Caracterização Hierárquica do Tráfego e Padrões de Comunicação de um Serviço de Blogs *

Fernando Duarte O. Castro (aluno), Virgílio Augusto F. Almeida (orientador)

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais

{fernando,virgilio}@dcc.ufmg.br

Abstract. *We present a thorough characterization of the access patterns in a blogosphere. Our characterization of over 32 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. We observe that access in blogosphere could be conceived as part of an interaction between an author and its readership, and we classify blogs into three groups: "broadcast-type", "registration-type", and "parlor-type". We characterize the main features of the blogosphere workload, and we investigate the differences between it and typical web server workloads.*

Resumo. *Neste trabalho apresentamos uma caracterização dos padrões de acesso a um serviço de blogs. Nossa caracterização de mais de 32 milhões de requisições de leitura, de escrita e administrativas, enviadas em um período de 28 dias, foi feita sob três diferentes pontos de vista. 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. Consideramos os acessos à blogosfera como parte de interações entre os donos e os leitores dos blogs, e classificamos os blogs em três grupos: broadcast, livro de visitas e fórum. Apresentamos as principais características da carga de trabalho de uma blogosfera e as comparamos com características de típicas cargas de trabalho de servidores da Web.*

1. Introdução

Com o surgimento de serviços de blogs, que facilitam tanto a criação quanto a atualização de blogs, esse novo modelo de disponibilização de conteúdo passou a receber cada vez mais adeptos. Blogs são sítios da Web que possuem a aparência de diários pessoais, onde as opiniões dos autores estão bem delimitadas e separadas em publicações. Um conjunto de blogs com interações sociais entre seus usuários forma uma blogosfera. Os serviços de blogs oferecem ferramentas para os usuários publicarem informações de maneira simples e, além disso, armazenam e disponibilizam os blogs na Web. Uma característica diferenciadora dos blogs é a possibilidade dos leitores enviarem comentários para as publicações, o que incentiva tanto comentários de outros leitores quanto a criação de novas publicações. Blogs se modificam através da adição de novas publicações e comentários, exibem as datas de criação de cada publicação e comentário, e organizam o conteúdo do mais recente para o mais antigo.

*A versão completa da dissertação está disponível em www.dcc.ufmg.br/~fernando/ctd. Nossos resultados foram publicados na conferência ICWSM e nosso trabalho foi indicado ao prêmio de melhor artigo [3]. Esta pesquisa recebeu apoio do UOL (www.uol.com.br), através do Programa UOL Bolsa Pesquisa, processo número 20060520221328a.

Neste trabalho, estudamos o tráfego e os padrões de comunicação entre usuários do serviço de blogs do UOL [1]. Dada a relevância e o crescimento da blogosfera, é natural questionar se seus padrões de acesso são similares aos de serviços já existentes na Web, e é importante prever o impacto do uso dos blogs nos servidores do serviço. Estudos sobre padrões de acesso ao conteúdo tradicional da Web descobriram propriedades fundamentais para explicar o tráfego [7], para construção de modelos de carga de trabalho e para geração de cargas de trabalho sintéticas [6]. Nós 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 [2] ou da evolução da estrutura de rede entre os blogs [4].

Principais Contribuições: Utilizando uma carga de trabalho do serviço de blogs do UOL, analisamos como os usuários lêem os blogs, como enviam comentários e como os donos atualizam seus blogs. Abaixo apresentamos os principais resultados da caracterização do tráfego e da comunicação da blogosfera:

- 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.

- 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. Verificamos que blogs mais freqüentemente atualizados não necessariamente recebem mais visitantes.

- 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. Mostramos que o tráfego possui alta variabilidade ao longo do tempo, com picos de acesso em diversos momentos causados pelo assunto, qualidade dos comentários e quantidade de acessos de outros blogs.

- As distribuições de popularidade dos blogs seguem uma lei de potência para diversas métricas: 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 a maioria dos arquivos transferidos são menores do que 12 KB.

- 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.

- Mostramos que 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. Embora os blogs mais populares recebam mais comentários, muitos visitantes desses blogs somente lêem as publicações e não enviam comentários.

- 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.

Tabela 1. Sumário da carga de trabalho

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

2. Descrição da Carga de Trabalho

Nós analisamos três cargas de trabalho anonimizadas do serviço de blogs do UOL: as requisições de leitura, os comentários enviados e as atividades administrativas. 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 horário e data em que a requisição foi recebida. Na carga de leituras, o campo *requisição* contém o objeto requisitado para leitura. 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. O campo *tamanho* indica a quantidade de bytes transferidos. O campo *origem* mostra a URL de onde se originou a requisição do visitante. O último campo, *agente*, identifica o navegador e o sistema operacional utilizado para enviar a requisição. A tabela 1 apresenta um sumário da carga de trabalho. 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.

