Universidade Federal do Amazonas
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Programa de Pós-Graduação em Informática

Classificação Automática de Consultas em Sistemas de Busca da Web

Mauro Rojas Herrera

Manaus – Amazonas
Fevereiro de 2010

Mauro Rojas Herrera

Classificação Automática de Consultas em Sistemas de Busca da Web

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Departamento de Ciência da Computação da Universidade Federal do Amazonas, como requisito parcial para obtenção do Título de Mestre em Informática. Área de concentração: Recuperação de Informação e Banco de Dados.

Orientador: Prof. Edleno Silva de Moura, Doutor

Mauro Rojas Herrera

Classificação Automática de Consultas em Sistemas de Busca da Web

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Departamento de Ciência da Computação da Universidade Federal do Amazonas, como requisito parcial para obtenção do Título de Mestre em Informática. Área de concentração: Recuperação de Informação e Banco de Dados.

Banca Examinadora

Prof. Dr. Edleno Silva de Moura – Orientador
Departamento de Ciência da Computação – UFAM/PPGI

Prof. Dr. Adriano Alonso Veloso
Departamento de Ciência da Computação – UFMG

Prof. João Marcos Bastos Cavalcanti, Ph.D.
Departamento de Ciência da Computação – UFAM/PPGI

Manaus – Amazonas
Fevereiro de 2010

Agradecimentos

Resumo

Consultas submetidas a máquinas de busca da web podem ser classificadas de acordo com a finalidade da consulta do usuário dentro de três distintas categorias: navegacional, informacional e transacional. Tal classificação pode ser utilizada como informação adicional em sistemas de seleção de anúncios e em funções de ordenação de respostas de máquinas de busca, dentre outras possíveis aplicações. Neste trabalho, apresentamos um estudo sobre o uso de características extraídas da coleção de documentos e dos logs de consultas de uma máquina de busca na tarefa de classificar, automaticamente, as consultas de acordo com sua finalidade. Propomos o uso de novas características não mencionadas em trabalhos publicados anteriormente na literatura e estudamos o impacto dessas características na qualidade dos resultados de classificação de consultas. Além disso, também apresentamos um estudo sobre a eficácia de cada característica proposta em diferentes coleções de documentos web, mostrando que a escolha do melhor conjunto de características pode depender da coleção de documentos adotada.

Os resultados obtidos indicam que o conjunto de características proposto neste trabalho melhoram os resultados de classificação de consultas quando comparados aos resultados obtidos em trabalhos anteriores. Reportamos experimentos com duas coleções de documentos web onde alcançamos 82.5% e 77,67% de acurácia na classificação de consultas dentro das três finalidades de busca estudadas.

Palavras-chave: Recuperação de Informação, Busca na Web, Classificação de consultas, Aprendizagem Automática, SVM.

Abstract

Queries submitted to search engines can be classified according to the user goals into three distinct categories: navigational, informational, and transactional. Such classification may be useful, for instance, as additional information for advertisement selection algorithms and for search engine ranking functions, among other possible applications. This paper presents a study about the impact of using several features extracted from the document collection and query logs on the task of automatically identifying the users' goals behind their queries. We propose the use of new features not previously reported in literature and study their impact on the quality of the query classification task. Further, we study the impact of each feature on different web collections, showing that the choice of the best set of features may change according to the target collection.

The results obtained indicate the new proposed set of features improves the quality of the classification task when compared to previous proposals. We report experiments with two web collections where we were able to obtain 82.5% and 77.67% of overall accuracy when classifying queries according to the three distinct user goals studied.

Keywords:Information Retrieval, Web Search, Query Classification, Machine Learning, SVM.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O sucesso dos sistemas de busca da Web está diretamente relacionado com a habilidade de satisfazer às necessidades dos seus usuários. Para que isso ocorra, é importante entender quais os tipos de busca que os usuários apresentam ao submeterem suas consultas a um sistema. Essas informações podem ser utilizadas para empregar, no sistema, mecanismos de recuperação e ordenação de documentos adaptados à finalidade de busca da consulta submetida pelo usuário, possibilitando que o conjunto final de documentos mostrado possua maior qualidade. Trabalhos na literatura têm mostrado que pode haver uma melhora na qualidade da ordenação de respostas em sistemas de busca quando a finalidade de busca do usuário é levada em consideração [Craswell et al., 2001, Li et al., 2006]. Além disso, conhecendo a intenção da consulta, pode-se exibir anúncios publicitários mais apropriados, o que contribui para melhorar o desempenho de mecanismos de busca patrocinada, onde um anunciante paga para aparecer na resposta a determinadas consultas. Neste trabalho, propomos uma nova maneira de representar consultas de usuários, a qual possibilita identificar a categoria de cada consulta (finalidade) dentro de uma taxonomia de classes.

De acordo com trabalhos recentes [Broder, 2002, Rose and Levinson, 2004], cada consulta submetida a uma máquina de busca da web pode ser classificada de acordo com sua finalidade em ao menos três classes distintas: navegacional, informacional e transacional. Em consultas navegacionais, o usuário está interessado em alcançar um sítio web específico

e usa a máquina de busca para encontrar seu endereço. Por exemplo, quando o usuário escreve a consulta "Ufam", a intenção principal provavelmente é encontrar a página principal do sítio da universidade na web. Em consultas informacionais, o usuário não tem um sítio particular em mente, mas está interessado em aprender mais sobre um determinado tópico. Por exemplo, na consulta "floresta amazônica", o usuário provavelmente está interessado em encontrar documentos sobre a Floresta Amazônica, de forma que ele possa aprender mais sobre esse tópico. Por fim, em consultas transacionais, o usuário está interessado em encontrar sítios que provêm algum tipo de recurso. O usuário tipicamente necessita realizar uma transação ou interagir com sítios que oferecem algum tipo de serviço. Exemplos de tais serviços são acesso a entretenimento, programas, músicas, filmes, fotos, envio de cartões virtuais e compras. Por exemplo, na consulta "cartões virtuais", o usuário provavelmente deseja encontrar diferentes sítios que o permitam enviar cartões virtuais.

Classificação automática de consultas é normalmente realizada através da representação das consultas usando características extraídas da coleção de documentos da máquina de busca e de seus logs de consultas. Mais especificamente, as principais fontes de informação adotadas na literatura para classificação de consultas são (i) o texto de âncora presente nos apontadores entre as páginas da web, (ii) a URL das páginas, (iii) o conteúdo textual das consultas, (iv) o conteúdo textual de documentos relacionado com a consulta, e (v) a informação de clique disponível em logs de consultas passadas da máquina de busca. Neste trabalho, focamos nosso estudo nas fontes listadas de (i) até (iv), apresentando formas alternativas de aplicá-las para determinar a finalidade de busca das consultas submetidas pelos usuários. Embora a informação de clique seja uma das fontes mais úteis adotadas em tarefas de classificação de consultas, não a incluímos neste trabalho, uma vez que não existem coleções de documentos web disponíveis publicamente que contenham essa informação. Contudo, nossos resultados são úteis para entender melhor como as outras quatro fontes podem ser exploradas. Além disso, as características extraídas de informações de clique podem ser combinadas com aquelas aqui apresentadas para obter

melhores resultados de classificação.

Nossas contribuições na área de classificação automática de consultas incluem o uso de novas características para serem adotadas durante o processo de classificação de consultas; e um estudo detalhado sobre o impacto de cada característica em diferentes coleções de documentos e tarefas de classificação. Uma das novas características exploradas aqui é baseada na idéia de que estatísticas sobre a ocorrência dos termos da consulta sobre direfentes domínios[1] são úteis para determinar a finalidade de cada consulta. Usamos essa evidência para incluir duas novas características não mencionadas anteriormente na literatura sobre classificação de consultas: *mqd* e *dda*. Outra característica incluída neste estudo é a popularidade da consulta. Como será mostrado, a popularidade tem impacto na tarefa de classificação e é uma importante nova característica que adicionamos para o processo de classificação.

1.1 Trabalhos Relacionados

Em sistemas de Recuperação de Informação clássicos, os usuários estão basicamente interessados em buscar informações. Além disso, essa é a principal motivação para o uso de máquinas de busca na web [Navarro-Prieto et al., 1999, Muramatsu and Pratt, 2001, Choo et al., 1999]. No entanto, esses sistemas também funcionam como ferramenta para auxiliar os usuários na localização e acesso a enorme quantidade de diferentes recursos disponíveis na web. O estudo de logs de consultas de máquinas de busca da web tem mostrado que existem diversas finalidades implícitas nas consultas dos usuários. Em [Broder, 2002], os autores mostraram que 48% das consultas dentro do log de consultas analisado foram informacionais, 30% transacionais e 20% navegacionais. Os 2% de consultas restantes não foram classificadas. Outro estudo ([Spink and Jansen, 2004]) apontou que 12% a 24% das consultas submetidas estão relacionadas com transações de comércio eletrônico. No estudo seguinte ([Jansen et al., 2005]), baseado em logs de consultas extraídos da

[1]Neste trabalho, consideramos que um domínio é uma string em um formato de três níveis ("servidor.organização.tipo"), usado para identificar uma entidade na internet, tal como uma companhia (ex, www.nhemu.com).

máquina de busca Altavista[2] em 2002, os autores mostraram que o sistema de busca foi largamente usado como uma ferramenta de navegação.

