EXPLORAÇÃO DA POPULARIDADE PARA BUSCA DE INFORMAÇÃO EM BLOGS

LUIZ GUILHERME PAIS DOS SANTOS
Orientador: Marcos André Gonçalves

EXPLORAÇÃO DA POPULARIDADE PARA BUSCA DE INFORMAÇÃO EM BLOGS

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

Belo Horizonte
Fevereiro de 2009

Santos, Luiz Guilherme Pais dos
S237e Exploração da Popularidade para Busca de Informação em Blogs / Luiz Guilherme Pais dos Santos. — Belo Horizonte, 2009
x, 37 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Marcos André Gonçalves

1. Recuperação de informação - Teses. 2. Banco de dados - Teses. 3. Blogs - Teses. I. Título.

CDU 519.6*73(043)

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

FOLHA DE APROVAÇÃO

Exploração da Popularidade para Busca de Informação em Blogs

LUIZ GUILHERME PAIS DOS SANTOS

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. MARCOS ANDRÉ GONÇALVES - Orientador
Departamento de Ciência da Computação - UFMG

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

PROF. EDLENO SILVA DE MOURA
Departamento de Ciência da Computação - UFAM

PROF. CLODOVEU AUGUSTO DAVIS JÚNIOR
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 13 de julho de 2009.

Aos meus pais Luiz e Fernanda pelo apoio incondicinal.

Agradecimentos

Agradeço primeiramente a Deus pela oportunidade.

À minha família, meus pais por sempre me apoiarem e estarem presentes mesmo nos momentos mais difíceis e minhas irmãs que sempre torceram por mim.

Ao Professor Marcos André Gonçalves e ao Professor Alberto H. F. Laender que me orientaram no desenvolvimento desse trabalho.

Agradeço também aos colegas do Laboratório de Banco de Dados pela companhia, pelos experimentos realizados indispensáveis para minha dissertação e pelas idéias que sem dúvida contribuíram para os resultados aqui alcançados.

Aos colegas do Synergia, que estiveram presentes durante todo período do mestrado. E aos companheiros do melhor time do campeonato de futebol do DCC, o Livernull!

Resumo

A *blogosfera* é um subconjunto da Web altamente dinâmico e conectado que despertou um grande interesse devido à sua natureza social. Nesta dissertação, é apresentado um estudo de um importante aspecto dos *blogs*, a popularidade. Os *blogs* mais populares de quatro importantes domínios brasileiros foram coletados por um período de tempo considerável para obter informações sobre aqueles mais populares. Os experimentos realizados, com a ajuda de vários voluntários, mostram que apesar de a blogosfera ser uma rede social, a popularidade tem sido subutilizada, pelo menos pelas máquinas de busca mais importantes no contexto de busca de *blogs*. Nos experimentos, consultas especificamente formuladas para recuperar esses *blogs* populares não foram capazes de trazê-los entre as primeiras posições (100 primeiros) nas principais máquinas de busca. Mais ainda, os valores de PageRank desses *blogs* populares são também muito baixos. É mostrado, ainda, que incorporar explicitamente a popularidade ao algoritmo de ordenação de consultas de uma máquina de busca produz resultados que foram considerados pelos voluntários, de modo geral, bastante relevantes.

Abstract

The blogosphere is a highly dynamic and interconnected subset of the Web that has triggered a lot of interest due to its social and personal nature. In this dissertation, we present a study of an important social aspect of these blogs, namely popularity. The most popular blogs from four important blog domains in Brazil were crawled for a considerable period of time in order to collect information about the most popular blogs. The experiments, conducted with several volunteers, show that despite the blogosphere being a social network, popularity has been underexplored by at least the most popular search engines in the context of blog search. In the experiments, queries specifically formulated for retrieving these popular blogs were not capable of ranking them in the top positions (top 100) of the most popular search engines, and their page ranks, as measured by the typical web graph topology of links, are very low. It is also shown that explicitly incorporating popularity in the search engine algorithm produces rankings which were considered by volunteers, in general, very relevant.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O crescimento da utilização de *blogs* na Internet criou um subconjunto da Web altamente conectada e dinâmica que é conhecida como *blogosfera*. Devido à sua inerente natureza pessoal e social, a *blogosfera* se tornou tema de um grande número de pesquisas que consideram tanto o seu conteúdo quanto a sua estrutura [Mishne e de Rijke, 2006]. Esses estudos procuram responder questões como: "É possível processar *blogs* automaticamente para descobrir opiniões ou sentimentos sobre algum produto?" ou "Qual é a dinâmica da *blogosfera*?".

O número de *blogs* cresceu exponencialmente desde o começo da década de noventa até a atualidade. Esse número que girava em torno de alguns milhares, atualmente chega a mais de uma centena de milhão[1]. Esse crescimento expressivo gerou a necessidade de mecanismos de acessos eficientes, como, por exemplo, através de máquinas de busca. Atualmente existem vários serviços de busca oferecidos por vários sítios na Web, alguns deles especializados em busca de *blogs* (e.g., GoogleBlogSearch[2] e Technorati[3]).

A *blogosfera* tem, por sua natureza, um conteúdo mais extrovertido e uma linguagem bastante informal se comparada com páginas da Web tradicional. Muitos *blogs* são criados por seus próprios autores não com o objetivo de atingir uma audiência considerável, mas sim, servir como um mecanismo de expressão pessoal ou divertimento.

[1] http://en.wikipedia.org/wiki/Technorati
[2] http://blogsearch.google.com
[3] http://technorati.com

Apesar disso, vários deles possuem um número de visitas elevado [Ounis et al., 2006], o que significa que existe um grande interesse sobre os mesmos. Isso mostra também que existe um nicho com um grande potencial de exploração comercial como a veiculação de peças publicitárias e sistemas de recomendação [Stewart et al., 2007].

As máquinas de busca tradicionais podem obviamente ser utilizadas para procurar por *blogs*, especialmente quando o usuário sabe exatamente as palavras-chave que definem o *blog* procurado. Entretanto, máquinas de busca especializadas podem potencialmente satisfazer melhor as necessidades dos usuários se elas disponibilizarem mecanismos específicos que aproveitem melhor as características intrínsecas da *blogosfera*, diferenciando-se, assim, das máquinas de busca tradicionais.

Uma análise prévia de mais de 35 milhões de requisições enviadas a um grande servidor de *blogs* no Brasil concluiu que uma grande parte (cerca de 46%) do tráfego para os *blogs* tem origem nas máquinas de busca [Duarte et al., 2007]. Nesse mesmo estudo, os autores observaram que a maioria dos *blogs* mais populares são geralmente acessados mais facilmente através de *links* de outros *blogs* e não através de máquinas de busca. Mesmo as máquinas de busca sendo responsáveis pela maioria do tráfego na *blogosfera*, elas não foram capazes de alcançar os *blogs* mais populares como era esperado. Em outras palavras, a intensidade do tráfego direcionado a um *blog* através de uma máquina de busca parece não condizer com a real popularidade do mesmo. Como os usuários tipicamente clicam nos primeiros resultados, isso pode ser uma grande evidência que as máquinas de busca não estão considerando a popularidade como uma característica para ordenar os resultados das consultas quando o objetivo da busca são *blogs*. Isso evidencia a necessidade de se desenvolver estratégias de ordenação de consultas adequadas ao contexto de busca especializada de *blogs* e a integração de conceitos sociais com técnicas já conhecidas de recuperação de informação pode ser utilizada como meio de melhorar consideravelmente a qualidade das buscas como sugerido por [Mislove et al., 2006].

Para ser mais preciso, popularidade é considerada aqui como uma relação intrínseca

entre o comportamento coletivo de uma dada comunidade e um determinado objeto (por exemplo, um blog). Isso significa que uma parcela significativa da comunidade gosta, aprova ou julga o objeto adequado em um determinado contexto. Assume-se que um indicador de popularidade pode ser associado com este relacionamento, o que nos permite quantificar o nível de popularidade de um determinado objeto e comparar vários objetos de acordo com a sua popularidade relativa. Exemplos desses indicadores incluem número de visitas, *downloads* e, até mesmo, aspectos socialmente orientados, como o número de anotações sociais em conteúdo gerado pelo usuário [Bao et al., 2007]. Para *blogs*, especificamente, outros indicadores de popularidade incluem o número de indivíduos que assinam o RSS (*Really Simple Syndication*), a taxa de cliques [Baehni et al., 2007] e, como aqui considerado, o número de vezes que o *blog* apareceu nas listas de destaque dos domínios.

O foco principal desta dissertação é a busca de *blogs*, partindo do ponto de vista que a *blogosfera* pode ser considerada uma rede social onde a popularidade de cada *blog* desempenha um importante papel [Ali-Hasan e Adamic, 2007]. Inicialmente é analisada a qualidade da busca de *blogs* nas máquinas de busca atuais. Em geral, esperaria-se que uma busca bem sucedida na *blogosfera* retornasse não somente documentos relevantes (i.e., *blogs*), mas idealmente os mais populares, como, no caso de uma busca em uma rede social. Verificou-se, no entanto, que isso não acontece atualmente. Quatro importantes domínios de *blogs* brasileiros foram monitorados por um período de tempo considerável para extrair o endereço de seus *blogs* mais populares. Nos experimentos, consultas especificamente formuladas por voluntários para recuperar esses *blogs* não foram capazes de retorná-los nas primeiras posições da lista de resultados da consulta em duas máquinas de busca amplamente utilizadas na Web. Mais ainda, os respectivos valores do *PageRank* desses *blogs*, como medidos para uma típica topologia de grafo da Web, foram considerados muito baixos.

Além disso, com o intuito de investigar o potencial de explorar a popularidade na busca de *blogs*, alguns experimentos foram propostos onde a popularidade foi explici-

tamente incorporada junto a técnicas de recuperação de informação já conhecidas. Ao fazer isso, a ordenação das consultas foi considerada muito relevante por voluntários e muito melhor (até 63% de melhora considerando-se a métrica *Normalized Discounted Cumulative Gain*) que a original onde não se usava o fator de popularidade.