3. Caracterização do Tráfego

A caracterização do tráfego do serviço de blogs foi feita de forma hierárquica, utilizando três diferentes pontos de vista: de como os usuários acessam a blogosfera, de como os blogs são acessados e de como é o tráfego agregado nos servidores.

3.1. Caracterização ao Nível de Usuários

Nesta seção, focamos no estudo dos usuários da blogosfera, em como eles utilizam o serviço de blogs através de requisições de leitura, de escrita e administrativas.

Definição e criação de sessões: Para analisar como os usuários utilizam o serviço de blogs, agrupamos as requisições em sessões [6]. Identificamos unicamente um usuário através dos campos *máquina* e *agente* e definimos uma sessão como o intervalo de tempo em que um usuário está ativamente acessando a blogosfera. Uma sessão se inicia com a primeira requisição enviada pelo usuário e termina quando o tempo desde a última requisição na sessão ultrapassa um valor limite de 30 minutos. Encontramos que quase 7 milhões de sessões, representando mais de 4 milhões de usuários, visitaram a blogosfera.

Origem das sessões: Para investigar como os usuários chegam à blogosfera, analisamos quantas sessões utilizam máquinas de busca ou sítios externos para acessar os blogs. A

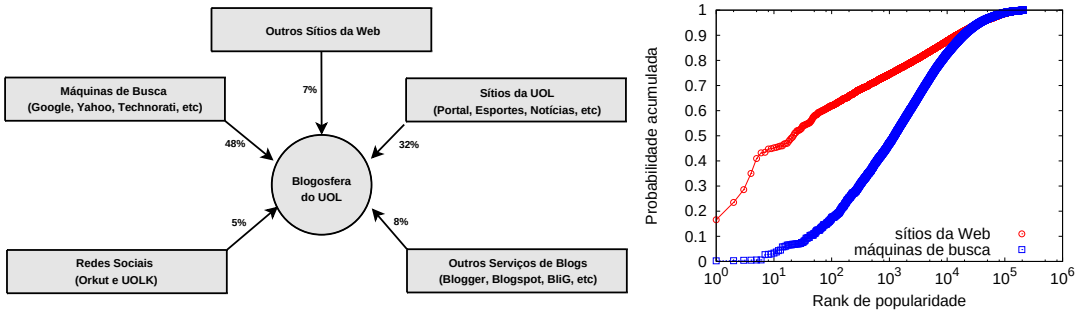

Figura 1. Diferentes formas de acesso à blogosfera (esquerda). Acessos através de máquina de busca ou de sítios da Web aos blogs populares (direita).

Figura 1 apresenta o resultado do estudo do campo *origem* da primeira requisição de cada sessão da carga de trabalho de leituras. Observamos 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 rede da Web, poderíamos esperar que blogs populares atraíssem uma desproporcional fração das sessões iniciadas através de máquinas de busca. Para verificar se isso ocorre em nosso serviço de blogs, mostramos na Figura 1 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, 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 [5] na blogosfera.

Acessos dos usuários: Durante o nosso período de observação, cada usuário poderia acessar a blogosfera várias vezes, seja para visitar, se expressar com o envio de comentários, ou administrar blogs. Para caracterizar a intensidade de interesse dos usuários pela blogosfera, examinamos a distribuição da quantidade de acessos dos usuários em número de requisições de leitura, escrita e administrativas, do usuário mais ativo ao usuário menos ativo. As distribuições das atividades dos usuários seguem uma lei de potência com expoente $\alpha = 0,83$ para a quantidade de requisições de leitura, $\alpha = 0,54$ para a quantidade de requisições de escrita e $\alpha = 0,53$ para a quantidade de requisições administrativas, todas as três regressões lineares com $R^2 = 0,99$.

3.2. 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.

Padrão temporal do acesso aos blogs: A Figura 2 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 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.