Uma vez que a qualidade das máquinas de busca da web está diretamente relacionada com quão bem o sistema está apto a atender os interesses dos usuários, muitos estudos têm se focado na questão de como classificar as consultas de acordo com a finalidade da busca de forma precisa e eficiente [Kang and Kim, 2003, Kang, 2005, Lee et al., 2005, Lu et al., 2006], bem como na questão de como utilizar essa informação para melhorar a qualidade dos resultados do sistema [Craswell et al., 2001, Rose and Levinson, 2004, Li et al., 2006].

Em particular, em [Craswell et al., 2001], os autores estudaram a possibilidade de utilizar a classificação das consultas como informacional e navegacional para melhorar a qualidade dos resultados dentro de uma máquina de busca, aplicando funções de ordenação de respostas especializadas. Em seus estudos, assumiram que as consultas já estavam previamente classificadas e concluíram que a informação de classe ou finalidade é útil, uma vez que melhores estratégias de ordenação podem ser aplicadas para cada tipo de consulta. Portanto, nosso trabalho e outros trabalhos de classificação de consultas podem ser empregados para proverem a informação de classe das consultas.

Outro trabalho que mostra uma possível vantagem de utilizar a informação de classe da consulta é apresentado em [Li et al., 2006]. Os autores propuseram um método para identificar automaticamente páginas web construídas com finalidade transacional. Também mostram que tal classificação pode ser usada para melhorar a qualidade dos resultados de busca quando consultas transacionais são processadas. Dessa forma, a classificação de páginas web e consultas pode ser usada como estratégia complementar para melhorar a qualidade dos resultados para consultas transacionais.

Considerando classificação automática de consultas, muitos trabalhos têm adotado a taxonomia introduzida por [Broder, 2002], onde as consultas podem ter finalidade navegacional, informacional e transacional. Por exemplo, [Kang and Kim, 2003] apresentaram

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

uma abordagem para classificar consultas como navegacional e informacional. Os autores realizaram experimentos na coleção TREC de documentos web [Hawking et al., 1999] e, em seu método, utilizaram uma combinação linear de quatro características. Para as duas primeiras características, os documentos da coleção são separados em dois conjuntos denominados documentos de tópico e documentos de página inicial. Essas características consistem na distribuição das respostas para a consulta e na distribuição dos termos da consulta nos dois conjuntos gerados. A terceira característica explora a taxa de ocorrências da consulta em textos de âncora. Finalmente, a última característica consiste na detecção de aspectos linguísticos dos termos da consulta, tais como a ausência de verbos em consultas navegacionais.

Em um trabalho seguinte, [Kang, 2005] levaram as consultas transacionais em consideração. O método proposto consiste na combinação do seguinte conjunto de características através de um método de aprendizagem automático: (1) a classe resultante pelo método proposto em [Kang and Kim, 2003]; (2) a primeira e última palavra da consulta; (3) a identificação da consulta como sendo o nome de um arquivo realizada por meio de expressões regulares simples; (4) pesos para apontadores, indicando a natureza da consulta como um sítio, sub-sítio, música, figura, texto, aplicação, serviço e arquivo. Os pesos para apontadores foram obtidos através do uso de dados de treinamento, onde os textos de âncora foram rotulados de acordo com a ação associada com o apontador (leitura, visita e download). Estas ações foram determinadas através da identificação de certos padrões usando expressões regulares. Das características listadas, nós também usamos a taxa de ocorrência dos termos da consulta nos textos de âncora (caracterísa *dda*). No entanto, nosso cálculo de similaridade entre consulta e textos de âncora é baseado no Modelo Vetorial. Neste trabalho, não empregamos características similares a aquelas baseadas em documentos previamente classificados e análise de aspectos linguísticos. O conjunto de características de (2) a (4) são aquelas relacionadas com nossa característica *terms*. Entretanto, especialmente as características (3) e (4) são baseadas em muitas heurísticas de detecção de padrões , enquanto que nós simplesmente usamos o conjunto

original de termos das consultas sem nenhuma modificação. Neste trabalho, realizamos experimentos com a mesma coleção de documentos empregada nos dois trabalhos descritos anteriormente para comparar esses métodos com a nossa proposta. No entanto, é importante mencionar que as características empregadas em [Kang and Kim, 2003, Kang, 2005] podem ser utilizadas como características adicionais em nossa representação de consultas.

Lee et al [Lee et al., 2005] propuseram o uso do comportamento de clique do usuário e da distribuição dos textos de âncora para classificação de consultas como informacional e navegacional. Eles realizaram experimentos com 30 consultas populares submetidas a máquinas de busca a partir de sua universidade, excluindo consultas onde não houve consenso quando classificadas por especialistas dentro de uma das duas categorias. O texto de âncora foi obtido através da coleta de 60 milhões de documentos do diretório ODP[3]. Os dados de clique dos usuários foram obtidos a partir do log de acesso dos usuários da universidade para a web. Baeza-Yates et al. também estudaram o uso de informações de log de consultas para classificação [Baeza-Yates et al., 2006], reportando que a informação de clique do usuário é uma excelente fonte de informação para classificação de consultas. Em ambos os casos, as coleções adotadas não estão disponíveis publicamente, o que torna difícil a reprodução dos seus experimentos. Além disso, as coleções empregadas em nossos experimentos não incluem informações de clique do usuário. Incluímos no estudo realizado aqui, a comparação com a característica distribuição dos textos de âncora proposta por Lee et al. e investigamos seu impacto quando combinada com novas características propostas neste trabalho.

Em [Lu et al., 2006], os autores estudaram o uso de muitos métodos de aprendizagem automática para identificar consultas navegacionais, não conduzindo experimentos para identificar consultas informacionais e transacionais. Eles usaram milhares de características extraídas de dados de clique dos usuários, logs de consultas e da coleção de documentos da máquina de busca. Os experimentos foram realizados sobre 2012 consultas selecionadas aleatoriamente de um log de consultas. Além de considerar somente

[3]Open Directory Project (http://www.dmoz.org/)

consultas navegacionais, os autores não mostraram uma lista detalhada das características utilizadas, nem mostram informações sobre a coleção de documentos adotada. Além disso, a coleção de documentos não está disponível publicamente.

Até onde sabemos, este é o primeiro trabalho que estuda a influência da popularidade da consulta em tarefas de classificação de consultas. Por fim, neste trabalho propomos e estudamos maneiras alternativas de usar fontes de evidências previamente estudadas na literatura.

1.2 Organização da Dissertação

Este trabalho está organizado da seguinte forma:

O Capítulo 2, formaliza o problema de identificar a finalidade das consultas dos usuários e descreve, em detalhes, a lista de características utilizadas para representar tais consultas. O Capítulo 3 apresenta os experimentos realizados, mostrando uma análise das características empregadas e dos resultados comparativos com outras abordagens encontradas na literatura. Por fim, o Capítulo 4 conclui a dissertação, indicando algumas direções para trabalhos futuros.

Capítulo 2

Classificação Automática de Consultas

Neste capítulo, apresentamos a formalização do problema de identificar a finalidade de busca implícita nas consultas dos usuários. Além disso, discutimos em detalhes as características propostas e utilizadas para representar as consultas.

2.1 Formalização do problema

Neste trabalho, modelamos o problema de determinar a finalidade das consultas submetidas por usuários a uma máquina de busca como um problema de classificação. Neste problema, temos um conjunto de n consultas $Q = \{q_1, q_2, ..., q_n\}$, onde cada consulta q_i é representada por um conjunto de m características, isto é, $q_i = \{f_1, f_2, ..., f_m\}$; e uma taxonomia, ou seja, um conjunto fixo de t categorias $C = \{c_1, c_2, ..., c_t\}$. Dada uma ou mais consultas de treinamento, nosso objetivo é aprender uma função de classificação $\gamma : Q \rightarrow C$ que mapeia consultas a suas categorias. Como taxonomias, consideramos quatro conjuntos. O primeiro, C_{todas} = {informacional, transacional, navegacional}, representa o conjunto completo de finalidades descrito em [Broder, 2002]. As taxonomias restantes, C_{inf} = {informacional, não informacional}, C_{tra} = {transacional, não transacional}, e C_{nav} = {navegacional, não navegacional}, representam os casos positivos e

negativos de classificação para cada finalidade de usuário em C_{todas}.