1.1 Objetivos e Contribuições

As principais contribuições dessa dissertação são: (a) investigar a capacidade das máquinas de busca atuais em explorar a popularidade dos *blogs* para evidenciar possíveis melhorias que podem ser feitas devido às características particulares dos *blogs*; (b) apresentar uma nova forma de ordenar o resultado das consultas feitas a uma máquina de busca para *blogs* utilizando a sua popularidade como mais um fator no cálculo da similaridade entre a consulta e o documento.

Para isso, foram monitoradas as lista dos *blogs* mais populares dos domínios estudados, listas essas que os próprios domínios disponibilizam em suas páginas iniciais. Em seguida, como forma de avaliar a importância que as máquinas de busca atribuem a esses *blogs*, analisou-se o *PageRank* de cada um deles que, de forma geral, e se mostrou muito baixo não ultrapassando o valor 4, em uma escala entre 0 e 10. Além do *PageRank*, também foi analisada a posição que cada um desses *blogs* aparecia nos resultados de consultas realizadas no Google e no Yahoo! (onde foram utilizadas palavras-chaves atribuídas por voluntários baseadas no conteúdo de cada *blog*).

Baseado nesses resultados, uma nova estratégia para ordenação de resultados de consultas foi proposta levando em consideração a popularidade de cada *blog*. Foi feita uma coleta da *blogosfera* da UOL e construída uma máquina de busca onde a cada *blog* é atribuído um fator de ajuste de acordo com a sua popularidade estimada. Os resultados foram validados com usuários e as métricas indicaram ganhos significativos na qualidade das respostas.

1.2 Trabalhos Relacionados

Existe um grande esforço de pesquisa relacionado à melhoria de métodos para busca de *blogs*, onde algumas iniciativas exploram o comportamento dos usuários através da coleta de dados da navegação e outras exploram as características particulares dos *blogs*. Aqui serão citados somente os trabalhos mais relacionados ao problema central abordado nesta dissertação.

Duarte et al. [2007] utilizaram uma grande quantidade de evidências para caracterizar o padrão de acesso à *blogosfera* a partir de três perspectivas diferentes. A perspectiva do servidor descreve o padrão de acesso de todos os usuários a todos os *blogs*, a perspectiva do usuário descreve como um indivíduo interage com os objetos da *blogosfera* e a perspectiva dos objetos descreve como cada *blog* é acessado. As principais conclusões são que a natureza das interações entre usuários e os objetos são diferentes das observadas na Web tradicional e que o padrão de acesso aos *blogs* é muito dependente das relações sociais. Esta dissertação vai além ao considerar explicitamente a questão da popularidade na qualidade da busca na *blogosfera*.

Mishne e de Rijke [2006] apresentaram uma análise do *log* de uma importante máquina de busca para *blogs*, tendo como foco os tipos de consulta que os usuários endereçavam ao domínio, o comportamento dos usuários no sentido da quantidade de consultas por páginas visitadas e a categoria das consultas. A conclusão foi que a busca de *blogs* é diferente em vários aspectos da busca na Web tradicional em termos do assunto de interesse, uma vez que, as buscas em *blogs* estão mais relacionadas a tecnologia, entretenimento e política, com um interesse particular em eventos atuais. Entretanto, ao navegar pelos resultados da consulta o comportamento é similar; os usuários estão interessados somente nas primeiras posições da lista de resultados retornada. A estratégia de ordenação para máquinas de busca para *blogs* proposta nessa dissertação explora esse principio, ao promover os *blogs* mais importantes (os mais populares) para o topo da ordenação.

Fujimura et al. [2006] propuseram uma nova máquina de busca que leva em consideração as características particulares dos *blogs*. A ferramenta apresenta três tipos diferentes de interface, cada uma com um foco específico: busca por tópico, busca por autor do *blog* e busca por reputação. Embora essa ferramenta represente uma evolução na busca de *blogs*, a busca por reputação não explora a popularidade para ordenar os resultados das consultas.

Mais relacionado a esta dissertação está um método de ordenação de resultados baseado na aplicação do *PageRank* [Brin e Page, 1998] para um grafo estendido com arestas que representam os autores e a semelhança entre o tema dos *blogs* [Kritikopoulos et al., 2006]. Além disso, de acordo com seu pedido de patente, o algoritmo do Google BlogSearch considera a popularidade do blog, avaliada pelo número de leitores de RSS, como um possível indicador positivo de sua qualidade [Baehni et al., 2007]. Outros indicadores incluem taxa de cliques e o *PageRank*. Um estudo recente (ainda preliminar) propôs um novo método de ordenção por popularidade (BRank), que explora diversas interconexões sociais entre os blogueiros [Lin et al., 2009].

Mishne [2007] explorou diversas propriedades dos *blogs* tais como informações temporais, nível de discussão (medida através da taxa de código HTML e XML) e nível de *spam*, com o intuito de melhorar a recuperação de opiniões em *blogs*. Os resultados mostram ganhos significativos ao utilizar as técnicas descritas. Juffinger et al. [2009] estudaram uma outra propriedade dos *blogs*, a credibilidade, para ordenar os resultados das consultas. A credibilidade foi estimada ao comparar aspectos estruturais dos *blogs*, assim como seu conteúdo, com uma base de dados de notícias. Esta dissertação foca em outras propriedades dos *blogs*, obtendo, também, ganhos significativos no contexto de busca de *blogs*.

Mislove et al. [2006] examinaram o potencial de se utilizar as características de redes sociais para melhorar as máquinas de busca. Foram analisadas as diferenças entre a Web e as redes sociais em termos dos mecanismos utilizados para publicar e localizar informações relevantes. Foram discutidos também os benefícios de se integrar os

mecanismos de procura de informação relevante da Web tradicional com os de redes sociais. Os experimentos realizados mostraram um aumento de aproximadamente 8% na qualidade dos resultados ao realizar essa integração. Os experimentos foram conduzidos tendo como foco a busca na Web tradicional, enquanto que nessa dissertação, utiliza-se os conceitos sociais (especificamente a popularidade) no contexto de busca de *blogs*.

Järvelin e Kekäläinen [2000] utilizam uma grande quantidade de informação de tráfego de dados para parcialmente validar o algoritmo de *PageRank* e, acima de tudo, mostrar que padrões navegacionais atuais não são capturados por esse algoritmo. Sugerem que dados disponíveis nos ISPs (*Internet Service Providers*) podem ser utilizados para induzir ordenações de consultas de acordo com a dinâmica do comportamento dos usuários da Web. Sugerem também, que máquinas de busca poderiam firmar parcerias com os ISPs para explorar os benefícios de integrar as informações de tráfego com os algoritmos de ordenação. Liu et al. [2008] vão na mesma direção ao demonstrar que utilizar um "grafo de navegação do usuário" criado a partir de dados de comportamento apresentam resultados melhores que o algoritmo original de *PageRank*. Em harmonia com esses trabalhos, esta dissertação investiga ainda uma outra propriedade, i.e., a popularidade que se incorporada ao algoritmo de ordenação pode melhorar a busca de *blogs*.

Finalmente vale ressaltar que algumas máquinas de busca utilizam o comportamento dos usuários para ordenar os resultados das consultas. Algumas das mais marcantes são A9[4] e Google personalisado[5]. A técnica utilizada é guardar o perfil do usuário e utilizar essa informação em consultas posteriores.

[4]http://www.a9.com
[5]http://www.google.com/psearch

1.3 Organização

Os capítulos seguintes desta dissertação estão organizados da seguinte forma. O Capítulo 2 apresenta dois experimentos importantes. Primeiro, uma análise dos valores do *PageRank* dos *blogs* mais populares de quatro importantes domínios da Web brasileira e, depois, a análise da ordenação desses *blogs* ao realizar consultas em máquinas de busca de uso geral disponíveis na Web. O Capítulo 3 mostra que ao se incorporar o fator de popularidade ao algoritmo de ordenação de consultas a qualidade dos resultados melhora significativamente. Finalmente, o Capítulo 4 apresenta as conclusões desta dissertação e sugere alguns temas para trabalhos futuros.

Capítulo 2

Análise da Popularidade na *Blogosfera*

Neste capítulo é feita uma investigação da qualidade dos resultados da busca por *blogs* em duas importantes máquinas de busca disponíveis na Web: Google e Yahoo!. Para isso, foi feito o monitoramento da lista dos *blogs* mais populares dos domínios estudados. Através dele foi possível analisar o *PageRank* de cada *blog*. Essa análise está descrita na Seção 2.2. Em seguida, atráves de palavras-chave atribuídas por voluntários aos *blogs* com base nos respectivos conteúdos, foi estudada a posição em que cada *blog* apareceu no resultado das consultas realizadas em cada máquina de busca. Essa análise está descrita na Seção 2.3.

2.1 Coleta de Dados

Durante trinta dias, entre 29/03/2008 e 29/04/2008, foi monitorada a lista dos *blogs* mais populares de quatro dos mais importantes e conhecidos domínios da Web brasileira: UOL[1], Blogger[2], BlogLog[3] e Terra[4]. O UOL e o Terra são dois importantes

[1] blog.uol.com.br
[2] www.blogger.com.br
[3] bloglog.globo.com
[4] blog.terra.com.br

portais onde os usuários podem criar seus *blogs* gratuitamente. Blogger Brasil é um domínio de *blogs* que pertence a uma grande empresa de comunicação. A criação de *blogs* no Blogger Brasil requer uma assinatura que é paga. O BlogLog é um domínio de *blogs* restrito onde somente artistas convidados podem manter um *blog*.

Cada domínio utiliza uma estratégia diferente para determinar os *blogs* mais populares. O UOL utiliza um sistema de votação onde os usuários atribuem aos *blogs* uma pontuação entre zero e dez pontos baseada em sua opinião. O Blogger Brasil e o Terra utilizam uma lista semanal determinada por funcionários especializados, onde os *blogs* mais acessados e com melhor recomendação dos usuários são selecionados. O BlogLog utiliza a quantidade de acessos. Todos os domínios disponibilizam essa lista em sua página principal.

A cada dia do período de coleta, a lista dos dez *blogs* mais populares foi coletada, o que resultou em um total de trinta listas de dez *blogs* populares no final do período. Para cada domínio, os *blogs* foram ordenados pelo número de vezes que apareceram na lista dos dez mais populares. Depois foram selecionados os dez *blogs* que mais apareceram na lista dos dez mais populares como sendo os *blogs* mais populares dos domínios. Assim, quarenta *blogs* foram selecionados para análise.

