FERNANDO DUARTE OLIVEIRA CASTRO

CARACTERÍSTICAS DO TRÁFEGO E PADRÕES DE COMUNICAÇÃO DE UM SERVIÇO DE BLOGS

Belo Horizonte
12 de julho de 2007

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

CARACTERÍSTICAS DO TRÁFEGO E PADRÕES DE COMUNICAÇÃO DE UM SERVIÇO DE BLOGS

Dissertação apresentada ao Curso 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.

FERNANDO DUARTE OLIVEIRA CASTRO

Belo Horizonte
12 de julho de 2007

Resumo

Neste trabalho apresentamos uma caracterização detalhada dos padrões de acesso a um serviço de blogs, uma nova forma de disponibilizar conteúdo na Web. Os blogs são compostos por uma série de textos escritos em publicações e comentários por um crescente número de usuários, que em conjunto constituem uma blogosfera. Nossa caracterização de mais de 35 milhões de requisições de leitura, de escrita e administrativas, enviadas em um período de 28 dias, foi feita sob três diferentes pontos de vista da blogosfera. Na visão do servidor, caracterizamos os padrões de acesso de todos usuários para todos os blogs; na visão dos usuários, caracterizamos como cada um dos usuários interagem com os blogs; e, na visão dos objetos, caracterizamos como cada um dos blogs são acessados. Nossos resultados sugerem duas importantes conclusões. Em primeiro lugar, mostramos que a natureza mais interativa da blogosfera gera padrões interessantes de tráfego e de comunicação que são diferentes dos observados em serviços estáticos da Web. Consideramos os acessos aos objetos da blogosfera como parte de interações entre os donos e os leitores dos blogs. Com base em nosso estudo sobre a conversação entre os usuários da blogosfera, classificamos os blogs em três grupos, que chamamos de *broadcast*, *livro de visitas* e *fórum*. As interações entre membros de grupos de interesse criam uma comunicação mais freqüente dos donos dos blogs para seus leitores em blogs do tipo *broadcast*, mais freqüente dos leitores para os donos dos blogs, em blogs do tipo *livro de visitas*, e mais distribuída em ambas direções em blogs do tipo *fórum*. Em segundo lugar, identificamos e caracterizamos novas propriedades da carga de trabalho de uma blogosfera e investigamos as similaridades e diferenças entre cargas de trabalho de servidores típicos da Web e cargas de trabalho de servidores de blogs.

Abstract

We present a thorough characterization of the access patterns in blogspace – a fast-growing constituent of the content available through the Internet – which comprises a rich interconnected web of blog postings and comments by an increasingly prominent user community that collectively define what has become known as the blogosphere. Our characterization of over 35 million read, write, and administrative requests spanning a 28-day period is done from three different blogosphere perspectives. The *server view* characterizes the aggregate access patterns of all users to all blogs; the *user view* characterizes how individual users interact with blogs; the *object view* characterizes how individual blogs are accessed. Our findings support two important conclusions. First, we show that the more-interactive nature of the blogosphere leads to interesting traffic and communication patterns, which are different from those observed in static web content. We observe that access to objects in blogspace could be conceived as part of an interaction between an author and its readership. As we show in our work, such interactions range from one-to-many *"broadcast-type"* and many-to-one *"registration-type"* communication between an author and its readers, to multi-way, iterative *"parlor-type"* dialogues among members of an interest group. Second, we identify and characterize novel features of the blogosphere workload, and we investigate the similarities and differences between typical web server workloads and blogosphere server workloads.

Agradecimentos

Aos meus familiares, um agradecimento especial, pois foram a fonte, inspiração e suporte para este trabalho. À Leninha, pelo apoio e por compreender as minhas várias horas de dedicação aos estudos. Aos colegas da graduação e do mestrado, pelos momentos tanto divertidos quanto instrutivos.

A todos os envolvidos nas várias edições da Maratona de Programação da ACM. Pelas divertidas participações como competidor, técnico e organizador, agradeço em especial aos lamadivers, pelos auxílios e conquistas, e ao DCC, pelo suporte aos times nas competições.

Aos renomeados professores, pela oportunidade que me proporcionaram, da convivência, suporte e ensinamentos. Ao professor Meira, pelas orientações na iniciação científica. À professora Jussara e ao professor Azer pelas fundamentais contribuições para o presente trabalho.

Ao professor Virgílio, meu orientador, pelos incentivos, idéias e ensinamentos, que me possibilitaram a experiência de escrever artigos e participar de importantes conferências, tanto no Brasil, quanto no exterior.

Por fim, também agradeço à CAPES, pelo suporte financeiro, através do Programa de Fomento à Pós-Graduação. Ao UOL, pelo reconhecimento deste projeto, através do programa Bolsa UOL Pesquisa e pelo acesso às cargas de trabalho que permitiram esta dissertação.

Sumário

Capítulo 1

Introdução

Nos últimos anos, a disponibilização de conteúdo em sítios diferenciados da Web, chamados de blogs, é cada vez mais popular. Blogs, também conhecidos por *weblogs*, são sítios da Web que possuem a aparência de diários pessoais, onde as opiniões dos autores estão bem delimitadas, separadas em publicações, e organizadas de forma cronológica. A idéia de blogs surgiu durante a década de 1990, e a maioria dos blogs eram de jornalistas, como dos pioneiros Justin Hall [3], Jorn Barger [2] e Rebecca Blood [1]. Entretanto, foi somente com o surgimento de serviços de blogs, que facilitaram tanto a criação quanto a atualização de blogs, que esse novo modelo de sítio passou a receber cada vez mais adeptos entre os usuários da Web. Os serviços de blogs oferecem ferramentas para os usuários publicarem informações em blogs de maneira simples e, além disso, armazenam e disponibilizam os blogs na Web.

A liberdade existente na Web para criar publicações sobre os mais diversos temas é um dos fatores desencadeadores do sucesso dos blogs. A maioria dos blogs tratam de assuntos de interesse de muitos usuários, como política, esportes ou tecnologia, contudo, também existem muitos blogs que tratam de assuntos voltados para um público mais restrito, como observações para alunos de um curso ou para funcionários de uma empresa. Os blogs pessoais são muito comuns, por causa da facilidade de criação oferecida pelos serviços de blogs. Além disso, assim como sítios tradicionais da Web, blogs usualmente combinam conteúdo textual com conteúdo multimídia e adicionam elos para outros blogs ou para outros sítios da Web.

Uma característica diferenciadora dos blogs, com relação a outras formas de disponibilização de conteúdo na Web, é a possibilidade dos leitores enviarem comentários para as publicações dos donos dos blogs. Os comentários podem incentivar tanto comentários de outros leitores quanto a criação de novas publicações no mesmo blog ou em blogs diferentes. Isso proporciona um ambiente para interações sociais entre os usuários participantes dos blogs, que são os leitores e os donos do blogs.

Uma característica única dos blogs é como seu conteúdo se modifica ao longo do tempo. Ao contrário de sítios da Web, que são em maioria estáticos e com modificações arbitrárias, tais como substituição ou remoção, difíceis de monitorar ao longo do tempo [55], blogs se modificam usualmente através da adição de novas publicações ou de novos comentários. Além disso, os blogs exibem as datas de criação de cada publicação e de cada comentário, e o con-

teúdo dos blogs é tipicamente mostrado em ordem cronológica, da publicação ou comentário mais recentes para os mais antigos.

O conjunto de blogs e as interações sociais entre os usuários que os acessam formam a blogosfera [37]. Neste trabalho, estudamos o tráfego do serviço de blogs do UOL [7], um provedor de conteúdo bastante popular no Brasil, e analisamos os padrões de comunicação entre os usuários dessa blogosfera.

1.1 Motivação

Nos últimos anos houve um considerável aumento no tamanho da blogosfera. Em 2002, a revista Newsweek [45] estimou que o número de blogs era de meio milhão, atribuindo essa explosão às facilidades de criação do serviço de blogs Blogger.com. No final de novembro de 2006, a blogosfera atingiu a marca de 60 milhões de blogs [9], um número de blogs 120 vezes maior em apenas quatro anos.

Dada a relevância e o contínuo crescimento da blogosfera, é natural questionar se suas características são similares a de serviços existentes da Web. Nos últimos anos, foram apresentados estudos que exploram vários aspectos da blogosfera. Por exemplo, há trabalhos [26, 35, 45] que descrevem o escopo, a estrutura e o padrão de crescimento da blogosfera, como também a rede social entre os participantes de conjuntos de blogs. Tais estudos são importantes porque permitem prever o impacto do uso dos blogs nos servidores do serviço e em outras aplicações, tais como máquinas de busca e sistemas de recomendação.

Uma importante característica diz respeito aos padrões de acesso à blogosfera e qual o impacto do tráfego gerado por esses padrões. Estudos sobre padrões de acesso ao conteúdo tradicional da Web descobriram propriedades fundamentais para explicar características observadas no tráfego [27], que serviram de base para construção de modelos de carga de trabalho e para geração de cargas de trabalho sintéticas [63]. Neste trabalho, focamos nessa dimensão da caracterização da blogosfera, com ênfase no impacto do tráfego e no estudo de padrões de comunicação, em oposto a uma visão de alto nível, tais como a de uma análise da difusão de informação na blogosfera [11] ou da evolução da estrutura de rede entre os blogs [45].

1.2 Definições Básicas

Neste trabalho, nós usamos o termo *blogosfera* para nos referirmos a um conjunto de blogs que induzem interações sociais entre os usuários que os acessam. Nós usamos o termo *dono do blog* para nos referirmos ao usuário que cria e atualiza um blog e usamos o termo *visitante* para nos referirmos aos leitores dos blogs. Definimos os textos criados pelos donos em seus blogs como *publicações* e as escritas criadas por visitantes em resposta a alguma publicação como *comentários*. Nós usamos o termo *requisição* para nós referirmos a um acesso ao servidor do serviço de blogs, tanto para leitura quanto para escrita, e o termo *sessão* para nos referirmos ao período de atividade de um visitante, que é quando acessa os blogs sem longos intervalos de tempo entre as requisições. Ao longo de todo texto, usamos o termo *popularidade* para

referenciar os blogs que recebem mais requisições ou os usuários que mais enviam requisições, sempre como uma métrica para referenciar os mais populares em tráfego. Usamos a expressão *atividades administrativas* para referenciar as facilidades oferecidas pelo serviço de blogs para a criação, edição e remoção de publicações.

1.3 Objetivos e Contribuições

Utilizando uma carga de trabalho do serviço de blogs do UOL, com mais de 32 milhões de requisições, para mais de 210 mil blogs, que transferiram aproximadamente 1 TeraByte de dados em um período de 4 semanas, nós apresentamos uma análise estatística sobre como os usuários lêem os blogs, como enviam comentários e como os donos atualizam seus blogs.

Nós caracterizamos o tráfego da blogosfera de forma hierárquica, utilizando três pontos de vista: ao nível de usuários analisamos como os usuários acessam a blogosfera; ao nível de objetos estudamos como os blogs são acessados; ao nível dos servidores analisamos a agregação das requisições de todos usuários para todos os blogs. Abaixo apresentamos de forma sucinta os principais resultados da análise do tráfego:

- Sessões iniciadas em máquinas de busca, ao contrário de sessões iniciadas em outros sítios da Web, direcionam-se mais para blogs com pouca popularidade do que para blogs com muita popularidade. Isso demonstra que existem sítios na Web que direcionam muitos usuários para os blogs mais populares e indica que as máquinas de busca estão falhando em identificar os blogs mais populares e mais interessantes para os usuários da blogosfera.

- Os donos dos blogs aparentam explorar todas as facilidades do serviço de blogs para manterem os seus blogs atualizados. Os usuários criam, editam e publicam novos textos durante as atualizações dos blogs.

- O tráfego de requisições de leitura, de escrita e administrativas apresentam um comportamento periódico, com maior intensidade durante períodos diurnos e menor intensidade durante períodos noturnos.

- A quantidade de acessos recebida por cada blog ao longo do tempo possui alta variabilidade, com picos de acesso em diversos momentos. Mostramos que não é a quantidade de publicações que influencia na variação de popularidade, mas sim o assunto das publicações, a qualidade dos comentários e a quantidade de acessos vindos de outros blogs.

- As distribuições de popularidade dos blogs seguem uma lei de potência para diversas métricas de popularidade: número de requisições, publicações, sessões e visitantes por blog. Isso mostra que o acesso à blogosfera é concentrado em poucos blogs.

- A distribuição de tamanho das transferências de arquivos possui cauda pesada e segue uma lei de potência. A maioria dos arquivos transferidos são menores do que 12 KB, embora existam arquivos maiores do que 100 KB que representam quase 40% do total de bytes transferidos do servidor.

- Verificamos que blogs com intensa atividade administrativa, em que o dono cria, remove ou edita publicações com freqüência, não necessariamente recebem mais visitantes.

- Em média, a cada 10 requisições de leitura ocorre a transição de um usuário de um blog para outro blog da blogosfera ou para outras partes de um mesmo blog. A maioria das requisições de leitura para os blogs vêm de máquinas de busca e de sítios do provedor de conteúdo que hospeda o serviço de blogs.

Nós também estudamos as interações sociais entre os participantes da blogosfera: os donos dos blogs e seus visitantes. Nós consideramos a blogosfera como um novo meio de comunicação, onde através de leituras, escritas e publicações, esses participantes interagem e dialogam. Na caracterização da comunicação entre os usuários encontramos os seguintes resultados:

- Caracterizamos o diálogo entre os participantes da blogosfera através do intervalo de tempo entre publicações, o intervalo de tempo entre comentários, o intervalo de tempo entre sessões e o intervalo de tempo entre a criação das publicações e os vários comentários que as publicações recebem de visitantes.

- Existe uma tendência que blogs mais populares recebam mais comentários, contudo, existem consideráveis variações na quantidade de comentários entre blogs que recebem uma mesma quantidade de visitantes.

- Existe uma relação inversa entre a popularidade dos blogs e a proporção de sessões que interagem com os blogs através do envio de comentários. Embora os blogs mais populares recebam mais comentários, muitos visitantes desses blogs somente lêem as publicações e não enviam comentários.

- A partir de nossas observações sobre a conversação entre usuários, classificamos os blogs em três grupos, que chamamos de *broadcast*, *livro de visitas* e *fórum*. Blogs do tipo *broadcast* recebem muitas sessões visitantes que somente lêem o blog e não enviam comentários. Blogs do tipo *livro de visitas*, apesar de não serem muito populares, recebem visitantes que em sua maioria enviam comentários. Blogs do tipo *fórum* favorecem a comunicação entre os usuários e recebem uma quantidade razoável de visitas e escritas.