A solução que propomos para este problema consiste em sugerir um conjunto de características $\{f_1, f_2, ..., f_m\}$ usadas para representar as consultas e então aplicá-las a um método de aprendizagem para realizar a tarefa de classificação. O algoritmo de aprendizagem selecionado é o Máquina de Vetores de Suporte (SVM - Support Vector Machine) [Joachims, 1998], método estado-da-arte já aplicadado em tarefas de classificação de consultas apresentando excelentes resultados [Lu et al., 2006]. Na Seção 2.1.1, descreve-se o método SVM com mais detalhes.

2.1.1 Máquina de Vetores de Suporte (SVM)

Máquina de Vetores de Suporte (SVM) é um método de aprendizagem automático estado-da-arte em tarefas de classificação. O método SVM busca um hiperplano que separe um conjunto de exemplos de treinamento rotulados como positivos e negativos. O hiperplano é definido por $w^T x + b = 0$, onde o parâmetro $w \in \mathbf{R}^m$ é um vetor ortogonal ao hiperplano e $b \in \mathbf{R}$ é o erro. A função de decisão é um classificador de hiperplano:

$$H(x) = sign(w^T x + b = 0) \tag{2.1}$$

O hiperplano é projetado de maneira que $y_i(w^T x_i + b = 0) \geq 1 - \xi_i, \forall_i = 1, ..., N$, onde $x_i \in \mathbf{R}^m$ é uma instância nos dados de treinamento e $y_i \in \{+1, -1\}$ denota a classe do vetor x_i. A margem é definida pela distância entre dois hiperplanos pararelos $w^T x + b = 1$ e $w^T x + b = -1$, isto é, $2/||w||_2$. O problema de treinamento do SVM é definido a seguir:

$$\begin{aligned} &minimize \quad (1/2) w^T w + \gamma \mathbf{1}^T \xi \\ &sujeito\ a \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad sendo \quad i = 1, ...N \quad e \quad \xi \geq 0 \end{aligned} \tag{2.2}$$

onde o escalar γ é chamado de parâmetro de regularização, e normalmente é selecionado de forma empírica para reduzir a taxa de erro na fase de teste.

A formalização básica do SVM pode ser estendida para casos não-lineares, usando kernel não-linear. A complexidade de um classificador SVM não depende do número

de características, mas sim do número de vetores de suporte (o número de exemplos de treinamento próximos ao hiperplano). Esta propriedade torna o SVM adequado para problemas de classificação com númerosas dimensões. Em nossos experimentos, adotamos o pacote de software LIBSVM [Chang and Lin, 2001] com kernel RBF(Radial Basis Function).

2.2 Representação das Consultas

Determinar que características devem ser usadas para representar uma consulta é uma decisão chave em tarefas de classificação de consultas. Neste trabalho, a palavra *característica* descreve uma estatística que representa a medição de algum aspecto de uma dada consulta de usuário ou algum termo pertencente a essa consulta. Para determinar o valor discriminativo das características, experimentamos muitas combinações, sempre verificando a utilidade de cada opção.

As características que adotamos neste estudo são descritas nas próximas seções, agrupadas de acordo com a fonte de informação de onde foram extraídas e descrevendo como computamos o valor de cada característica para cada consulta. Consideramos diversas maneiras alternativas para calcular os valores das características como, por exemplo, o uso de diferentes estratégias para casamentos entre títulos e URLs, e para caracterizar a inclinação de distribuições. Entretanto, as características resultantes não apresentaram ganhos em relação ao conjunto final de características consideradas nos experimentos.

2.2.1 Características baseadas em textos de âncora

Estas características foram extraídas da concatenação dos textos de âncora que apontam para documentos presentes no conjunto de respostas para uma consulta. Nossa intuição em relação a textos de âncora é que, geralmente, existe um pequeno conjunto de páginas com autoridade para consultas navegacionais, enquanto que um grande número de páginas existe para consultas informacionais. Dessa forma, a distribuição das ocorrências dos

termos da consulta nos textos de âncora das páginas é provavelmente mais inclinada para consultas navegacionais. Por exemplo, a palavra "Yahoo", a qual tende a ser navegacional, é provavelmente um termo muito comum em textos de âncora que apontam para a página inicial do Yahoo[1] e tem menos ocorrências em outras páginas. Por outro lado, o termo "câncer", o qual é uma consulta informacional, é provavelmente encontrado em textos de âncora de apontadores direcionados para muitas páginas com autoridade sobre câncer e com uma distribuição uniforme.

- **Distribuição das frequências nos textos de âncora** (af): Mede a distribuição das ocorrências da consulta nos textos de âncora dos documentos da coleção. A consulta é sempre processada como uma frase e a inclinação da distribuição é calculada através da métrica mediana. Esta característica foi proposta em [Lee et al., 2005]. Também consideramos a utilização de uma variação desta característica, onde a consulta foi processada de acordo com a especificação do usuário. Esta opção tornaria a característica mais flexível para ser aplicada na prática, no entanto, a proposta original apresentou melhores resultados em todos os experimentos realizados e, por esse motivo, decidimos manter a proposta original de [Lee et al., 2005].

- **Densidade de domínios nos textos de âncora mais similares** (dda): Neste caso, processamos a consulta e computamos a similaridade de cada texto de âncora com a consulta usando o Modelo Vetorial [Salton et al., 1975]. Fazendo isso, contornamos algumas das limitações da característica af proposta em [Lee et al., 2005], a qual somente considera textos de âncora que são idênticos a consulta. No Modelo Vetorial, documentos e consultas são representados como vetores em um espaço composto por termos. Dessa forma, um documento d_j é representado como um documento de t pesos para seus termos $d_j = (w_{1j}, w_{2j}, ..., w_{tj})$. Cada peso w_{ij} reflete a importância do termo k_i no documento d_j e é computado como $w_{ij} = tf_{ij} \times \log \frac{N}{ni}$, onde tf_{ij} é o número de vezes que o termo k_i apareceu no documento d_j, n_i é o número de documentos nos quais k_i apareceu, e N é o número total de documentos

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

na coleção. Para calcular a similaridade entre uma consulta q e um documento d_j, usamos o valor do cosseno do ângulo entre q e d_j, como:

$$sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \tag{2.3}$$

A partir das respostas ordenadas obtidas através da Equação. 2.3, consideramos as K primeiras respostas e verificamos se os resultados estão ou não concentrados em um único domínio. A densidade é computada como:

$$densidade = \frac{distintos}{K + max_{dominio}} \tag{2.4}$$

onde *distintos* é o número de domínios distintos encontrados entre as K primeiras respostas, e $max_{dominio}$ é o número de documentos no domínio mais comum encontrado entre as K primeiras respostas. O valor de K foi obtido por meio de experimentos usando validação cruzada no dados de treinamento extraídos das coleções WBR03 e WT10g. O valor selecionado foi $K = 3$ para ambas coleções.

Note que a função de similaridade usada por *dda* (em nossos experimentos, o cosseno no Modelo Vetorial) pode ser diferente daquela empregada em uma máquina de busca, embora ela use os mesmos dados para obter a característica e computar as respostas.

2.2.2 Características baseadas no conteúdo das páginas

As características listadas a seguir foram extraídas do conteúdo das páginas que estão nas respostas recuperadas pela consulta do usuário.

- **Densidade de domínios nos textos mais similares** (*ddt*): Análoga a densidade de domínios nos textos de âncora mais similares (*dda*), considerando, neste caso, o conteúdo textual dos documentos, ao invés de seus textos de âncora. A intuição desta característica é que diferentes distribuições de domínios são observadas entre

as diferentes categorias. O valor de K foi determinado seguindo a mesma estratégia experimental aplicada para *dda*. O valor selecionado foi $K = 50$ em ambas as coleções, sendo 50 o melhor valor para WT10g e um dos melhores valores na WBR03. Como em *dda*, a função de similaridade usada por *ddt* pode ser diferente daquela adotada pelo sistema de busca onde o classificador será aplicado.

- **Titlemr** (*title*): Esta característica é baseada na idéia original de [Lu et al., 2006], onde a URL dos documentos é utilizada como fonte de informação em classificação de consultas. Aqui, aplicamos a mesma idéia para extrair informações dos títulos dos documentos. Ela consite em considerar os 100 primeiros resultados obtidos pela máquina de busca e computar, para cada resposta, a razão entre a maior substring da consulta encontrada no título da página e o tamanho do título. O maior valor entre as 100 primeiras respostas é atribuído ao valor da característica.

2.2.3 Características baseadas em URL

Estas características foram extraídas do endereço URL das páginas que estão no conjunto de resultados recuperados pela consulta do usuário. A intuição geral desta característica é que termos em consultas navegacionais aparecem, mais frequentemente, em endereços URL.

- **Casamento entre consulta e domínio** (*mqd*): Neste caso, computamos o vocabulário de todas as palavras dentro de todos os nomes de domínios encontrados na coleção. Uma vez que a consulta é submetida, contamos o número de termos da consulta que casam no vocabulário e dividimos esse valor pelo número de termos na consulta. Além disso, somamos um, ao valor anterior, se a concatenação de todos os termos da consulta casam como uma palavra no vocabulário. Por exemplo, esse seria o caso da consulta "Banco do Brasil", que casa com o domínio *bancodobasil.com.br*.