Entre os quarenta *blogs* selecionados, existem alguns que não tinham sido atualizados recentemente no período da coleta, mas, mesmo assim, figuravam na lista dos mais populares dos domínios. Dentre eles pode-se citar Pai de Gêmeos (http://paidegemeos.zip.net), Diário de Letícia (http://diariodeleticia.zip.net) e Kátia Leal (http://katialealsousa.zip.net), nos quais as últimas atualizações ocorreram nos dias 26/06/2007, 20/11/2007 e 06/11/2007, respectivamente. Embora não esperado, esse comportamento pode ser explicado pelas características desses *blogs*. Por exemplo, nos *blogs* Pai de Gêmeos e Diário de Letícia os autores narram durante os nove meses a gravidez pela qual passaram. Ao final desse período, encerraram a narrativa, porém não se pode afirmar que somente pelo fato de não existirem atualizações recentes no *blog* ele deixou de ser popular. Outros usuários interessados em gravidez

podem continuar consultando o *blog* e dando boas notas por considerar uma narrativa interessante e bem escrita. Já o *blog* Katia Leal possuía a última atualização há cerca de três meses do período da coleta. Esse tempo, provavelmente, não foi suficiente para que os usuários percebessem que o *blog* tinha deixado de ser atualizado e parassem de consultá-lo de modo que, durante algum tempo ele ainda figurou entre os dez mais populares do seu domínio.

2.2 Análise do *PageRank*

Umas das estratégias mais eficientes para se determinar a relevância ou importância de uma página na Web é o *PageRank* [Brin e Page, 1998]. O PageRank é um algoritmo baseado na ligação entre as páginas da Web utilizado pelo Google para determinar a relevância de uma página específica. Se uma página possui o valor do *PageRank* baixo significa geralmente que ela não é considerada importante, pelo menos no contexto do grafo da Web. A patente do Google BlogSearch menciona explicitamente o *PageRank* como um possível indicador positivo para a qualidade de um *blog* [Baehni et al., 2007]. Assim, nesse primeiro experimento são analisados os valores do *PageRank* dos *blogs* mais populares dos domínios selecionados (i.e., *blogosferas*). O objetivo é verificar se existe uma correlação entre a popularidade e a importância do *blog* medida utilizando o algoritmo de *PageRank*.

Os *blogs* selecionados estão listados na Tabela 2.1. A maioria deles corresponde a *blogs* de diários pessoais. Desses, dez são de artistas famosos no Brasil. Outros temas tratados são: cinema, opinião, poesia, imagens, humor, crônicas e esportes.

O valor do *PageRank* foi medido para cada *blog* utilizando o *plugin* Google Toolbar[5]. Uma vez instalado no *browser*, esse *plugin* disponibiliza o valor do *PageRank* de cada página visitada. Os valores de *PageRank* medidos por esse *plugin* variam em uma escala entre zero (menos importante) e dez (mais importante). O valor especial -1 é utilizado quando uma página não possui valor para o *PageRank*, indicando que a mesma

[5]toolbar.google.com

Domínio	Endereço do *blog*
Blogger	anotacoescinefilo.blogger.com.br
	bolsinhademulher.blogger.com.br
	ledramaqueen.blogger.com.br
	soltanomundo.blogger.com.br
	vizinhodojefferson.blogger.com.br
	copacabanadetoledo.blogger.com.br
	homensdopantano.blogger.com.br
	espacocubo.blogger.com.br
	h18.blogger.com.br
	juntandooscacos.blogger.com.br
BlogLog	bloglog.com.br/angelica
	bloglog.com.br/astridfontenelle
	bloglog.com.br/cleopires
	bloglog.com.br/fanipacheco
	bloglog.com.br/joanabalaguer
	bloglog.com.br/anamariabraga
	bloglog.com.br/samarafelippo
	bloglog.com.br/carolinadieckmann
	bloglog.com.br/brunodeluca
	bloglog.com.br/carolcastro
Terra	zocorelio.blog.terra.com.br
	vittini.blog.terra.com.br
	contonton.blog.terra.com.br
	betocruz.blog.terra.com.br
	maluasia.blog.terra.com.br
	ofantasmadaopera.blog.terra.com.br
	cronicasdecelsocruz.blog.terra.com.br
	cousas.blog.terra.com.br
	joselmanoal.blog.terra.com.br
	tonemportugal.blog.terra.com.br
UOL	augustocorradini.blog.uol.com.br
	diariodeleticia.zip.net
	jessy.valim.zip.net
	deborahbllom.zip.net
	taticats.blog.uol.com.br
	espallhamerda.zip.net
	paz.amor.e.magiia.zip.net
	laion.zip.net
	paidegemeos.zip.net
	katialealsousa.zip.net

Tabela 2.1. Os dez *blogs* mais populares de cada domínio.

não foi indexada pela máquina de busca Google. Antes de mostrar os resultados, é necessário enfatizar que o valor do *PageRank* da página principal da UOL é 8, enquanto o valor do *PageRank* de sua página inicial de *blogs* é 6. Da mesma forma, o valor do *PageRank* da página de *blogs* dos outros três domínio estudados é 6, enquanto que os respectivos valores de suas páginas principais são 7 (www.terra.com.br) e 6

(www.globo.com). Esses valores implicam que a cobertura do Google sobre os domínios estudados é satisfatória, de modo que não é esperado um resultado tendencioso devido à falta de cobertura dessa máquina de busca.

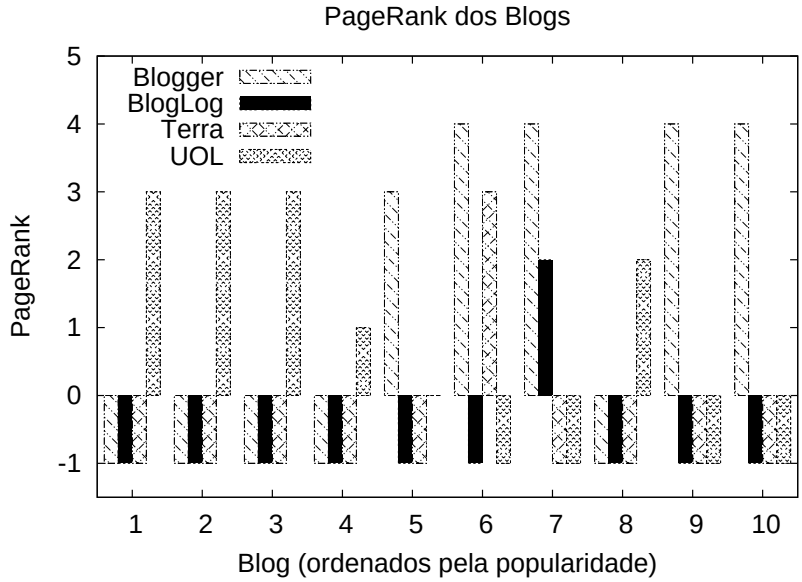


Figura 2.1. O PageRank dos quarenta *blogs* mais populares dos quatro domínios estudados.

A Figura 2.1 mostra os valores para cada *blog* analisado, ordenada, no eixo x, por sua popularidade em seus respectivos domínios. É possível notar que poucos *blogs* populares possuem valores de *PageRank* pouco significativos (em torno de 3 e 4) dado o valor associado ao seu domínio. Assim, apesar das diferentes formas que os domínios utilizam para estimar a popularidade e focando somente no valor do *PageRank*, isso parece uma evidência de que esses *blogs* são realmente alguns dos mais populares de seus respectivos domínios. Por outro lado, em uma visão mais ampla, o maior valor absoluto do *PageRank* foi 4, o que pode ser considerado baixo, uma vez que se trata dos *blogs* mais populares de importantes domínios. Mais ainda, a grande maioria deles, 27, não possuem valores de *PageRank*. Na verdade, os quatro *blogs* mais populares do Blogger, BlogLog e Terra não possuem valor de *PageRank*, enquanto que os quatro *blogs* mais populares da UOL possuem valores abaixo de 3.

Em suma, os resultados acima indicam que existe uma baixa correlação entre a im-

portância dos *blogs* no grafo da Web e suas popularidades relativas. Apesar de aspectos desse problema já terem sido discutidos em cenários mais restritos [Kritikopoulos et al., 2006], aqui apresentam-se evidências mais claras sobre o assunto através de medidas quantitativas aplicadas especificamente para o caso dos *blogs* populares, onde poderia se pensar que existisse uma conectividade maior do que em *blogs* não populares.

2.3 Análise da Ordenação dos Resultados das Consultas

Como o *PageRank* é apenas uma primeira evidência da hipótese que a popularidade dos *blogs* não está sendo explorada pelas máquinas de busca, um segundo experimento foi elaborado para que houvesse mais evidências sobre esse fato.

Foram recrutados cinco voluntários para cada um analisar vinte *blogs* escolhidos aleatoriamente entre os quarenta mais populares. Cada voluntário deveria assinalar seis palavras-chave para cada *blog* analisado. As palavras-chave deveriam ser aquelas que ele utilizaria se fosse procurar aquele *blog* usando uma máquina de busca disponível na Web. Os voluntários deveriam assinalar as palavras por ordem de importância: as mais importantes primeiro e as menos importantes por último. Cada *blog* foi examinado por dois voluntários diferentes. Das doze palavras-chave assinaladas para cada *blog*, foram selecionadas seis. Foram priorizadas as palavras assinaladas por ambos os voluntários desde que tivessem a mesma ordem de importância. Quando não existiam palavras-chave em comum, foi feita uma escolha aleatória entre as palavras dos dois voluntários, respeitando-se a ordem de importância assinalada por eles. É importante ressaltar que em alguns casos as palavras-chave escolhidas não estavam presentes nos textos dos *blogs* (e.g., "diário", "vídeo" e "crianças") embora representassem adequadamente o seu contexto.

