruzada, que é descrita a seguir.

2.3.1 Validação Cruzada de k partições

Um valor estimado a partir de uma amostra de uma população representa o valor real (válido para toda a população) na medida em que a amostra é representativa. Para garantir a representabilidade, as amostras são tomadas aleatoriamente. Ainda assim, há sempre o risco que elas sejam viciadas. Para minimizar esta possibilidade, valores são tomados de múltiplas amostras aleatórias e não apenas de uma. O risco é menor ainda se as amostras usadas são independentes.

Uma metodologia de experimentação comumente usada em aprendizado de máquina e também utilizada por Yih et al. [2006] é a validação cruzada de k partições.

Nesta metodologia, valores são estimados a partir de k amostras aleatórias totalmente independentes entre si. Para conseguir isto, neste método os documentos são aleatoriamente divididos em k partições mutuamente exclusivas, de tamanho aproximadamente igual a n/k, onde n é o número de documentos na coleção. Com isso, são realizados k experimentos. Em cada experimento uma partição diferente é escolhida para o teste e as $k-1$ partições restantes são escolhidas para treino. A medida de eficiência é a média das medidas de eficiência calculadas para cada uma das partições. A maior vantagem deste método é que todos os documentos são usados tanto para treino quanto para teste.

No nosso trabalho consideramos k igual a 10 e utilizamos uma coleção composta por 300 páginas resultando em partições de tamanho 30. Estes valores representam um bom compromisso entre a garantia da obtenção de resultados estatisticamente confiáveis e a possibilidade de realização dos experimentos, na prática

2.4 Trabalhos Relacionados

A construção de sistemas de publicidade de busca e publicidade contextual motivaram uma série de pesquisas relacionadas aos seus vários componentes: subsistema de seleção e sugestão de termos, casamento de termos e propagandas, seleção de termos e veiculação de propagandas. As seções seguintes descrevem trabalhos em cada uma dessas linhas de pesquisa.

2.4.1 Casamento de termos e propagandas

Os primeiros problemas abordados foram aqueles relacionados ao casamento de termos e propagandas. Ribeiro-Neto et al. [2005] investigaram novas estratégias de casamento entre páginas e propagandas diminuindo a discrepância (*impedance mistatch*) entre o vocabulário da página e o vocabulário dos anúncios através da expansão do vocabulário da página apartir de páginas similares. Além destes métodos que utilizam expansão de página com páginas similares, eles propuseram métodos que utilizam o conteúdo da página. O melhor deles foi chamado AAK (*ads and keywords*). Esse método consiste no casamento entre o conteúdo da página (e somente com as palavras pertencentes à página) e o conteúdo do anúncio, incluindo título, descrição e palavra-chave. O método inclui um filtro adicional que considera relevantes apenas os pares (página-anúncio) em que a página possui todas as palavras encontradas na palavra-chave associada ao anúncio

Broder et al. [2007] tenta resolver este problema propondo um mecanismo de casamento que combina uma frase semântica com o casamento tradicional de palavras-chave, ou seja, uma frase sintática. A frase semântica classifica a página e os anúncios em uma taxonomia de tópicos e usa a proximidade de classe do anúncio e da página como um fator no *ranking* de anúncios. Dando sequência à pesquisa, Broder et al. [2008b] propõem um método para expandir as consultas e as propagandas utilizando conhecimentos adicionais a fim de criar atributos mais informativos para melhorar o casamento entre consulta e propaganda. Desta vez, eles utilizam resultados inicialmente retornados para a consulta e, a partir dela, criam um conjunto das palavras que ocorrem dentro das páginas respostas ou dos anúncios. Deste conjunto, selecionam os mais relevantes, utilizando o esquema de ponderação $TF - IDF$.

Ainda nesta linha de pesquisa, Radlinski et al. [2008] focou um outro aspecto do problema de casamento que é a seleção de anúncios com maiores valores de lance. Eles propuseram um método híbrido onde os maiores lances foram coletados de forma *offline* para então serem usados de forma *online* através do casamento tradicional entre anúncio e consulta.

Outra questão importante sobre publicidade *online* é selecionar anúncios considerando consultas raras e decidir se um anúncio é interessante o suficiente para ser mostrado numa página. Esse problema foi estudado por Broder et al. [2008a], Yih & Meek [2008] e Lacerda et al. [2006].

Broder et al. [2008a] investigaram o problema de predizer automaticamente se um anúncio ou um conjunto de anúncios é relevante o suficiente para ser mostrado em uma página. Dada uma consulta (ou página alvo) onde se deseja colocar anúncio, o objetivo é recuperar um conjunto de candidatos a anúncios ordenado pelos seus valores de similaridade. Para determinar se um anúncio é relevante ou não, eles utilizam duas abordagens. Na primeira, assumem que anúncios com alta similaridade têm mais chances de serem relevantes e aplicam um limiar de corte. Na segunda usam classificação supervisionada, onde avaliadores julgam a relevância dos anúncios, com base em algumas características e a média da relevância é calculada e comparada a um limiar.

Yih & Meek [2008] buscaram fornecer medidas de relevância consistentes para uma frase presente ou não no documento e ordená-las pela sua relevância para o documento. Alguns problemas existem ao considerar documentos como consulta e aplicar funções de *ranking* tradicionais, como o cosseno. Por exemplo, se um documento contém a frase "*Major League Baseball*" mas não "MLB", essas frases podem ter pontuações diferentes, mesmo sendo sinônimas. Para resolver estes problemas eles propuseram duas abordagens. A primeira é um método baseado em similaridade cujo objetivo é usar expansão de consulta para identificar um conjunto de palavras relacionadas para

representar a semântica da frase que servirá de consulta. Um inconveniente desta abordagem é a impossibilidade de explorar características importantes da frase e que proporcionam medidas mais exatas, como por exemplo: se a frase está no texto de âncora ou se está no título. Para lidar com este inconveniente, a segunda abordagem propõe pegar como entrada a relevância estimada de algumas frases no documento e usar regressão de processo gaussiano para melhorar a precisão ao predizer a pontuação de uma determinada frase fora do documento. As abordagens foram avaliadas no cenário de publicidade *online* onde foi julgada a relevância da palavra-chave para a página contendo uma determinada propaganda. Para medir a relevância dos métodos, foram utilizadas várias métricas, tais como precisão@k e acurácia. Eles concluiram que as duas abordagens têm ganhos significativos (entre 10 e 20%) sobre o método utilizado como linha de base, mas destacaram que os resultados da regressão têm melhor desempenho em relação à abordagem baseada em similaridade.

Já Lacerda et al. [2006] propuseram um método de aprendizado para determinar quando um anúncio deve ser mostrado. Esse método é baseado em programação genética e um dos aspectos a ser aprendido é justamente o nível de relevância do anúncio para a página. Os resultados mostraram que ao utilizar programação genética houve uma melhora significativa quando comparado com os melhores métodos (sem expansão de vocabulário) relatados em Ribeiro-Neto et al. [2005].