- **URLmr** (*url*): Esta característica foi proposta em [Lu et al., 2006] e consiste em extrair os 100 primeiros resultados obtidos pela máquina de busca e computamos,

para cada resposta, a razão entre o tamanho da maior substring da consulta encontrada na URL e o tamanho da URL. Somente o maior valor entre as 100 primeiras respostas é assinalado como o valor da característica.

2.2.4 Características baseadas nas consultas

Estas características foram extraídas das consultas submetidas por usuários a máquina de busca. Nossa intuição é que o tamanho e o conteúdo das consultas são bons indicadores de sua natureza.

- **Número de termos** (*#terms*): Nossa intuição é que consultas navegacionais (ex: "ufam") são, na média, mais curtas que consultas transacionais (ex: "letra da música garota de ipanema") e informacionais (ex: "como é o tratamento do câncer de pulmão"). Além disso, esta característica é sugerida em [Kang and Kim, 2003, Kang and Kim, 2004], mas até onde sabemos, esta fonte de evidência não foi aplicada em classificação de consultas em trabalhos anteriores.

- **Termos** (*terms*): Neste caso, representamos as consultas pelo seus termos. É importante mencionar que esta não uma característica isolada, mas um conjunto delas. O uso dos termos é particularmente motivado pela observação que algumas palavras ocorrem mais frequentemente em certas categorias de consultas. Por exemplo, sufixos de domínios (ex: "edu") são mais comuns em consultas navegacionais, enquanto que termos que indicam perguntas e preposições (ex: "como", "de" e "para") são mais comuns em consultas informacionais; por fim, palavras como "mp3", "letras", "jogos", "fotos", entre outras, são comuns em consultas transacionais.

2.2.5 Características baseadas em logs de consultas

Estas características foram extraídas de logs de consultas passadas da máquina de busca.

- **Popularidade da consulta** (*qpop*): O número de ocorrências da consulta em um log de consultas. Esta característica é, normalmente, pouco discriminativa

quando usada isoladamente, porém, muito útil quando combinada com outras características. Nossa intuição é que estatísticas tal como a distribuição nos textos de âncora e padrões de ocorrências dos termos podem variar de acordo com a popularidade da consulta. O período do log adotado para computar a popularidade das consultas foi de um mês.

Capítulo 3

Experimentos

Neste capítulo, apresentamos todos os experimentos realizados para avaliar a eficácia da nossa abordagem para classificação automática de consultas. Descrevemos uma análise detalhada de cada uma das características empregadas neste trabalho em experimentos realizados sobre duas coleções de documentos web. Além disso, para fortalecer nossos achados, reportamos experimentos comparativos com trabalhos já publicados na literatura, mostrando que nosso método alcança resultados significativamente melhores.

3.1 Base de Dados

Neste trabalho, utilizamos duas coleções para experimentos. A primeira é a coleção WT10g, a qual foi adotada na Web TREC 2001 [Bailey et al., 2003]. A segunda é a coleção WBR03, uma base de dados extraída da web brasileira contendo consultas submetidas ao TodoBR[1], uma máquina de busca real.

A coleção WT10g não contém informações de logs de consultas e foi empregada, neste trabalho, principalmente para facilitar a comparação dos nossos resultados com outras abordagens de classificação de consultas, particularmente aquelas apresentadas em [Kang and Kim, 2003, Kang and Kim, 2004]. A WT10g contém aproximadamente

[1]TodoBR é uma marca registra da Akwan Information Technologies, a qual foi adquirida pelo Google em Júlio de 2005.

7.5Gb de texto em 1,692,098 documentos e 2.5 milhões de apontadores conectando suas páginas.

O conjunto de consultas usado nos experimentos com a WT10g é o empregado usado em [Kang and Kim, 2003, Kang and Kim, 2004]. Dessa forma, como informacionais, usamos as consultas de tópico de 451 a 500 da TREC-2000 e as consultas de tópico 501 a 550 da TREC-2001, somando um total de 100 consultas. Como consultas transacionais, usamos as 100 consultas de serviço extraídas do arquivo de log da Lycos[2], também empregadas em [Kang, 2005]. Como consultas navegacionais, usamos as 245 consultas de página inicial da TREC-2001. Note que a distribuição de classes usada em [Kang and Kim, 2003, Kang, 2005] é um pouco arbitrária, já que estudos anteriores sobre a distribuição dos tipos de consultas indicam que não é esperado que sejam submetidas mais consultas navegacionais do que outros tipos [Baeza-Yates et al., 2006]. Para evitar que o processo de aprendizagem sofra influência da distribuição alterada da classe navegacional, selecionamos aleatoriamente 100 consultas do conjunto de 245 consultas navegacionais utilizadas em [Kang and Kim, 2003].

A coleção WBR03 foi empregada para que possamos estudar o efeito da popularidade em tarefas de classificação de consultas. Esta coleção contém 12,020,513 páginas da web, com aproximadamente 140 milhões de apontadores conectando suas páginas, e cerca de 60.3GB de texto. Extraímos 600 consultas do log de consultas do TodoBR, o qual é composto de 11,246,351 consultas, através do processo descrito abaixo.

Primeiro, selecionamos aleatoriamente um conjunto de consultas do log, as quais foram classificadas até obtermos 200 consultas de cada categoria. Um total de 2564 consultas foram classificadas nessa primeira etapa. Todas as consultas da WBR03 foram classificadas por pessoas, tal que cada uma foi classifacada por três avaliadores humanos. Embasados nas definições de classes de consultas descritas no Capítulo 1, os avaliadores foram questionados a indicar a finalidade mais provável para cada consulta. Neste tipo de tarefa, a possibilidade de uma consulta possuir múltiplas finalidades é muito comum, o

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

que dificulta determinar a finalidade de busca original da consulta. Por exemplo, dada a consulta "MP3", o usuário pode estar interessado em (a) aprender sobre o tópico MP3, o que torna a consulta informacional, (b) encontrar sites que contêm músicas no formato MP3, nesse caso sendo uma consulta transacional, ou (c) em acessar o sítio MP3.com ("http://www.mp3.com"), sendo uma consulta navegacional. Para contornar esses casos, solicitamos aos usuários que avaliaram as consultas para assinalarem, a cada consulta, a categoria mais apropriada de acordo com suas opiniões.

Na segunda etapa, extraímos somente 200 consultas de cada categoria usando o mesmo processo aleatório. Nesta etapa, cada consulta foi considerada como pertencente a uma classe se ao menos uma pessoa a assinalou para essa classe. Dessa maneira, espera-se que o conjunto final de consultas tenha uma distribuição de finalidades de busca similar àquela encontrada no log. No conjunto selecionado existem tanto consultas populares como não populares, dado que a popularidade das consultas no log segue a distribuição de Zipf [Saraiva et al., 2001].

A Tabela 3.1 apresenta várias estatísticas, sobre alguns exemplos, relacionadas as coleções WBR03 e WT10g. As colunas "Total" e "Tam. Médio" mostram, respectivamente, o número de consultas e o tamanho médio dessas consultas em cada classe e coleção. A coluna "Multi" mostra o percentual de consultas que receberam duas categorias distintas na WBR03. Analisando a tabela, notamos que as consultas mais curtas são as navegacionais. Além disso, as consultas com mais termos são as informacionais na WBR03 e as transacionais na WT10g. O fato das consultas informacionais serem maiores na WBR03 é explicado, em parte, pelo frequente uso de proposições na língua Portuguesa.

3.2 Metodologia de Avaliação

Para realizar os experimentos, usamos o método de validação cruzada em 10 partes, da forma como é descrito em [Mitchell, 1997]. Além disso, em todas as comparações reportadas neste trabalho, utilizamos o teste estatístisco de Wilcoxon [Wilcoxon, 1945] para determinar se as diferenças em questão são estatisticamente significantes. Wilcoxon

Classe da Consulta	Total	Multi (%)	Tam. Médio	Exemplo
		WBR03		
Navegacional	200	39.5	2.61	Discovery Channel
Transacional	200	61.0	3.35	Músicas de Rap MP3
Informacional	200	60.5	4.25	vida pessoal de Van Gogh
		WT10g		
Navegacional	245	-	3.05	American Chemical Society
Transacional	100	-	4.21	Picture of the Gulf War
Informacional	100	-	3.20	Estrogen Why Needed

Tabela 3.1: Estatísticas dos conjuntos de consultas usados nas coleções WBR03 e WT10g.

é um teste pareado não paramétrico que não assume nenhuma distribuição particular dos valores testados. Em todos os casos, somente apresentamos conclusões dos resutados que foram significantes ao menos no nível de 5%.