Foram definidos então três tipos diferentes de consulta. A primeira considera utilizar as duas palavras-chave mais importantes, a segunda as três palavras-chave mais

importantes e a última todas as palavras-chave. Para os dois primeiros tipos de consulta foi feita uma escolha conservadora de retirar as palavras-chave que apareciam na URL ou no título do *blog*, porque as máquinas de busca utilizam essa informação como sendo um forte indício que é o *blog* procurado. Em outras palavras, o foco para os cenários de duas e três palavras-chave foi procurar os *blogs* populares de um assunto específico, i.e., as consultas poderiam ser consideradas como pesquisas pelo conteúdo informacional dos *blogs*. No último caso, foram utilizadas todas as seis palavras-chave independentemente de onde apareciam no *blog*.

O primeiro passo foi tentar efetuar consultas em máquinas de busca oferecidas pelos próprios domínios dos *blogs*. Somente dois domínios dos quatro estudados ofereciam um serviço de busca, sendo que um deles (Terra) utiliza o Google. Nesse caso, o cenário para esses domínios foi capturado no último conjunto de experimentos desta seção. Portanto, serão apresentados aqui somente os resultados para o domínio UOL.

Os resultados obtidos para os três tipos de consultas estão apresentados na Figura 2.2 e foram sumarizados na Tabela 2.2. A Tabela mostra o percentual de *blogs* populares que aparecem na primeira página de resultados (i.e., entre as dez primeiras posições) para o domínio UOL. O gráfico mostra os *blogs* ordenados pela popularidade no eixo x. Se o *blog* apareceu após a centésima posição ele recebeu a posição 100 para que o gráfico não perdesse a escala. Na tabela, pode-se notar que mesmo para o melhor resultado (com seis palavras-chave) 80% dos *blogs* populares do domínio UOL não foram retornados na primeira página. O gráfico mostra ainda que em somente sete dentre os trinta resultados o *blog* correspondente aparece antes da centésima posição. Mais ainda, isso aconteceu para somente três dos *blogs* mais populares do domínio sendo que em um deles o *blog* foi retornado após a posição 50.

Em seguida foram realizadas consultas em duas das maiores e mais utilizadas máquinas de busca para *blogs*: Google *blog* Search e Technorati. Ambas não permitem que a busca seja restringida a um domínio específico. Assim a busca foi realizada na porção da *blogosfera* indexada por essas máquinas de busca. Nenhum dos *blogs* desejados foi

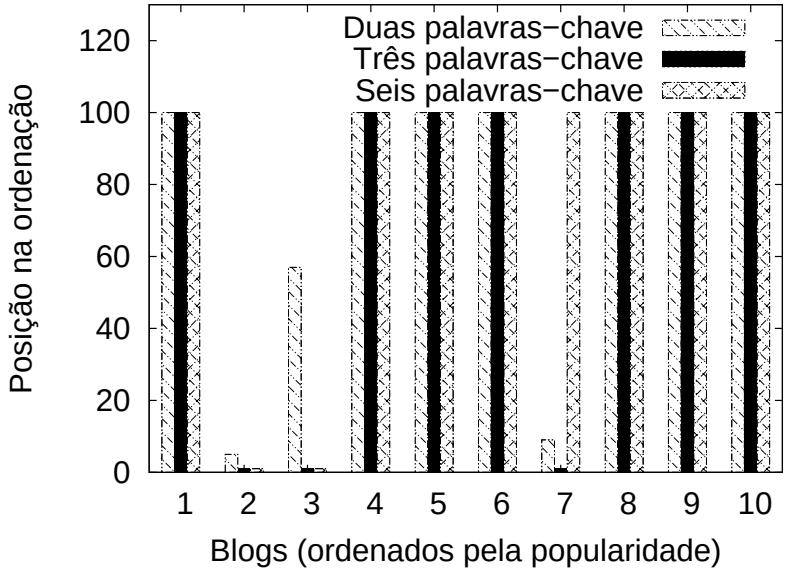


Figura 2.2. Resultado da ordenação das consultas na máquina de busca do UOL para os *blogs* do UOL

	Primeira página
2 Palavras-chave	10.0%
3 Palavras-chave	30.0%
6 Palavras-chave	20.0%

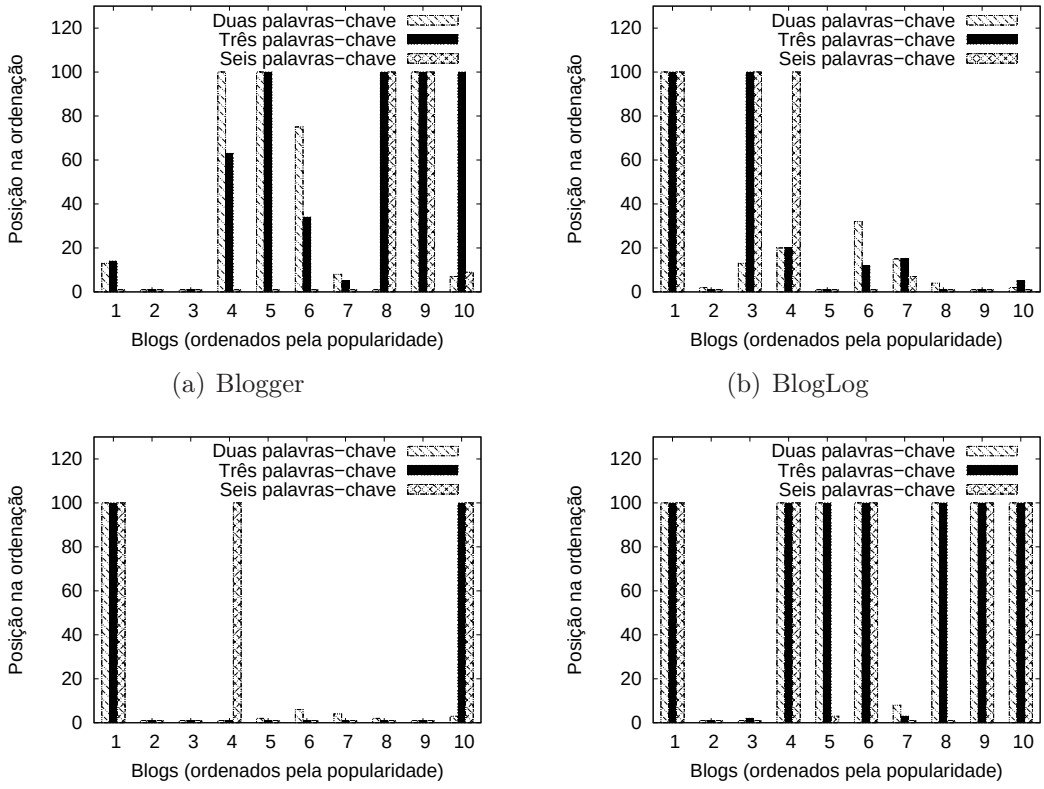
Tabela 2.2. Porcentagem de *blogs* que aparecem nas primeiras páginas de resultados da máquina de busca do UOL

encontrado entre as cem primeiras posições nas duas máquinas de busca, com nenhum dos três tipos de consulta.

Finalmente, os experimentos foram focados em duas grandes máquinas de busca de uso geral da Web, que, em tese, também indexam grande parte da *blogosfera*. Elas são, geralmente, o ponto de entrada de usuários não especializados. As consultas, como mencionado anteriormente, foram restringidas aos domínios de cada *blog*. Isso quer dizer que os *blogs* do BlogLog, por exemplo, foram procurados somente no domínio BlogLog. Esse experimento permite comparar os resultados entre os diferentes domínios com a mesma metodologia (i.e., a mesma máquina de busca e a mesma coleção).

Os gráficos da Figura 2.3 representam a posição em que cada *blog* popular foi recuperado pelo Google para cada um dos domínios considerados, enquanto a Figura

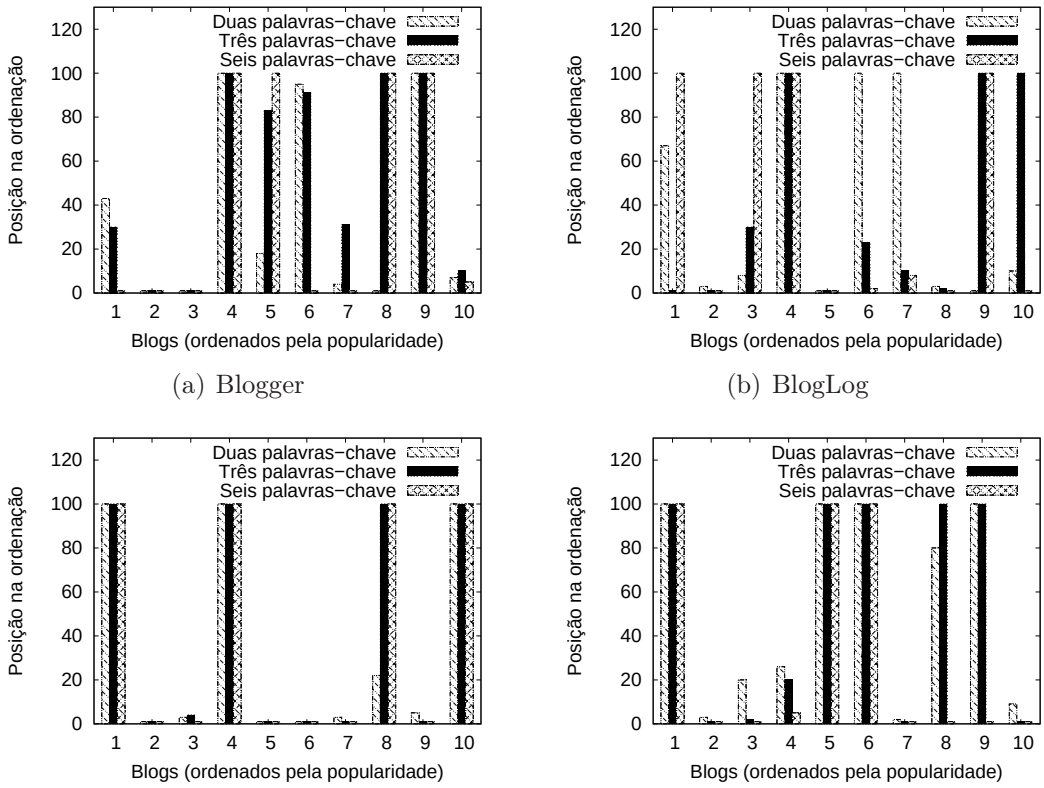
2.4 apresenta resultados similares para o Yahoo!. Todos os gráficos mostram os *blogs* ordenados por sua popularidade no eixo x, ou seja, o *blog* que apareceu mais vezes na lista dos mais populares aparece na posição 1 do gráfico. Se o *blog* foi recuperado além da centésima posição, ele recebeu a posição cem para que o gráfico não perdesse a escala. É possível notar que em vários casos (Figuras 2.3(b), 2.3(c), 2.3(d), 2.4(c) e 2.4(d)) os *blogs* mais populares aparecem somente a partir da centésima posição. Isso acontece para os três tipos de experimento. Os usuários estão interessados normalmente somente nos *blogs* que aparecem nas primeiras posições da página de respostas, porém dos quatro domínios estudados, considerando-se duas máquinas de busca diferentes, somente em alguns poucos casos o *blog* mais popular foi retornado na primeira página e, mesmo assim, somente para um tipo de consulta. Em geral, uma fração significante dos *blogs* populares só aparece em posições muito distantes na ordenação dos resultados das consultas.