2.4.2 Subsistema de seleção e sugestão de termos

Entre os trabalhos relacionados à sugestão de termos temos o de Chen et al. [2008]. Eles tentam diminuir a distância entre as palavras-chave escolhidas pelos anunciantes e as consultas realizadas nas máquinas de busca, encontrando novas palavras-chave relevantes baseadas em dados estatísticos, por exemplo, na co-ocorrência de uma palavra-chave. Eles propõem um novo método de sugestão de palavra-chave que explora o conhecimento semântico entre conceito de hierarquia. Dada uma palavra-chave, primeiramente são combinados alguns conceitos relevantes. Em seguida, os conceitos são utilizados com os seus superiores hierárquicos para gerar os significados das palavras. Finalmente, novas palavras-chave são sugeridas de acordo com o conceito. Como contribuição, o trabalho apresenta uma nova abordagem de sugestão de palavra-chave em que se tenta evitar termos ambíguos. Já Wu & Bolivar [2008] selecionam um conjunto de características que eles julgam ser útil para determinar a importância das palavras-chave candidatas. Por último, usam regressão logística para determinar de forma automática as palavras-chave. Depois disso, usam um método para eliminar ambigüidade entre as palavras.

2.4.3 Seleção de termos e veiculação de propagandas

Nosso trabalho está mais relacionado ao problema de seleção de termos, mais especificamente aos trabalhos propostos por Yih et al. [2006] e Goodman & Carvalho [2005] que focam a tarefa de selecionar palavras-chave de textos de mensagens de correio eletrônico e páginas web com o objetivo de selecionar anúncios. Yih et al. [2006], descrevem um sistema que aprende como extrair palavras-chave de páginas web para publicidade direcionada. Para tanto, os autores usaram regressão logística para o aprendizado. O sistema proposto utiliza-se de uma variedade de características para determinar a importância de uma palavra-chave para a publicidade da página, tais como: a freqüência de cada termo candidato a palavra-chave, o quão raro é o termo na coleção, a presença ou não na seção de *meta* dados e a freqüência que o termo ocorre no *log* de consulta da pesquisa. O sistema foi treinado com um conjunto de 828 páginas web com publicidade contextual, coletadas aleatoriamente, com no máximo uma página por domínio. As páginas foram divididas para 5 avaliadores que rotularam manualmente as palavras-chave relevantes (usando uma abordagem humana). Uma palavra só poderia ser relevante se ela aparecesse na página. Anteriormente a Yih et al. [2006], Goodman & Carvalho [2005] aplicaram técnicas similares para sistemas de consulta implícita para mensagens de correio eletrônico com o objetivo de encontrar boas palavras ou frases no texto da mensagem e depois enviar para uma máquina de busca e obter boas publicidades. O diferencial deste trabalho é que foram utilizadas características específicas para mensagens de correio eletrônico, como, a ocorrência da palavra no assunto da mensagem de correio eletrônico, a quantidade de vezes que uma palavra ocorre no texto da mensagem de correio eletrônico ou no assunto, quantas vezes a palavra ocorre no início de uma frase, se o assunto começa com "FW" ou "RE", entre outras. O sistema foi treinado com 1143 mensagens (de correio eletrônico) e 5 avaliadores. Para cada avaliador, foi pedido que rotulasse manualmente as palavras, presentes no texto, que eles julgavam interessantes para uma máquina de busca. Para fazer a seleção automática foi utilizado o método de regressão logística.

Ao contrário destes trabalhos e de outras abordagens encontradas na literatura para selecionar palavras-chaves de páginas web, guiamos o processo de aprendizagem usando a abordagem de seleção de palavras-chave baseada em precisão. Além desta importante diferença, exploramos novas características derivadas da coleção de anúncios que não foram consideradas por Yih et al. [2006] e Goodman & Carvalho [2005].

No próximo capítulo são descritas as características utilizadas neste trabalho, tanto as derivadas da coleção de páginas - se a palavra está em maíusculo, se faz parte da url da página, a frenquência do termo no documento, entre outras - quanto as

derivadas da coleção de anúncios, como por exemplo, frequência do termo no título, descrição e palavra-chave do anúncio. Além de apresentar a abordagem utilizada para selecionar as palavras-chave candidatas e as estratégias de definição de relevância de uma candidata a palavra-chave para publicidade contextual.

Capítulo 3

Seleção de Palavras-chave

Neste capítulo apresentamos as abordagens adotadas para selecão e definição de relevância de uma candidata a palavra-chave para publicidade contextual, bem como as características utilizadas para representar tais palavras.

3.1 Regressão Logística como um Problema de Classificação

Neste trabalho, a exemplo de Yih et al. [2006], tratamos o problema de determinar a relevância de uma palavra-chave (sequência de termos) para publicidade como um problema de classificação. Com isso, seja $K = \{k_1, k_2, ..., k_n\}$ um conjunto de palavras-chave candidatas. Cada palavra-chave k_i é representada por um conjunto de m características $F = \{F_1, F_2, ..., F_m\}$, tais que $k_i = (f_{i1}, f_{i2}, ..., f_{im})$ é um vetor que representa k_i, onde cada f_{ij} é o valor da característica F_j na palavra-chave k_i. A Figura 3.1 ilustra esta forma de representação da base de dados onde cada uma das 300 páginas Web pertencentes a nossa base de dados é composta por um conjunto de n palavras-chave candidatas (identificadas na definição pela letra K, do inglês *keyword*) que, por sua vez, são representadas por um conjunto de m características que estão identificadas na definição pela letra F (do inglês *features*).

É importante ressaltar que o termo *característica* descreve uma estatística que representa a medida de algum indicador de relevância de publicidade para uma palavra-chave candidata. Por exemplo, se acreditamos que a ocorrência de um nome próprio aumenta as chances de uma candidata ser uma palavra-chave, um indicador de capitalização do termo pode ser tomado como característica de interesse. A intuição é que se o termo está capitalizado, tem mais chances de ser nome próprio e, portanto, ser palavra-

chave. Ou ainda, se acreditamos que o fato de uma palavra pertencer a URL da página pode aumentar as chances desta ser uma candidata a palavra-chave, um indicador que guarde esta informação pode ser, também, uma característica interessante.

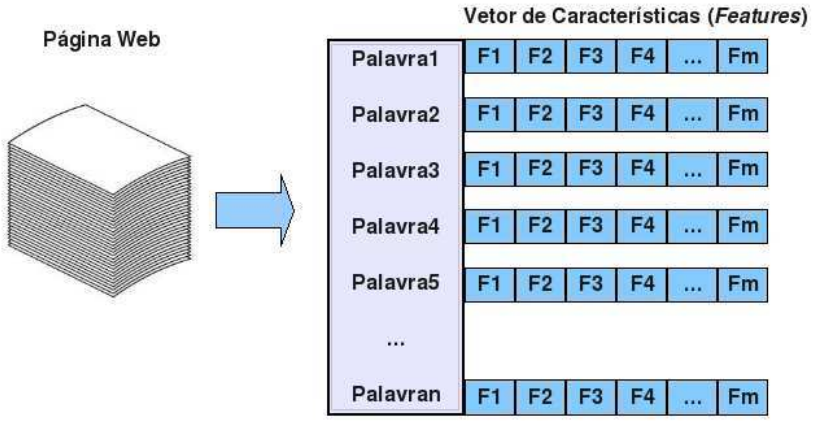

Figura 3.1. Forma de representação das páginas Web. Cada página é representada por um conjunto de n palavras-chave candidatas e cada palavra-chave candidata é descrita por um conjunto de m características.

Assumimos que temos acesso a *dados de treino* na forma de $\{(k_1, r_1), (k_2, r_2), ..., (k_n, r_n)\} \subset K \times \{0, 1\}$, onde cada par (k_i, r_i) representa uma candidata a palavra-chave e seu valor de relevância correspondente, de forma que, se $r_i = 1$, então a candidata k_i é uma palavra-chave. Caso contário, não é uma palavra-chave.