O desempenho dos métodos nas diversas tarefas de classificação foi avaliado usando a medida convencional de acurácia, a qual é definida como a proporção de exemplos classificados corretamente. Para comparação com outros métodos, foram também usadas as métricas de precisão, revocação e Medida F1. A precisão p é definida como a proporção de exemplos classificados corretamente no conjunto de todos os exemplos assinalados para a classe alvo. A revocação r é definida como a proporção de exemplos classificados corretamente em relação a todos os exemplos que são da classe alvo. A Medida F1 é a combinação da precisão e revocação e é definida como $\frac{2pr}{p+r}$.

3.3 Resultados

A seguir, descrevemos e analisamos todos os resultados obtidos em nossos experimentos.

3.3.1 Análise das características

Nesta Seção, avaliamos o impacto de cada característica na acurácia do classificador. Para realizar essa tarefa, foram conduzidos dois experimentos distintos. No primeiro, aplicamos cada característica ao classificador de modo que possamos determinar seu impacto individual. Visto que algumas características podem não ser mais úteis individualmente,

estudamos o impacto de cada característica quando removidas do conjunto aplicado ao classificador.

Dessa forma, para avaliar o impacto da característica *qpop*, por exemplo, representamos a consulta usando apenas *qpop* no primeiro experimento e usamos todas as características, exceto *qpop*, no segundo experimento. Em todos os experimentos, a característica *terms* é considerada como uma única característica, que representa os termos da consulta.

Para todas as bases de dados, apresentamos a acurácia obtida na classificação para as taxonomias C_{inf}, C_{nav}, C_{tra}, e C_{todas}, as quais implicam em quatro diferentes tarefas de classificação: distinguir consultas informacionais de não informacionais; consultas navegacionais de não navegacionais; consultas transacionais de não transacionais; e identificar a finalidade da consulta do usuário entre as três possibilidades. Em todas as tabelas com resultados, os números entre parêntesis indicam a importância relativa de cada característica, sendo que quanto menor esse número, melhor o desempenho da característica. As características estão ordenadas de acordo com a acurácia obtida na taxonomia C_{todas}. Além disso, para referência, mostramos a acurácia obtida quando todas as características são consideradas. É importante destacar que não é possível incluir a característica *qpop* nos experimentos com a WT10g, já que essa informação não está disponível para essa coleção. As Tabelas 3.2 e 3.3 apresentam os resultados obtidos aplicando cada característica isolada, respectivamente, nas coleções WBR03 e na WT10g.

As Tabelas 3.4 e 3.5 apresentam os resultados obtidos, para as coleções WBR03 e WT10g, após a remoção de cada característica proposta do conjunto utilizado para representar as consultas. É importante notar que, neste caso, quanto menor for a acurácia obtida depois da remoção da característica, maior é sua capacidade de auxiliar o classificador a tomar uma decisão correta. Nestas duas tabelas, os resultados iguais ou piores àqueles obtidos com todas as características são mostrados em negrito.

Das Tabelas 3.2 a 3.5, podemos observar que a característica *terms* é muito efetiva para classificação de consultas. Isso provavelmente acontece pelos padrões específicos dos

Característica Usada	Taxonomia			
	C_{inf}	C_{nav}	C_{tra}	C_{todas}
todas	83.67	90.67	90.33	82.50
terms	73.67 (2)	83.33 (1)	89.33 (1)	71.83 (1)
dda	77.33 (1)	76.50 (4)	67.00 (4)	59.67 (2)
mqd	70.33 (3)	80.74 (2)	65.67 (7)	56.17 (3)
url	66.17 (6)	77.00 (3)	70.00 (2)	54.00 (4)
title	66.67 (4)	72.00 (5)	69.67 (3)	52.17 (5)
af	66.67 (4)	71.00 (6)	66.67 (6)	49.33 (6)
#terms	66.50 (5)	70.83 (7)	67.83 (4)	48.83 (7)
ddt	66.17 (6)	66.17 (9)	66.83 (5)	44.17 (8)
qpop	66.67 (4)	66.67 (8)	66.67 (6)	42.33 (9)

Tabela 3.2: Acurácia obtida na classificação em diferentes taxonomias e características individuais na coleção WBR03. A linha *todas* representa a combinação de todas as características estudadas.

Característica Usada	Taxonomia			
	C_{inf}	C_{nav}	C_{tra}	C_{todas}
todas	83.67	92.67	82.00	77.67
terms	77.67 (1)	76.33 (3)	82.00 (1)	64.67 (1)
ddt	69.67 (4)	81.00 (1)	66.67 (4)	57.00 (2)
title	66.67 (6)	74.00 (5)	71.33 (2)	54.67 (3)
url	66.67 (6)	75.33 (4)	68.67 (3)	50.67 (4)
dda	71.67 (2)	80.00 (2)	66.67 (4)	50.33 (5)
mqd	69.33 (5)	72.00 (6)	66.67 (4)	43.67 (6)
#terms	71.00 (3)	66.67 (7)	66.67 (4)	42.00 (7)
af	66.67 (4)	67.33 (8)	66.67 (4)	36.00 (8)

Tabela 3.3: Acurácia obtida na classificação em diferentes taxonomias e características individuais na coleção WT10g. A linha *todas* representa a combinação de todas as características estudadas.

vocabulários de cada tipo de consulta. Por outro lado, o número de termos na consulta (*#terms*) introduz, em alguns casos, perdas na acurácia da classificação. Também verificamos que o impacto de *terms* é muito mais significante na WBR03 do que na WT10g. Isso acontece devido ao fraco desempenho desta característica para distinguir consultas navegacionais na WT10g, fato evidenciado na Tabela 3.5, uma vez que a remoção de *terms* proporciona uma melhora na acurácia.

Características baseadas em textos de âncora são boas para identificar consultas navegacionais na coleção WT10g, conclusão indêntica à reportada em [Kang and Kim, 2003]. Apesar de introduzir ruído no reconhecimento de consultas transacionais, sua contri-

Característica Removida	Taxonomia			
	C_{inf}	C_{nav}	C_{tra}	C_{todas}
nenhuma	83.67	90.67	90.33	82.50
terms	80.33 (2)	85.67 (1)	80.33 (1)	71.50 (1)
mqd	79.33 (1)	87.17 (2)	89.50 (3)	77.33 (2)
qpop	80.33 (2)	90.00 (4)	89.33 (2)	80.00 (3)
url	81.33 (3)	89.50 (3)	89.67 (4)	81.50 (4)
dda	82.17 (4)	89.50 (3)	89.33 (2)	81.83 (5)
af	**83.67** (6)	**90.83** (7)	90.17 (5)	82.00 (6)
ddt	**84.33** (7)	**90.67** (6)	**90.33** (6)	82.17 (7)
title	**83.50** (5)	**90.83** (7)	**90.33** (6)	82.33 (8)
#terms	**83.50** (5)	90.17 (5)	89.83 (2)	**82.67** (9)

Tabela 3.4: Acurácia obtida na classificação usando diferentes taxonomias na WBR03. Cada linha representa a combinação de características propostas, removendo uma característica do conjunto. A linha *nenhuma* representa o caso em que nenhuma característica é removida.

Característica Removida	Taxonomia			
	C_{inf}	C_{nav}	C_{tra}	C_{todas}
Nenhuma	83.67	92.67	82.00	77.67
terms	78.67 (1)	**93.00** (6)	71.33 (1)	71.00 (1)
ddt	81.67 (3)	86.87 (1)	**83.00** (5)	73.67 (2)
title	83.33 (5)	91.00 (3)	80.67 (2)	76.00 (3)
url	83.00 (4)	91.67 (4)	81.33 (3)	76.33 (4)
#terms	**84.00** (7)	**92.67** (5)	**82.00** (3)	76.67 (5)
af	**83.67** (6)	**92.67** (5)	**82.00** (3)	**77.67** (6)
dda	79.00 (2)	90.33 (2)	**83.00** (5)	**78.00** (7)
mqd	**83.67** (6)	**93.33** (7)	**82.67** (4)	**78.00** (7)

Tabela 3.5: Acurácia obtida na classificação usando diferentes taxonomias na WT10g. Cada linha representa a combinação de características propostas, removendo uma característica do conjunto. A linha *nenhuma* representa o caso em que nenhuma característica é removida

buição é grande o suficiente para melhorar os resultados em relação a taxonomia C_{todas} na WBR03. Ganhos em relação a taxonomia C_{todas} na WT10g não foram significantes. Como podemos observar, em geral, *dda* é mais útil que *af* em ambas as coleções.

Entre as características baseadas no conteúdo dos documentos, *title* foi útil para identificar consultas navegacionais e transacionais na WT10g. Por sua vez, *ddt* foi muito útil na WT10g, na tarefa de distinguir consultas navegacionais. Note, no entanto, que ela introduziu algum ruído quando aplicada para identificar consultas transacionais na mesma coleção. Em geral, para a WBR03, ambas as características tiveram pouco impacto ou

introduziram ruído.