(a) Blogger (b) BlogLog

(c) Terra (d) UOL

Figura 2.3. Posição que os *blogs* foram recuperados pelo Google



(a) Blogger (b) BlogLog

(c) Terra (d) UOL

Figura 2.4. Posição que os *blogs* foram recuperados pelo Yahoo!

A porcentagem de *blogs* populares que apareceram na primeira página de resultados (i.e., entre os dez primeiros) para as consultas realizadas no Google e no Yahoo! podem ser visualizadas, respectivamente, nas Tabelas 2.3 e 2.4. Os resultados foram divididos em duas categorias: *blogs* que aparecem na primeira página e *blogs* que aparecem na segunda página em diante. A fração de *blogs* populares que não aparecem na primeira página de resultados em ambas as máquinas de busca é bastante significativa (mais de 52%). Na verdade, mais de 57% dos *blogs* não aparecem na primeira página de resultados do Yahoo! nos dois primeiros experimentos. Mesmo ao utilizar as seis palavras-chave, que deveria ser a situação mais fácil, já que as palavras-chave podem aparecer na URL ou no título do *blog*, não foram retornados nem um terço dos *blogs* populares na primeira página de resultados das duas máquinas de busca. É um resultado surpreendente já que as consultas estavam restritas ao próprio domínio do *blog*.

	Primeira página
Duas palavras-chave	52,5 %
Três palavras-chave	42,5 %
Todas palavras-chave	62,5 %

Tabela 2.3. Porcentagem de *blogs* que aparecem na primeira página de resultados do Google.

	Primeira página
Duas palavras-chave	47,5 %
Três palavras-chave	37,5 %
Todas palavras-chave	52,5 %

Tabela 2.4. Porcentagem de *blogs* que aparecem na primeira página de resultados do Yahoo!.

Blog	**Palavras-chave**	**Pos Google**	**Pos Yahoo!**
paidegemeos.zip.net	Diário, Gêmeos	Após 100ª	80
soltanomundo.blogger.com.br	Pensamentos, Escritora	Após 100ª	Após 100ª
anotacoescinefilo.blogger.com.br	Cinema, Festival	13ª	43
diariodeleticia.zip.net	Diário, Crianças	Após 100ª	26

Tabela 2.5. Quatro exemplos de *blogs* com palavras-chave relevantes mas retornados em posições ruins

Na Tabela 2.5 são mostrados quatro exemplos de *blogs* que os voluntários atribuíram palavras-chave relevantes (i.e. foi verificado manualmente que as palavras-chave realmente capturam o assunto do *blog*) mas a posição na ordenação das consultas está ruim. A terceira coluna mostra a posição na ordenação do resultado da consulta para o Google e a quarta para o Yahoo!. Alguns deles estão muito longe do topo, aparecendo na quinta, oitava e décima terceira página de resultados, muito longe para prender a atenção dos usuários.

2.4 Resumo dos Resultados

Os resultados dos dois experimentos podem ser resumidos e contrastados nas Tabelas 2.6 and 2.7. A análise teve como ênfase principal o pior caso, aqui definido quando um *blog* possui um valor de *PageRank* muito baixo (menor que 3) e não aparece na primeira página de resultados. Considerando cada experimento separadamente, pelo menos 30% e 35% dos *blogs* caem no pior caso para o Google e para o Yahoo!, respectivamente. Considerando todos os três tipos de consulta conjuntamente, a média de *blogs* que cai no pior caso é 36% para o Google e 41% para o Yahoo!. Em suma, uma parcela significante de *blogs* populares em todos os quatro domínios está sendo negligenciada pelas atuais estratégias de ordenação de resultados das máquinas de busca atuais.

	Após primeira página e PR < 3
2 Palavras-chave	37,5%
3 Palavras-chave	42,5%
6 Palavras-chave	30,0%

Tabela 2.6. Comparação entre a posição na ordenação dos resultados no Google e o valor do PageRank

	Após primeira página e PR < 3
2 Palavras-chave	42.5%
3 Palavras-chave	47.5%
6 Palavras-chave	35.0%

Tabela 2.7. Comparação entre a posição na ordenação dos resultados no Yahoo! e o valor do PageRank

Capítulo 3

Ordenação de Resultados com Base na Popularidade dos *Blogs*

Neste Capítulo é proposta uma nova estratégia de busca para *blogs* baseada em sua popularidade. Será mostrado como utilizar essa importante característica para ordenar melhor os resultados das consultas e melhorar a experiência dos usuários nas buscas por *blogs*. A idéia principal é incorporar a popularidade como um fator na fórmula de ordenação dos resultados das consultas. Os resultados foram contrastados considerando-se uma máquina de busca que utiliza o fator de popularidade e uma máquina de busca que não utiliza o fator de popularidade para verificar a melhoria na posição de ordenação de cada *blog* desejado. Os resultados foram validados ainda com a ajuda de voluntários que atribuíram graus de relevância para cada um deles. É necessário enfatizar que o objetivo aqui não é propor a "melhor estratégia possível" de ordenar consultas utilizando a popularidade, mas, prover evidências de que a popularidade pode realmente ser benéfica na busca por *blogs* e melhorar a experiência do usuário como um todo.

3.1 Configuração Experimental

Devido à falta de informação sobre a popularidade dos *blogs* em coleções padrões, como a TREC *blog* Track [Ounis et al., 2006], bem como pela falta de informação

proveniente de *logs* de consultas de máquinas de busca reais, mais uma vez foi utilizado o conjunto de *blogs* populares coletados e as palavras-chave associadas a eles. Para esses experimentos também foi coletada uma amostra do domínio UOL. Esse domínio foi escolhido principalmente porque sua estratégia para estimar a popularidade de um *blog*, conforme mencionado na Seção 2.2, utiliza a opinião dos próprios usuários, através de uma escala de votação que varia de 0 a 10 pontos.

Foram coletados cerca de 15.000 *blogs* do domínio UOL. Esses *blogs* foram indexados utilizando-se a API Lucene[1]. A popularidade dos *blogs* foi incorporada ao índice utilizando ferramentas disponíveis na própria API Lucene. Foi necessário coletar e indexar uma coleção própria para facilitar a avaliação experimental, uma vez que é muito difícil conduzir esse tipo de experimento utilizando uma máquina de busca comercial.

Um fator de popularidade (FP) foi definido para cada *blog* da coleção que é proporcional à sua importância no domínio. A importância de cada *blog* foi estimada de acordo com a quantidade de vezes que o *blog* apareceu na lista dos dez mais populares durante os 30 dias de monitoramento descrito na Seção 2.2.

Esse fator de popularidade é calculado utilizando a Equação 3.1, onde N representa o número de dias que o *blog* apareceu na lista dos 10 mais populares e M o número máximo de dias que um *blog* apareceu nessa lista. Ele varia entre 1, i.e., o *blog* não apareceu na lista dos dez mais populares nenhum dia, e 20, i.e., o *blog* apareceu o número máximo de vezes, onde 20 é um escalar escolhido empiricamente. Para definir este escalar, a coleção de *blogs* foi indexada utilizando quatro valores diferentes: 15, 20, 50 e 100. Em seguida os resultados foram avaliados utilizando-se a métrica NDCG (*Normalized Discounted Cumulative Gain*) [Järvelin e Kekäläinen, 2000] definida na Seção 3.3. Os valores 50 e 100 promovem excessivamente os *blogs* populares para as primeiras posições dos resultados, enquanto ao se utilizar o valor 15 os resultados do NDCG não foram tão satisfatórios quanto ao se utilizar 20.

[1] lucene.apache.org

$$FP = \frac{20 \times N}{M} + 1 \tag{3.1}$$

Lucene utiliza as estratégias tradicionais de cálculo de relevância de páginas para uma consulta: *Term Frequency-Inverse Document Frequency* (TF-IDF) e o modelo vetorial [Baeza-Yates e Ribeiro-Neto, 1999]. O fator TF é baseado na freqüência que uma palavra-chave aparece em um documento. A idéia é que palavras-chave mais freqüentes podem capturar melhor o conteúdo de um documento. O fator IDF considera a freqüência de uma palavra-chave perante toda a coleção de documentos. A idéia é que palavras-chave que aparecem com muita frequência na coleção não descriminam de forma satisfatória os documentos. Esses dois fatores são multiplicados para obter o peso de uma palavra-chave. A similaridade entre a consulta e o *blog* é obtida ao calcular o produto interno da representação vetorial do texto do *blog* e da consulta, onde os pesos dos vetores são calculados utilizando o esquema TF-IDF. Ao final o FP é utilizado como um fator multiplicativo adicional na equação de similaridade para obter a pontuação do *blog*.