1.4 Organização do Texto

Nosso trabalho está organizado da seguinte forma: no capítulo 2 descrevemos a carga de trabalho do serviço de blogs; no capítulo 3 apresentamos os resultados da caracterização do tráfego da blogosfera sob o ponto de vista dos usuários, dos blogs e dos servidores; no capítulo 4 apresentamos os resultados da caracterização das interações entre os usuários e apresentamos uma classificação de blogs fundamentada nessas diferentes interações; no capítulo 5 discutimos trabalhos disponíveis na literatura sobre caracterização de carga e sobre blogs; e finalmente, apresentamos no capítulo 6 nossas conclusões e indicamos linhas de pesquisa para futuros trabalhos em blogs.

Capítulo 2

Descrição da Carga de Trabalho

Em nosso estudo, nós analisamos três cargas de trabalho do serviço de blogs do UOL [7], um provedor de conteúdo bastante popular no Brasil. A primeira, que chamaremos de *carga de trabalho de leituras*, contém as requisições para o conteúdo dos blogs. A segunda, que chamaremos de *carga de trabalho de escritas*, contém os comentários enviados para os blogs. Finalmente, a terceira, que chamaremos de *carga de trabalho de administração*, contém as requisições correspondentes às atividades administrativas dos donos dos blogs.

2.1 Formato

Cada uma das linhas das cargas de trabalho representa uma requisição enviada por um usuário ao serviço de blogs. As seguintes informações estão disponíveis para cada requisição:

máquina data requisição status tamanho origem agente

O campo *máquina* é o endereço IP que gerou a requisição. O campo *data* indica o segundo, minuto, hora, dia, mês e ano em que a requisição foi recebida pelo servidor. Na carga de leituras, o campo *requisição* contém o objeto requisitado para leitura por um usuário. Na carga de escritas, esse campo contém o comentário escrito por um usuário, mostrando para qual blog e para qual publicação a escrita se destina. Na carga de administração, esse campo indica qual o blog manipulado pelo dono do blog. O campo *status* mostra o código de resposta do protocolo HTTP para a requisição. O campo *tamanho* indica a quantidade de bytes transferidos do servidor pela requisição. O campo *origem* mostra a URL de onde se originou a requisição do visitante. Por exemplo, se um usuário estiver na página A de um sítio qualquer e clicar em um elo que direciona para um blog B, o campo *requisição* conterá a página requisitada do blog B e o campo *origem* conterá a página A. O último campo, *agente*, identifica o navegador e o sistema operacional utilizado para enviar a requisição. Os campos *origem* e *agente* não são obrigatórios, um usuário pode removê-los para aumentar sua privacidade, e o campo *origem* pode não ocorrer, como quando um usuário digita o endereço do blog que vai acessar. Um traço como valor desses campos indica que eles não ocorreram ou estão indisponíveis.

As três cargas de trabalho disponibilizadas para nossa pesquisa foram anonimizadas pelo provedor de conteúdo. Isso foi feito para proteger a privacidade dos usuários do serviço de blogs. A anonimização não impediu que estudássemos as características do tráfego e o comportamento dos usuários. Os campos anonimizados foram *máquina*, *requisição* e *origem*. A anonimização foi feita transformando os IPs dos usuários e as URLs identificadoras dos blogs para números. Durante a anonimização, por exemplo, uma requisição para blog `http://pessoa.blog.uol.com.br/pesquisa.html` seria anonimizada para o formato `http://anon_blog_x/pesquisa.html`, sendo x um número único utilizado para identificar o blog `http://pessoa.blog.uol.com.br` nas três cargas de trabalho.

2.2 Limpeza

Nas cargas de trabalho existem requisições feitas de forma automática, como as enviadas por máquinas de busca, requisições que não foram completadas com sucesso e requisições mal formatadas. Nós eliminamos essas requisições e, portanto, elas não foram utilizadas para a obtenção dos resultados apresentados neste trabalho.

2.2.1 Requisições Feitas de Forma Automática

As requisições presentes nas cargas de trabalho foram feitas tanto por usuários reais, em navegação pelos blogs, quanto por processos automáticos do serviço de blogs ou de máquinas de busca. Esses processos automatizados são programas conhecidos como robôs. O serviço de blogs pode executar um robô para analisar o desempenho do servidores, para coletar páginas com a finalidade de elaborar um mecanismo de busca ou para verificar se há alguma inconsistência nas páginas, como algum elo para uma página inexistente. Os robôs de máquinas de busca são programas feitos para coletar o conteúdo dos blogs e analisar a estrutura de elos entre sítios da Web. Essas duas informações são necessárias para os mecanismos de busca tradicionais.

Robô	Empresa	Quantidade de requisições
Todobr Robot	Akwan e UOL	4.456.198
FAST Enterprise Crawler	UOL	4.323.093
Yahoo! Slurp	Yahoo	1.976.175
Blogshares	Santa Cruz Tech	1.274.341
GoogleBot	Google	1.078.911
MSNBot e MSRBot	Microsoft	458.845
Bloglines	IAC Search & Media	133.616
GigaBot	Gigablast	177.543
Exabot	Exalead	184.958
Outros	-	513.264
Total	-	14.576.944

Tabela 2.1: Requisições feitas por robôs e empresas responsáveis diretamente ou indiretamente pelos robôs durante nosso período de coleta entre janeiro e fevereiro de 2006.

As máquinas de busca identificam seus robôs no campo *agente*, como *Googlebot* e *Yahoo! Slurp* para identificar, respectivamente, os robôs das máquinas de busca Google[5] e Yahoo[6]. Os robôs do próprio provedor de conteúdo são identificados pelos valores *Todobr_ Robot* ou *FAST Enterprise Crawler* no campo *agente*. Embora não tenhamos certeza de como o serviço de blogs utiliza os dados coletados por esses robôs, tivemos que eliminar essas requisições pois estamos interessados em estudar somente o tráfego e o comportamento dos usuários.

A tabela 2.1 mostra que eliminamos mais de 14 milhões de requisições feitas por robôs e apresenta os robôs mais ativos em nossas cargas de trabalho. Essas coletas feitas por robôs têm um impacto alto na infra-estrutura do serviço de blogs. Uma solução seria, por exemplo, utilizar os melhores recursos para atender os usuários e direcionar as requisições de robôs para servidores com menor capacidade de processamento.

Descobrimos que a maioria dos robôs coletam apenas as páginas iniciais dos blogs, poucos coletam as páginas com os históricos de publicações e raramente as páginas com comentários são coletadas. Como exemplo, o robô *Yahoo! Slurp*, que visitou 174.005 blogs diferentes, coletou comentários em apenas 40 desses blogs. Com proporção parecida, o robô *GoogleBot* coletou comentários em apenas 83 dos 139.135 blogs diferentes que visitou. Isso mostra que máquinas de busca não estão coletando informações dos blogs que podem ser úteis para melhorar algoritmos que mostrem para os usuários as melhores publicações sobre um tema.

2.2.2 Requisições com Erros ou mal Formatadas

Em nosso trabalho nós também excluímos as requisições com erro ou mal formatadas. As requisições mal formatadas são aquelas que, talvez por erro dos servidores, não possuem todos os campos ou possuem valores inválidos em algum campo. As requisições com erro são aquelas para páginas inexistentes ou inválidas. Encontramos e excluímos 1.060.216 requisições, sendo 305.045 requisições mal formatadas e 755.171 requisições com erro.

Para descobrir as requisições inválidas ou para páginas inexistentes, nós analisamos os códigos do campo *status* das requisições. Esses códigos, definidos por Fielding *et alia* [31], são números que indicam se o servidor conseguiu interpretar e satisfazer as requisições. Um método comum é classificar os códigos de status em 4 grupos: *com sucesso*, *sem sucesso*, *encontrado* e *não modificado*. O servidor Web retorna um código de status *com sucesso* se a requisição é para um documento válido e o servidor foi capaz de enviar o documento ao cliente. O código *não modificado* é retornado se o cliente já possui a versão atual do documento requisitado em uma cache. Se o documento requisitado possui um novo endereço, o servidor retorna o código *encontrado* além do endereço em que o objeto se encontra. O código *encontrado* também é utilizado pelos servidores para redirecionar os clientes em caso de uma requisição inválida ou para um documento inexistente. Finalmente, o código *sem sucesso* é retornado se um erro ocorre no cliente ou no servidor, como uma requisição inválida, para um documento inexistente ou para um documento sem permissão de acesso. A tabela 2.2 mostra o código de status para as requisições de nossa carga de trabalho. Nós eliminamos as que não foram satisfeitas pelo servidor, que são requisições com código *encontrado* ou com código *sem sucesso*.

Grupo	Código de status	Quantidade de requisições
com sucesso	200	30.982.151
não modificado	304	4.669.030
encontrado	301 e 302	25.177
sem sucesso	4xx e 5xx	729.994

Tabela 2.2: Análise do código de resposta das requisições de nossas cargas de trabalho.

2.3 Sumário

Característica	Valor
Duração	28 dias
Data de início	12/01/2006
Total de bytes transferidos em GB	992,79
Número de requisições de leitura	32.369.178
Número de requisições de escrita (comentários)	277.709
Número de requisições de administração	3.004.294
Número de blogs na carga de leituras	210.738
Número de blogs na carga de administração	74.405
Número de blogs na carga de escritas	30.145
Número de publicações comentadas na carga de escritas	81.561

Tabela 2.3: Sumário da carga de trabalho, excluindo requisições feitas de forma automática, com erros ou mal formatadas.

A tabela 2.3 apresenta um sumário com estatísticas sobre as cargas de trabalho. Podemos observar que os servidores que analisamos neste trabalho recebem uma considerável quantidade de requisições. Nosso estudo foi feito sobre mais de 32 milhões de requisições de leitura e cerca de 278 mil comentários. As requisições foram feitas no período de 4 semanas, de 12 de janeiro a 9 de fevereiro de 2006. Durante esse período de tempo, aproximadamente 992 GB de dados foram transferidos pelos usuários, cerca de 210 mil blogs distintos foram acessados e mais de 81 mil publicações de mais de 30 mil blogs receberam pelo menos um comentário.

Observamos que nossa carga de trabalho e, conseqüentemente, nossa caracterização, contém todos comentários enviados por visitantes, incluindo aqueles que foram removidos ou não autorizados pelo dono do blog de serem exibidos na página do blog. Portanto, nossa análise de comentários, especialmente no que se refere à popularidade de blogs, publicações e interações entre usuários, é mais precisa do que uma análise que utiliza robôs para, por exemplo, caracterizar a distribuição de comentários por publicação. Um outro ponto importante é que o serviço de blogs requer que os usuários respondam a um *captcha* a cada envio de comentário. O *captcha* é um teste que solicita ao usuário que escreva uma série de letras ou números que aparecem em uma imagem, que é geralmente um pouco distorcida ou ofuscada para evitar o reconhecimento por máquinas. Com isso, acreditamos que nossa análise de comentários não é distorcida pela presença de *spams*, que são comentários com publicidade ou propaganda, normalmente com fins comerciais, enviados de forma automática por empresas ou usuários.

Capítulo 3

Caracterização do Tráfego

3.1 Metodologia

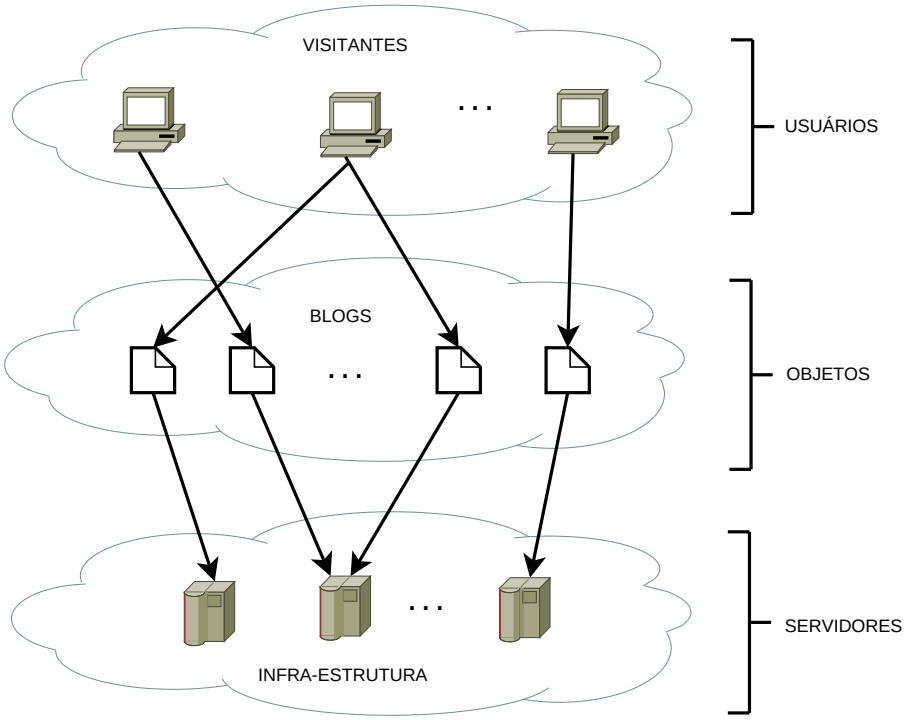

Figura 3.1: Três diferentes visões utilizadas para caracterizar a blogosfera: usuários, objetos e servidores.

A caracterização do tráfego do serviço de blogs foi feita de forma hierárquica, utilizando os três diferentes pontos de vista exibidos na figura 3.1. Primeiramente, analisamos na seção 3.2 como os usuários acessam a blogosfera. Em seguida, investigamos na seção 3.3 como os blogs são acessados e finalizamos na seção 3.4 com o estudo do tráfego recebido pelos servidores, a agregação das requisições de todos usuários para todos os blogs.

3.2 Caracterização ao Nível de Usuários

Nesta seção nós focamos no estudo dos usuários da blogosfera, ou seja, investigamos como usuários utilizam o serviço de blogs através de requisições de leitura, de escrita e administrativas.