Características baseadas em URL apresentaram resultados diferenciados, sendo mais úteis na tarefa de identificar consultas navegacionais na WBR03. Particularmente, *url* melhorou ligeiramente os resultados em ambas as coleções em todas as tarefas de classificação. Diferentemente de *url*, *mqd* foi melhor na WBR03 do que na WT10g, sendo claramente mais efetiva que *url* na WBR03. O inverso foi observado na WT10g, na qual intruduziu perdas.

A popularidade da consulta *qpop* foi efetiva em todas as tarefas de classificação de consultas estudadas aqui. Esta característica sempre melhorou os resultados na tarefa de identificar consultas navegacionais. É importante notar que sua efetividade apenas acontece quando ela é combinada com outras características, aumentando sua natureza discriminativa. Quando usada isoladamente, não apresenta impacto na acurácia. Para entender melhor como *qpop* relaciona-se com outras características, a Tabela 3.6 apresenta os resultados obtidos quando combinamos cada uma das características com *qpop* na WBR03 usando a taxonomia C_{todas}. A partir desta tabela, podemos observar que as características baseadas em textos de âncora e conteúdo dos documentos são aquelas que tiram mais vantagens de *qpop*. Estas características exploram diferenças na distribuição das ocorrências dos termos da consulta nos textos de âncora e conteúdo dos documentos observados em consultas navegacionais, quando comparados com consultas informacionais. Essas diferenças na distribuição dos termos são reforçadas quando a popularidade da consulta é levada em consideração como ilustrada na Figura 3.1. Esta figura retrata o efeito de combinar *qpop* com *ddt* e *dda*. Note que, apesar de *ddt* ter sido a característica que mais ganho obteve após a combinação com *qpop*, *dda* foi a que mais ganhou entre as características mais efetivas na classificação. Como se pode observar, os valores de *dda* são menores para consultas transacionais e maiores para consultas informacionais, com os valores para consultas navegacionais aparecendo entre os dois tipos anteriores.

Quando a popularidade é considerada, os resultados melhoram porque consultas navegacionais são normalmente mais populares que os outros tipos de consultas, tornando-se

mais fácil de distinguir. Isto também é observado na combinação de *ddt* com *qpop*. Neste caso, no entanto, é muito mais difícil classificar consultas quando somente se considera os valores da distribuição de *ddt*, o que explica os ganhos significativos obtidos depois da combinação.

qpop +	terms	dda	mqd	url	title	af	#terms	ddt
acurácia	72.50	65.67	58.33	58.50	56.33	54.83	50.50	53.33
ganho	1%	10%	4%	8%	8%	11%	3%	21 %

Tabela 3.6: Acurácia na classificação obtida usando a taxonomia C_{todas} e cada uma das características propostas combinadas com *qpop* na coleção WBR03. Os ganhos foram calculados em relação ao uso isolado das características.

(a) (b)

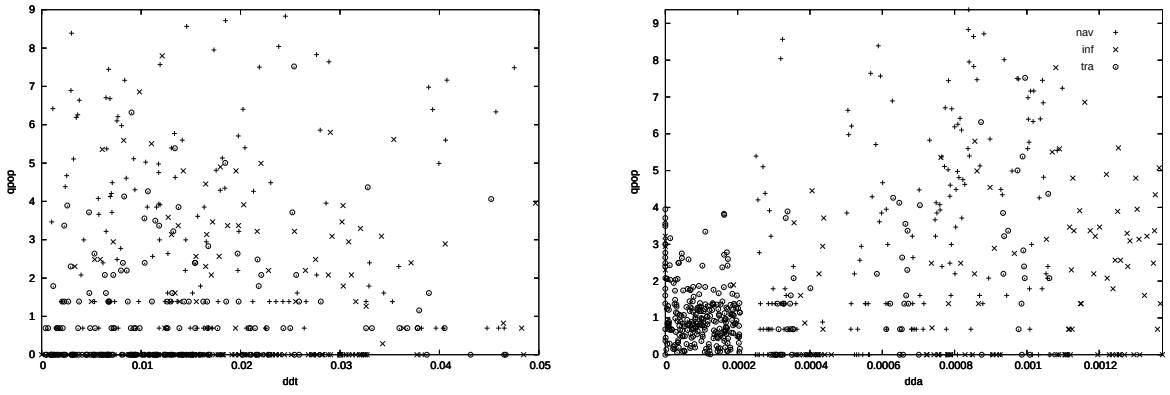

Figura 3.1: Efeito da combinação da popularidade com as características (a) *ddt* e (b) *dda*. Para melhorar a visualização, os eixos estão em escala logarítmica.

A partir desses resultados podemos notar que, em geral, *af* e *#terms* não são úteis para identificar consultas informacionais. Em particular, *af* também apresenta fraco desempenho quando aplicada a consultas navegacionais. Além disso, alguns resultados são muito diferentes entre as coleções, o que pode ser explicado pela fonte de onde foram obtidas as consultas, a qualidade da informação de texto e conteúdo dos apontadores disponível. Diferentemente da coleção WT10g, as consultas utilizadas da WBR03 foram extraídas de um único log de máquina de busca e representam o comportamento real dos

usuários, fato que tem forte impacto, principalmente, em consultas transacionais. Além disso, informações de apontadores são mais comuns na WBR03 do que na WT10g, o que leva a diferentes desempenhos para as características *terms*, *title*, *ddt* e *mqd*.

3.3.2 Comparação com outros trabalhos

Nesta seção, comparamos nosso método utilizando diferentes representações de consultas com outros métodos anteriormente propostos, mais especificamente, aqueles apresentados em [Kang and Kim, 2003, Kang and Kim, 2004] e [Kang, 2005]. A partir de agora, nos referimos a estes métodos, respectivamente, como KANG1 e KANG2. É importante notar que as implementações originais destes métodos não estão disponíveis publicamente. Além disso, não foi possível implementá-los devido a falta de uma descrição detalhada de tais abordagens. Dessa forma, nossa comparação é baseada nos resultados destes métodos sobre a mesma coleção de documentos, tarefas de classificação e conjunto de consultas.

Adotamos dois conjuntos de consultas para estas comparações. O primeiro, foi utilizado nos experimentos reportados em [Kang and Kim, 2003, Kang and Kim, 2004] na tarefa de classificar consultas informacionais e navegacionais. Esse conjunto contém 150 consultas de treinamento (100 navegacionais e 50 informacionais) e 195 consultas de teste (145 navegacionais e 50 informacionais). O segundo conjunto de consultas é aquele utilizado em [Kang, 2005] na tarefa de classificar consultas navegacionais, informacionais e transacionais. Este conjunto consiste de 200 consultas de treinamento (100 navegacionais, 50 informacionais e 50 transacionais) e 245 consultas de teste (145 navegacionais, 50 informacionais e 50 transacionais).

A Tabela 3.7 mostra valores de precisão, revocação e Medida F1 para KANG1, KANG2 e nosso classificador SVM usando diferentes representações de consultas. A comparação foi realizada para as mesmas tarefas de classificação apresentadas em [Kang and Kim, 2003, Kang and Kim, 2004] e [Kang, 2005], isto é, classificando consultas como (a) navegacional ou informacional e (b) navegacional, informacional ou transacional. Para ambas tarefas, mostramos os resultados alcançados com todas as características estudadas e algumas

combinações sem *terms*. Ao fazermos isso, podemos analisar o impacto da característica *terms* no conjunto de consultas, a qual é tendenciosa para consultas navegacionais.

Note que a primeira tarefa é aquela semelhante a tarefa de distinguir consultas navegacionais de não navegacionais que estudamos anteriormente. Dessa forma, para a primeira tarefa de classificação, também reportamos os resultados considerando a representação das consultas sem *af* e *mqd*, as características que apresentaram os piores resultados nesta tarefa na coleção WT10g. De maneira similar, para a segunda tarefa de classificação, reportamos os resultados considerando a representação das consultas sem *#terms*, *af* e *mqd*, as quais foram as características que apresentaram piores resultados em classificar todos os tipos de consultas na WT10g.

Método	Precisão	Revocação	Medida F1
Tarefa 1: navegacional x informacional			
KANG1	91.70	61.50	73.62
SVM TC	90.77	90.77	90.77
SVM TC exceto *terms*	91.79	91.79	91.79
SVM TC exceto *#terms*, *af*, *mqd*	90.26	90.26	90.26
SVM TC exceto *terms*, *#terms*, *af*, *mqd*	**94.87**	**94.87**	**94.87**
Tarefa 2: navegacional x informacional x transacional			
KANG2	78.00	78.00	78.00
SVM TC	69.79	69.79	69.79
SVM TC exceto *terms*	78.37	78.37	78.37
SVM TC exceto *dda*, *af*, *mqd*	65.71	65.71	65.71
SVM TC exceto *terms*, *dda*, *af*, *mqd*	**79.18**	**79.18**	**79.18**

Tabela 3.7: Comparação entre propostas anteriores e nossa proposta usando SVM com diferentes representações de consultas. TC indica que o conjunto de todas as características foi empregado

Para ambas tarefas de classificação, os melhores resultados foram obtidos com a remoção da característica *terms* junto com aquelas que apresentaram os piores resultados em nossa análise. Dessa forma, para a primeira tarefa, a melhor representação de consultas não contém *terms*, *#terms*, *af* e *mqd*. Podemos notar que nosso classificador supera KANG1 com grande margem, devido a sua melhor revocação, já que o método consegue classificar todas as consultas.