3.2 Eficiência do Fator de Popularidade

Nesta seção é analisada a eficiência do fator de popularidade proposto ao comparar os resultados obtidos utilizando-o com os resultados da ordenação original. A idéia é não só investigar se os *blogs* populares foram considerados relevantes e promovidos às primeiras posições da nova ordenação mas também averiguar o impacto dessas modificações no resultado final. Em outras palavras, deseja-se verificar se a estratégia realmente melhora o resultado de uma forma geral ao promover os *blogs* populares (quando esses possuem similaridade com a consulta) e não retira das primeiras posições outros *blogs* não populares mas com grande similaridade com a consulta. Como algumas palavras-chave são muito genéricas (veja a Tabela 2.5), essa é uma situação muito provável.

Foram utilizadas as mesmas palavras-chave previamente definidas pelos de voluntários (Seção 2.3) para os dez *blogs* mais populares do UOL para se realizar consultas em duas máquinas de busca: uma que foi indexada utilizando o fator de popularidade e outra sem o fator de popularidade. Da mesma forma que na Seção 2.3, três tipos de consultas foram feitas em cada máquina de busca, utilizando duas, três e seis palavras-chave.

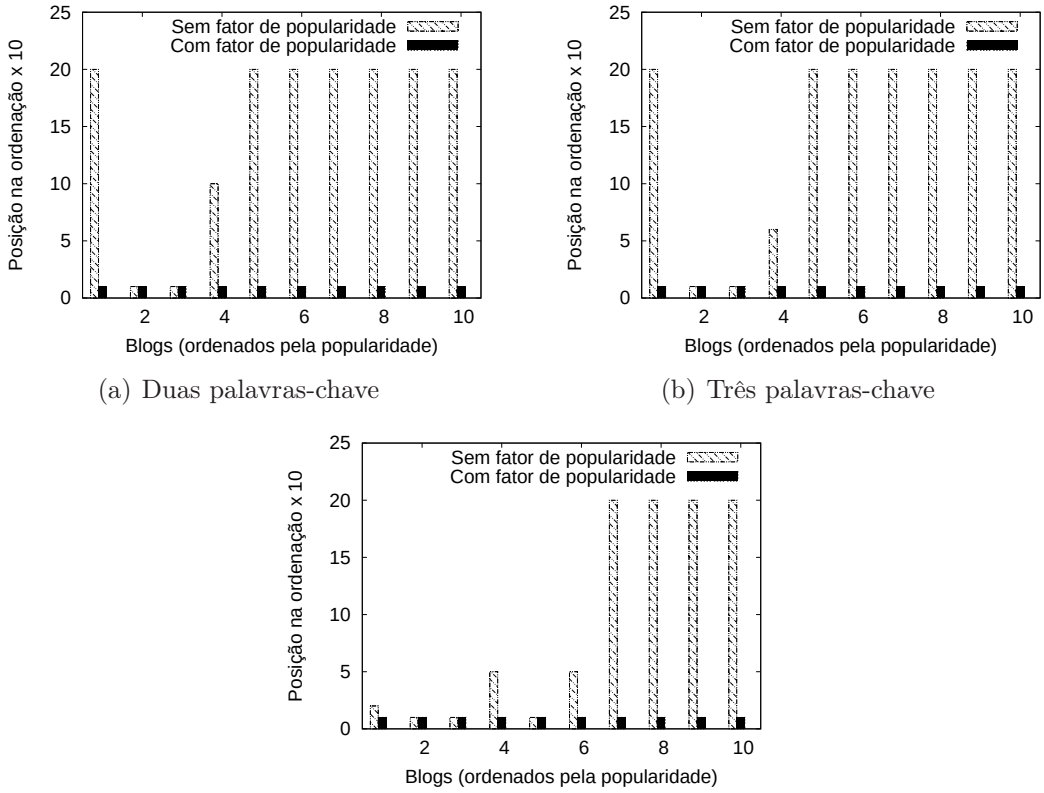
No Capítulo 2 assume-se que os *blogs* recuperados pelas máquinas de busca são considerados relevantes quando eles aparecem na primeira página de resultados, i.e., entre as dez primeiras ocorrências. Os *blogs* recuperados pelas máquinas de busca considerando-se o fator de popularidade e sem considerá-lo foram classificados em duas categorias: os que aparecem na primeira página e os que aparecem após a primeira página. A Figura 3.1 mostra em qual página cada *blog* apareceu em cada uma das máquinas de busca. Se um *blog* apareceu após a vigésima página ele recebeu o valor 20.

Como é possível notar, para todas as consultas realizadas na máquina de busca que incorpora o fator de popularidade, todos os *blogs* foram recuperados entre os 10 primeiros resultados. Já no caso da máquina de busca que não utiliza o fator de popularidade, entre as trinta consultas realizadas somente sete *blogs* desejados foram recuperados entre as dez primeiras posições.

A Tabela 3.1 sumariza esses resultados. Na maioria dos casos, o fator de popularidade melhorou os resultados da máquina de busca, possibilitando que os *blogs* mais populares fossem recuperados entre as primeiras dez posições.

3.3 Validação com Voluntários

Na seção anterior foi mostrado que a máquina de busca que utiliza o fator de popularidade recuperou os *blogs* mais populares entre as primeiras posições no resultado da ordenação da consulta. Agora será analisado se essas páginas são consideradas rele-



(a) Duas palavras-chave

(b) Três palavras-chave

(c) Seis palavras-chave

Figura 3.1. Comparação de buscas realizadas não utilizando o fator de popularidade e utilizando o fator de popularidade para uma consulta com duas, três e seis palavras-chave.

vantes por usuários não somente por sua popularidade, mas também por sua relação entre as palavras-chave e o conteúdo dos *blogs*.

Para isso um outro experimento foi então proposto. As palavras-chave designadas pelos voluntários foram utilizadas para efetuar consultas na máquina de busca que incorpora o fator de popularidade. Novamente três tipos de consulta foram formuladas, com duas, três e seis palavras-chave. Os dez primeiros resultados foram apresentados para um conjunto diferente de voluntários (diferente daqueles que assinalaram as palavras-chave para os *blogs*). Os novos voluntários deveriam classificar cada *blog* em três categorias: muito relevante, relevante e irrelevante, considerando-se uma consulta específica e o conteúdo do *blog*. Cada par (*blog*, consulta) foi avaliado por exatamente dois voluntários diferentes. É possível que essa configuração experimental possa favore-

Consulta	Máq. de Busca	Primeira página	Após primeira página
2	Sem fator de popularidade	2	8
	Com fator de popularidade	10	0
3	Sem fator de popularidade	2	8
	Com fator de popularidade	10	0
6	Sem fator de popularidade	3	7
	Com fator de popularidade	10	0

Tabela 3.1. Comparação entre a máquina de busca que não utiliza o fator de popularidade e a que utiliza o fator de popularidade para consultas com duas, três e seis palavras-chave

cer de certa forma a versão do sistema com o fator de popularidade, mas isso pode ser contrabalanceado pelo caráter amplo de algumas consultas, principalmente nas consultas com duas palavras-chave, dentre elas "viagem diário", "gêmeos pais", "cinema festival" e "escritor pensamentos", que refletem interesses gerais e podem recuperar um número grande de *blogs* não somente os populares.

Esse experimento produziu sessenta resultados: 10 *blogs* $\times$ 3 consultas $\times$ 2 voluntários por *blog*. Para cada categoria (muito relevante, relevante e irrelevante) foi atribuído três, dois ou um ponto(s) respectivamente, baseado na categoria que foi assinalada pelo voluntário. Ao somar os pontos para cada par (*blog*, consulta) foi obtida a classificação do *blog* com seis, cinco, quatro, três ou dois pontos. Os resultados estão mostrados na Tabela 3.2.

Como é possível notar, um total de 28 dentre os 30 resultados foram considerados pelo menos relevante pelos voluntários, sendo que 24 desses foram considerados muito relevantes e somente 2 foram considerados irrelevantes pelos voluntários. Mais ainda, ao considerar as consultas com seis palavras-chave, todos os resultados foram considerados muito relevantes. Esses resultados evidenciam o alto grau de satisfação dos voluntários com os resultados e o potencial de eficiência do fator de popularidade para buscas na *blogosfera*.

Os resultados também foram avaliados utilizando-se a métrica NDCG, definida na Equação 3.2.

	Número de palavras-chave		
Pontos	**Duas**	**Três**	**Seis**
6	7	7	10
5	0	1	0
4	2	1	0
3	1	0	0
2	0	1	0

Tabela 3.2. Número de *blogs* classificados como Muito Relevante, Relevante e Irrelevante

$$NDCG = N_i \sum_{i=1}^{k} \frac{2^{label(j)} - 1}{\log_2 (1 + i)} \tag{3.2}$$

Nessa equação, N_i é uma constante de normalização calculada com base na ordenação perfeita dos resultados para uma consulta q_i e $label(j)$ é o ganho de valor associado ao documento na j_{th} posição. Por exemplo, $label(j)$ é igual a 3 se o documento é considerado muito relevante, igual a 2 se considerado relevante e igual a 1 se irrelevante. Na equação $log_b(1+i)$ é uma função de desconto que reduz o ganho de um documento à medida que esse sobe na posição de ordenação. A base do logaritmo, b, controla o grau de redução. Foi utilizado $b = 2$ nos experimentos, o que corresponde a uma redução leve.