3.2.1 Definição e Criação de Sessões

Para analisar como os usuários utilizam o serviço de blogs nós agrupamos as requisições em sessões. Nossa definição de sessões e o método que utilizamos para encontrá-las foram propostos em estudos sobre usuários de servidores tradicionais da Web[47, 58, 63]. Identificamos unicamente um usuário através do par formado pelos campos *máquina* e *agente* das requisições e definimos uma sessão como o intervalo de tempo em que um usuário está ativamente utilizando a blogosfera. Sessões são separadas por um período de inatividade do usuário. Uma sessão inicia com a primeira requisição enviada por um usuário e termina quando o tempo desde a última requisição na sessão ultrapassar um valor limite de τ minutos. Após esse tempo, uma nova sessão é iniciada para o mesmo usuário.

É importante escolher um bom valor para o parâmetro τ que separa as sessões de um mesmo usuário. Por um lado, se esse valor limite for muito curto, a visitação de um usuário à blogosfera poderá ser incorretamente dividida em várias sessões, principalmente se um usuário permanecer inativo por algum tempo lendo publicações de um blog ou escrevendo comentários mais elaborados. Por outro lado, se o valor limite for muito longo, diferentes sessões de um mesmo usuário, em diferentes parte do dia, poderão ser agrupadas em uma única sessão. Além disso, se o valor for muito longo, será mais provável a agregação de diferentes visitantes que enviam requisições através uma mesma máquina e utilizando um mesmo navegador.

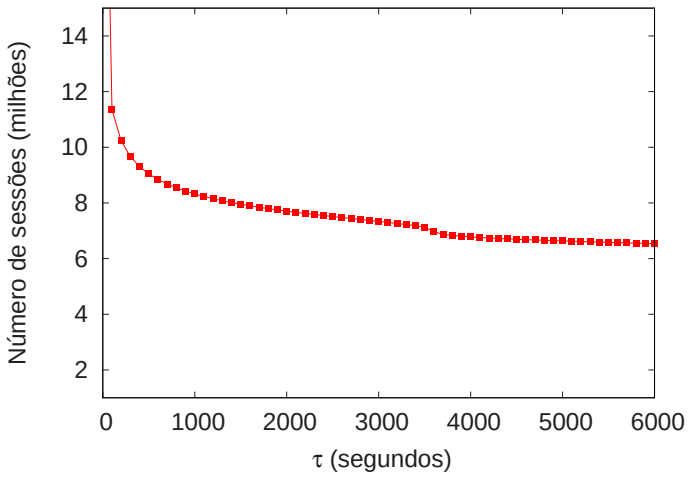

Figura 3.2: Número de sessões variando o intervalo de tempo entre sessões τ.

A figura 3.2 permite avaliar o efeito da utilização de diferentes valores para o parâmetro τ em nossa carga de trabalho. Agregamos as três cargas de trabalho para dividir os acessos em sessões, pois, em uma mesma sessão, um usuário pode ler um blog, enviar um comentário e

administrar seu blog. Podemos notar que de 0 a 1000 segundos, o número de sessões geradas cai rapidamente. De 1000 segundos em diante, o decrescimento no número de sessões é lento. Isso indica que a maioria das sessões duram menos do que 1000 segundos. A partir dessa análise resolvemos escolher o valor de 1800 segundos, 30 minutos, para o parâmetro τ.

Tipo	Número de usuários	Número de sessões
Carga de Trabalho Agregada	4.235.557	6.968.140
Carga de Trabalho de Leituras	4.193.371	6.818.510
Carga de Trabalho de Escritas	117.150	149.439
Carga de Trabalho de Administração	187.982	268.310

Tabela 3.1: Quantidade de sessões e usuários das cargas de trabalho.

A tabela 3.1 apresenta a quantidade de usuários e de sessões encontradas na carga de trabalho agregada. Além disso, essa tabela mostra quantos usuários e sessões da carga de trabalho agregada fazem parte das nossas três diferentes cargas de trabalho: leituras, escritas e administração. O total da carga de trabalho agregada é diferente da soma das outras três cargas pois existem sessões em que os usuários enviam mais de um tipo de requisição. Podemos ver que quase 7 milhões de sessões, representando o acesso de mais de 4 milhões de usuários, acessaram a blogosfera no período analisado. Também percebemos que a grande maioria das sessões visitantes lê o conteúdo dos blogs.

3.2.2 Origem das Sessões

Para investigar como os usuários chegam à blogosfera, nós analisamos a quantidade de sessões que utilizam máquinas de busca ou sítios externos para acessar os blogs. A figura 3.3 apresenta o resultado do estudo do campo *origem* da primeira requisição de cada sessão da carga de trabalho de leituras. Do total de sessões, 29% não possuíam o campo *origem* preenchido na primeira requisição e, por esse motivo, não foram utilizadas para essa análise. Além disso, ignoramos 4% das sessões que vieram da própria blogosfera. Isso ocorre quando o usuário utiliza mais de um IP durante seu acesso aos blogs, como por causa da interrupção de uma conexão discada, ou quando o usuário fica inativo por um longo período sem fechar o navegador.

É interessante observar na figura 3.3 que uma grande parte das sessões acessam a blogosfera através de máquinas de busca. Como máquinas de busca costumam ordenar seus resultados baseado na popularidade, tais como algoritmos que utilizam a estrutura de elos entre sítios da Web [24, 41], poderíamos esperar que blogs populares atraíssem uma desproporcional fração das sessões iniciadas através de máquinas de busca. Para verificar se isso ocorre em nosso servidor de blogs, nós contamos o número de requisições que tiveram início, em máquinas de busca ou em sítios da Web externos à blogosfera, que foram para 5% dos blogs mais populares e para 95% dos blogs menos populares em quantidade de acessos. A tabela 3.2 apresenta o resultado e indica que, ao contrário do esperado, máquinas de busca direcionam mais tráfego para blogs menos populares do que para blogs mais populares. Essa é uma observação importante, pois sugere que o uso de máquinas de busca tem um efeito igualitário

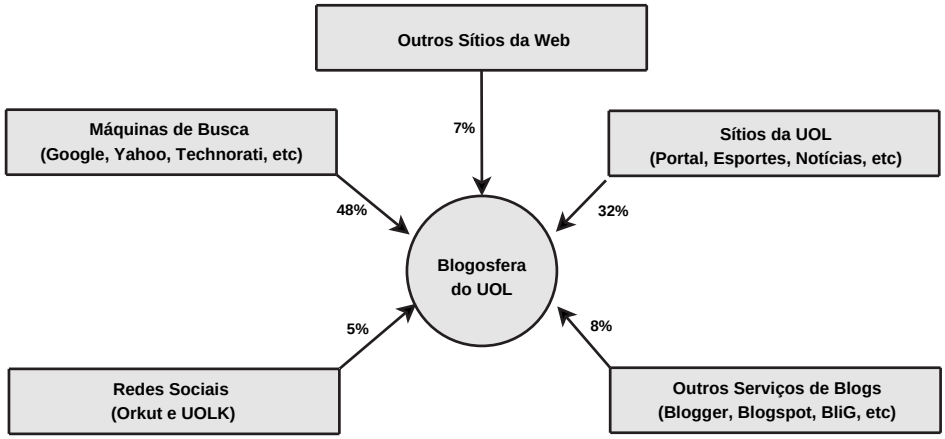

Figura 3.3: Diferentes formas de acessar a blogosfera: fração de sessões que tem origem em máquinas de busca e em diferentes sítios da Web.

[48] na blogosfera. Uma outra maneira de interpretar esse resultado é que a popularidade dos blogs mais acessados não é uma influência da utilização de máquinas de busca, porém o resultado do direcionamento de usuários para os blogs através da estrutura dos blogs, da Web e da rede social entre os usuários da blogosfera.

Origem da requisição	**Fração de requisições para**	
	5% blogs mais populares	**outros 95% blogs menos populares**
Máquinas de Busca	0.46	0.54
Sítios da Web	0.63	0.37

Tabela 3.2: Diferentes taxas com que máquinas de busca e outros sítios da Web direcionam tráfego para os blogs mais populares e para blogs menos populares.

A figura 3.4 quantifica essa observação mostrando a probabilidade acumulada que uma sessão iniciada através de máquina de busca ou de sítios da Web acessam blogs com popularidade maior do que um certo valor, isto é, blogs com *rank* de popularidade menor do que um certo valor. Essa figura mostra claramente que sessões iniciadas através de máquinas de busca são menos prováveis de acessar blogs mais populares do que aquelas sessões iniciadas através de sítios da Web.

3.2.3 Quantidade de Atividades dos Usuários

Durante o nosso período de observação, cada usuário poderia acessar a blogosfera várias vezes, seja para visitar os blogs, se expressar com o envio de comentários, ou administrar um ou mais blogs. Para caracterizar a intensidade de interesse dos usuários pela blogosfera, a figura 3.5 apresenta a quantidade de acessos dos usuários em número de requisições de leitura, escrita e administrativas. As curvas são mostradas do usuário mais ativo para o usuário menos ativo. As distribuições das atividades dos usuários seguem uma lei de potência, sendo a quantidade de atividades do i-ésimo usuário mais ativo proporcional a $i^{-\alpha}$. Encontramos o expoente $\alpha = 0.83$ para a quantidade de requisições de leitura, $\alpha = 0.54$ para a quantidade de

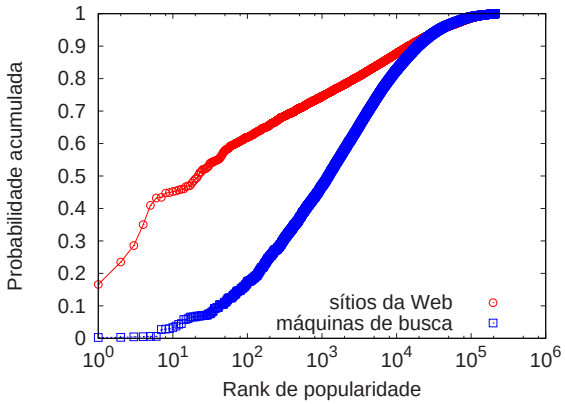

Figura 3.4: Probabilidade acumulada que uma sessão originada através de máquina de busca ou sítios da Web irá acessar um blog com rank de popularidade menor do que um certo valor.

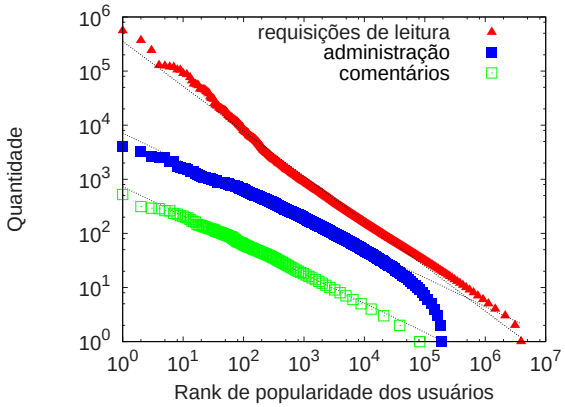

Figura 3.5: Freqüência de acesso dos usuários de acordo com o rank de interesse dos usuários.

requisições de escrita e $\alpha = 0.53$ para a quantidade de requisições administrativas, todas as três regressões lineares com $R^2 = 0.99$. Para o cálculo do expoente da curva de administração não consideramos a cauda da curva, por ela possuir um decaimento provocado por usuários com poucas requisições administrativas.

3.2.4 Identificação de Atividades Administrativas

A carga de trabalho de administração contém uma série de requisições que representam as atividades administrativas dos donos dos blogs. As atividades administrativas podem ser: salvar sem publicar um texto editado ou novo; salvar e publicar um texto editado ou novo; remover uma publicação já existente; editar uma publicação já existente; e publicar um texto que foi salvo mas não publicado.

Para poder analisar o comportamento dos donos dos blogs, nós criamos uma heurística que identifica as atividades administrativas a partir da análise das requisições. Durante a navegação pelo sítio de administração de blogs, cada atividade administrativa é realizada com o envio de uma seqüência de requisições. A tabela 3.3 apresenta assinaturas das atividades administrativas que obtivemos após minucioso estudo do sítio de administração de blogs. As

assinaturas são descritas usando os campos *requisição* e *origem* das requisições. Repare que algumas atividades são identificadas com mais de uma requisição. Nesses casos, as requisições devem aparecer na ordem indicada e sem requisições intermediárias. Note também que, com as informações disponíveis na carga de trabalho, não é possível distinguir uma operação de remoção de uma operação de publicação.

Salvar
requisição: POST `showposts.html` *origem:* - *requisição:* GET `showposts.html?paramCase=listPosts&publishStatus=0` *origem:* -
Salvar e publicar
requisição: POST `showposts.html` *origem:* - *requisição:* GET `showposts.html?paramCase=listPosts&publishStatus=2` *origem:* -
Editar
requisição: GET `showposts.html?paramCase=showOnePost&id_da_publicacao` *origem:* `showposts.html?paramCase=listPosts`
Remover ou publicar
requisição: POST `showposts.html` *origem:* `showposts.html?paramCase=listPosts`

Tabela 3.3: Seqüência de requisições que identificam as atividades administrativas. Todas requisições são para o servidor *http://blog.uol.com.br*.

Nós utilizamos nossa heurística para identificar as atividades administrativas de cada sessão de usuário da carga de trabalho de administração. A tabela 3.4 apresenta um sumário das atividades dos donos dos blogs, mostrando o número de vezes que cada atividade foi realizada e o número de sessões que realizaram pelo menos uma atividade. Aproximadamente 30% das sessões não tiveram nenhuma atividade identificada além das requisições de navegação pelo sítio de administração. É possível que haja atividades não identificadas devido à existência de requisições sem os endereços necessários no campo *origem*. Podemos perceber que a atividade mais freqüente dos donos do blogs é salvar e publicar um texto editado ou novo. O alto valor de ocorrência das atividades indica que os usuários usufruem das facilidades do sítio de administração de blogs, criando, editando e publicando novos textos.

3.3 Caracterização ao Nível de Objetos

Nesta seção nós investigamos a blogosfera no nível de blogs, ou seja, apresentamos características dos blogs e como eles são acessados pelos usuários.

Tipo de atividade	Número de realizações	Número de sessões
Salvar	95.126	69.682
Salvar e publicar	290.207	99.940
Editar	188.154	62.932
Remover ou publicar	207.020	113.702
Total	780.507	178.149

Tabela 3.4: Sumário das atividades administrativas. Número de vezes que cada atividade foi realizada e número de sessões que realizaram pelo menos uma atividade.