Usando essa abordagem de aprendizagem, a solução desse problema consiste em:

- determinar o conjunto de características $\{F_1, F_2, ..., F_m\}$ usado para representar uma palavra-chave candidata em K; e
- aplicar o método de classificação para encontrar a melhor combinação de características para predizer o valor de relevância r_i para uma palavra-chave k_i.

3.2 Seleção de Palavras-Chave Candidatas

Para selecionar palavras-chave candidatas, usamos a mesma configuração do seletor de candidatas monolítico combinado, descrito em Yih et al. [2006], visto que este foi o melhor seletor de palavras-chave encontrado naquele trabalho.

No seletor de candidatas monolítico combinado, uma palavra-chave candidata é qualquer palavra ou frase (composta por até 5 palavras consecutivas) que pertença

ao conteúdo da página, incluindo as seções de título, de corpo e das *meta-tags* da página. Neste seletor, ainda que uma palavra apareça em fragmentos diferentes da página, ela é considerada uma única candidata. Mesmo que, por exemplo, a palavra "tintos e brancos" apareça no título e no corpo de uma página, ela será considerada uma única candidata. Mas, se mudar a ordem das palavras na frase estas são consideradas diferentes. Por exemplo, se no título da página aparece a candidata "tintos e brancos" e no corpo da página, aparece "brancos e tintos", as duas ocorrências são consideradas candidatas diferentes.

Frases não são selecionadas como candidatas se possuem palavras que estejam no fim de uma sentença e no inicio de outra, ou seja, se ultrapassam limites de sentenças ou de blocos. Por exemplo, ao analisar o seguimento de texto "A uva Malbec é muito melhor na Argentina do que na França. Em seu país de origem...", a frase "França em seu pais" não será relacionada como candidata. Isto ocorre porque a palavra "França" pertence a uma sentença enquanto "em seu pais" pertence a outra sentença da mesma página.

Além disso, as frases são vistas como entidades individuais, tal que características não descrevem estatísticas sobre as palavras que constituem a frase, mas sim sobre a frase como um todo.

Outro ponto importante é que não tratamos palavras reduzidas a seu radical (*stemming*) nem tampouco foram retiradas as *stopwords*. O único tratamento de exclusão de *stopword* feito foi garantir que uma frase não é selecionada se ela começar ou terminar com uma *stopword*. Ainda considerando o trecho da página "A uva Malbec é muito melhor na Argentina do que na França. Em seu país de origem...", a frase "a uva malbec" não será relacionada como candidata pois começa com a palavra "a" que é uma *stopword*. Já a frase "argentina do que na franca", apesar de possuir *stopwords* pode ser uma candidata pois não começa nem termina com *stopword*.

3.3 Palavras-Chave Relevantes

Como mencionado anteriormente, assumimos que os dados de nossa coleção de treino têm a forma $\{(k_1, r_1), (k_2, r_2), ..., (k_n, r_n)\}$, onde cada par (k_i, r_i) representa uma palavra-chave candidata e seu valor de relevância correspondente. Para obter o valor de relevância r usamos duas estratégias.

Na primeira estratégia, pessoas escolhem as palavras-chave, como proposto em [Yih et al., 2006]. Esta primeira estratégia é usada como linha de base de nossa pesquisa. Já na segunda, as candidatas foram tomadas como palavras-chave de acordo

com sua capacidade de associar anúncios relevantes. Essa segunda estratégia é a nossa proposta para selecionar palavras-chave relevantes.

Nas duas estratégias utilizamos uma coleção de teste composta por um conjunto de 300 páginas web extraídas de um portal de notícias brasileiro contendo vários assuntos, tais como: cultura, notícias nacionais e internacionais, esporte, economia, carro, infantil e computadores.

Na primeira abordagem, solicitamos que 60 usuários rotulassem as palavras ou frases pertencentes a cada página como relevante ou não para publicidade na coleção de teste. Cada página foi avaliada por três usuários diferentes. E, como já mencionado, uma candidata pode ser qualquer palavra ou frase com até 5 palavras que estejam presentes na página. A Figura 3.2 ilustra este processo de avaliação de uma palavra-chave candidata, onde cada palavra pode receber nenhum, um, dois ou três votos de relevância. Com isso, dada uma candidata k_i, ela é considerada relevante ($r_i = 1$) se, pelo menos, um usuário a julgou como palavra-chave. Caso contrário, ela é considerada não relevante ($r_i = 0$). De agora em diante, nos referiremos às palavras-chave selecionadas com essa estratégia como *seleção manual*.

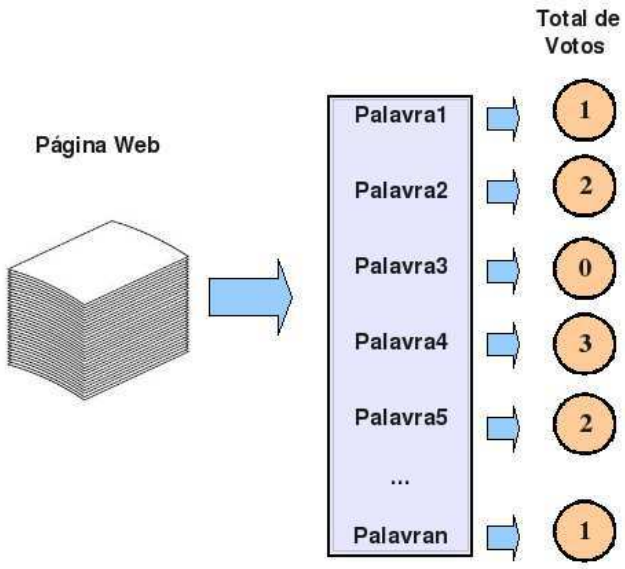

Figura 3.2. Processo de avaliação da abordagem de seleção manual. Os usuários avaliam diretamente a relevância de uma palavra-chave candidata para a página.

A segunda abordagem, que é a nossa proposta, está sendo chamada *seleção por precisão*. Nela, nós recuperamos para cada candidata a palavra-chave nas páginas de treino, os anúncios mais similares a ela. Para isso, nós indexamos os anúncios presentes na coleção de anúncios e realizamos uma consulta sobre esse índice para cada candidata

encontrada na coleção de páginas. A partir de então, selecionamos os cinco anúncio mais similares a cada candidata da página.

Uma vez que recuperamos os anúncios mais similares a cada página, pedimos que os usuários avaliassem a relevância de cada anúncio para a página de onde a candidata a palavra-chave foi extraída. Novamente cada uma das 300 páginas da coleção de teste foi avaliada por 3 usuários garantindo, com isso, que cada anúncio recebesse três avaliações. Mais especificamente, para uma determinada candidata a palavra-chave k_i, são retornados cinco anúncios de acordo com sua similaridade com k_i. A Figura 3.3 ilustra os anúncios retornados para cada palavra da página e a precisão desses anúncios, ou seja, se pelo menos um avaliador julgou pelo menos um anúncio relevante para aquela página a precisão da palavra que retornou aquele anúncio é $\frac{1}{5}$, porém, se os 5 anúncios foram considerados relevantes por, pelo menos, um avaliador a precisão desta palavra é 1. A candidata é considerada como relevante ($r_i = 1$) se pelo menos um dos anúncios retornados é considerado relevante para ser mostrado na página por, no mínimo, um avaliador. Caso contrário, ela é considerada irrelevante ($r_i = 0$).