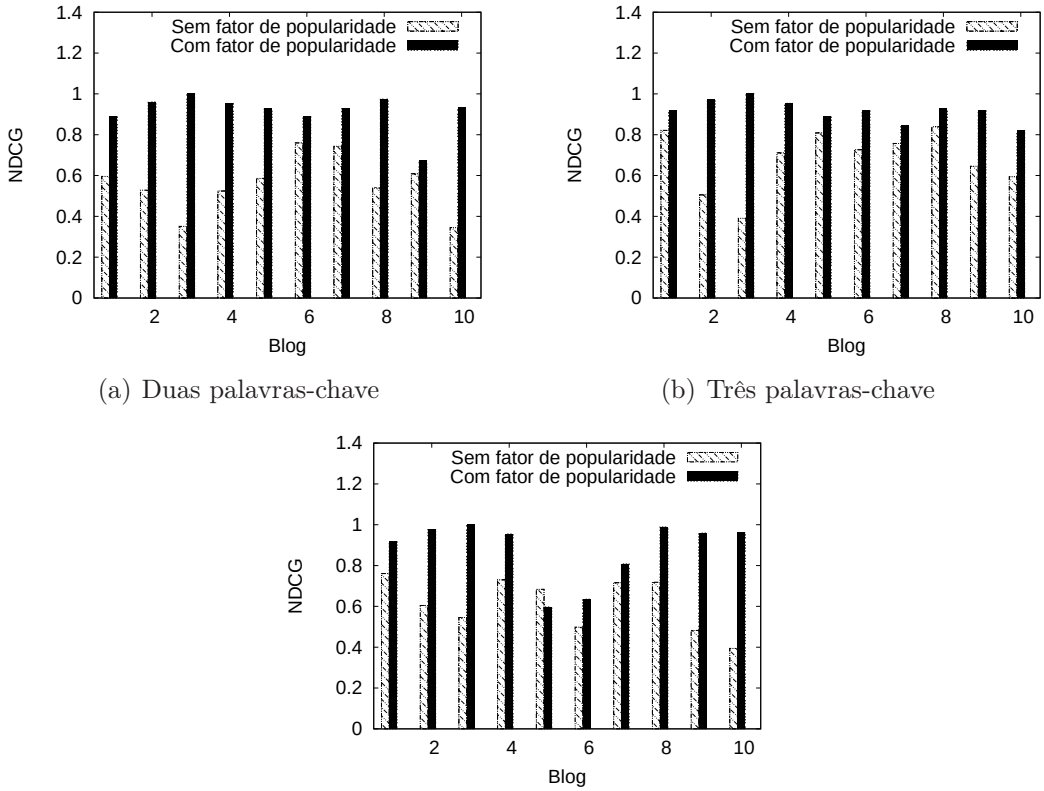
No contexto deste experimento, um valor mais elevado de NDCG para a versão com o fator de popularidade, por exemplo, significa que os *blogs* menos relevantes nas primeiras posições da lista de resultados da consulta estão sendo substituídos por outros mais relevantes, permitindo, assim, mensurar o impacto do fator de popularidade sobre esses resultados. Outras vantagens do NDCG inclui o fato que essa métrica lida naturalmente com diversos níveis de relevância ao considerar a posição do *blog* melhor colocado na lista de resultados e descontar logariticamente o valor à medida que as posições na ordenação vão diminuindo. É importante ressaltar que o NDCG é normalizado pelo melhor resultado possível, representado pelo fator N_i. Para esse experimento o melhor resultado possível foi calculado com base na avaliação de relevância obtida

com o julgamento dos voluntários para ambos os tipos de consulta, com e sem o fator de popularidade. O mesmo fator de normalização foi utilizado para os dois tipos de consulta. Um exemplo do cálculo do NDCG pode ser visto ao se considerar as duas ordenações apresentadas na Tabela 3.4 produzida pela consulta *'viagem diário'* com e sem o fator de popularidade e seus respectivos julgamentos de relevância. Nesse caso, considerando todos os *blogs* retornados pelas duas consultas e o respectivo julgamento, a melhor ordenação possível é a apresentada pela versão que utiliza o fator de popularidade, ou seja, o NDCG é igual a 1 (para $N_i = 15{,}32$) e 0,427 para a versão sem o fator de popularidade, utilizando a Equação 3.2.

A Figura 3.2 mostra a média do NDCG dos dois voluntários para consultas com duas, três e seis palavras-chave respectivamente, considerando os dez primeiros resultados para cada tipo de consulta. Pode-se notar que para todos os casos exceto um (consulta para o *blog* 5 com seis palavras-chave) houve melhoras ao utilizar o fator de popularidade. De fato, em diversos casos o NDCG das consultas sem o fator de popularidade foi muito baixo (menos que 0,6) quando comparado com a ordenação ideal, evidenciando a dificuldade de se efetuar a busca por *blogs* com estratégias tradicionais de recuperação de informação. Ignorando o resultado do *blog* 5 da Figura 3.2(c) (único caso onde a estratégia que utiliza o fator de popularidade perdeu), a melhora varia entre 9,65% e 184,91%.

Os valores médios do NDCG, quando todos os *blogs* são considerados com os diferentes tipos de consulta estão mostrados na Tabela 3.3. Os ganhos gerais são da ordem de 63% para consultas com duas palavras-chave, 34% para consultas com três palavras-chave e 43% para consultas com seis palavras-chave. Todos os resultados são estatisticamente significantes com 99,9% de confiança (teste t).

Os gráficos das Figuras 3.3, 3.4 e 3.5 mostram o NDCG acumulado para cada posição do resultados da ordenação de consultas que obtiveram melhores resultados ao se utilizar o fator de popularidade. Como pode-se observar, em todos os casos o ganho acumulado ao se utilizar o fator de popularidade é muito superior ao da ordenação



(a) Duas palavras-chave

(b) Três palavras-chave

(c) Seis palavras-chave

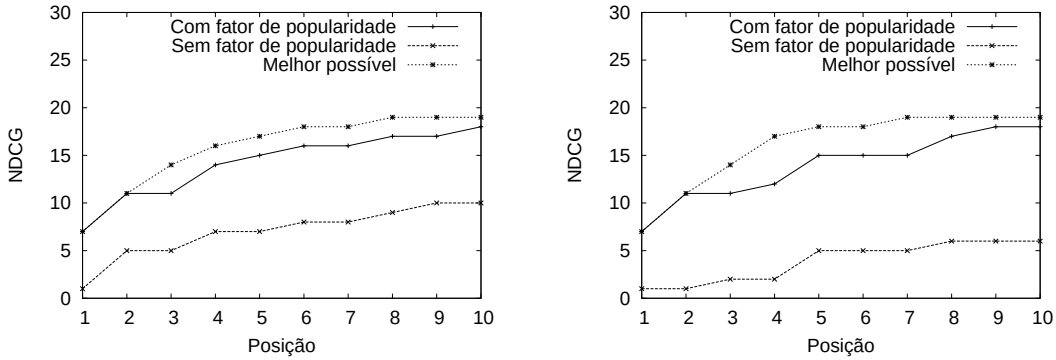
Figura 3.2. O valor médio do NDCG com e sem o fator de popularidade para duas, três e seis palavras-chave

	2 palavras-chave	3 palavras-chave	6 palavras-chave
Com fator de popularidade	0,912	0,915	0,879
Sem fator de popularidade	0,558	0,679	0,613

Tabela 3.3. Resultados globais para o NDCG

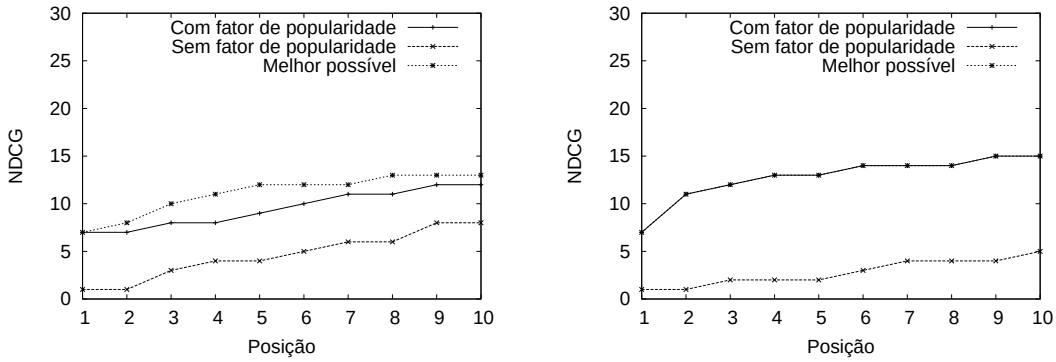
original, sendo que em alguns casos ele se iguala ao melhor ganho possível para aquela consulta. É preciso enfatizar que melhorias no NDCG podem ser obtidas somente se de fato forem substituídos *blogs* menos relevantes pelos mais relevantes, nas primeiras posições da lista de resultados. Assim, os resultados sugerem que, se existe alguma similaridade textual entre uma consulta e um *blog* popular, em vários casos, pelo menos naqueles estudados aqui, é interessante promover os populares. Entretanto, o balanceamento entre o nível de similaridade e a força da popularidade para coleções específicas é algo que deve ser melhor estudado em trabalhos futuros. Como mencionado an-

teriormente, aqui o interesse maior é prover evidências do potencial de se utilizar a popularidade na busca de *blogs*.



(a) deborahbllom.zip.net, 2 palavras-chave (b) espallhamerda.zip.net, 2 palavras-chave

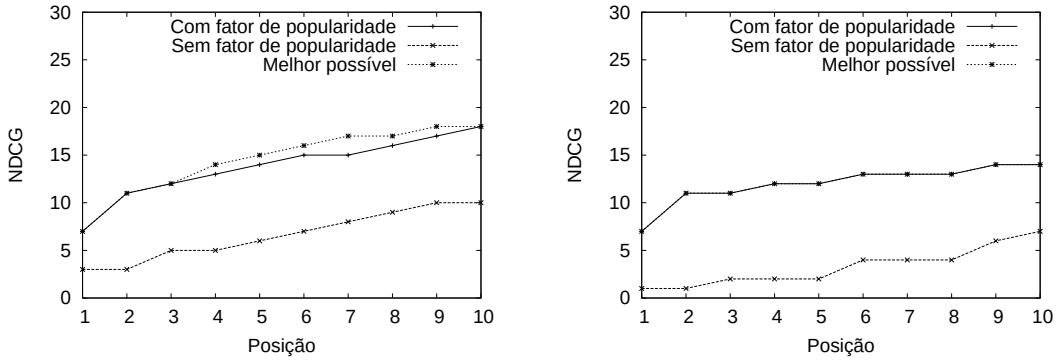
Figura 3.3. NDCG acumulado para as consultas como os melhores ganhos, para 2 palavras-chave



(a) jessy.valim.zip.net, 3 palavras-chave (b) diariodeleticia.zip.net, 3 palavras-chave

Figura 3.4. NDCG acumulado para as consultas como os melhores ganhos, para 3 palavras-chave