3.3.1 Padrão Temporal do Acesso aos Blogs

A figura 3.6 mostra a quantidade de blogs sendo requisitados e comentados ao longo do tempo, em intervalos de quinze minutos. As curvas apresentam padrões periódicos, com maior intensidade de acesso durante o dia e menor intensidade durante a noite. Aproximadamente 3000 blogs são requisitados para leitura durante o dia em horários de maior movimento, enquanto que no período noturno uma média de 500 blogs diferentes são requisitados. Durante o período mais intenso do dia, em média cerca de 100 blogs recebem comentários e, durante períodos de menor tráfego, menos de 10 blogs recebem comentários. É esperado que o tráfego de escritas seja menos intenso do que o de leituras, pois é bem mais trabalhoso enviar um comentário do que somente acessar e ler um blog.

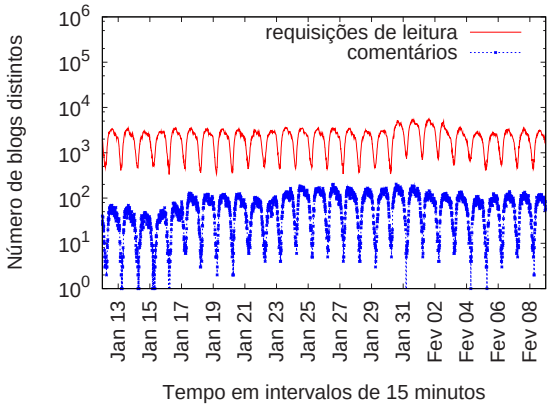

Figura 3.6: Comportamento periódico dos acessos à blogosfera: número de blogs distintos acessados com requisições de leitura e comentários.

3.3.2 Variabilidade na Intensidade dos Acessos

O número de acessos recebidos por um sítio da Web está relacionado à popularidade do conteúdo disponível nas páginas e, essa popularidade, pode variar ao longo do tempo. Para o conteúdo tradicional da Web, as mudanças na popularidade de uma página geralmente não ocorrem rapidamente, o que resulta em pouca variabilidade na intensidade dos acessos, por exemplo, sendo mais perceptíveis somente variações de intensidade entre o dia e a noite e entre dias úteis e fins de semana [63]. Na blogosfera, a popularidade de um blog ao longo do

tempo varia mais em função do conteúdo das publicações e dos comentários, da quantidade de referências de outros blogs populares, e do renome do dono do blog.

As únicas exceções, no caso das páginas com conteúdo tradicional, ocorrem para sítios relacionados a notícias, os quais possuem uma variabilidade nos acessos similar a que ocorre em blogs. Nesse tipo de sítios, a popularidade pode variar em um período muito curto de tempo devido a fatores externos, como guerras, crises, notícias de celebridades etc. Como nós mostraremos na seção 4.2, blogs com conteúdo direcionado a notícias representam uma classe distinta de blogs, que apesar de serem muito populares, não possuem muita interação com os visitantes.

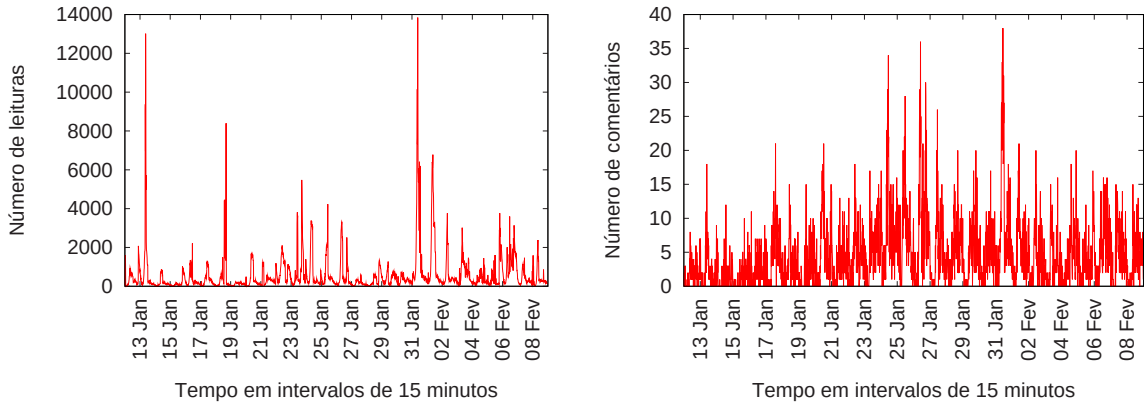

Figura 3.7: Variabilidade na popularidade do blog mais popular: número de requisições de leitura (direita) e número de comentários (esquerda).

Para ilustrar a intensa variabilidade na popularidade dos blogs, a figura 3.7 apresenta a quantidade de requisições de leitura e de escritas enviadas para o blog mais popular em tráfego de nossa carga de trabalho de leituras. Fica clara a variação na intensidade dos picos de acesso, que chega a ser maior do que uma ordem de grandeza, e a falta de diferenciação entre dias úteis e fins de semana. A curva apresenta mais um comportamento com alta variabilidade de intensidade do que periódico. Na visitação ao blog mais popular, podemos observar picos de acesso com alta intensidade, influenciados pelo comportamento diurno, próximo aos dias 13, 18, 24 de janeiro e 1 de fevereiro. Nós analisamos o tráfego de atividades administrativas para esse blog e concluímos que aumentos na quantidade de leituras não coincidem com aumentos de intensidade na atividade do administrador do blog. Para ficar mais claro, a correlação entre o número de leituras e o número de atividades administrativas por dia é de apenas 0,25 para esse blog, e de 0,20, em média, para os 10 blogs mais populares. Na verdade, percebemos que o aumento de visitantes incentiva novas atividades administrativas e que, contudo, o contrário nem sempre é verdade. Isso nos permite supor que não é o número de publicações, porém o assunto das publicações, a qualidade dos comentários e a quantidade de acessos vindos de outros blogs é que geram os picos de acessos.

3.3.3 Popularidade dos Blogs

Vários trabalhos [15, 28, 64] mostram que a distribuição de popularidade dos objetos disponíveis na Web segue uma lei de potência. Esses estudos mostram o número de acessos aos objetos em função do rank de popularidade dos objetos, do objeto mais popular para o objeto menos popular, sendo a quantidade de acessos ao i-ésimo objeto mais popular proporcional a $i^{-\alpha}$. Uma distribuição que segue uma lei de potência aparece como uma reta quando essa análise é feita em um gráfico que esteja em escala logarítmica nos dois eixos.

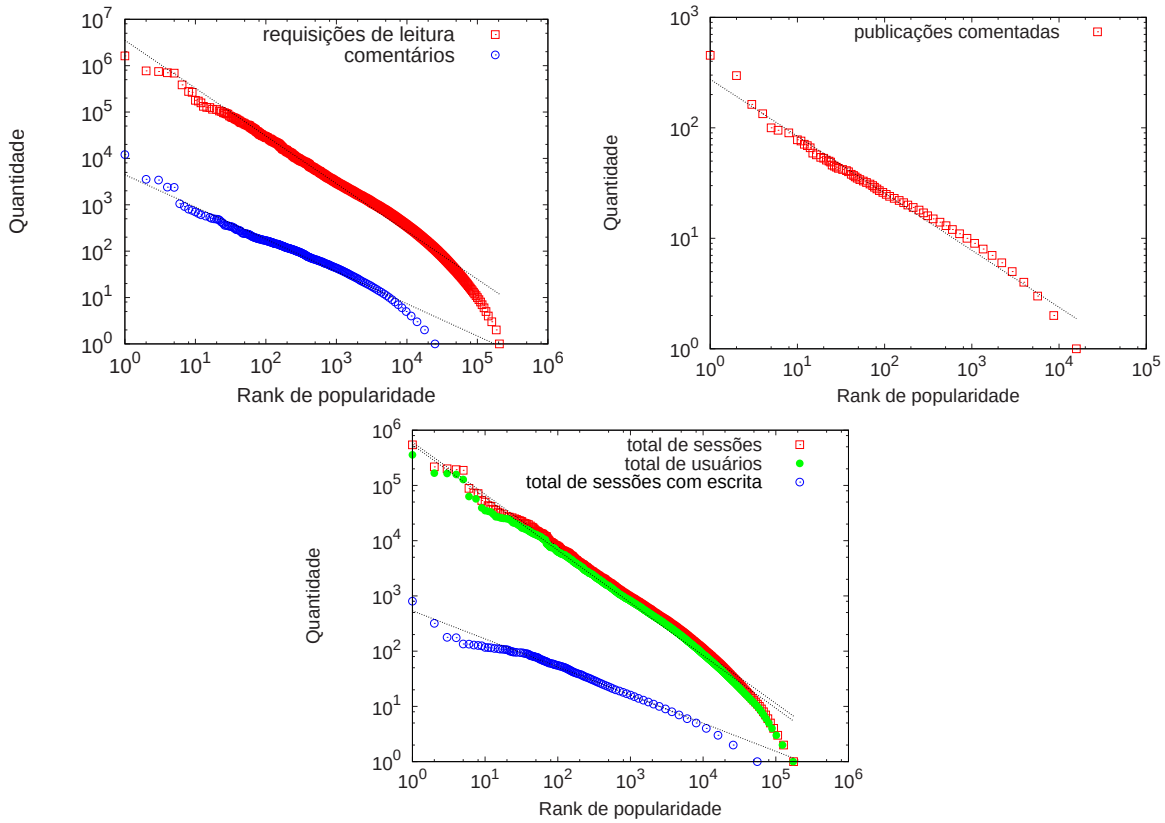

Figura 3.8: Popularidade dos blogs em diferentes métricas: quantidade de requisições de leitura e de escrita (superior à esquerda), publicações comentadas (superior à direita), sessões, sessões com escrita e usuários (inferior).

Em nossa análise da blogosfera, nós encontramos que a popularidade de acesso aos blogs também segue uma lei de potência. A figura 3.8 mostra a popularidade dos blogs usando diferentes métricas de popularidade: requisições, publicações, usuários e sessões. Os gráficos mostram o perfil de popularidade dos blogs utilizando uma escala logarítmica nos dois eixos e exibindo a resultado do blog mais popular para o blog menos popular. A análise das requisições indica que o acesso é concentrado nos blogs mais populares, sendo que aproximadamente 90% das leituras e 60% dos comentários são enviados para 10% dos blogs mais populares. Essa concentração fica mais clara quando observamos que 21 blogs, 0,01% do total de blogs, concentram 7,5 milhões das requisições de leitura, cerca de 23% do total de requisições de leitura. A figura 3.8 mostra que a quantidade de requisições de leituras e de escritas em função do rank de popularidade do blog segue uma lei de potência com parâmetro α. Para o

total de requisições de leitura como indicador de popularidade encontramos $\alpha = 0.97$ ($R^2 = 0.96$). Encontramos uma menor concentração de requisições de escrita enviadas aos blogs, com $\alpha = 0.70$ ($R^2 = 0.97$). A figura 3.8 também mostra que o mesmo perfil de popularidade, uma lei de potência, ocorre quando consideramos a quantidade de publicações que receberam pelo menos um comentário, o número de usuários distintos que acessaram o blog, o total de sessões ou o total de sessões com escrita de comentários, como métricas de popularidade. Esse resultado é importante para o planejamento da infra-estrutura do serviço de blogs. Pode ser interessante alocar recursos para os blogs mais populares e tratar os blogs mais populares de forma diferenciada em um mecanismo de *caching*.

3.3.4 Impacto da Atividade do Administrador na Popularidade

Como blogs possuem diferentes níveis de popularidade e de interações com os visitantes, uma pergunta que podemos fazer é se essas características estão relacionadas com o nível de atividade do dono do blog. Queremos responder se a intensidade de atividades administrativas em um blog influencia na intensidade de requisições enviadas pelos visitantes.

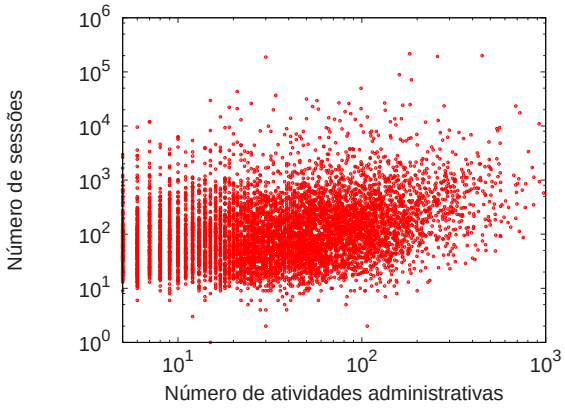

Figura 3.9: Pouca correlação entre o total de sessões e o total de atividades administrativas.

Na figura 3.9, cada blog é representado por um ponto, e as coordenadas representam o total de sessões que o acessaram e o número de atividades administrativas de seu dono. Podemos observar que a correlação entre a quantidade de atividades administrativas e sessões visitantes é muito pequena, praticamente inexistente. Isso é confirmado pelo coeficiente de correlação de valor 0,26. Por este motivo, acreditamos que a popularidade de um blog depende mais do conteúdo e visibilidade das publicações, do relacionamento entre os usuários participantes, donos e visitantes dos blogs, do que da quantidade de atividades dos donos dos blogs.

3.4 Caracterização ao Nível de Servidores

Nesta seção analisamos a carga de trabalho que chega aos servidores do serviço de blogs. Nesse nível, nós investigamos a agregação das requisições enviadas por todos os usuários, para todos os blogs.

3.4.1 Tipos de Arquivos Requisitados

Utilizamos o campo *requisição* das requisições da carga de trabalho de leituras para identificar a extensão dos arquivos requisitados e verificar quais os tipos de arquivos mais solicitados dos blogs.

Tipo	% requisições	% bytes
HTML	61.55	96.76
Imagem	00.58	00.87
Java Script	35.57	00.87
CSS	00.14	00.02
Áudio	00.00	00.05
Vídeo	00.42	00.70
Diretório	00.06	00.00
Outros	01.68	00.73
Total	100.00	100.00

Tabela 3.5: Sumário dos tipos de arquivos transferidos dos blogs.

A tabela 3.5 mostra que a maioria das requisições são para arquivos HTML e Java Script. O Java Script é utilizado principalmente para formatar as páginas dos blogs. O serviço de blogs, por exemplo, mantém as páginas dos blogs estáticas e, para cada página, cria um arquivo contendo o número de comentários enviados para cada um das publicações do blog. O Java Script é utilizado para ler o arquivo com a quantidade de comentários enviados para cada publicação e exibi-los na página do blog no momento da visualização pelo navegador.