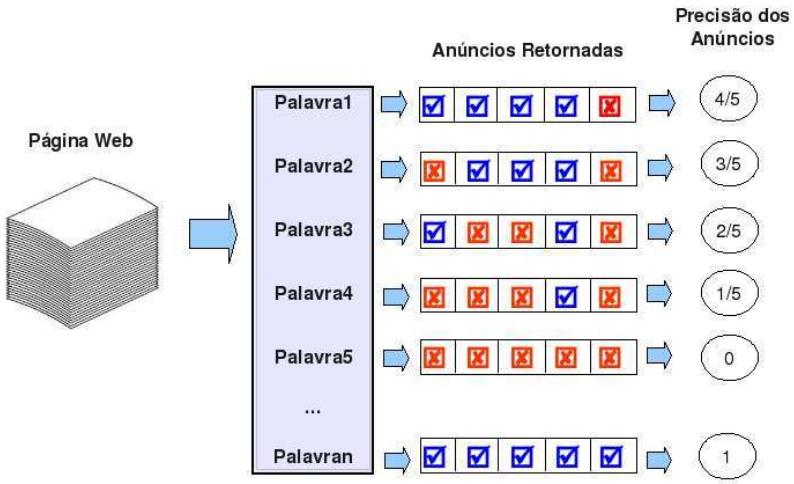

Figura 3.3. Abordagem de seleção por precisão. Para cada palavra-chave candidata são retornados os cinco anúncios mais relevantes. Estes anúncios são avaliados por três avaliadores diferentes e em seguida é calculada a precisão para cada candidata.

Apesar do método que estamos propondo ser, também, baseado no aprendizado da avaliação de relevância humana, estamos utilizando outro tipo de informação: os anúncios mais similares às palavras ou frases que compõem a página. Nossa intuição é que é mais fácil para os seres humanos selecionarem os anúncios que eles acham relevantes para serem mostrados numa página do que determinarem palavras-chave relevantes para aquela página.

Baseado em uma análise amostral dos nossos dados, acreditamos que alguns fatores podem contribuir para erros quando humanos julgam a relevância de uma palavra-chave, como segue:

- É natural que pessoas leiam um texto e não dêem importância às candidatas que estejam em fragmentos de menor importância para a página;
- Algumas páginas podem descrever assuntos que não fazem parte do conhecimento do avaliador. Com isso, é possível que o mesmo não tenha conhecimento suficiente para escolher as melhores palavras-chave para a página. Apesar de que, o fato do usuário não conhecer o assunto da página pode dificultar também na escolha de boas propagandas;
- O avaliador não tem conhecimento sobre a base de dados de anúncio, em especial, quanto ao seu vocabulário e às oportunidades de publicidade.

Além disso, uma vantagem natural desta nova abordagem é que as coleções de referência adotadas para avaliar o desempenho dos sistemas de anúncios contém as informações necessárias para o treino, embora para criar tais coleções de referência seja necessário avaliar a relevância dos anúncios dados a uma página web. Coleções de referência são também avaliadas se o sistema de anúncios usa abordagens de *learn-to-advertise* [Lacerda et al., 2006]. Assim, em casos reais, a mudança do foco na seleção de palavras-chave pode reduzir o custo do treino. Além do que, a informação de taxa de clique pode ser usada como uma aproximação da relevância do julgamento humano para os anúncios, uma vez que, para a maioria das palavras-chave, esta informação já está disponível, especialmente para as empresas que operam com sistemas de busca patrocinada.

Ainda assim, alguém poderia argumentar que, se uma coleção de referência não estivesse disponível, o custo de treinar utilizando o seletor de palavras-chave por precisão seria maior que rotular diretamente palavras-chave, como ocorre no sistema que estamos utilizando como linha de base. Porém, ainda que esta situação fosse real, este custo não seria maior já que na seleção manual um grande número de candidatas a palavra-chave não casa com as palavras-chave da coleção de anúncios, um fenômeno que é conhecido na literatura como impedância entre o vocabulário da página e os termos encontrados nos sistemas de anúncio [Ribeiro-Neto et al., 2005; Yih et al., 2006].

Na abordagem de seleção de palavras-chave por precisão é necessário recuperar os anúncios mais similares a cada candidata a palavra-chave. Para esta tarefa, usamos o Modelo de Espaço Vetorial [Salton et al., 1974]. Como citado anteriormente, nesse

modelo, anúncios e palavras-chave candidatas são representadas como vetores em um espaço de termos indexados.

Apesar de podermos adotar um método de ordenação de anúncios mais sofisticado, optamos por utilizar o Modelo de Espaço Vetorial por já ter sido usado na literatura para o mesmo problema. Além disso, nosso objetivo maior é validar nosso método de seleção de palavras-chave e não encontrar a forma mais eficiente de ordenar anúncios para uma página.

Nas seções anteriores vimos que as páginas são representadas por palavras-chave candidatas. Cada candidata é representada por um conjunto de características. A seção seguinte descreve cada uma das características utilizadas neste trabalho.

3.4 Representação de Palavra-Chave

Nessa seção, descreveremos as características usadas para representar uma palavra-chave. Tais características estão divididas em dois grupos, de acordo com sua origem: página & *log* e coleção de anúncios.

3.4.1 Características da Página & *Log*

Essas são características extraídas do conteúdo textual das páginas e do *log* de consulta. Foram originalmente propostas e estudadas por Yih et al. [2006]. Do conjunto de características usado naquele trabalho nós não utilizamos as informações linguísticas. Como observado pelos autores [Yih et al., 2006], tais características não ajudam na tarefa de extrair palavras-chave pois fornecem informações redundantes quando usadas em conjunto com outras características, como por exemplo, a capitalização e a presença no *log* de consultas.

As características da página e do *log* estão organizadas em diversos grupos, como descritos nas seções a seguir.

3.4.1.1 Capitalização

Esta característica indica se a palavra-chave está capitalizada. A capitalização pode indicar se a palavra-chave faz parte de um nome próprio, se é a primeira palavra de uma frase ou se é uma palavra importante para o texto.

3.4.1.2 Hipertexto

Esta característica diferencia se a frase ou palavra faz parte do texto de âncora.

3.4.1.3 Características da Seção Meta

O cabeçalho de um documento HTML pode fornecer informações adicionais presentes na *tag* meta. Embora o texto presente nesta região normalmente não possa ser visto pelos navegadores, se uma candidata aparece nesta seção tem grande chance de ser importante para a página. Por isso, uma das características utilizadas é se uma candidata faz parte da seção de metadados do documento HTML.

3.4.1.4 Título

O único texto presente no cabeçalho HTML que o usuário pode ler é o título, que normalmente é mostrado na página pelo navegador. A característica utilizada é se a candidata faz parte do título da página.

3.4.1.5 Características Meta

Assim como o título, diversas *meta tags* estão potencialmente relacionadas à palavra-chave, e são usadas para obter características, como a *meta-description*, *meta-keywords* ou *meta-title*.

3.4.1.6 URL

Um documento HTML tem o seu nome como uma propriedade amplamente utilizada, conhecido, também, como URL. Como característica estamos considerando se uma candidata faz parte do URL da página.

3.4.1.7 Características de Recuperação de Informação

Usamos as duas estatísticas mais comumente usadas em recuperação de informação, ou seja, o valor da frequência do termo (TF, *term frequency*) e da frequência do documento (DF, *document frequency*) da candidata. A frequência do documento é o número de documentos na coleção de páginas web que contém a candidata. Este valor foi obtido através de consulta a uma máquina de busca e o número de páginas retornadas foi utilizado como DF. Além destas características, utilizamos também seus logarítimos na forma $log(TF + 1)$ e $log(DF + 1)$.