Na segunda tarefa de classificação, nossa representação de consultas que contém todas as características menos *terms*, *dda*, *af*, *mqd* foi a que apresentou melhor desempenho,

melhorando ligeiramente os resultados em relação ao KANG2. Para ambas as tarefas, o uso do vocabulário das consultas (característica *terms*) não melhora os resultados de classificação. Isso pode ser atribuído a grande quantidade de consultas navegacionais no conjunto de teste que foi utilizado. Como verificamos na Tabela 3.5, a característica *terms* apresenta um fraco desempenho para distinguir consultas navegacionais na WT10g. Além disso, para estas comparações, usamos a mesma metodologia e o mesmo conjunto de consultas de treinamento usados em [Kang, 2005]. Como consequência, o vocabulário utilizado nas comparações com KANG1 e KANG2 é menor em relação a WBR03, fato que tem forte impacto negativo em *terms*. Além disso, é importante destacar que o conjunto de consultas utilizado poderia ter desempenho melhor se características adicionais, como a popularidade, tivessem sido empregadas.

3.3.3 Análise de erros

Nesta seção, avaliamos os erros de classificação de consultas observados usando nossa taxonomia mais geral, C_{todas}. Particularmente, analisamos 105 consultas na WBR03 e 67 consultas na WT10g, as quais não foram classificadas nas classes escolhidas pelos avaliadores humanos. O objetivo desta análise é evidenciar as possíveis razões para as decisões dos classificadores automáticos. Os erros estudados são apresentados nas Tabelas 3.8 e 3.9. Nas tabelas, os termos "nav", "inf" e "tra" representam, respectivamente, as consultas navegacionais, informacionais e transacionais. Além disso, as palavras *consulta*, *âncora*, *url*, *conteúdo* e *log* representam o conjunto de características usado, onde as características estão agrupadas pela fonte de informação de onde foram extraídas.

Classe Correta	nav		inf		tra		total de
Classe Assinalada	inf	tra	nav	tra	nav	inf	erros
todas	26%	2%	18%	10%	4%	**40%**	105
consulta	11%	2%	**49%**	8%	12%	18%	171
âncora	4%	23%	13%	**42%**	16%	1%	226
url	8%	10%	15%	**43%**	10%	14%	240
conteúdo	15%	5%	**29%**	14%	17%	20%	281
log	**38%**	18%	2%	6%	5%	31%	340

Tabela 3.8: Erros na coleção WBR03 usando a taxonomia C_{todas}.

Classe Correta	nav		inf		tra		total de
Classe Assinalada	inf	tra	nav	tra	nav	inf	erros
todas	3%	15%	10%	**42**%	15%	15%	67
consulta	2%	10%	**46**%	6%	31%	6%	101
âncora	0%	40%	0%	**55**%	1%	4%	151
url	28%	10%	8%	13%	8%	**34**%	156
conteúdo	7%	12%	2%	29%	2%	**48**%	109

Tabela 3.9: Erros na coleção WT10g usando a taxonomia C_{todas}.

Como resultados das análises, observamos que nas Tabelas 3.8 e 3.9 os erros mais comuns envolvem consultas informacionais, independentemente do grupo de características usado, demostrando que este tipo de consultas é claramente mais ambíguo. Em particular, quando utilizamos todas as características, o erro mais comum é confundir consultas transacionais com consultas informacionais. Quando analisamos os valores das características para consultas informacionais e transacionais, percebemos que tais valores são semelhantes em muitos casos. Por exemplo, o número médio de termos na consulta é alto em ambas categorias quando comparadas com consultas navegacionais e a informação de textos de âncora relacionada com as consultas em ambos os casos (informacionais e transacionais) é normalmente encontrada em domínios distintos. É importante destacar que informações de clique do usuário provavelmente não reduziriam esse tipo de erros de forma significativa, uma vez que os padrões de clique para consultas informacionais e transacionais são similares. Em ambas categorias, os cliques dos usuários estão distribuídos entre as diversas páginas alvo possíveis.

As características baseadas no conteúdo das páginas apresentam um comportamento diferente entre as coleções, com os erros mais distribuídos na WBR03 e muito concentrados na WT10g, onde a metade dos erros consiste em classificar consultas transacionais como consultas informacionais. Quando verificamos os valores das características, percebemos que isso é consequência da diferença de informação disponível em cada coleção. Por exemplo, a estratégia de coleta empregada para criar as coleções é diferente, com a WBR03 sendo criada dando prioridade ao número de domínios cobertos, enquanto que WT10g foi criada por um coletor que tentou obter uma boa cobertura de cada domínio. Como

resultado, a coleção WT10g contém, em média, cerca de 144 documentos por domínio, enquanto que a WT10g contém apenas 12 documentos por domínio. Essa melhor cobertura de cada domínio na WT10g contribuiu para permitir o melhor desempenho das características como *ddt*, visto que existe mais informação de texto por domínio nessa coleção.

Outro exemplo das diferenças entre as duas coleções é a pouca disponibilidade de informação de textos de âncora na WT10g quando comparada a WBR03. A coleção WBR03 contém, em média, 200 textos de âncora por documento, enquanto que na WT10g essa média é aproximadamente 5. Esta diferença afeta negativamente a qualidade das características baseadas em textos de âncora na WT10g. É importante deixar claro que, conclusões sobre características propostas e experimentos com outras coleções, tal como aquelas reportadas em [Lee et al., 2005] e [Lu et al., 2006], são similares as conclusões obtidas quando experimentamos estas características na WBR03.

Após a inspeção manual dos erros de classificação de consultas, notamos que muitas dessas consultas poderiam ser classificadas em mais de uma classe. Visto que cada consulta na WBR03 foi avaliada por três usuários, podemos estudar, nesta coleção, os casos em que consultas podem ter multi-classes. A Tabela 3.10 mostra o número de erros em duas situações. A primeira, consideramos como correta somente a classe majoritária assinalada pelos avaliadores. Na segunda situação, consideramos como correta qualquer uma das classes assinaladas. Podemos verificar que dos 105 erros originais observados para a configuração de classe majoritária (com acurácia de aproximadamente 82%), 69 erros podem ser considerados como acertos na configuração multi-classe (com acurácia de 93%). Para as 36 consultas restantes, o classificador automático não conseguiu assinalar a classe escolhida por alguns dos avaliadores.

A Tabela 3.10 também apresenta exemplos destas consultas. Uma análise cuidadosa dos erros indica que não existe uma razão específica para que esses erros aconteçam. Por exemplo, a consulta "TIM telecomunicações" foi classificada incorretamente como informacional, sendo navegacional a classe correta. Esta companhia é normalmente re-

Classe Real	Classe Assinalada	erros classe única	erros multi-classe	Exemplos
nav	inf	27	7	Celeranet, TIM telecomunicações
nav	tra	2	2	Anhembi Veículos
inf	nav	19	10	Paulo Zulu, Aranha, Células Eucariontes
inf	tra	11	4	Como adornar paredes
tra	nav	4	2	Papel de parede do São Paulo
tra	inf	42	11	Driver EP320XS, mapa de Goiás

Tabela 3.10: Erros na coleção WBR03, usando a taxonomia C_{todas} para as configurações de classe única e multi-classe.

ferenciada apenas com a palavra "TIM" e a inclusão da palavra "telecomunicações" fez com que a distribuição de domínios para a consulta seja similar àquelas encontradas em consultas informacionais. Acreditamos que neste caso, a inclusão de novas características, tal como a informação de clique do usuários em logs de consultas passadas, pode contribuir para reduzir essa taxa de erros. No entanto, quando examinamos os erros, podemos concluir que é praticamente impossível prover sempre a classe correta para cada consulta usando um classificador automático.

3.3.4 Questões sobre performance

Uma vez que a classificação é realizada no momento do processamento da consulta pela máquina de busca, o bom desempenho do sistema de classificação é uma questão importante. O projeto de tal sistema vai depender, fundamentalmente, da taxa de novas consultas sendo submetidas, já que as categorias de consultas anteriores podem ser armazenadas, reduzindo o custo de classificação para menos da metade [Saraiva et al., 2001]. Além disso, o processo de aprendizagem, chamado também de fase de treinamento, pode ser realizado separadamente, não afetando o tempo de processamento da consulta, nem a experiência do usuário com o sistema. Portanto, o processo mais crítico consiste em determinar o valor das características que serão usadas para representar as consultas submetidas. De nossa lista de características, *#terms*, *terms* e *qpop* não afetam o tempo de processamento de modo significante. O cálculo da distribuição de palavras para cada nome de domínio em *mqd* também pode ser feito separadamente, armazenando o resultado

em estruturas de dados que permitam um acesso rápido e eficiente.