Uma outra observação sobre a tabela 3.5 é que, diferentemente de outros estudos sobre o tráfego da Web [15, 64], a carga de trabalho possui poucas requisições para imagens. Isso ocorre porque o serviço de blogs armazena imagens comuns a todos os blogs, tais como o logotipo do serviço de blogs e imagens para formatação padrão das páginas, em um servidor separado, ao qual nós não tivemos acesso a carga de trabalho. Entretanto, requisições para imagens adicionadas pelo dono do blog e armazenadas em sua conta do serviço de blogs aparecem em nossa carga de trabalho. Portanto, podemos supor que os blogs possuem muitas publicações somente com texto ou publicações que apontam para imagens armazenadas em outros servidores.

3.4.2 Distribuição de Tamanho das Transferências de Arquivos

Para analisar o tamanho dos arquivos transferidos pelos visitantes dos blogs, nós utilizamos o campo *tamanho* das requisições da carga de trabalho de leituras. A tabela 3.6 apresenta um sumário sobre o tamanho dos arquivos requisitados. Observe que os blogs não possuem objetos grandes, sendo que a mediana do tamanho dos arquivos transferidos é de 12 KB. Talvez, por existir uma limitação de tamanho para os blogs armazenados no servidor, que varia entre 6 e 50 MB, o maior arquivo transferido possui aproximadamente 21 MB. Embora a maioria dos arquivos tenha tamanho pequeno, devido ao intenso tráfego do serviço de blogs,

em média mais de 32 GB são transferidos por dia e quase 1 TB de dados foi transferido dos servidores no período de 4 semanas.

Característica	Valor
Total de requisições	32.369.178
Total de GB transferidos	992,79
Média de requisições por dia	1.156.042
Média de GB transferidos por dia	35,46
Média em KB	32,16
Mediana em KB	12,07
Tamanho máximo em KB	22.056,96

Tabela 3.6: Sumário sobre o tamanho dos arquivos transferidos da blogosfera.

A figura 3.10 mostra a distribuição acumulada complementar do tamanho das transferências realizadas pelos visitantes. A distribuição possui uma cauda pesada que segue uma lei de potência. Ela é melhor aproximada por uma distribuição Pareto com expoente $\kappa \approx 1$ e, logo, por uma lei de potência com $\alpha = \kappa + 1$ [10]. Este resultado é similar ao encontrado em estudos sobre o tráfego de servidores da Web [27, 64]. A tabela 3.6 mostra que 50% das requisições são para arquivos menores do que 12 KB e podemos observar na figura 3.10 que 94% das requisições são para objetos menores do que 100 KB. Entretanto, embora não freqüêntes, os arquivos maiores do que 100 KB representam 36% do total de bytes transferidos do servidor. Entre os arquivos maiores que 5 MB, encontramos arquivos comprimidos (extensão zip), arquivos de vídeo (extensão mpg) e arquivos de áudio (extensão mp3).

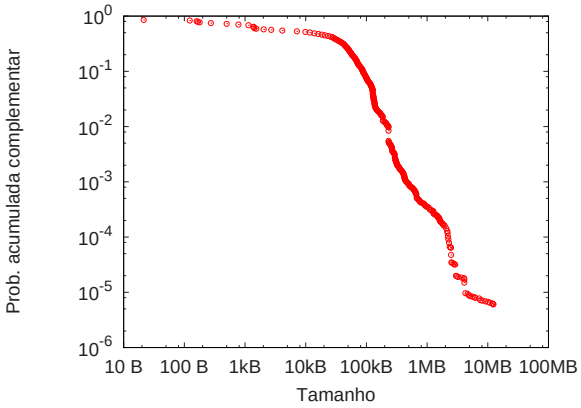

Figura 3.10: Probabilidade acumulada complementar (CCDF) do tamanho das transferências de arquivos.

3.4.3 Padrão Temporal do Tráfego de Requisições

Analisando o tráfego de requisições ao longo do tempo, nós observamos que o acesso agregado de todos usuários para todos os blogs é periódico. Assim como discutido na seção 3.3.1, sobre os acessos ao nível de blogs, o tráfego possui maior intensidade durante o dia e menor intensidade durante a noite, e é similar ao descrito em estudos sobre servidores tradicionais

da Web [63]. A figura 3.11 mostra a intensidade do tráfego em duas diferentes granularidades: medida em número de bytes transferidos em leituras e em número de requisições de leituras e de escritas. Comparando com a figura 3.7, percebemos que a agregação do tráfego ocasiona a perda de informações sobre blogs que possuem comportamento distinto, com alta variabilidade de tráfego.

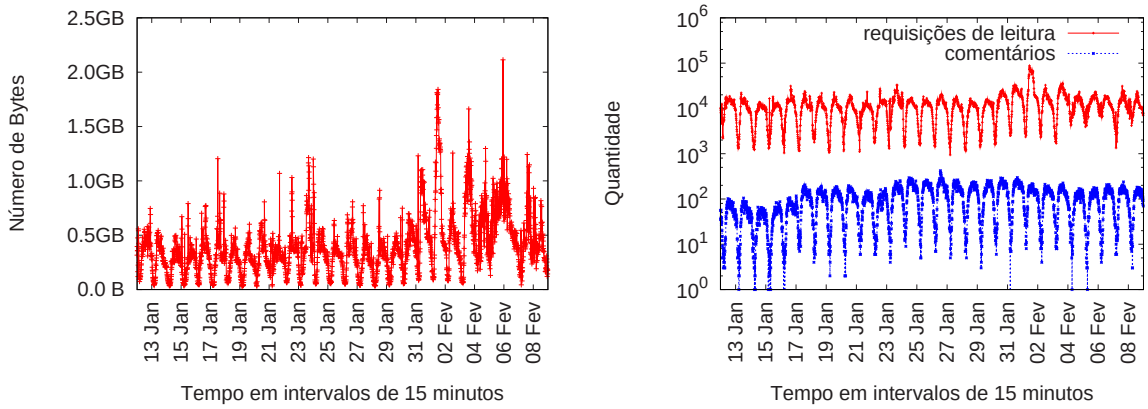

Figura 3.11: Padrão temporal do tráfego: número de bytes transferidos em requisições de leitura e quantidade de requisições de leitura e de escrita.

Podemos observar na figura 3.11 que, em média, 500 MB são transferidos do servidor a cada 15 minutos durante o período diurno. Isso mostra a alta taxa de utilização dos servidores pelos usuários dos blogs. Além disso, no decorrer do tempo, percebemos que a quantidade de comentários enviados foi aproximadamente duas ordens de grandeza menor do que a quantidade de requisições de leitura. Também podemos ver uma grande variabilidade na intensidade dos picos de acesso. Existem períodos em que 2 GB são transferidos do servidor em quinze minutos, um valor 4 vezes maior do que a média. Embora as variações das curvas de requisições de escrita e leitura estejam amenizadas com a escala logarítmica, podemos ver, por exemplo, períodos de quinze minutos em que o tráfego passa de 10 mil requisições de leitura para 100 mil requisições de leitura. Na seção 3.3.2 argumentamos que essa variabilidade no tráfego ocorre como uma conseqüência das interações sociais entre os membros da blogosfera.

Também é interessante observar na figura 3.11, que o tráfego de requisições de leitura é intenso mesmo no período noturno. Em média, acima de 1000 requisições são enviadas aos servidores mesmo durante a madrugada. Isso indica que existem usuários noturnos ou usuários acessando de outros países, em diferentes fusos horários. O horário que estamos analisando é o de recebimento das requisições pelos servidores de blogs e, além disso, os endereços IP das máquinas que enviam as requisições estão anonimizados. Por esses motivos, não foi possível investigarmos a localização geográfica dos usuários e o impacto de fusos horários em nosso trabalho. Entretanto, como o serviço de blogs possui muita popularidade no Brasil e o conteúdo dos blogs é escrito em português, acreditamos que a grande maioria dos acessos vem de usuários localizados no Brasil, e que nosso estudo sofre pouca influência de fusos horários.

3.4.4 Origem das Requisições

Além de analisar o tráfego do serviço de blogs ao longo do tempo, nós também investigamos a origem do tráfego, de forma semelhante ao que foi feito para sessões de usuários na seção 3.2.2. Agora, analisamos o campo *origem* das requisições carga de trabalho de leituras. Nós ignoramos 30% das requisições que não possuíam esse campo preenchido. Classificamos a origem das outras requisições como internas ou externas. Requisições internas são originadas pelos próprios blogs do serviço que estamos analisando, através de elos disponibilizados em comentários ou publicações. Requisições externas são originadas de redes sociais, outros serviços de blogs, máquinas de busca, sítios tradicionais da Web ou de sítios do UOL.

Origem	**% de Requisições**
Requisições Externas	
Máquinas de Busca (Google, Yahoo, Technorati, etc)	34
Redes Sociais (Orkut e UOLK)	3
Outros Serviços de Blog (Blogger, Blogspot, BliG,etc)	8
Sítios do UOL (Portal, Esportes, Notícias, etc)	25
Outros Sítios da Web	6
Total	76
Requisições Internas	
Dentro de um mesmo Blog	14
Entre Blogs Diferentes	10
Total	24

Tabela 3.7: Sumário sobre a origem das requisições

A tabela 3.7 apresenta o resultado da análise do campo *origem* das requisições internas e externas. Podemos entender cada uma dessas requisições como uma transição feita pelo usuário, que explicitamente, como através de um clique em uma figura que lhe interesse, passa de uma página para outra página. A análise das requisições internas indica como os usuários estão navegando dentro da blogosfera. Percebemos que em apenas 10% das requisições ocorreu a transição de um usuário de um blog para outro blog da blogosfera e que em 14% das requisições ocorreu a navegação de um usuário visitando diferentes parte de um mesmo blog. A maioria das requisições são externas. Isso indica que muitos usuários acessam os blogs por influência externa, como de máquinas de busca (34%) ou sítios do provedor de conteúdo que hospeda o serviço de blogs (25%).

Capítulo 4

Padrões de Comunicação

Neste capítulo estudamos a interação entre os participantes da blogosfera: os donos dos blogs e seus visitantes. Nós consideramos a blogosfera como um novo meio de comunicação, onde, através das leituras, escritas e publicações, esses participantes interagem e dialogam.

4.1 Interações entre os Participantes da Blogosfera

Uma das características mais marcantes da blogosfera são as interações entre os usuários através de publicações e comentários. Essas interações formam diálogos entre os participantes da blogosfera. Esses diálogos representam uma nova forma de comunicação na Web, e ocorrem entre os donos do blogs e seus visitantes.

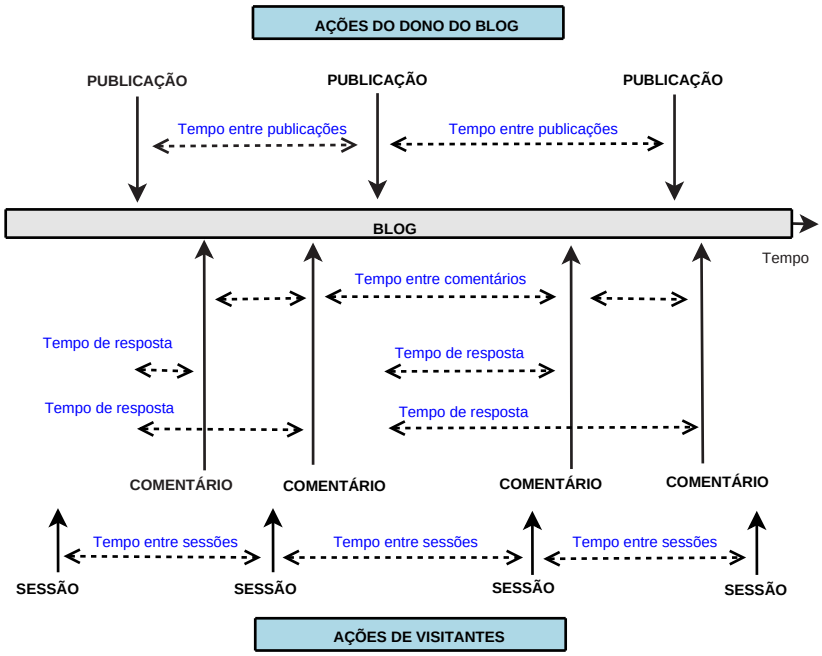

Figura 4.1: A estrutura das interações induzidas por um dado blog é definida pelas de ações do dono do blog e dos visitantes através de publicações, sessões e comentários.

Para analisar as características da comunicação entre os participantes da blogosfera, nós

propomos a estrutura de interações mostrada da figura 4.1. Nós representamos o diálogo utilizando a seqüência de publicações criadas pelo dono de um blog e a seqüência de comentários enviada pelos seus visitantes. As publicações representam mensagens enviadas pelo dono do blog aos visitantes e os comentários representam as respostas dos visitantes às mensagens enviadas pelo dono do blog. Além disso, sessões de usuários também representam respostas dos visitantes às publicações de um blog. Com esse ponto de vista, fundamentado nas atividades do dono de um blog e de seus visitantes, nós podemos definir e e quantificar o nível de interação entre os vários participantes da rede social de um blog.

A figura 4.1 mostra uma série de atributos que utilizamos para caracterizar o nível de interação entre os usuários de um blog. Para analisar as ações do dono do blog, nós estudamos o intervalo de tempo entre a criação de novas publicações. Para analisar a participação dos visitantes nós caracterizamos o intervalo de tempo entre entre chegada de comentários e o intervalo de tempo entre sessões. Para mostrar a velocidade em que as publicações recebem as respostas dos visitantes nós também caracterizamos o tempo de resposta. O tempo de resposta é definido como o tempo entre a criação de uma publicação pelo dono de um blog e os vários comentários que a publicação recebe de visitantes.