Variabilidade na intensidade dos acessos: Na blogosfera, a popularidade de um blog ao longo do tempo varia em função do conteúdo das publicações e dos comentários, da

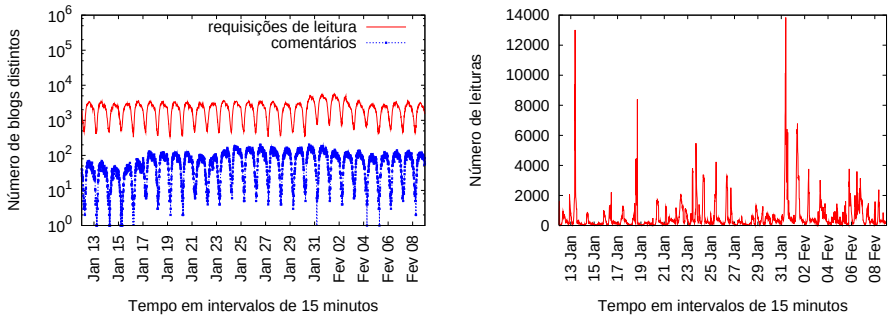

Figura 2. Comportamento periódico dos acessos à blogosfera (esquerda). Variabilidade no acesso ao blog mais popular (direita)

quantidade de referências de sítios populares, e do renome do dono do blog. Para ilustrar a intensa variabilidade na popularidade dos blogs, a Figura 2 apresenta as requisições de leitura do blog mais popular em tráfego de nossa carga de trabalho. 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 uma clara diferenciação entre dias úteis e fins de semana. 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. 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, a qualidade dos comentários e a quantidade de acessos vindos de outros sítios é que geram os picos de acessos.

Popularidade dos Blogs: A Figura 3 mostra a popularidade usando diferentes métricas: 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 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 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 infraestrutura do serviço de blogs, para alocar recursos para os blogs mais populares e tratar os blogs mais populares de forma diferenciada em um mecanismo de *caching*.

Impacto da Atividade do Administrador na Popularidade: Verificamos que o nível de atividade do dono influencia pouco na popularidade do blog. Calculamos a correlação entre o total de sessões visitantes e o total de atividades administrativas e encontramos um valor baixo de 0,26 para o coeficiente de correlação.

3.3. Caracterização ao Nível de Servidores

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

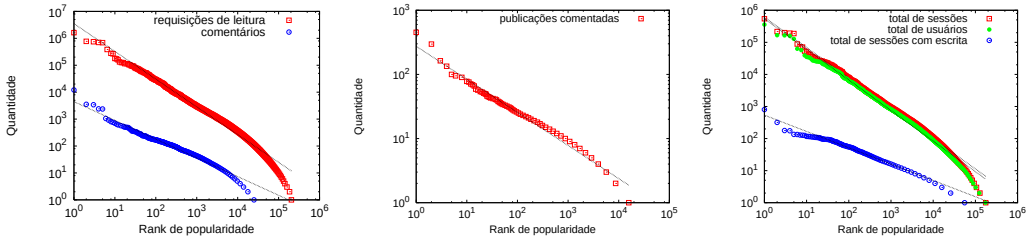

Figura 3. Popularidade dos blogs: leituras e comentários (esquerda), publicações comentadas (centro), sessões e usuários (direita).

Distribuição de Tamanho das Transferências de Arquivos: Para analisar o tamanho dos arquivos transferidos pelos visitantes, utilizamos o campo *tamanho* das requisições. Blogs não possuem objetos grandes, sendo que a mediana do tamanho dos arquivos transferidos é de 12 KB. 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 foi transferido dos servidores em 4 semanas. A distribuição acumulada complementar do tamanho das transferências possui uma cauda pesada melhor aproximada por uma distribuição Pareto com expoente $\kappa \approx 1$. Este resultado é similar ao encontrado em estudos sobre o tráfego de servidores Web [7]. Embora não freqüentes, 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 de vídeo e arquivos de áudio.

Padrão Temporal do Tráfego de Requisições: Analisando o tráfego de requisições ao longo do tempo, observamos que o acesso agregado de todos usuários para todos os blogs é periódico. Assim como na Figura 2, sobre os acessos ao nível de blogs, o tráfego medido em número de bytes, requisições de leitura e de escritas possui maior intensidade durante o dia e menor intensidade durante a noite. 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, encontramos que a quantidade de comentários enviados foi aproximadamente duas ordens de grandeza menor do que a quantidade de requisições de leitura. Também encontramos 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, e existem períodos de quinze minutos em que o tráfego passa de 10 mil requisições de leitura para 100 mil requisições de leitura. Argumentamos que essa variabilidade no tráfego ocorre como uma conseqüência das interações sociais entre os membros da blogosfera.

4. Padrões de Comunicação

Nesta seção estudamos a interação entre os participantes da blogosfera: os donos dos blogs e seus visitantes.