3.4.1.8 Localização da candidata no conteúdo da página

O início de um documento geralmente contém uma introdução com palavras e frases importantes. Portanto, o local da ocorrência da candidata é extraído como uma

característica. Visto que o comprimento de um documento varia substancialmente, usamos a posição relativa considerando um documento normalizado de tamanho igual a 1. Quando a candidata é uma frase, a primeira palavra é utilizada como localização. Foram utilizadas três características diferentes relativas à localização:

- *wordRatio*: a localização relativa da candidata na frase;
- *sentRatio*: a localização da frase a candidata dividida pelo total de sentenças no documento;
- *wordDocRatio*: a localização relativa da candidata no documento.

Em adição a estas três características, nós utilizamos, também, seus logarítmos. Mais especificamente $log(1+wordRatio)$, $log(1+sentRatio)$ e $log(1+wordDocRatio)$.

3.4.1.9 Tamanho da sentença e do documento

Também consideramos características pertinentes ao tamanho tanto da frase quanto do documento. Neste sentido foram utilizadas duas características:

- *sentLen*: tamanho (em palavras) da frase onde ocorre a candidata;
- *docLen*: tamanho do documento, excluindo as palavras que pertencem ao cabeçalho;

Similarmente, $log(1 + sentLen)$ e $log(1 + docLen)$ também são utilizados.

3.4.1.10 Tamanho da frase candidata

O tamanho da frase candidata ($phLen$) em palavras e $log(1 + phLen)$ são incluidas como características.

3.4.1.11 *Log* de Consulta

O *log* de consulta de uma máquina de busca reflete a distribuição de palavras-chave que as pessoas acham mais interessantes. Nós usamos a informação para definir três características:

- Se a frase aparece no *log* de consulta;
- A frequência em que a palavra aparece no *log* de consulta e;
- O logarítmo da frequência $log(1 + frequency)$.

Neste trabalho, nós usamos o *log* de consulta descrito no Capítulo 4, Seção 1.

3.4.2 Características da Coleção de Anúncios

Características da coleção de anúncios sao extraídas de uma base de dados de anúncios. Até onde sabemos, esta é a primeira vez que características da coleção de anúncios são utilizadas em método de aprendizagem aplicado à seleção de palavras-chave.

Neste trabalho, consideramos que um anúncio é composto por três partes estruturais: um título, uma descrição textual e um hiperlink.

Afim de melhorar a representação de uma palavra-chave candidata e utilizar os benefícios das estatísticas presentes na base de anúncios, estamos propondo as características descritas nas seções seguintes. Essas características foram escolhidas devido ao sucesso da utilização da frequência dos termos em objetos (TF) e ao número de objetos onde um termo ocorre (DF) como características em trabalhos anteriores relacionadas à tarefa de recuperação de informação [Lacerda et al., 2006] [A. A. Veloso & Jr., 2008].

3.4.2.1 TF na Seção do Anúncio

Frequência da candidata em cada uma das seções estruturais de um anúncio. Visto que um anúncio tem três seções, nós usamos três características para representá-las:

- TF no título do anúncio;
- TF na descrição do anúncio e;
- TF na palavra-chave do anúncio.

3.4.2.2 TF Máximo na Seção do Anúncio

Frequência máxima da candidata em cada uma das seções estruturais do anúncio. Quanto ao *TF na seção do anúncio*, usamos três características para representar cada seção:

- TF máximo no título do anúncio;
- TF máximo na descrição do anúncio e;
- TF máximo na palavra-chave do anúncio.

3.4.2.3 TF Médio na Seção do Anúncio

Frequência média da candidata nas três seções do anúncio:

- TF médio no título do anúncio;

- TF médio na descrição do anúncio e;
- TF médio na palavra-chave do anúncio.

3.4.2.4 DF na Seção do Anúncio

Número de anúncios onde a candidata ocorre em uma determinada seção do anúncio. As três características são:

- DF no título do anúncio;
- DF na descrição do anúncio e;
- DF na palavra-chave do anúncio.

3.4.2.5 TF Máximo na Seção da Campanha

Frequência máxima da candidata na seção estrutural de todos os anúncios da campanha. As três características neste caso são:

- TF máximo no título da campanha;
- TF máximo na descrição da campanha e;
- TF máximo nas palavras-chave da campanha.

3.4.2.6 TF Médio na Seção da Campanha

Frequência média da candidata em cada uma das seções estruturais de todos os anúncios da campanha. As três características neste caso são:

- TF médio no título da campanha;
- TF médio na descrição da campanha e;
- TF médio nas palavras-chave da campanha.

3.4.2.7 DF na Seção da Campanha

Número de campanhas em que a candidata ocorre em uma determinada seção do anúncio. As três características neste caso são:

- DF no título da campanha;

- DF na descrição da campanha e;
- DF nas palavras-chave da campanha.

Até então conhecemos as abordagens de seleção de palavras-chave, como as páginas são representadas e as características que compõem cada candidata a palavra-chave. No próximo capítulo descreveremos as coleções de dados que estamos utilizando, nossa metodologia e os resultados obtidos com as abordagens de seleção de palavras-chave.

Capítulo 4

Experimentos

Nesse capítulo descreveremos nosso conjunto de dados, a metodologia experimental usada para conduzir nosso estudo empírico e os resultados obtidos.

4.1 Bases de Dados usadas nos Experimentos

Para treinar e avaliar nosso modelo de seleção de palavras-chave para escolher bons anúncios, usamos uma coleção construída com um conjunto de 300 páginas extraídas de um portal de notícias brasileiro. Como não tivemos preferência por um tópico específico, essas páginas possuem diversos assuntos tais como cultura, notícias locais, notícias internacionais, economia, esporte, política, agricultura, carro, infantil, computadores e Internet.

A informação de IDF *(Inverse Document Frequency)* usada foi obtida de uma máquina de busca comercial. Submetemos cada palavra-chave candidata como uma consulta à máquina de busca e o número de documentos retornados foi considerado como DF (*Document Frequency*). Então, computamos o IDF como $log(\frac{DF(kw)}{N})$, onde N é o total de documentos encontrados na coleção da máquina de busca e kw é a palavra-chave candidata.

Os anúncios utilizados em nossos experimentos foram obtidos de uma coleção real de anúncios composto por 93.972 anúncios agrupados em 2.029 campanhas fornecidas por 1.744 anunciantes. A esses anúncios, anunciantes associaram um total de 68.238 palavras-chave. Nessa coleção somente uma palavra-chave é associada para cada anúncio.

A coleção adotada não contém informações de taxa de clique e também não contém uma lista de anúncios relevantes que possa ser associada a cada página. Assim, foi necessário realizar uma avaliação de anúncios para fornecer tais informações, per-

mitindo aplicar nosso método e avaliar o desempenho do nosso sistema de seleção de anúncios nessa coleção.

A fim de preparar o treino para nosso método, indexamos os anúncios da coleção e realizamos uma consulta sobre esse índice para cada conjunto distinto de palavras-chave candidatas encontrada nas 300 páginas. As cinco primeiras respostas foram utilizadas para compor o conjunto de pares constituído de páginas e anúncios para ser avaliado por 75 voluntários. Para cada par (*p, ad*), foi solicitado a três julgadores que informassem se o anúncio *ad* é relevante para a página *p*. É importante ressaltar que esse esforço seria desnecessário se a informação de taxa de clique estivesse disponível para a coleção de anúncios adotada ou mesmo se nós tivéssemos uma coleção de referência com um conjunto completo dos anúncios relevantes para cada página da coleção.