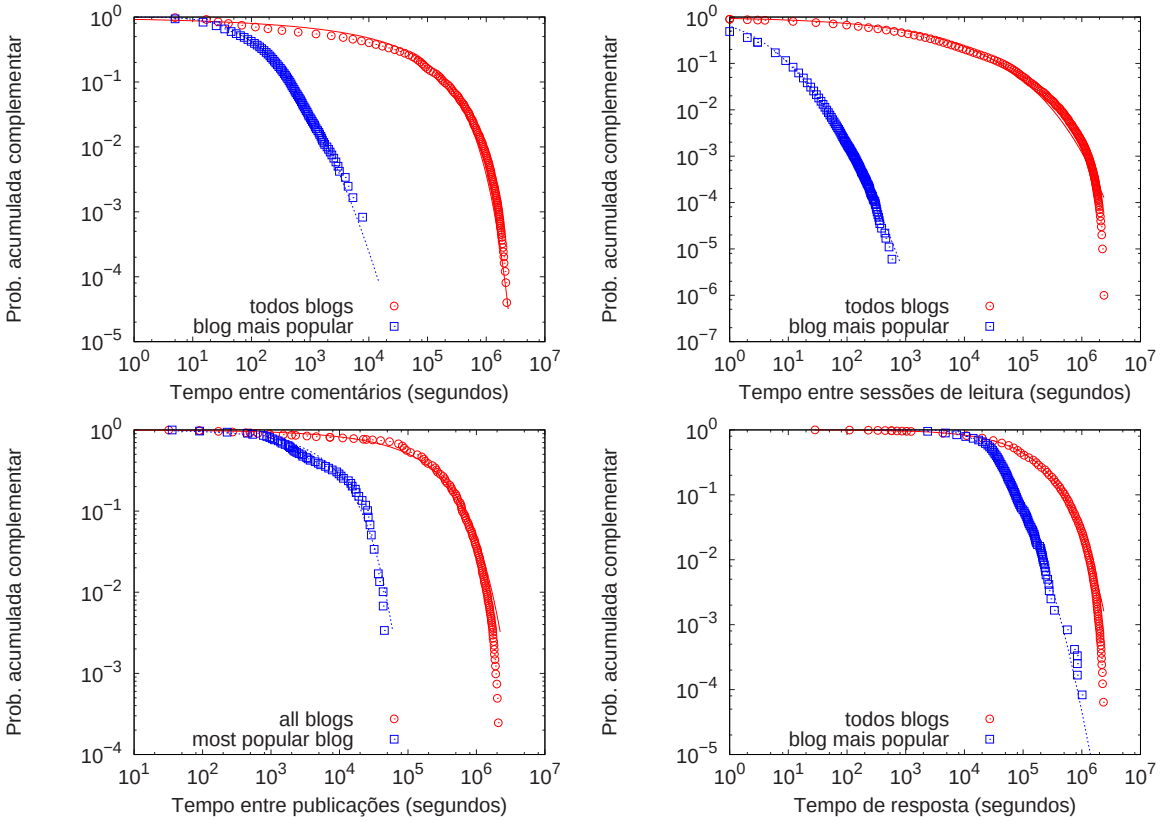

Figura 4.2: Distribuição do intervalo de tempo entre comentários (superior à esquerda), do intervalo de tempo entre sessões (superior à direita), do intervalo de tempo entre publicações (inferior à esquerda) e do tempo de resposta (inferior à direita).

Apresentamos na figura 4.2 a probabilidade acumulada complementar para os intervalos de

tempo que representam as interações entre os usuários de nossa blogosfera. Duas distribuições são mostradas em cada gráfico, uma para o blog mais popular em termos de tráfego, número de requisições de leitura recebidas, e outra para todos os blogs. Para encontrar o resultado agregado, primeiramente, os intervalos de tempo são calculados para cada um dos blogs, assim como foi feito para o blog mais popular. Em seguida, agregamos os valores de intervalo de tempo de todos os blogs e calculamos a distribuição. Para esse estudo nós consideramos apenas as publicações criadas no período de nossa carga de trabalho, pois não temos acesso aos comentários enviados e as visitas realizadas fora desse período.

A distribuição do intervalo de tempo entre a chegada de comentários é mostrada na figura 4.2 (superior à esquerda). Podemos ver, por exemplo, que a probabilidade do intervalo de tempo entre comentários ser maior do que 7 minutos, 420 segundos, é de 10% para o blog mais popular, ou seja, que, para esse blog, 90% dos intervalos são menores do que 7 minutos. Quando analisamos o resultado agregado encontramos intervalos de tempo maiores, sendo 97% dos intervalos maiores do que 7 minutos.

Percebemos na figura 4.2 (superior à direita) que o tráfego do blog mais popular é bastante intenso. Esse blog é visitado por uma nova sessão em intervalos que normalmente não passam de 10 segundos, pois 90% dos intervalos de tempo entre sessões são menores do que 10 segundos. Para o resultado agregado, 70% dos intervalos são maiores do que uma hora e poucos maiores do que 1 semana.

A figura 4.2 (inferior à esquerda) mostra que 50% dos intervalos de tempo entre publicações são maiores do que 2 dias. Verificamos que esse resultado é valido para vários blogs. Diferentemente de blogs pouco atualizados, o blog mais popular não costuma demorar mais do que 7 horas para publicar uma nova entrada, pois apenas 10% dos intervalos de tempo são maiores do que 7 horas.

Podemos analisar na figura 4.2 (inferior à direita) quanto tempo os visitantes demoram para responder as publicações. Para o blog mais popular, é interessante observar que a maioria dos comentários, aproximadamente 90% do total, são enviados no mesmo dia em que as publicações foram criadas. Para o resultado agregado, metade dos comentários foram enviados no mesmo dia da criação da publicação. Além disso, a distribuição do tempo de resposta nos informa sobre o tempo de vida das publicações, pois observamos que dificilmente comentários são enviados uma semana após a criação das publicações.

Para cada atributo que nós caracterizamos na figura 4.2, nós também mostramos as distribuições que melhor aproximam nossos dados experimentais. As curvas das distribuições são representadas nas figuras através de linhas pontilhadas ou sólidas. A tabela 4.1 apresenta as distribuições e os parâmetros que melhor representam nossos dados experimentais. Os parâmetros informados são para distribuição Lognormal dada por $(1/\sigma x\sqrt{2\pi})e^{-(log(x)-\mu)^2/2\sigma^2}$, para distribuição Gamma dada por $(1/\beta^{\alpha}\Gamma(\alpha))x^{\alpha-1}e^{(x/\beta)}$ e para distribuição Weibull dada por $\beta\alpha^{-\beta}x^{\beta-1}e^{-(x/\alpha)^{\beta}}I_{(0,\infty)(x)}$.

Atributo da interação	Todos blogs Distribuição (parâmetros)	Blog mais popular Distribuição (parâmetros)
Tempo de resposta	Weibull ($\alpha = 0.000469$, $\beta = 0.64892$)	Weibull ($\alpha = 0.000015$, $\beta = 1.04838$)
Tempo entre sessões	Weibull ($\alpha = 0.069633$, $\beta = 0.33081$)	Lognormal ($\mu = 4.310535$, $\sigma = 1.40456$)
Tempo entre publicações	Gamma ($\alpha = 0.462894$, $\beta = 528,047$)	Gamma ($\alpha = 0.642546$, $\beta = 12,624$)
Tempo entre comentários	Gamma ($\alpha = 0.208459$, $\beta = 328,572$)	Lognormal ($\mu = 4.310535$, $\sigma = 1.40456$)

Tabela 4.1: Distribuições e parâmetros que melhor representam os valores observados para os atributos das interações.

4.2 Classificação de Blogs Baseada no Tipo de Interação

Os acessos a blogosfera são influenciados pelos diálogos ou interações entre os participantes dos blogs. Uma pergunta que pode ser feita é se existem diferenças na forma de interação entre os usuários em diferentes blogs.

Podemos caracterizar as interações entre os usuários de um blog usando a intensidade em que comentários são enviados para o seu dono. A taxa de leitores que visitam um blog e comentam alguma publicação é uma medida do envolvimento dos visitantes com o blog. Um blog com uma taxa pequena de comentários por visita reflete uma comunicação ou interação em uma única direção. Nesse tipo de blog, o dono do blog escreve para seus visitantes de maneira similar a comunicação unidirecional de um editorial de jornal com os seus leitores. Por outro lado, um blog com uma alta taxa de comentários por visita apresenta uma comunicação em várias direções e, nesse caso, tanto o dono do blog quanto seus visitantes estão envolvidos na conversação.

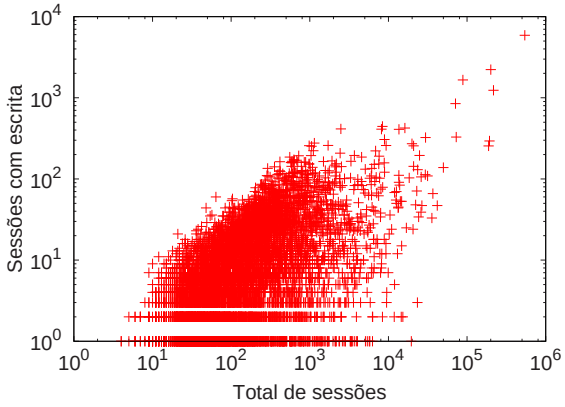

Figura 4.3: Correlação entre o número de sessões que visitam os blogs e o número de sessões que, além de visitarem os blogs, enviam pelo menos um comentário.

A figura 4.3 apresenta a correlação entre o número de sessões que visitam os blogs e o número de sessões que, além de visitarem os blogs, também comentam pelo menos uma publicação. Cada ponto representa um blog e as coordenadas refletem o número de sessões

que visitaram o blog e o número de sessões que visitaram e deixaram comentários. Percebemos uma correlação positiva entre o número de sessões que acessaram o blog e o número de sessões que escreveram comentários, ou seja, que existe uma tendência dos blogs mais populares também receberem mais sessões com envio de comentários. Calculamos e encontramos um coeficiente de correlação igual a 0,87. Contudo, o gráfico mostra que existem diferenças entre blogs que receberam a mesma quantidade de sessões visitantes. Por exemplo, entre blogs que foram acessados por cerca de 10.000 sessões, existem blogs em que apenas 2 sessões deixaram comentários e existem blogs em que mais de 1.000 sessões participaram do blog com o envio de comentários. Isso indica que existem blogs onde a conversação ou interação entre os usuários têm intensidades diferentes.

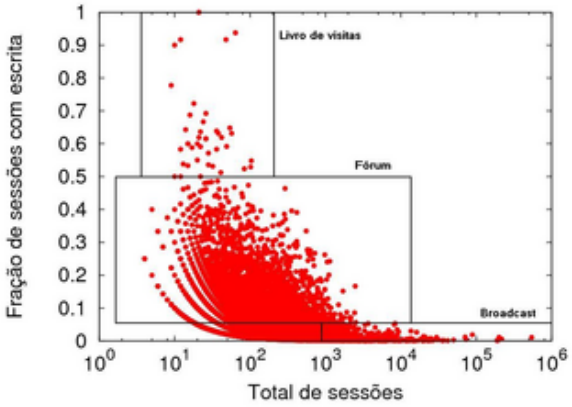

Figura 4.4: Classificação fundamentada na fração de sessões com escrita e no número de sessões visitantes: blogs classificados como do tipo broadcast, fórum ou livro de visitas.

A figura 4.4 mostra nossa metodologia de classificação de blogs fundamentada nos diferentes tipos de interações entre os usuários e os blogs. Cada ponto nessa figura continua representando um blog. Mostramos novamente no eixo x o total de sessões visitantes e agora mostramos no eixo y a fração de sessões que enviaram comentários. Percebemos que existe uma relação inversa entre a popularidade do blog e a proporção de sessões que interagem com o blog através do envio de comentários. No extremo direito da curva estão os blogs que recebem um considerável número de sessões visitantes que, entretanto, em maioria não enviam comentários. Esses são blogs semelhantes a meios de comunicação de notícias do tipo *broadcast*, onde a comunicação é em uma única direção, do dono do blog para os leitores. No outro extremo da curva estão os blogs que, apesar de não serem muito populares, recebem visitantes que em sua maioria enviam comentários quando visitam o blog. Nesses blogs, do tipo *livro de visitas*, a comunicação dos leitores com o dono do blog ocorre com maior probabilidade. Entre os dois extremos da curva estão os blogs do tipo *fórum*. Esses blogs são os que recebem uma quantidade razoável de leituras, uma quantidade significativa de escritas e neles ocorrem interações entre todos participantes dos blogs.

Com base nas observações apresentadas nesta seção, nós classificamos os blogs presentes em nossa blogosfera em categorias, de acordo com a popularidade dos blogs e a fração de sessões com escritas. Como ilustrado na figura 4.4, blogs do tipo *broadcast* são aqueles acessa-

Tipo de blog	Percentagem		
	todos blogs	**todas sessões**	**sessões com escrita**
Broadcast	7%	74%	25%
Fórum	55%	12%	63%
Livro de visitas	1%	0%	1%
Não classificados	37%	14%	11%

Tabela 4.2: Percentagem de blogs e de sessões em cada classificação de blog.

dos por mais de 1.000 sessões, em que 5% ou menos dessas sessões enviaram comentários para pelo menos uma publicação. Blogs do tipo *fórum* são aqueles em que mais de 5% e menos de 50% das sessões enviaram algum comentário e blogs do tipo *livro de visitas* são aqueles em que o número de sessões com escrita supera o número de sessões somente com leitura. A tabela 4.2 apresenta o resultado da classificação em nossa blogosfera. Por acreditarmos que não temos observações suficientes sobre blogs pouco populares, nós desconsideramos blogs que receberam menos do que 50 sessões. Vemos que blogs do tipo *broadcast* recebem a maioria das sessões visitantes, que blogs do tipo *fórum* são os mais freqüêntes e recebem a maioria das sessões com escrita e que blogs do tipo *livro de visitas* não são tão comuns na blogosfera, sendo visitados por menos de 0,5% das sessões. É importante ressaltar que, embora os valores delimitadores das classificações e a quantidade de blogs de cada tipo possam ser diferentes para outra blogosfera, as nossas observações básicas e a nossa metodologia continuam válidas e podem ser aplicadas.

Capítulo 5

Trabalhos Relacionados

Neste capítulo apresentamos estudos relacionados à nossa pesquisa. Nós discutimos não somente trabalhos da literatura sobre caracterização de cargas de servidores da Web, como também trabalhos sobre os diversos aspectos de uma blogosfera.