Interações entre os Participantes da Blogosfera: Uma das características mais marcantes da blogosfera são as interações entre os usuários através de publicações e comentários. A Figura 4 mostra os atributos que utilizamos para caracterizar o nível de interação entre usuários. Para analisar as ações do dono do blog, estudamos o intervalo de tempo entre a criação de novas publicações. Para analisar a participação dos visitantes 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, caracterizamos o tempo de resposta. Calculamos a probabilidade acumulada complementar para cada atributo e, para exemplificar, apresentamos na Figura 5 a CCDF para os intervalos de tempo que representam quanto tempo os visitantes demoram para responder as publicações. Duas distribuições são mostradas no gráfico, uma

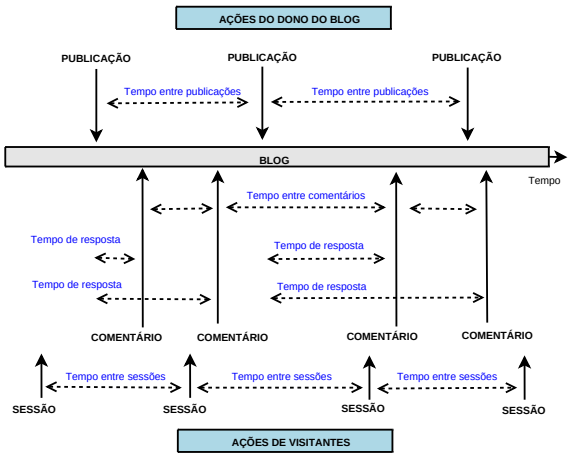

Figura 4. Estrutura das interações entre os participantes da blogosfera.

para o blog mais popular em termos de número de leituras e outra para todos blogs. Para o blog mais popular, aproximadamente 90% dos comentários 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. As curvas indicam que dificilmente comentários são enviados uma semana após a criação das publicações. As distribuições que melhor aproximam nossos dados experimentais e resultados para os outros intervalos de tempo podem ser encontrados na versão completa da dissertação.

Classificação de Blogs Baseada no Tipo de Interação: Os acessos a blogosfera são influenciados pelas 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 usuários de um blog usando a intensidade em que comentários são enviados para seu dono. Encontramos uma correlação de 0,87 entre o número de sessões visitantes e o número de sessões que, além de visitarem os blogs, também comentam pelo menos uma publicação. Entretanto, encontramos diferenças entre blogs que receberam a mesma quantidade de sessões visitantes. Por exemplo, entre blogs 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 deixaram comentários. Isso indica que existem blogs onde a conversação entre os usuários tem intensidade diferente.

A Figura 5 mostra nossa metodologia de classificação de blogs fundamentada nos diferentes tipos de interações entre os usuários. Cada ponto nessa figura representa um blog, mostramos no eixo X o total de sessões visitantes e no eixo Y a fração de sessões que enviaram comentários. Percebemos uma relação inversa entre a popularidade do blog e a proporção de sessões que interagem 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. Como resultado da classificação em nossa blogosfera, encontramos que blogs do tipo *broadcast* recebem 74% das sessões visitantes, que 55% dos blogs são do tipo *fórum* e recebem 63% 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

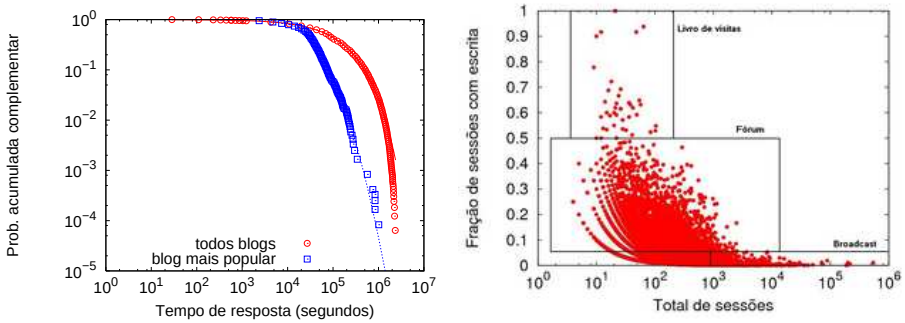

Figura 5. Distribuição do tempo de resposta (esquerda). Classificação de blogs fundamentada nas interações sociais entre usuários (direita).

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.

5. 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 úteis para o projeto de novos serviços de blogs e para planejamento de capacidade, como 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*. 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. Fundamentados 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.

Referências

[1] Serviço de blogs do UOL. `http://blog.uol.com.br`.

[2] E. Adar, L. Zhang, L. Adamic, and R. Lukose. Implicit Structure and the Dynamics of Blogspace. In *Workshop on the Weblogging Ecosystem*, maio 2004.

[3] F. Duarte, B. Mattos, A. Bestavros, V. Almeida, and J. Almeida. Traffic Characteristics and Communication Patterns in Blogosphere. In *Intl. Conf. on Weblogs and Social Media*, março 2007.

[4] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the Bursty Evolution of Blogspace. In *WWW*, pages 568–576. ACM Press, 2003.

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

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

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