Para obter o conjunto de anúncios relevantes para nossa coleção de teste, usamos o mesmo conjunto de métodos usado para avaliar a coleção baseada na Web TREC (*Text Retrieval Conference*) [Hawking & Craswell, 1998]. Em nosso trabalho, para cada uma das 300 páginas de teste, nós selecionamos os cinco primeiros anúncios retornados para cada palavra-chave candidata encontrada na página.

Como resultado do método que gerou o conjunto de anúncios para serem avaliados, selecionamos um total de 95.327 pares de anúncio-página distintos. Assim, obtivemos cerca de 317 anúncios avaliados por página. Eles foram inseridos em conjuntos correspondentes a cada página de teste.

Todos os anúncios foram submetidos à avaliação manual. A cada voluntário foi pedido para avaliar os anúncios de acordo com a relevância para a página. A avaliação foi realizada de forma que cada par recebesse três avaliações. Um par (*p, ad*) foi considerado relevante se pelo menos um voluntário avaliou o anúncio como relevante. Nós também apresentamos o cenário em que um anúncio é considerado relevante quando recebe voto de pelo menos dois voluntários. O número médio de anúncios relevantes por página foi 41,83 e o número médio de palavras-chave candidatas encontradas foi 279,76.

Nossa característica referente ao *log* de consulta foi derivada do *log* de consulta da coleção WBR03, uma base de dados extraída da web brasileira que contém consultas submetidas ao TODOBR, uma máquina de busca real. O *log* contem 12.795.101 consultas submetidas em 2002, para um total de 2.987.745 consultas distintas.

4.2 Metodologia de Avaliação

Para executar os experimentos, foi usado o método de validação cruzada de 10-partições [Mitchell, 1997; Stone, 1974]. Nossa coleção de 300 páginas foi dividida em 10 partições diferentes e foram realizados 10 experimentos distintos. Para cada experimento, uma partição diferente foi usada para teste. Desta forma, todos os resultados relatados são valores médios da execução das 10 partições.

Além de utilizar validação cruzada para evitar vícios na aprendizagem do método utilizamos, em todas as comparações relatadas nesse trabalho, o teste T de *Student* [Mitchell, 1997] para determinar se a diferença observada no desempenho é estatisticamente significante. Consideramos estatisticamente significantes resultados com valor-p $\leq 0{,}02$.

Medimos a qualidade da palavra-chave selecionada em dois experimentos distintos. Primeiro, medimos a qualidade do nosso classificador de palavras-chave. Usamos a medida de precisão, que é definida como a proporção de exemplos classificados corretamente para esse propósito (conforme Seção 2.3).

Nosso segundo experimento foi avaliar a qualidade dos anúncios retornados pelas palavras-chave para cada abordagem. Cada conjunto de palavras-chave retornado para cada método é usado como uma consulta submetida ao sistema que retorna um *ranking* dos anúncios baseados na palavra-chave dada como entrada. O algoritmo de *ranking* que adotamos em nossos experimentos é o algoritmo AAK, proposto por Ribeiro-Neto et al. [2005].

Esse segundo experimento visa avaliar o impacto das duas estratégias de seleção de palavra-chave estudadas, seleção manual e baseada em precisão, quando usadas no sistema de seleção de anúncios. Os experimentos da pesquisa foram avaliados usando P@3, que expressa o percentual médio de anúncios relevantes nos três primeiros anúncios retornados para cada método. Essa métrica é adotada também em experimentos anteriores de seleção de anúncios como Ribeiro-Neto et al. [2005].

Como a avaliação para julgar a relevância dos anúncios foi feita sobre os anúncios retornados pelas palavras ou frase candidatas individualmente, no experimento final é possível que apareçam novos pares de anúncios e páginas não avaliados em nosso conjunto inicial usado para obter anúncios relevantes, visto que, desta vez submetemos uma consulta com o conectivo "OR" com todos os termos retornados pelo classificador. Deste modo, uma segunda rodada de avaliação foi necessária para completar o conjunto de anúncios relevantes. Nessa segunda rodada encontramos uma média de 1,82 anúncios não avaliados por página. Após a avaliação, encontramos uma média de 0,83 anúncios relevantes por página.

4.3 Resultado dos Experimentos

Nesta seção, apresentamos os resultados dos experimentos realizados para avaliar o método proposto e a comparação com o trabalho encontrado na literatura que estamos usando como base para nossos experimentos.

Nos resultados que serão apresentados nesta seção comparamos os seguintes métodos:

- **selecaoPorPrecisaoPL:** abordagem de seleção de palavras-chave baseada em precisão onde as palavras-chave são representadas usando características da página e do *log* de consulta;
- **selecaoPorPrecisaoPLAI:** equivalente ao anterior, mas com palavras-chave sendo representadas por características da coleção de anúncios, além das derivadas da página e do *log*;
- **selecaoManualPL:** abordagem de seleção de palavras-chave baseada no julgamento humano onde as palavras são representadas usando características da página e do *log* de consulta;
- **selecaoManualPLAI:** equivalente ao anterior, mas com palavras-chave sendo representadas por características da coleção de anúncio, assim como características derivadas da página e do *log* de consulta;
- **selecaoPorPrecisaoIDEAL:** abordagem de seleção de palavras-chave que classifica como palavras-chave todos e somente os termos considerados relevantes pela abordagem baseada em precisão;
- **selecaoManualIDEAL:** abordagem de seleção de palavras-chave que classifica como palavras-chave todos e somente os termos rotulados como relevantes pelos avaliadores humanos.

Quando nos referimos ao termo "ideal", queremos dizer que a abordagem atinge a precisão máxima, visto que, considera como palavras-chave todos e somente os termos retornados pelo método, desconsiderando qualquer falha do classificador.

Para confirmar nossa intuição inicial realizamos dois experimentos. O primeiro, com o objetivo de avaliar a qualidade dos métodos de seleção, tendo o julgamento humano como balisador para considerar uma palavra-chave relevante. O segundo tem por objetivo mostrar a precisão média para os três primeiros anúncios retornados. Estes experimentos são descritos em detalhe nas próximas seções.

4.3.1 Qualidade dos métodos considerando o julgamento humano

Nestes experimentos medimos a qualidade do anúncio de acordo com os termos que os avaliadores julgaram ser relevantes para uma página.

Considerando este cenário, é esperado que o método manual alcance uma melhor precisão, visto que, neste método as palavras-chave são aprendidas utilizando como conjunto de treino exatamente as candidatas rotuladas como relevantes pelos avaliadores. Já na segunda abordagem, a seleção por precisão, as palavras foram selecionadas com base em seu desempenho ao escolher anúncios, razão que nos leva a acreditar que, neste experimento, o método manual terá um desempenho superior.

Nossos experimentos mostraram que o conjunto de palavras-chave gerado pelo método manual é bem diferente do conjunto de palavras-chave gerado pelo método de seleção por precisão. Isto é, os conjuntos de palavras-chave resultantes do treino têm pouca interseção. Das 10444 candidatas a palavras-chave usadas para treinar os métodos somente 10% delas foram consideradas palavras-chave por ambos os métodos.