Para entender melhor os ganhos obtidos, os resultados das ordenações das consultas e suas respectivas avaliações de relevância foram manualmente verificadas concluindo-se que de fato o fator de popularidade foi capaz de promover os *blogs* específicos que eram esperados para as primeiras posições. Além disso, esses *blogs* foram considerados muito relevantes pelos voluntários e que, em geral, eles substituíram ou removeram dos resultados das consultas *blogs* que foram considerados irrelevantes. Mais ainda, foi verificado que outros *blogs* populares com "similaridade" textual com a consulta



(a) deborahbllom.zip.net, 6 palavras-chave (b) diariodeleticia.zip.net, 6 palavras-chave

Figura 3.5. NDCG acumulado para as consultas com os melhores ganhos, para 6 palavras-chave

Posição	Com Popularidade	$label(j)$	Sem Popularidade	$label(j)$
1	**katialealsousa.zip.net**	3	valmap.blog.uol.com.br	2
2	**paidegemeos.zip.net**	3	edotani.blog.uol.com.br	1
3	valmap.blog.uol.com.br	2	necadalmolin.zip.net	1
4	edotani.blog.uol.com.br	1	blog.uol.com.br	1
5	necadalmolin.zip.net	1	microdoc.zip.net	1
6	blog.uol.com.br	1	luciene.sa.blog.uol.com.br	1
7	microdoc.zip.net	1	goncalves.leandro.blog.uol.com.br	1
8	devaneiosdocotidiano.zip.net	1	maiaraborges.zip.net	1
9	luciene.sa.blog.uol.com.br	1	cw.schulze.zip.net	1
10	goncalves.leandro.blog.uol.com.br	1	josedito.blog.uol.com.br	1

Tabela 3.4. Exemplo de um julgamento de relevância

também foram promovidos pelo fator de popularidade e considerados muito relevantes pelos voluntários mesmo para consultas que não foram especificamente formuladas para eles. Isso aconteceu provavelmente devido à natureza abrangente das palavras-chave especificadas. Um exemplo dessa situação é mostrado na Tabela 3.4, que apresenta a ordenação original e a ordenação modificada pelo fator de popularidade com suas respectivas avaliações de relevância para a consulta "viagem diário". Os *blogs* em negrito são aqueles promovidos. A maioria dos resultados expressivos foi obtida em situações semelhantes. Entretanto, o único caso em que houve perda é exatamente quando uma quantidade excessiva de *blogs* foi promovida (para a consulta "paz amor magia imagens religião Jesus"). Apesar de só existir um caso dentre os trinta resultados, isso indica que deve-se investigar situações em que talvez não seja adequado utilizar o fator de

popularidade, isto é, simplesmente utilizar um fator escalar alto na fórmula do fator de popularidade para todos os casos indiscriminadamente não seria útil, já que muitos *blogs* populares irrelevantes sempre estariam nas primeiras posições do resultado das consultas, independente da consulta. Isso é deixado para trabalhos futuros.

Capítulo 4

Conclusão e Trabalhos Futuros

O foco principal desta dissertação foi explorar o potencial de características sociais, mais especificamente a popularidade, para melhorar a busca por *blogs*.

A partir da lista dos dez *blogs* mais importantes de quatros domínios brasileiros foi possível estudar propriedades da busca por *blogs*, fazer um paralelo com as máquinas de busca atuais e descobrir características que podem ser utilizadas para melhorar as estratégias de ordenação dos resultados das consultas. Ao medir, por exemplo, o *PageRank* desses quarenta *blogs* observou-se que em nenhum dos casos o seu valor foi superior a quatro em uma escala de 0 a 10. Mas ainda, muitos deles (27) possuem *PageRank* igual a -1, indicando que não foram indexados pelo Google. Por se tratar dos *blogs* mais importantes dos domínios em questão, pode-se dizer que os valores estão muito baixos, indicando que não são páginas consideradas importantes para o Google.

Ainda com o intuito de avaliar a importância dos *blogs* populares perante as máquinas de busca atuais, um segundo experimento foi realizado. O objetivo foi analisar a posição que esses *blogs* eram retornados ao se fazer uma consulta utilizando palavras-chave adequadas para cada *blog*. Essas palavras-chave foram atribuídas por voluntários e as consultas realizadas no Google no Yahoo!, sempre restringindo o domínio. Os resultados mostram que mais de 52% dos *blogs* populares não foram retornados na primeira página de resultados (entre as 10 primeiras posições). Mais uma vez, esse é um forte

indício que as métricas utilizadas pelas máquinas de busca disponíveis na Web não são adequadas ao contexto de busca de *blogs*.

Ao calcular o *Mean Reciprocal Rank* (MRR), uma métrica utilizada para avaliar o quanto uma ordenação de resultados se aproxima do ideal, obteve-se uma média de apenas 0,42 e 0,34 respectivamente para o Google e o Yahoo! em uma escala que varia entre 0 e 1. São resultados muito ruins para os *blogs* mais populares do domínio.

Dessa forma, uma nova máquina de busca foi construída incorporando um fator de popularidade com o intuito de adequar as métricas de ordenação de consultas ao contexto de busca de *blogs*. As consultas realizadas nessa máquina de busca com as palavras-chave previamente atribuídas pelos voluntários foram avaliadas utilizando a métrica NDCG. Os resultados apontam ganhos de 63% ao utilizar duas palavras-chave na consulta, 34% ao utilizar três palavras-chave e 43% ao se utilizar seis palavras-chave, todos com 99,9% de confiança (teste t).

Como trabalho futuro pretende-se definir melhores estratégias para calcular o fator de popularidade. Existem vários desafios para esse objetivo já que cada domínio possui a sua própria maneira de determinar a popularidade de um *blog*. Pretende-se também definir um protocolo similar ao *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH) [Lagoze e de Sompel, 2001], que, ao se informar um *blog*, retorne a sua popularidade dentro de um período de tempo estipulado, para que dessa forma as máquinas de busca possam utilizar essa informação para melhorar a ordenação das consultas. Além disso, é necessário investigar as situações em que o fator de popularidade pode influenciar negativamente a ordenação ao promover uma quantidade desnecessária de *blogs* populares.

Referências Bibliográficas

Ali-Hasan, N. e Adamic, L. A. (2007). Expressing social relationships on the blog through links and comments. In *Proceedings of the 1st International Conference on Weblogs and Social Media*, Boulder, Colorado, USA. Retrieved January 19, 2009 from: http://www.icwsm.org/papers/2–Ali-Hasan–Adamic.pdf.

Baehni, S.; Guerraoui, R.; Koldehofe, B. e Monod, M. (2007). Towards fair event dissemination. In *Proceedings of the 27th International Conference on Distributed Computing Systems Workshops*, p. 63, Toronto, Ontario. IEEE Computer Society.

Baeza-Yates, R. e Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley.

Bao, S.; Xue, G.; Wu, X.; Yu, Y.; Fei, B. e Su, Z. (2007). Optimizing web search using social annotations. In *Proceedings of the 16th International Conference on World Wide Web*, pp. 501–510, Banff, Alberta, Canada. ACM.

Brin, S. e Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.

Duarte, F.; Mattos, B.; Bestavros, A.; Almeida, V. e Almeida, J. (2007). Traffic characteristics and communication patterns in blogosphere. In *Proceedings of the 1st International Conference on Weblogs and Social Media*, Boulder, Colorado, USA. Retrieved January 19, 2009 from: http://www.icwsm.org/papers/2–Duarte-Mattos-Bestavros-Almeida-Almeida.pdf.

Fujimura, K.; Toda, H.; Inoue, T.; Hiroshima, N.; Kataoka, R. e Sugizaki, M. (2006). Blogranger-a multi-faceted blog search engine. *Institute of Electronics, Information and Communication Engineers Technical Report*, 105(650):19–24.

Järvelin, K. e Kekäläinen, J. (2000). Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 41–48, Athens, Greece. ACM.

Juffinger, A.; Granitzer, M. e Lex, E. (2009). Blog credibility ranking by exploiting verified content. In *Proceedings of the 3rd Workshop on Information Credibility on the Web*, pp. 51–58, Madrid, Spain. ACM.

Kritikopoulos, A.; Sideri, M. e Varlamis, I. (2006). Blogrank: ranking weblogs based on connectivity and similarity features. In *Proceedings of the 2nd international workshop on Advanced architectures and algorithms for internet delivery and applications*, p. 8, New York, NY, USA. ACM.

Lagoze, C. e de Sompel, H. V. (2001). The open archives initiative: building a low-barrier interoperability framework. In *Proceedings of the 2001 Joint International Conference on Digital Libraries*, pp. 54–62, Roanoke, Virginia, USA. ACM.

Lin, C.-L.; Tang, H.-L. e Kao, H.-Y. (2009). Utilizing social relationships for blog popularity mining. In *Proceedings of the 5th Asia Information Retrieval Symposium*, pp. 409–419, Sapporo, Japan. Springer.

Liu, Y.; Gao, B.; Liu, T.-Y.; Zhang, Y.; Ma, Z.; He, S. e Li, H. (2008). Browserank: letting web users vote for page importance. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 451–458, Singapore. ACM.

Mishne, G. (2007). Using blog properties to improve retrieval. In *Proceedings of the 1st International Conference on Weblogs and Social Media*, Boulder, Colorado, USA. Retrieved January 19, 2009 from: http://www.icwsm.org/papers/3–Mishne.pdf.

Mishne, G. e de Rijke, M. (2006). A study of blog search. In *Proceedings of the 28th European Conference on Information Retrieval*, pp. 289–301, London, UK. Springer.

Mislove, A.; Gummadi, K. P. e Druschel, P. (2006). Exploiting social networks for internet search. In *Proceedings of the 5th Workshop on Hot Topics in Networks*, pp. 79–84, Irvine, California, USA.

Ounis, I.; de Rijke, M.; Macdonald, C.; Mishne, G. e Soboroff, I. (2006). Overview of the trec-2006 blog track. In *Proceedings of the Fifteenth Text REtrieval Conference*, pp. 15–27, Gaithersburg, Maryland, USA. National Institute of Standards and Technology (NIST).

Stewart, A.; Chen, L.; Paiu, R. e Nejdl, W. (2007). Discovering information diffusion paths from blogosphere for online advertising. In *Proceedings of International Workshop on Data Mining and Audience Intelligence for Advertising*, pp. 46–54, San Jose, California, USA.