5.1 Sobre Caracterização de Servidores da Web

Caracterização de carga é fundamental para o entendimento e criação de sistemas para a Internet. Muitos estudos focaram na caracterização do tráfego da Internet e de cargas de trabalho de servidores da Web [13, 15, 19, 23, 28, 34, 57]. Entre as importantes contribuições desses trabalhos estão o estabelecimento de uma lei de potência para descrever a popularidade dos objetos da Web, de uma distribuição de cauda pesada para descrever o tamanho dos objetos e transferências, e da localidade espacial e temporal do fluxo de requisições. Uma discussão sobre as várias características apresentadas em estudos sobre o conteúdo tradicional da Web está fora do escopo deste trabalho. Desse modo, nas próximas seções nós somente apresentamos estudos direcionados à modelagem e caracterização de aspectos da blogosfera.

5.2 Sobre Blogs

Nesta seção, nós discutimos os principais trabalhos existentes na literatura sobre diversos aspectos da blogosfera. Acreditamos que nosso trabalho [29, 30] seja o primeiro a caracterizar uma base de dados de um serviço de blogs.

5.2.1 Redes de Blogs

Muitos trabalhos analisaram redes que representam o relacionamento entre blogs. Nessas redes, os vértices representam blogs, e as arestas representam relacionamentos explícitos ou implícitos entre os blogs. Relacionamentos explícitos são aqueles expressos através de listas de blogs favoritos criadas pelos donos dos blogs ou em citações para outros blogs inseridas nas publicações. Relacionamentos implícitos são aqueles criados pela interação entre usuários, por exemplo, conectando dois blogs se o dono de um blog comenta o outro blog.

O primeiro trabalho sobre redes de blogs foi apresentado por Kumar *et alii* [45]. Nesse trabalho os autores analisaram a evolução temporal de uma rede formada por blogs, revelando padrões de evolução com rápidas alterações e ressaltando a possibilidade de identificação automática de comunidades. Shi *et alii* [60] compararam redes de blogs extraídas de duas bases de dados, mostrando que, apesar de cobrirem conjuntos diferentes de blogs, as redes possuíam propriedades estruturais semelhantes. Outro trabalho discutiu como agrupar blogs sabendo como um dono de blog faz referências a outros blogs em suas publicações e fez uma análise da topologia de cascata que surge das seqüências de referências entre publicações [46]. O relacionamento entre blogs e sítios da Web foi investigado por Bhagat *et alii* [22], que utilizaram outras informações disponíveis nos blogs, como localização geográfica, idade, amigos e números de comunicadores instantâneos, para completarem a análise. Herring *et alii* [37] mostraram que as redes de blogs podem ser utilizadas para caracterizar o relacionamento entre blogs e para inferir conversações e comunidades.

Noor Ali-Hasan e Lada Adamic [12] analisaram a estrutura de rede social formada por três comunidades de blogs e descobriram, através de entrevistas com membros das comunidades, que poucas das interações que ocorrem por meio dos blogs refletem uma proximidade de relacionamentos na vida real e que muitos relacionamentos entre usuários surgem através do uso de blogs. Nessa direção, Furukawa *et alii* [32] analisaram vários tipos de redes de um serviço de blogs japonês e relacionaram características das redes com padrões de leitura de alguns donos de blogs. Eles mostraram que os usuários lêem outros blogs com regularidade e que as relações expressas nas redes estão correlacionadas com o padrão de leitura dos usuários.

Redes de blogs também foram construídas para analisar a propagação de informação na blogosfera. Adar *et alii* [11] propuseram um algoritmo para descobrir blogs que mais influenciam a blogosfera. Tais blogs publicam novas opiniões ou notícias que são discutidas ou comentadas em outros blogs. Em outro trabalho, Gruhl *et alii* [35] investigaram a dinâmica da propagação de informação através da identificação e rastreamento de dois tipos de publicações: com assuntos normalmente discutidos pelos donos dos blogs e com assuntos estimulados por eventos externos, como de temas atuais da mídia. Esses autores usaram modelos já conhecidos de propagação de infecções biológicas para acompanhar a difusão da discussão sobre tais assuntos. Kolari *et alii* [43] estudaram o uso de blogs em uma empresa para melhorar a colaboração e o compartilhamento de experiências entre funcionários. Eles apresentaram características da rede de blogs, analisaram o alcance de publicações e discussões na hierarquia da organização e discutem sobre os blogs mais influentes.

Um trabalho interessante seria investigar como a estrutura das redes está relacionada com nossos resultados sobre o tráfego de requisições e sobre a comunicação entre usuários. Neste trabalho não analisamos o conteúdo dos blogs e por isso não analisamos a rede da blogosfera.

5.2.2 Palavras-chaves das Publicações

Em muitos serviços de blog os usuários podem adicionar palavras-chaves a cada uma de suas publicações. As palavras-chaves devem descrever a publicação e são usadas para

máquinas de busca encontrarem conteúdo relevante, para organização de conteúdo e para sistemas de recomendação.

Nos últimos anos surgiram vários trabalhos sobre como aproveitar as palavras-chaves das publicações. Christopher Brooks e Nancy Montanez [25] analisaram a efetividade do uso de palavras-chaves para classificação de publicações. Nesse trabalho, os autores coletaram as 350 palavras-chaves mais populares do sítio da Web Technorati e mediram a similaridade entre publicações que compartilhavam uma mesma palavra-chave. Eles descobriram que palavras-chaves são úteis para classificar publicações em categorias, porém são menos eficientes para indicar o conteúdo das publicações. Nessa direção, Bettina Berendt e Christoph Hanser [21] argumentaram que as palavras-chaves são utilizadas mais como complemento do assunto da publicação do que como um sumário. Eles sugeriram que, para refletir melhor o conteúdo da publicação, informações disponíveis no texto da publicações, como substantivos, devem ser combinados com as palavras-chaves. Conor Hayes e Paolo Avesani [36] argumentaram que palavras-chaves não são eficientes para separar os blogs em grupos, porém, após separar os blogs usando técnicas tradicionais de agrupamento, eles mostram que o uso de palavras-chaves pode servir para identificar os blogs mais relevantes dentro de cada agrupamento.

Um dos problemas do uso de palavras-chaves é que seu uso é arbitrário e os usuários podem não ter experiência suficiente para escolher boas palavras-chaves. Para tentar amenizar esse problema, Sood *et alii* [62] propuseram um sistema que, além de permitir que os usuários indiquem as palavras-chaves, automaticamente sugere palavras-chaves para novas publicações fundamentado nas palavras-chaves das publicações já existentes.

5.2.3 Opinião e Sentimento Expressos nas Publicações

Vários estudos apresentaram técnicas para identificar a opinião dos donos dos blogs sobre um determinado tema, como opinião positiva, negativa ou neutra, e mesmo para determinar o sentimento expresso pelo dono do blog nas publicações, como raiva, alegria ou tristeza.

A linguagem utilizada nas publicações dos blogs é normalmente a referência para a investigação da personalidade dos donos dos blogs. Um estudo comparativo entre usuários dos sexos masculino e feminino foi feito para blogs de adolescentes [38], para comentários enviados para blogs [40] e também para usuários de diferentes idades [59]. Utilizando uma pequena base de dados, com cerca de 100 usuários, Scott Nowson e Jon Oberlander [56] apresentaram uma metodologia para classificação da personalidade de donos dos blogs a partir dos textos publicados nos blogs. Em uma continuação desse estudo, os autores apresentaram resultados para uma base de dados maior, porém obtiveram uma menor taxa de acerto em suas inferências [54]. Comparando técnicas mais sofisticadas de lingüística, Benamara *et alii* [20] sugeriram o uso de advérbios e adjetivos, e não só adjetivos, para análises de sentimento.

Os textos das publicações dos donos dos blogs também foram utilizados para extrair opinião sobre determinados temas. Yang *et alii* [65] apresentaram uma metodologia dividida em duas etapas: recuperação de publicações sobre um tópico seguida da classificação da opinião expressa nas publicações. Através da análise do texto de blogs e de sítios de notícias, Godbole *et alii* [33] relacionaram uma opinião positiva ou negativa dos autores às entidades,

tais como pessoas, lugares e acontecimentos. Nessa direção, Andreevskaia *et alii* [14] propuseram e analisaram o desempenho de um sistema que indica automaticamente o sentimento positivo, negativo ou neutro expresso em sentenças dos textos.

Através da combinação de análise de texto e de informações obtidas em fontes externas, como do serviço para encontrar produtos do sítio de vendas da Amazon, Gilad Mishne e Maarten de Rijke [51] apresentaram um método para analisar blogs e recomendar livros para os donos dos blogs. Em outro trabalho, Gilad Mishne e Natalie Glance [53] argumentaram que a opinião positiva expressa em blogs sobre filmes que estão para ser lançados pode ser um bom indicador de um futuro sucesso desses filmes.

Alguns serviços de blogs, como o LiveJournal [8], permitem que donos de blogs, além de adicionar palavras-chaves às publicações, adicionem o humor que estejam no momento da escrita. O serviço de blogs LiveJournal permite que os usuários escolham entre 132 estados, tais como feliz ou nervoso. Muitos donos de blogs utilizam essa opção e muitas publicações com a indicação do humor são criadas todos os dias. Gilad Mishne *et alii* [4, 50] criaram a ferramenta MoodViews para rastrear e analisar o humor das publicações criadas no serviço de blogs LiveJournal, fornecendo, por exemplo, os estados de humor mais freqüentes entre todas publicações. Krisztian Balog e Maarten de Rijke [18] argumentaram que o humor mais freqüentemente associado a um tópico não necessariamente é o humor mais apropriado para o tópico, dado que o humor representa mais o estado do dono do blog do que o significado do conteúdo. Um outro trabalho usou técnicas de análise de texto para tentar adivinhar o humor escolhido nas publicações [49], enquanto que outros investigaram eventos que provocam variações no nível de utilização de estados de humor [16, 17].

5.2.4 Comentários Enviados por Visitantes

Neste trabalho, nós apresentamos uma análise mais elaborada dos comentários do que a apresentada por Gilad Mishne e Natalie Glance [52] em um estudo sobre um pequeno conjunto de 724 blogs de diferentes serviços de blogs. Nesse trabalho, os autores argumentaram que há correlação entre popularidade, representada pelo grau dos vértices de uma rede de blogs ou pelo número de visualizações de página obtida em contadores de acesso, com o número de comentários enviados para os blogs. Eles também notaram a presença de pontos diferentes, como blogs populares que recebem uma quantidade pequena de comentários, e atribuem essas diferenças à moderação do dono dos blogs. Como nossa carga de trabalho contém todos os comentários enviados por visitantes, em oposição ao uso apenas dos comentários que apareceram na página dos blogs, tornou-se possível argumentar que a presença de blogs muito populares com poucos comentários não é uma conseqüência da moderação do dono do blog, porém uma característica das interações que ocorrem em blogs do tipo *broadcast*. Além disso, nós mostramos que existem exceções na correlação entre popularidade e número de comentários mesmo para blogs pouco populares, que recebem uma quantidade bem maior de comentários do que a esperada em blogs com interações do tipo *fórum* ou *livro de visitas*.

5.2.5 Outros Aspectos

Uma das questões da blogosfera é como ter conhecimento da numerosa quantidade de publicações disponibilizadas todos os dias. Uma estratégia é criar uma base de dados centralizada e atualizá-la a cada nova publicação, informando qual o blog atualizado e passando informações básicas sobre o conteúdo da publicação. Já existem sítios na Web que fornecem esse tipo de serviço e alguns serviços de blogs que oferecem a seus usuários a possibilidade de informar esses serviços centralizados a cada nova publicação. Esses serviços mostram quais os blogs mais freqüentemente atualizados, exibem as publicações mais recentes e utilizam essas informações em suas máquinas de busca. Pranam Kolari *et alii* [44] caracterizaram o uso indevido desse serviço por usuários que forjam atualizações somente para aumentar a publicidade de seus blogs e enganarem máquinas de busca. Em um estudo mais recente, esse mesmo grupo de pesquisa publicou um trabalho sobre como filtrar essas mensagens com falsas atualizações fundamentado na assinatura das ferramentas mais comumente utilizadas para enviar tais mensagens automaticamente [42]. O serviço de blogs que analisamos não possui essa facilidade e por isso não investigamos esse aspecto da blogosfera neste trabalho.

Uma outra forma de facilitar a descoberta de novas publicações é através da representação do conteúdo dos blogs em documentos estruturados. Tais documentos, comumente conhecidos como *feeds*, existem para cada um dos blogs e são textos estruturados de forma padronizada que contêm no mínimo as publicações mais recentes do blog e a data e título de cada publicação. Os visitantes podem utilizar leitores de *feeds* e adicionar os blogs que mais lhe interessem. Quando um usuário adiciona blogs, o leitor de *feeds* copia e armazena o texto estruturado de cada blog. Nas próximas vezes que os usuários acionarem o leitor, o programa irá novamente requisitar o *feed* de cada blog adicionado e verificar se houve alguma mudança, isto é, se houve alguma nova publicação. Dessa forma, os visitantes são informados de diversas atualizações sem terem que visitar manualmente todos os blogs. Akshay Java *et alii* [39] analisaram a lista de blogs lidos através do *feed* de vários usuários de um serviço de blogs. Eles propuseram um método para encontrar a lista de *feeds* mais interessantes para determinados categorias, como esportes ou política. Como a lista de blogs adicionados a um leitor de *feeds* pode ser muito grande, um usuário pode ainda assim ter trabalho em encontrar publicações que sejam de seu interesse. Para amenizar esse problema, Ka Cheung Sia *et alii* [61] propuseram um sistema que auxilia o usuário a organizar seus *feeds* fundamentado em seu padrão de leitura de publicações e de navegação pela Web. Os blogs que estudamos possuem *feeds*, porém não tivemos acesso as requisições feitas pelos leitores de *feeds*. Entretanto, sempre que um usuário que usa *feeds* quiser ler o conteúdo de uma publicação, ele deve fazer uma requisição de leitura, e nós temos esse tipo de requisição em nossa carga de trabalho. De qualquer forma, seria interessante saber como os usuários exploram essa facilidade e quais os blogs mais lidos através dos *feeds*.

Edith Cohen e Balachander Krishnamurthy [26] argumentaram que blogs provêem um paradigma de comunicação diferente do existente em sítios da Web. Mostramos em nosso trabalho que, diferentemente do acesso a sítios estáticos da Web, o acesso aos objetos da

blogosfera pode ser visto como parte de interações entre os donos e os leitores dos blogs. Eles analisaram um conjunto de blogs populares de um serviço de blogs e mostraram que a taxa de mudança em blogs é diferente da taxa de mudança em outros sítios. Além disso, eles apresentaram uma heurística simples para inferir se um sítio é um blog ou não e argumentam que acompanhar um conjunto de blogs pode ser útil para identificar interesses emergentes ou diálogos entre os participantes da blogosfera.