A Tabela 4.1 apresenta os resultados da precisão nas abordagens de detecção de palavras-chave estudadas. Nela podemos observar que apesar de haver pouca interseção entre as palavras-chave rotuladas como relevantes nos conjuntos de treino usados pelos métodos, o desempenho apresentado é muito parecido. Após inspecionar as palavras-chave usadas para treinar e as selecionadas pelo método no teste, observamos que no teste havia duas vezes mais palavras em comum entre os métodos que no treino. Apesar disso, a diferença nas saídas dos métodos continua grande. Por exemplo, considerando o conjunto de palavras-chave retornadas pela abordagem *manualPL* e o conjunto de palavras-chave retornadas pela abordagem *porPrecisaoPL*, 80% são diferentes. Entre os exemplos de palavras-chave selecionados somente pela abordagem manual, podemos citar "carroça", "menina prodígio" e "tarso genro". Essas palavras-chave, em geral, não foram encontradas ou tiveram pouca importância na coleção de anúncios. Da mesma forma, exemplos de palavras-chave selecionadas somente pela abordagem por precisão são "relacionamento", "time" e "obra de arte". Essas palavras normalmente estão em áreas de menor importância nas páginas, de forma que, o usuário não percebeu sua relevância.

No próximo experimento mostraremos que, muitas dessas candidatas não selecionadas pela abordagem manual, mas que foram selecionadas pela abordagem *baseada em precisão* retornam anúncios interessantes.

O desempenho similar apresentado pelas abordagens nos leva a pensar que a abordagem de seleção por precisão pode ser usada como um método mais geral para

Método	Precisão
*manual*PL	29,07%
*porPrecisao*PL	29,06%
*manual*PL-AI	29,16%
*porPrecisao*PL-AI	27,63%
manualIDEAL	100%
porPrecisaoIDEAL	22,34%

Tabela 4.1. Precisão para cada método na tarefa de selecionar boas palavras-chave, de acordo com o julgamento humano.

encontrar palavras-chave em páginas web. Embora isso não seja nosso objetivo, pretendemos estudar essas possibilidades como trabalho futuro.

Um segundo ponto a ser observado na Tabela 4.1 é que a abordagem *porPrecisaoIDEAL*, o método que considera relevante exatamente as palavras-chaves retornadas pela abordagem de seleção por precisão, resulta num classificador com menos palavras-chave relevantes quando comparado com o desempenho da avaliação humana. Entretanto, como mostra o próximo experimento, tais resultados não necessariamente implicam em seleção de boas palavras-chave para um sistema de anúncios. Esses resultados reforçam nossa intuição inicial de que a abordagem de seleção de palavras-chave por precisão pode ser melhor que a abordagem manual em sistemas de anúncios.

4.3.2 Precisão nos três primeiros anúncios retornados

A Tabela 4.2 mostra a precisão média para os três primeiros anúncios retornados para cada método. Para cada método apresentamos o resultado P@3, ou seja, consideramos como relevante anúncios escolhidos por, pelo menos, um usuário (P@3-1) e, por, pelo menos, dois usuários (P@3-2). Diferenças entre valores obtidos pelo método de seleção por precisão quando comparado com o correspondente na abordagem manual são estatisticamente significante em todos os casos, de acordo com o teste T de Student, com 98% de confiança.

É possível notar na Tabela 4.2 que todos os resultados obtidos pela abordagem de seleção por precisão foram melhores que os correspondentes na abordagem manual, confirmando nossa intuição de que nossa abordagem é bem melhor para recuperar anúncios relevantes. Por exemplo, a melhor precisão alcançada pela abordagem manual foi 0.128 (*manualPL*), enquanto a melhor precisão obtida pela abordagem por precisão foi 0.269 (*porPrecisaoPL-AI*), um ganho de aproximadamente 110% na P@3-1. Quando consideramos P@3-2, o melhor resultado obtido pela abordagem manual foi 0,086 (*manualPL*), enquanto que o melhor resultado obtido pela abordagem por precisão foi 0,154

Método	P@3-1	P@3-2
*manual*PL	0,128	0,086
*porPrecisao*PL	0,234	0,128
*manual*PL-AI	0,123	0,081
*porPrecisao*PL-AI	0,269	0,154
manualIDEAL	0,420	0,220
porPrecisaoIDEAL	0,654	0,284

Tabela 4.2. P@3 para cada método. Onde um anúncio é considerado relevante para a página se pelo menos um usuário votou como relevante (P@3-1) ou se pelo menos dois usuários julgaram relevante (P@3-2).

(*porPrecisaoPL-AI*), uma melhora de 79%. Esses resultados mostram a importância de escolher palavras-chave levando em consideração a qualidade dos anúncios retornados e não somente o que os humanos acreditam ser boas palavras-chave.

A Tabela 4.2 mostra, também, que a utilização de características da coleção de anúncios não resultou em ganho para a abordagem manual. Por outro lado, resultou num ganho de 20,3% para a abordagem de seleção por precisão. Tal comportamento é esperado visto que a abordagem por precisão seleciona palavras-chave com base na sua capacidade de selecionar anúncios relevantes, ou seja, anúncios associados ao conteúdo da página onde ela ocorre, sendo assim fortemente relacionada às características da coleção de anúncios. Por outro lado, a abordagem de seleção manual é focada em encontrar o que é uma palavra-chave boa para uma pessoa, não tirando proveito das vantagens da coleção de anúncios. Como consequência disso, boas características da coleção de anúncio são perdidas pela abordagem manual. Por exemplo, para uma página sobre futebol em nossa coleção de teste, vários anúncios oferecendo uniformes estão disponíveis. Esses anúncios estão diretamente relacionados à palavra "time", cuja importância foi capturada pelo conjunto de características relacionado ao anúncio, embora "time" somente apareça num fragmento periférico do conteúdo da página. De fato, pertencer a um fragmento de menor importância na página provavelmente contribuiu para a palavra "time" não ser selecionada pela abordagem manual.

Podemos observar, ainda, na Tabela 4.2 a precisão que os classificadores manual ideal e por precisão ideal poderiam obter. Se tais classificadores fossem usados, obteríamos um resultado na abordagem por precisão duas vezes melhor que nosso melhor resultado (P@3-1). Similarmente, os resultados da abordagem manual seriam três vezes melhor (P@3-1). Tais observações indicam que temos muito espaço para melhorar, aumentando a precisão dos nossos classificadores automáticos. Além disso, as diferenças obtidas entre os resultados do método manual ideal (*manualIDEAL*) e do método de seleção por precisão ideal (*porPrecisaoIDEAL*) foram de 55,7% quando con-

sideramos o P@3-1 e 29,1% quando consideramos o P@3-2. Estes resultados indicam que palavras-chave selecionadas usando a abordagem por precisão têm potencial para produzir melhores resultados ao selecionar anúncios no nosso sistema experimental.

O próximo capítulo relata os resultados alcançados pelos métodos e as considerações sobre o que se pretendia com este trabalho mostrando as dificuldades e trabalhos futuros.

Capítulo 5

Conclusão e Trabalhos Futuros

Nesse trabalho propusemos uma nova abordagem para selecionar palavras-chave de páginas web em sistema de publicidade baseada em conteúdo. Nossa principal contribuição foi a mudança na estratégia para compor uma coleção de treino para guiar o processo de aprendizagem. Em vez de pedir aos usuários que rotulassem as boas palavras-chaves encontradas nas páginas de treino, verificamos os anúncios correspondentes a cada palavra encontrada nas páginas de treino e pedimos que os usuários avaliassem suas relevâncias para estas palavras-chave.

Verificamos que esta estratégia nos dá resultados bastante competitivos quando comparada a um método anterior proposto na literatura, aqui chamado de abordagem manual.