As características restantes, *af*, *dda*, *ddt*, *title* e *url*, são as que demandam um custo maior para serem computadas. Para que as estatísticas dessas características sejam calculadas, é necessário que primeiro seja recuperado um conjunto de respostas. Dessa forma, o custo é proporcional ao tamanho desse conjunto. Da lista de características estudadas, a que demanda maior custo é a característica *af*, uma vez que a consulta sempre é processada como uma frase.

Todas as fontes de informação de onde estamos extraindo características já têm sido mencionadas na literatura como fontes possíveis e úteis para computar respostas de máquinas de busca [Joachims, 2002, Liu et al., 2007, Silva et al., 2009]. Dessa forma, os valores dessas características podem ser extraídos simultaneamente para os propósitos de classificação de consultas e cálculo das respostas, reduzindo o custo demandado pelo sistema de classificação em máquinas de busca reais. Portanto, o impacto final no desempenho do sistema pode não ser elevado na prática.

Capítulo 4

Conclusão

A evolução das máquinas de busca da web depende fortemente de sua habilidade para lidar apropriadamente com as diferentes finalidades de busca dos usuários. Nesta dissertação, propomos um método para *aprender* automaticamente a utilizar, de forma efetiva, diversas fontes de informação para determinar a finalidade da consulta do usuário entre três possibilidades: navegar na web, encontrar informações e realizar uma transação. Modelamos essa tarefa com um problema de classificação e estudamos o impacto de um conjunto de características na acurácia do classificador.

Experimentamos com sucesso novas maneiras de calcular as estatísticas de características previamente propostas. Através de nossos experimentos foi possível mostrar algumas diferenças na escolha do melhor conjunto de características para as coleções WBR03 e WT10g. É importante destacar que o melhor conjunto de características pode depender de acordo com a coleção de documentos adotada. No entanto, salienta-se que as conclusões obtidas neste trabalho sobre o desempenho do conjunto de características nos experimentos na WBR03 são similares àquelas obtidas em trabalhos anteriores com outras coleções de documentos web [Lee et al., 2005, Lu et al., 2006], o que indica que o comportamento das características na WT10g diverge da WBR03 e de outras coleções adotadas na literatura.

Nossos experimentos indicam que a popularidade da consulta, a qual foi proposta neste trabalho, é uma característica importante que melhora significativamente os re-

sultados de classificação. Além disso, também mostramos que características baseadas em domínios podem ser usadas com sucesso em tarefas de classificação. A característica *terms*, estudada neste trabalho, provou ser especialmente útil para identificar consultas transacionais. Finalmente, também estudamos o desempenho da característica número de termos (*#terms*), a qual foi mencionada em trabalhos anteriores, mas não aplicada em classificação de consultas. Os resultados indicaram que esta característica introduz perdas em algumas tarefas de classificação. Contudo, quando usamos nosso melhor conjunto de características, alcançamos melhores resultados quando comparados a abordagens propostas anteriormente.

Como direções futuras, pode-se fazer um estudo incluindo o uso outras fontes de informação não incluídas neste trabalho. Um exemplo é a informação de clique dos usuários, que de acordo com trabalhos anteriores, serve para ajudar a identificar o tipo de uma consulta após um certo número de vezes que a mesma é submetida. Outro exemplo que consideramos interessante é o uso de informação existente sobre coleções de anúncios publicitários como fonte de informação para ajudar na identificação de conteúdo transacional.

Também pode-se, ainda como trabalho futuro, analisar a aplicação do nosso método em tarefas de classificação de consultas em taxonomias distintas das estudadas aqui, como, por exemplo, a classificação de consultas de acordo com o tópico mais relacionado à mesma. Pode-se estudar o impacto das características propostas aquí quando empregadas a esses novos cenários ou ainda buscar novas características adequadas às necessidades específicas de tais aplicações.

Referências Bibliográficas

[Baeza-Yates et al., 2006] Baeza-Yates, R. A., Calderon-Benavides, L., and Gonzalez-Caro, C. N. (2006). The intention behind web queries. In *Proceedings of SPIRE*, pages 98–109.

[Bailey et al., 2003] Bailey, P., Craswell, N., and Hawking, D. (2003). Engineering a multi-purpose test collection for web retrieval experiments. *Inf. Process. Manage.*, 39(6):853–871.

[Broder, 2002] Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, 36(2):3–10.

[Chang and Lin, 2001] Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[Choo et al., 1999] Choo, C., Detlor, B., and Turnbull, D. (1999). Information Seeking on the Web–An Integrated Model of Browsing and Searching. In *Proceedings of the ASIS Annual Meeting*, volume 36, pages 3–16.

[Craswell et al., 2001] Craswell, N., Hawking, D., and Robertson, S. (2001). Effective site finding using link anchor information. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 250–257, New York, NY, USA. ACM Press.

[Hawking et al., 1999] Hawking, D., Voorhees, E., Craswell, N., and Bailey, P. (1999). Overview of the trec8 web track. In *8th Text REtrieval Conference*.

[Jansen et al., 2005] Jansen, B., Spink, A., and Pederson, J. (2005). Trend analysis of AltaVista Web searching. *Journal of the American Society for Information Science and Technology*, 56(6):559–570.

[Joachims, 1998] Joachims, T. (1998). Text categorization with suport vector machines: Learning with many relevant features. In *ECML '98: Proceedings of the 10th European Conference on Machine Learning*, pages 137–142, London, UK. Springer-Verlag.

[Joachims, 2002] Joachims, T. (2002). Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, New York, NY, USA. ACM.

[Kang, 2005] Kang, I.-H. (2005). Transactional query identification in web search. In Lee, G. G., Yamada, A., Meng, H., and Myaeng, S.-H., editors, *AIRS*, volume 3689 of *Lecture Notes in Computer Science*, pages 221–232. Springer.

[Kang and Kim, 2004] Kang, I.-H. and Kim, G. (2004). Integration of multiple evidences based on a query type for web search. *Information Processing and Management*, 40(3):459–478.

[Kang and Kim, 2003] Kang, I.-H. and Kim, G. C. (2003). Query type classification for web document retrieval. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 64–71, New York, NY, USA. ACM Press.

[Lee et al., 2005] Lee, U., Liu, Z., and Cho, J. (2005). Automatic identification of user goals in web search. In *Proceedings of World Wide Web Conference*, pages 391–400.

[Li et al., 2006] Li, Y., Krishnamurthy, R., Vaithyanathan, S., and Jagadish, H. V. (2006). Getting work done on the web: Supporting transactional queries. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 557–564, New York, NY, USA. ACM Press.

[Liu et al., 2007] Liu, T.-Y., J. Xu, T. Q., Xiong, W.-Y., and Li, H. (2007). Letor: Benchmark dataset for research on learning to rank for information retrieval. In *SIGIR 2007 Workshop on learning to rank for information retrieval*.

[Lu et al., 2006] Lu, Y., Peng, F., Li, X., and Ahmed, N. (2006). Coupling feature selection and machine learning methods for navigational query identification. In *CIKM '06: Proceedings of the 15th international conference on Information and knowledge management*, pages 682–689. ACM Press.

[Mitchell, 1997] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York.

[Muramatsu and Pratt, 2001] Muramatsu, J. and Pratt, W. (2001). Transparent Queries: investigation users' mental models of search engines. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 217–224. ACM New York, NY, USA.

[Navarro-Prieto et al., 1999] Navarro-Prieto, R., Scaife, M., and Rogers, Y. (1999). Cognitive strategies in web searching. In *Proceedings of the 5th Conference on Human Factors & the Web*.

[Rose and Levinson, 2004] Rose, D. E. and Levinson, D. (2004). Understanding user goals in web search. In *Proceedings of World Wide Web Conference*, pages 13–19.

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

[Saraiva et al., 2001] Saraiva, P. C., de Moura, E. S., Ziviani, N., Meira, W., Fonseca, R., and Riberio-Neto, B. (2001). Rank-preserving two-level caching for scalable search engines. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 51–58, New York, NY, USA. ACM.

[Silva et al., 2009] Silva, T. P. C., de Moura, E. S., Cavalcanti, J. M. B., da Silva, A. S., de Carvalho, M. G., and Goncalves, M. A. (2009). An evolutionary approach for com-

bining different sources of evidence in search engines. *Information Systems*, 34(2):276–289.

[Spink and Jansen, 2004] Spink, A. and Jansen, B. (2004). *Web search: Public searching on the Web*. Springer.

[Wilcoxon, 1945] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, (1):80–93.