Neste trabalho, focamos no impacto do tráfego e no estudo de padrões de comunicação, em oposto a uma visão de alto nível, tais como a de uma análise da difusão de informação na blogosfera ou da evolução da estrutura de rede entre os blogs.

Capítulo 6

Conclusão

Nesse trabalho, utilizamos uma significativa carga de trabalho para caracterizar os padrões de acesso à blogosfera sob três diferentes pontos de vista: dos usuários, dos blogs e dos servidores. Fornecemos modelos estatísticos para várias características, como popularidade dos blogs, intervalos de tempo entre comentários e distribuição de tamanho de arquivos transferidos, úteis para o projeto de novos serviços de blogs e para planejamento de capacidade. Nossas distribuições de probabilidade podem ser usadas para geração de cargas de trabalho sintéticas e para encontrar a infra-estrutura que proporcione uma melhor qualidade de serviço, como um menor tempo de resposta no atendimento às requisições e um maior período de disponibilidade dos servidores. Como mostramos que existe uma concentração de acessos em poucos blogs, pode ser interessante para o serviço de blogs reservar recursos para os blogs mais populares em tráfego e explorar mecanismos de *caching* que favoreçam objetos desses blogs. As distribuições identificadas para a carga de trabalho de blogs podem também ser usadas para construir modelos de desempenho para arquiteturas Web que atendam o trafego da blogosfera.

Nossos resultados indicam que as máquinas de busca não capturam as propriedades sociais e temporais da blogosfera e em geral não direcionam os usuários para os blogs mais populares em tráfego. Verificamos, por exemplo, que os robôs das máquinas de busca raramente coletam os comentários dos blogs e que, portanto, não estão coletando informações sobre os diálogos entre os usuários da blogosfera.

Encontramos que o acesso aos blogs é influenciado pela publicidade do blog em sítios da Web e pelas interações sociais entre os participantes da blogosfera. Mostramos que, diferentemente dos acessos aos serviços estáticos da Web, os acessos aos objetos da blogosfera são influenciados pelas interações entre os donos e os leitores dos blogs. Fundamentado nos diferentes tipos de interações entre os participantes da blogosfera, propusemos uma classificação que separa os blogs em três grupos: broadcast, fórum e livro de visitas.

Sugerimos três tópicos como direções para trabalhos futuros. A primeira idéia seria a elaboração de algoritmos de recuperação de informação que levem em consideração as características sociais e temporais da blogosfera, em oposição a somente a estrutura de rede entre os blogs. A segunda idéia seria um estudo do comportamento dos usuários ao longo do tempo, que analisasse a freqüência de leitura e escrita dos usuários e pudesse agrupá-los em classes de

maneira similar ao que propusemos para os blogs. Para isso seria necessária uma identidade permanente do usuário, uma informação não disponível nas cargas de trabalho que utilizamos. Finalmente, uma outra idéia seria o estudo do impacto da utilização de *feeds* na blogosfera. Seria interessante saber como os usuários exploram essa facilidade e quais os blogs mais lidos através dos *feeds*.

Referências Bibliográficas

[1] Blog da Rebecca Blood. `http://www.rebeccablood.com/`.

[2] Blog do Jorn Barger. `http://www.robotwisdom.com`.

[3] Blog do Justin Hall. `http://www.links.net`.

[4] MoodViews: ferramentas para análise de humor em blogs. `http://moodviews.com`.

[5] Robô da máquina de busca Google. `http://www.google.com/bot.html`.

[6] Robô da máquina de busca Yahoo. `http://help.yahoo.com/help/us/ysearch/slurp`.

[7] Serviço de blogs da UOL. `http://blog.uol.com.br`.

[8] Serviço de blogs LiveJournal. `http://www.livejournal.com`.

[9] Sítio da Web Technorati. `http://www.technorati.com`.

[10] L. Adamic. Zipf, Power-law, Pareto - A Ranking Tutorial. `http://www.hpl.hp.com/research/idl/papers/ranking`.

[11] E. Adar, L. Zhang, L. Adamic e R. Lukose. Implicit Structure and the Dynamics of Blogspace. *Workshop on the Weblogging Ecosystem, International World Wide Web Conference*, maio 2004.

[12] N. Ali-Hasan e L. Adamic. Expressing Social Relationships on the Blog through Links and Comments. *International Conference on Weblogs and Social Media*, março 2007.

[13] V. Almeida, A. Bestavros, M. Crovella e A. Oliveira. Characterizing Reference Locality in the WWW. *Conference on Parallel and Distributed Information Systems*, dezembro 1996.

[14] A. Andreevskaia, S. Bergler e M. Urseanu. All Blogs Are Not Made Equal: Exploring Genre Differences in Sentiment Tagging of Blogs. *International Conference on Weblogs and Social Media*, março 2007.

[15] M. Arlitt e C. Williamson. Web Server Workload Characteristics: The Search for Invariants. *IEEE/ACM Transactions on Networking*, 5(5), 1997.

[16] K. Balog, G. Mishne e M. Rijke. Why Are They Excited? Identifying and Explaining Spikes in Blog Mood Levels. *Meeting of the European Chapter of the Association for Computational Linguistics*, abril 2006.

[17] K. Balog e M. Rijke. Decomposing Bloggers' Moods: Towards a Time Series Analysis of Moods in the Blogosphere. *Workshop on the Weblogging Ecosystem, International World Wide Web Conference*, maio 2006.

[18] K. Balog e M. Rijke. How to Overcome Tiredness: Estimating Topic-Mood Associations. *International Conference on Weblogs and Social Media*, março 2007.

[19] P. Barford, A. Bestavros, A. Bradley e M. Crovella. Changes in Web Client Access Patterns: Characteristics and Caching Implications. *World Wide Web, Special Issue on Characterization and Performance Evaluation*, 2(1):15–28, 1999.

[20] F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato e V. Subrahmanian. Sentiment Analysis: Adjectives and Adverbs are Better than Adjectives Alone. *International Conference on Weblogs and Social Media*, março 2007.

[21] B. Berendt e C. Hanser. Tags are not Metadata, but Just More Content - to Some People. *International Conference on Weblogs and Social Media*, março 2007.

[22] S. Bhagat, G. Cormode, S. Muthukrishnan, I. Rozenbaum e H. Xue. No Blog is an Island - Analyzing Connections Across Information Networks. *International Conference on Weblogs and Social Media*, março 2007.

[23] L. Breslau, P. Cao, L. Fan, G. Phillips e S. Shenker. Web Caching and Zipf-like Distributions: Evidence and Implications. *INFOCOM*, abril 1999.

[24] S. Brin e L. Page. The Anatomy of a Large-scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 33, 1998.

[25] C. Brooks e N. Montanez. Improved Annotation of the Blogopshere via Autotagging and Hierarchical Clustering. *International Conference on World Wide Web*, maio 2006.

[26] E. Cohen e B. Krishnamurthy. A Short Walk in the Blogistan. *Computer Networks*, 50(5):615–630, 2006.

[27] M. Crovella e A. Bestavros. Self-similarity in World Wide Web Traffic: Evidence and Possible Causes. *IEEE/ACM Transactions on Networking*, 5(6), 1997.

[28] C. Cunha, A. Bestavros e M. Crovella. Characteristics of WWW Client-based Traces. Technical Report BU-CS-95-010, Computer Science Department, Boston University, abril 1995.

[29] F. Duarte, B. Mattos, A. Bestavros, V. Almeida e J. Almeida. Traffic Characteristics and Communication Patterns in Blogosphere. Technical Report 2006-033, Computer Science Department, Boston University, dezembro 2006.

[30] F. Duarte, B. Mattos, A. Bestavros, V. Almeida e J. Almeida. Traffic Characteristics and Communication Patterns in Blogosphere. *International Conference on Weblogs and Social Media*, março 2007.

[31] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach e T. Berners-Lee. Hypertext Transfer Protocol – HTTP/1.1. IETF RFC 2616.

[32] T. Furukawa, Y. Matsuo, I. Ohmukai, K. Uchiyama e M. Ishizuka. Social Networks and Reading Behavior in the Blogosphere. *International Conference on Weblogs and Social Media*, março 2007.

[33] N. Godbole, M. Srinivasaiah e S. Skiena. Large-Scale Sentiment Analysis for News and Blogs. *International Conference on Weblogs and Social Media*, março 2007.

[34] S. Gribble e E. Brewer. System Design Issues for Internet Middleware Services: Deductions from a Large Client Trace. *Symposium on Internet Technologies and Systems*, dezembro 1997.

[35] D. Gruhl, R. Guha, D. Liben-Nowell e A. Tomkins. Information Diffusion Through Blogspace. *International World Wide Web Conference*, pages 491–501. ACM Press, 2004.

[36] C. Hayes e P. Avesani. Using Tags and Clustering to Identify Topic-Relevant Blogs. *International Conference on Weblogs and Social Media*, março 2007.

[37] S. Herring, I. Kouper, J. Paolillo, L. Scheidt, M. Tyworth, P. Welsch, E. Wright e N. Yu. Conversations in the Blogosphere: An Analysis From the Bottom Up. *Hawaii International Conference on System Sciences*, 2005.

[38] D. Huffaker e S. Calvert. Gender, Identity e Language Use in Teenage Blogs. *Journal of Computer-Mediated Communication*, 10(2), 2005.

[39] A. Java, P. Kolari, T. Finin, A. Joshi e T. Oates. Feeds That Matter: A Study of Bloglines Subscriptions. *International Conference on Weblogs and Social Media*, março 2007.

[40] T. Kennedy, J. Robinson e K. Trammell. Does Gender Matter? Examining Conversations in the Blogosphere. *Internet Research 6.0: Internet Generations*, outubro 2005.

[41] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[42] P. Kolari, T. Finin, A. Java e A. Joshi. Towards Spam Detection at Ping Servers. *International Conference on Weblogs and Social Media*, março 2007.

[43] P. Kolari, T. Finin, K. Lyons, Ye. Yesha, S. Perelgut Ya. Yesha e J. Hawkins. On the Structure, Properties and Utility of Internal Corporate Blogs. *International Conference on Weblogs and Social Media*, março 2007.

[44] P. Kolari, A. Java e T. Finin. Characterizing the Splogosphere. *Workshop on the Weblogging Ecosystem, International World Wide Web Conference*, maio 2006.

[45] R. Kumar, J. Novak, P. Raghavan e A. Tomkins. On the Bursty Evolution of Blogspace. *International World Wide Web Conference*, pages 568–576. ACM Press, 2003.

[46] M. McGlohon, J. Leskovec, C. Faloutsos, M. Hurst e N. Glance. Finding Patterns in Blog Shapes and Blog Evolution. *International Conference on Weblogs and Social Media*, março 2007.

[47] D. Menascé, V. Almeida, R. Riedi, F. Ribeiro, R. Fonseca e W. Meira. A Hierarchical and Multiscale Approach to Analyze E-business Workloads. *Performance Evaluation*, 54(1):33–57, 2003.

[48] F. Menczer, S. Fortunato, A. Flammini e A. Vespignani. Googlearchy or Googlocracy? *IEEE Spectrum Online*, fevereiro, 1999.

[49] G. Mishne. Experiments with Mood Classification in Blog Posts. *Workshop on Stylistic Analysis Of Text For Information Access*, 2005.

[50] G. Mishne, K. Balog, M. Rijke e B. Ernsting. MoodViews: Tracking and Searching Mood-Annotated Blog Posts. *International Conference on Weblogs and Social Media*, março 2007.

[51] G. Mishne e de M. Rijke. Deriving Wishlists from Blogs: Show us Your Blog e We'll Tell you What Books to Buy. *International Conference on World Wide Web*, maio 2006.

[52] G. Mishne e N. Glance. Leave a Reply: An Analysis of Weblog Comments. *Workshop on the Weblogging Ecosystem, International World Wide Web Conference*, maio 2006.

[53] G. Mishne e N. Glance. Predicting Movie Sales from Blogger Sentiment. *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[54] S. Nowson e J. Oberlander. Identifying more Bloggers: Towards Large Scale Personality Classification of Personal Weblogs. *International Conference on Weblogs and Social Media*, março 2007.

[55] A. Ntoulas, J. Cho e C. Olston. What's New on the Web? The Evolution of the Web from a Search Engine Perspective. *International World Wide Web Conference*, pages 1–12. ACM Press, 2004.

[56] J. Oberlander e S. Nowson. Whose Thumb is it Anyway? Classifying Author Personality from Weblog Text. *International Conference on Computational Linguistics*, julho 2006.

[57] V. Paxson e S. Floyd. Wide-Area Traffic: The Failure of Poisson Modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, 1995.

[58] P. Pirolli e J. Pitkow. Distributions of Surfers' Paths Through the World Wide Web: Empirical Characterizations. *World Wide Web*, 2(1-2):29–45, 1999.

[59] J. Schler, Moshe Koppel, S. Argamon e J. Pennebaker. Effects of Age and Gender on Blogging. *AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, 2006.

[60] X. Shi, B. Tseng e L. Adamic. Looking at the Blogosphere Topology through Different Lenses. *International Conference on Weblogs and Social Media*, março 2007.

[61] K. Sia, J. Cho, K. Hino, Y. Chi, S. Zhu e B. Tseng. Monitoring RSS Feeds Based on User Browsing Pattern. *International Conference on Weblogs and Social Media*, março 2007.

[62] S. Sood, S. Owsley, K. Hammond e L. Birnbaum. TagAssist: Automatic Tag Suggestion for Blog Posts. *International Conference on Weblogs and Social Media*, março 2007.

[63] E. Veloso, V. Almeida, W. Meira, A. Bestavros e S. Jin. A Hierarchical Characterization of a Live Streaming Media Workload. *IEEE/ACM Transactions on Networking*, 14(1):133–146, 2006.

[64] A. Williams, M. Arlitt, C. Williamson e K. Barker. Web Workload Characterization: Ten Years Later. *Web Content Delivery*. Springer, 2005.

[65] K. Yang, N. Yu, A. Valerio, H. Zhang e Weimao Ke. Fusion Approach to Finding Opinions in Blogosphere. *International Conference on Weblogs and Social Media*, março 2007.