Vimos, ainda, que a abordagem de seleção por precisão que estamos propondo nos conduziu a ganhos significativos sobre a abordagem manual. Estes ganhos foram de 82,8% usando P@3, quando pelo menos um usuário disse que o anúncio é relevante (P@3-1) e de 48,8% com P@3 quando pelo menos dois usuários julgaram o anúncio relevante (P@3-2). Estes resultados foram obtidos quando consideramos somente as características propostas por [Yih et al., 2006].

Além disso, nós também propusemos e estudamos o uso de características extraídas da coleção de anúncios como uma fonte de informação para selecionar palavras-chave de uma página web. Quando adicionamos tais características, não melhoramos os resultados de palavras-chave selecionadas com a abordagem manual. Já esperávamos tal resultado visto que a abordagem manual não seleciona palavras-chave considerando estatísticas da base de anúncio e sim, considerando a opinião do avaliador. Por outro lado, elas melhoraram os resultados obtidos pela abordagem de seleção por precisão em aproximadamente 15% tanto para o P@3-1 quanto para o P@3-2.

Quando consideramos estas novas características, a abordagem de seleção por

precisão obteve uma melhora de 110% no P@3-1 e 79% no P@3-2.

5.1 Trabalhos Futuros

Como trabalhos futuros, pretendemos fornecer uma análise mais abrangente do nosso modelo e expandi-lo a fim de adicionar novas evidências. Neste trabalho, estamos propondo algumas características sem avaliar o impacto individual ou de cada grupo de características. A exemplo do que é feito no trabalho de Yih et al. [2006], temos o interesse de estudar a contribuição de cada uma destas características, sejam trabalhadas individualmente ou em grupo, no resultado final do método considerando a abordagem de aprendizado que propusemos. Estas comparações podem ser realizadas utilizando técnicas de seleção de atributos. Como base neste estudo, pretendemos propor um conjunto mínimo de características levando em consideração a eficiência e o custo de obtenção destas.

Além disso, pretendemos analisar mais cuidadosamente as campanhas, de forma que, possamos medir seu impacto na escolha das características. E, ainda, propor novos atributos que estejam relacionados à tarefa de classificação de anúncios e conteúdo.

Um outro contexto que temos interesse de trabalhar é publicidade em vídeo. Pretendemos utilizar o método de seleção por precisão adicionando características típicas de vídeos como por exemplo o texto associado ao vídeo, no caso de serviços de vídeo na web e, o texto de roteiros, no caso de vídeo em TV digital. Ao contrário de páginas web, o conteúdo textual de fontes como roteiros e legendas possui muito ruído, o que torna interessante um estudo mais específico.

Podemos, ainda, estudar o desempenho de outros métodos de aprendizagem de máquina buscando obter resultados mais próximos ao valor ideal descrito neste trabalho. Técnicas comumente usadas para melhorar o processo de classificação envolvem combinações de classificadores, aplicações de técnicas de meta-aprendizagem (como *boosting* [Y. Freund & Singer, 1998]), uso de técnicas de ponderação ou mesmo aplicação de métodos não lineares. Como se observa, há muito o que se pode tentar para melhorar nosso classificador logístico.

Um outro aspecto que também pode ser explorado é a forma de ranquear os anúncios. Uma forma de melhorar é, ao invés de submeter todas as palavras-chave retornadas para cada método a uma consulta com o conectivo "OR" e pegar os 5 anúncios mais relevantes, submeter cada palavra-chave da página à consulta individual e depois fazer a união dos anúncios retornados para cada página. Após isto, os anúncios podem ser reordenados de acordo com a similaridade obtida usando como chave secundária o

identificador do anúncio (desconsiderando anúncios e campanhas repetidas).

Referências Bibliográficas

A. A. Veloso, H. M. Almeida, M. A. G. & Jr., W. M. (2008). Learning to rank at query-time using association rules. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 267--274, New York, NY, USA. ACM.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley-Longman, 1st edição.

Broder, A.; Ciaramita, M.; Fontoura, M.; Gabrilovich, E.; Josifovski, V.; Metzler, D.; Murdock, V. & Plachouras, V. (2008a). To swing or not to swing: learning when (not) to advertise. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pp. 1003--1012, New York, NY, USA. ACM.

Broder, A.; Fontoura, M.; Josifovski, V. & Riedel, L. (2007). A semantic approach to contextual advertising. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 559--566, New York, NY, USA. ACM.

Broder, A. Z.; Ciccolo, P.; Fontoura, M.; Gabrilovich, E.; Josifovski, V. & Riedel, L. (2008b). Search advertising using web relevance feedback. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pp. 1013--1022, New York, NY, USA. ACM.

Chen, Y.; Xue, G.-R. & Yu, Y. (2008). Advertising keyword suggestion based on concept hierarchy. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pp. 251--260, New York, NY, USA. ACM.

Goodman, J. & Carvalho, V. R. (2005). Implicit queries for email. In *Second Conference on Email and Anti-Spam*, `http://www.ceas.cc/papers-2005/141.pdf`.

Hastie, T.; Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Hawking, D. & Craswell, N. (1998). Overview of TREC-7 very large collection track. In *Proc. of the Seventh Text Retrieval Conf.*, pp. 91--104.

Hosmer, D. & Lemeshow, S. (2000). *Applied Logistic Regression, 2nd Edition*. Wiley, New York.

Jin, X.; Li, Y.; Mah, T. & Tong, J. (2007). Sensitive webpage classification for content advertising. In *ADKDD '07: Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising*, pp. 28--33, New York, NY, USA. ACM.

Lacerda, A.; Cristo, M.; Gonçalves, M. A.; Fan, W.; Ziviani, N. & Ribeiro-Neto, B. (2006). Learning to advertise. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 549--556, New York, NY, USA. ACM.

Manning, C. D.; Raghavan, P. & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mitchell, T. (1997). *Machile Learning*. McGraw-Hill.

Radlinski, F.; Broder, A.; Ciccolo, P.; Gabrilovich, E.; Josifovski, V. & Riedel, L. (2008). Optimizing relevance and revenue in ad search: a query substitution approach. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 403--410, New York, NY, USA. ACM.

Ribeiro-Neto, B.; Cristo, M.; Golgher, P. B. & Silva de Moura, E. (2005). Impedance coupling in content-targeted advertising. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 496--503, New York, NY, USA. ACM.

Salton, G.; Wong, A. & Yang, C. S. (1974). A vector space model for automatic indexing. Technical report, Cornell University, Ithaca, NY, USA.

Stone, M. (1974). Cross-validation choices and assessment of statistical predictions. *Journal of the Royal Statiscial Society*, pp. 111--147.

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2st edição.

Wu, X. & Bolivar, A. (2008). Keyword extraction for contextual advertisement. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pp. 1195--1196, New York, NY, USA. ACM.

Y. Freund, R.D. Iyer, R. E. S. & Singer, Y. (1998). An efficient boosting algorithm for combining preferences. In *In Proceedings of the 15th International Conference on Machine Learning*, pp. 170--178.

Yih, W.-t.; Goodman, J. & Carvalho, V. R. (2006). Finding advertising keywords on web pages. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pp. 213--222, New York, NY, USA. ACM.

Yih, W.-t. & Meek, C. (2008). Consistent phrase relevance measures. In *ADKDD '08: Proceedings of the 2nd International Workshop on Data Mining and Audience Intelligence for Advertising*, pp. 37--44, New York, NY, USA. ACM.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo