Caracterização e Modelagem do Tráfego e da Navegação dos Usuários do Apontador

Theo Lins[1], **Helen Costa**[1], **Fabrício Benevenuto**[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Ouro Preto (UFOP)
Campus Universitário Morro do Cruzeiro
Ouro Preto, MG, Brasil

{theosl, helen.c.s.costa, benevenuto}@gmail.com

Abstract. *Recently, there has been a large popularization of location-based social networks, such as FourSquare and Gowalla, where users can share locations and upload tips and information about locations on the system. Part of this popularity is due to facility access to the internet through mobile devices with GPS. Despite the innumerous efforts towards understanding characteristics of these systems, little is known about the access pattern of users in these systems. Providers of this kind of services need to deal with different challenges that could benefit of such understanding, such as content storage, performance and scalability of servers, personalization and service differentiation for users. This paper aims at characterizing and modeling the patterns of requests that reach a server of a location-based social network. To do that, we use a dataset obtained from Apontador, a Brazilian system with characteristics similar to FourSquare and Gowalla, where users share information about their locations and can navigate on existent system locations. As results, we identified models that describe unique characteristics of the user sessions on this kind of system, patterns in which requests arrive on the server as well as the access profile of users in the system.*

Resumo. *Recentemente, tem ocorrido uma grande popularização das redes sociais baseadas em geo-localização, como o FourSquare e o Gowalla, onde usuários podem compartilhar suas localizações e adicionar informações e dicas sobre localizações do sistema. Parte dessa popularidade é devida a facilidade de acesso à internet através de dispositivos móveis dotados de GPS. Apesar de inúmeros esforços no sentido de entender as características desses sistemas, pouco se sabe sobre o padrão de acesso dos usuários desses sistemas. Provedores desse tipo de serviço lidam com diferentes desafios que poderiam se beneficiar desse entendimento, tais como armazenamento de conteúdo, desempenho e escalabilidade dos servidores, personalização e diferenciação de serviços para os usuários. Este trabalho visa caracterizar e modelar os padrões das requisições que chegam ao servidor da rede social baseada em geo-localização. Para isso, utilizamos um conjunto de dados obtidos junto ao Apontador, um sistema brasileiro com características semelhantes à do FourSquare e Gowalla, onde usuários compartilham informações sobre localizações e podem navegar por essas localizações. Como resultados, foram identificados modelos que descrevem características únicas das sessões de usuários, padrões com os quais*

requisições chegam ao servidor, além do perfil de acesso de usuários no sistema.

1. Introdução

Redes sociais baseadas em geolocalização são um novo tipo de sistema da Web 2.0 que vem atraindo cada vez mais novos usuários. Redes como Foursquare e Gowalla permitem que o usuário compartilhe a sua localização geográfica com sua rede social (check-in) através de algum tipo de smartphone dotado de GPS. Esses sistemas vem crescendo exponencialmente em termos de popularidade, principalmente devido à popularidade dos dispositivos móveis com acesso a internet, por onde usuários podem facilmente compartilhar suas localizações com seus amigos.

No Brasil, um sistema com as principais características de uma rede social baseda em geolocalização é o Apontador[1]. O Apontador permite que usuários busquem por localizações, cadastrem localizações e postem dicas e comentários sobre localizações existentes. Devido a grande popularidade desses sistemas, vários esforços surgiram na tentativa de caracterizar e entender esses sistemas [Scellato 2011, Vasconcelos et al. 2012a], identificando tipos de usuários [Vasconcelos et al. 2012b], aspectos topológicos do grafo social [Noulas et al. 2011b], bem como a mobilidade dos usuários desses sistemas [Scellato et al. 2011].

Apesar de extremamente úteis, todos os esforços apontados acima são calcados em dados coletados dos sistemas Web e não abordam questões relacionadas ao tráfego interno percebido pelos servidores desses sistemas. Este trabalho visa preencher essa lacuna, buscando caracterizar e identificar modelos que descrevem o padrão de acesso dos usuários que utilizam tais serviços.

A caracterização e modelagem desse tipo de dados é importante por várias razões. Primeiro, os provedores desses serviços enfrentam diversos tipos de problemas que dependem de entendimento da carga nos servidores para serem abordados. Como exemplos podemos citar problemas relacionados a armazenamento de conteúdo, desempenho e escalabilidade dos servidores, personalização e diferenciação de serviços para os usuários, detecção de conteúdo ilegal, etc. Segundo, modelos que descrevem como é a carga nesses servidores permite a reprodução dessa carga de forma sintética [Benevenuto et al. 2009, Benevenuto et al. 2012], de forma a viabilizar a construção de simuladores e sistemas de testes de redes sociais baseadas em localização. Finalmente, entender como as características de sistemas sociais podem ser úteis para a elaboração da infra-estrutura da Internet e de sistemas de distribuição de conteúdo no futuro [Pujol et al. 2010, Lins et al. 2012].

Para realizar tal estudo, utilizamos um conjunto de dados obtidos junto ao Apontador, um sistema brasileiro com características semelhantes à do FourSquare e Gowalla. Nossa base de dados consiste em um *stream* de cliques dos usuários que navegam pelo sistema e contém um total de 62.192.459 requisições, coletadas ao longo de um mês de log. Como resultados da caracterização desse log, foram identificados modelos que descrevem características únicas das sessões de usuários, padrões com os quais requisições chegam ao servidor, além do perfil de acesso de usuários no sistema. A seguir sumarizamos nossas

[1] www.apontador.com.br.

principais descobertas:

- As distribuições de popularidade de acesso a locais seguem distribuições de cauda longa.
- Uma sessão de usuários típica de sistemas de busca de local dura cerca de 10 minutos, correspondendo a um valor bem similar ao que foi obtido para sessões de sistemas tradicionais da *Web*.
- Os *rankings* de atividades dos usuários em termos do número de requisições enviadas ao sistema, seguem distribuições de cauda longa e logarítmica, já em termos do número de sessões criadas no sistema seguem distribuições de cauda longa e exponencial.
- A taxa de chegada de requisições ao sistema apresenta um padrão periódico, com maior intensidade de acessos durante o dia e menor intensidade durante a noite, e com maior intensidade durante os dias da semana e com menor intensidade durante os finais de semana e feriados.
- As distribuições do tempo entre chegadas de requisições e do tempo entre chegada de sessões ao sistema podem ser modeladas por distribuições de cauda longa.
- A origem da maior parte das requisições e dos usuários que acessam o serviço de busca locais do Apontador são de outros *sites* de buscas.
- Nossas análises revelam a necessidade de um sistema de recomendação eficiente que mantenha a navegação dos usuários no próprio sistema.

O restante do trabalho está organizado da seguinte forma. A seção 2 descreve os trabalhos relacionados. A seção 3 apresenta as estatísticas sobre a carga de trabalho do Apontador. A seção 4 explica a caracterização das requisições, sessões dos usuários e também uma análise das distâncias entre os locais nas sessões. Finalmente, a seção 5 oferece conclusões e direções para trabalhos futuros.

2. Trabalhos Relacionados

O processo de caracterização de carga é importante para o entendimento e aprimoramento de sistemas Web. Há vários estudos que apresentam caracterizações de carga de trabalho de diferentes tipos, tais como servidores *Web* [Arlitt and Williamson 1996], de comércio eletrônico [Menasce and Almeida 2000,Pereira et al. 2006], de blogs [Duarte et al. 2007], de vídeo sob demanda [Costa et al. 2004] e de compartilhamento de vídeo [Benevenuto et al. 2010]. Dentre as várias contribuições desses trabalhos, destacamos a criação de valiosos modelos capazes de descrever a carga que chega nesses servidores, essenciais para a geração de carga sintética que, por sua vez, possibilita a realização de experimentação e simulação baseadas em distribuições realistas. Em nosso trabalho, apresentamos uma caracterização da carga de um servidor de busca local do ponto de vista do servidor.

No contexto das redes sociais, Benevenuto e colaboradores [Benevenuto et al. 2009] utilizaram dados de cliques de usuários do Orkut de forma a caracterizar a navegação e as formas de interação dos usuários nesses sistemas. De forma semelhante, Schneider e colaboradores [Schneider et al. 2009] apresentaram um estudo da navegação dos usuários no Facebook. Em um estudo mais recente, Benevenuto e colaboradores [Benevenuto et al. 2012] mediram a distância física e topológica das interações entre os usuários do Orkut, mostrando que o conteúdo nesses sistemas é em sua maioria produzido e consumido localmente.

Existem vários trabalhos que caracterizam diferentes aspectos de sistemas de buscas locais. Em [Scellato 2011], os autores analisam três redes sociais de busca local e discutem como os usuários on-line são afetados de forma heterogênea pela distância geográfica. Em [Vasconcelos et al. 2012a], os autores apresentam uma caracterização de como os usuários interagem entre si utilizando *tips* e *dones*, através da coleta de seus perfis do Foursquare. *Tips* são dicas sobre um determinado local e podem ser marcadas como *dones* se um usuário concorda com seu conteúdo. Noulas e colaboradores [Noulas et al. 2011a] utilizaram um algoritmo de clusterização espectral para agrupar os usuários baseado nos padrões de check-ins. Baseados nos atributos das regiões e usuários de duas cidades metropolitanas, puderam identificar grupos de usuários que visitam categorias similares de lugares e caracterizar o tipo de atividade que acontece em cada região da cidade.

Diferentemente de todos esses esforços, nosso trabalho visa caracterizar e entender como as requisições chegam a um servidor de busca local, mas também investigar e identificar as distâncias entre os locais navegados pelos usuários.

3. Carga de Trabalho

Em nosso estudo, analisamos a carga de trabalho do *site* Apontador[2]. O Apontador é um *site* de busca local brasileiro que possui uma base georreferenciada com aproximadamente sete milhões de locais. Cada local possui uma página no *site* onde são apresentadas informações, tais como: o nome, endereço, latitude, longitude, categoria e telefone do local. Os usuários que acessam estas informações podem fazer isto de forma anônima ou registrada (logados). Além de procurar e visualizar as informações destes locais, os usuários também podem recomendar, avaliar, inserir fotos e cadastrar novos locais. No entanto, para que um usuário possa cadastrar um novo local, avaliar um existente ou associar uma foto ao local, é preciso estar logado ao *site*. Os mesmos locais disponíveis no *site* também estão disponíveis nas aplicações para dispositivos móveis das plataformas iPhone, Android ou BlackBerry. Nestas aplicações, um usuário cadastrado pode fazer check-in num lugar, tirar uma foto e associá-la ao lugar.

Descrição	**Únicos**	**Requisições**
Usuários Logados	41.084	482.195
Usuários Não Logados	51.839.306	61.710.264
Total	51.880.390	62.192.459
Locais acessados	**2.584.028**	**27.050.742**

Tabela 1. Características da base de dados do Apontador

Os logs utilizados correspondem ao período de um mês, de 01/09/2011 a 30/09/2011, contabilizando um total de **62.192.459** de requisições, vindas de 51.880.390 de usuários diferentes. Cada registro da carga de trabalho representa uma requisição enviada por um usuário ao Apontador. As seguintes informações estão disponíveis para cada requisição: *timestamp*, *usuário*, *objeto*, *tipo* e *local*. O campo timestamp é o momento em que a requisição foi recebida pelo servidor. O campo usuário corresponde a um identificador do *cookie* do usuário que gerou a requisição. O objeto é o código único para

[2]http://www.apontador.com.br

identificar a requisição. O campo tipo são as ações que uma pessoa pode realizar em um local. O campo local é o local solicitado na requisição pelo usuário.

Grupo	**Tipo de Ação**	**# Requisições**	Porcentagem
Visit	Acessar a página de um local	54.704.630	87,96 %
Phone	Clicar no telefone do local	6.763.124	10,87 %
Site	Clicar no botão ir para o site do local	529.387	0,85 %
Outros	demais ações do usuário	195.318	0,32 %

Tabela 2. Tipos de Ações

São várias as ações que uma pessoa pode realizar em um local e que são monitoradas pelo sistema de log. Estas ações são: acessar a página de um local (*visit*); clicar no telefone do local (*phone*)[3]; clicar no botão "recomendo" do local (*thumbs up*); clicar no botão "não recomendo" do local (*thumbs down*); clicar no botão ir para o site do local (*site*); fazer o upload de uma foto relacionada com o local (*send photo*); clicar no link que compartilha o local no Facebook (*facebook*); clicar no link que compartilha o local no Orkut (*orkut*); clicar no link que compartilha o local no Twitter (*twitter*); clicar no e-mail do local (*email*), e; quando a pessoa solicita o widget com o mapa do local (*widget*). Além das ações descritas acima, existem outras ações que são monitoradas quando o local é patrocinado. Estas ações são: quando a pessoa solicita a impressão de um cupom promocional (*focus coupon*); quando a pessoa visualiza o telefone do local (*focus phone*), e; quando a pessoa visualiza o e-mail do local (*focus email*).

3.1. Coleta de Locais

Os dados com cliques dos usuários obtidos junto ao apontador contém apenas o identificador dos locais armazenados no sistema. Sendo assim, informações como endereço, geolocalização e categoria do local não estão disponíveis nos logs dos servidores do Apontador. Entretanto, a partir do identificador do local é possível coletar tais informações através da API do Apontador[4].

Para realizar tal coleta desenvolvemos um coletor em Python que recuperou as informações de todos os locais disponíveis em nossa base de cliques dos usuários. A seguir, apresentamos as características dos locais coletados.

Descrição	**Únicos**
Locais acessados	2.584.028
Locais coletados em XML com sucesso	2.579.320
Locais em XML com erros	567
Locais coletados válidos	**2.578.753**

Tabela 3. Características da Base de Dados Coletada

A Tabela 3 apresenta as características da base de dados coletada. No total, foi possível recuperar informações de 99,8% dos locais únicos acessados. Cada local no

[3]Propositadamente o número do telefone do local é parcialmente ocultado. Para que a pessoa possa visualizar o número completo do telefone ela precisa clicar no número.

[4]http://api.apontador.com.br/pt/

formato XML possui as seguintes informações: identificação única, nome, descrição, contador de *clicks*, número de avaliações, número de recomendações, categoria do local, endereço, telefone, latitude, longitude, endereço do *site* do local e informações do usuário criador do local.

Através do campo endereço, conseguimos listar os Estados mais frequentes dos locais únicos acessados no período de um mês, conforme a Tabela 4. Observamos que três dos seis Estados mais frequentes pertencem à região sudeste do país e os outros três à região sul.

Estado	**Número de Locais Únicos**
São Paulo	790.580
Minas Gerais	271.628
Rio de Janeiro	245.360
Rio Grande do Sul	211.131
Paraná	189.480
Santa Catarina	140.713
Bahia	113.254
Ceará	72.005
Outros	544.602

Tabela 4. Estados com maior número de locais acessados

O campo categoria identifica qual é o tipo de estabelecimento ou serviço oferecido pelo local. A Tabela 5 mostra as categorias mais frequentes dos locais únicos acessados no período de um mês.

Categoria	**Número de Locais Únicos**
Endereços Empresariais	369.312
Automóveis e Veículos	130.479
Médicos e Consultórios	129.641
Construção	121.854
Restaurantes	100.427
Serviços Gerais	91.438
Lojas Diversas	83.153
Confeccções e Vestuário	80.115
Alimentos	77.866
Escolas	77.283
Beleza e Estética	66.969
Móveis e Decoração	65.247

Tabela 5. Categorias Mais Frequentes

4. Caracterização da Carga de Trabalho

Nesta seção, apresentamos uma caracterização da carga de trabalho do Apontador sob diferentes perspectivas, mostrando vários aspectos e distribuições.

4.1. Popularidade dos Locais

Primeiramente avaliamos a popularidade dos locais, com o objetivo de verificar se a mesma segue uma lei de potência. Leis de potência estabelecem a seguinte relação: $P(E_n) \propto n^{-\alpha}$, onde $P(E_n)$ é a probabilidade de referência ao n-ésimo elemento mais popular. Para verificar a acurácia dos modelos propostos, medimos o fator R^2 da regressão linear [Trivedi 2002] para cada distribuição analisada. Em todos os modelos apresentados no trabalho, os valores de R^2 estão acima de 0,96. Sendo que quando o valor de R^2 é igual a 1 significa que não há diferenças entre o modelo e a carga de trabalho real.

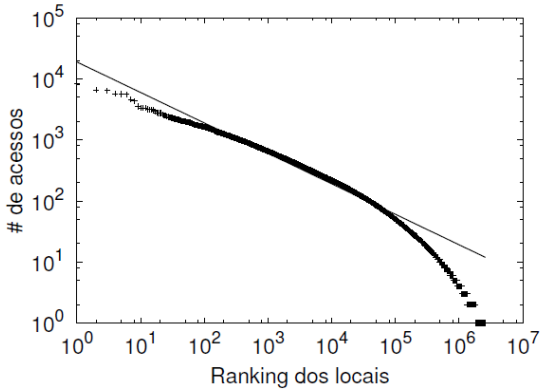


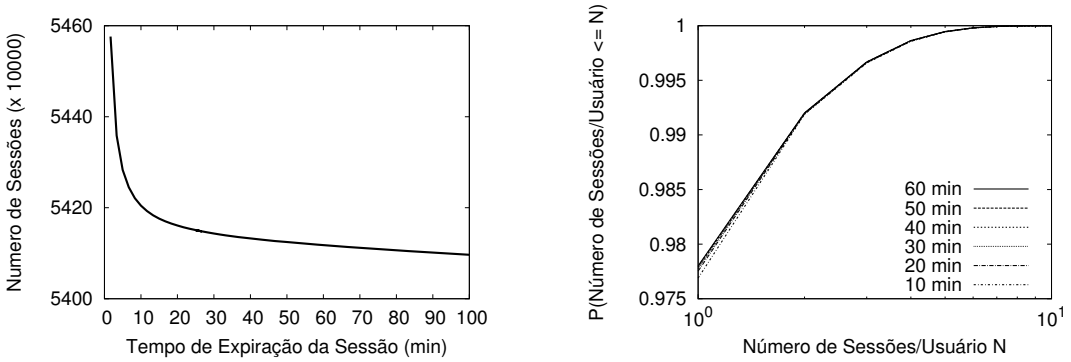
Figura 1. Popularidade dos Locais

A seguir analisamos se as distribuições de popularidade dos locais seguem leis de potência. A figura 1 mostra o *ranking* dos locais ordenados de forma decrescente pelo número de acessos. Podemos notar que existe uma pequena quantidade de locais com muitos acessos e uma grande quantidade locais com poucos. Por exemplo o primeiro local possui 8.273 acessos. Tal observação é importante pois mostra o grande potencial para *caching* de locais que o sistema possui. De fato, essa distribuição é bem modelada com uma distribuição que segue uma lei de potência, com $\alpha = 0,498$, e $R^2 = 0,96$.

4.2. Definição de Sessões

Uma sessão de um usuário é definida como um série de requisições realizadas pelo usuário a um *site* durante um determinado período de tempo [Menascé et al. 1999, Arlitt 2000]. Em ambientes de busca de locais, uma sessão de usuário inclui acesso ao local, acesso ao *site*, acesso ao telefone e todas as ações citadas na sessão 3. Tais tipos de requisições diferem bastante das sessões de usuários de *sites* convencionais, os quais não dispõem do mesmo grau de interação dos usuários de sistemas da *Web 2.0*.

A determinação do início e término de uma sessão em aplicações de busca de locais requer uma análise específica dos tempos entre requisições a fim de medir a inatividade do usuário, uma vez que a maioria das sessões não apresenta um registro explícito de operações de *login* e *logout*. Portanto, é necessário realizar uma análise para identificar um valor limite de tempo entre requisições para que sejam consideradas como sendo de uma mesma sessão. Assim, duas requisições consecutivas são consideradas da mesma sessão se o tempo entre elas é menor do que esse limite, denominado tempo de expiração da sessão.



(a) Tempo de Expiração das Sessões (b) Número de Sessões por Usuário-CDF

Figura 2. Definição de Sessões

É importante escolher um tempo de expiração adequado para não gerarmos sessões que não representam o uso do serviço pelos usuários, evitando unir diferentes momentos de uso do serviço ou fragmentar uma navegação realizada pelo usuário. Seguindo a metodologia proposta em [Menascé et al. 1999], realizamos uma avaliação do tempo de expiração da sessão mais adequado para nossa aplicação.

A figura 2(a) apresenta o número total de sessões para diferentes valores de tempo de expiração. Um valor extremamente pequeno (ex., 1 minuto) resulta em um volume de sessões extremamente alto (mais de 54 milhões de sessões), gerando praticamente somente sessões com uma requisição. À medida que o valor do tempo de expiração aumenta, o número de sessões reduz continuamente, até que essa diminuição se estabiliza. Essa estabilidade ocorre por volta dos 10 minutos, indicando que esse valor é um limite adequado para ser adotado como tempo de expiração da sessão.

A fim de testar essa hipótese geramos a distribuição de probabilidade acumulada (CDF) do número de sessões por usuário para vários valores de tempo de expiração de sessão, conforme ilustra a figura 2(b). A diferença entre as distribuições para os diferentes valores de tempo de expiração é maior para os valores menores, tornando-se mais consistente a partir de 10 minutos. Sendo assim, adotamos 10 minutos como tempo de expiração das sessões para nossas análises, obtendo um total de 54.204.556 sessões de usuários em nossa carga de trabalho.

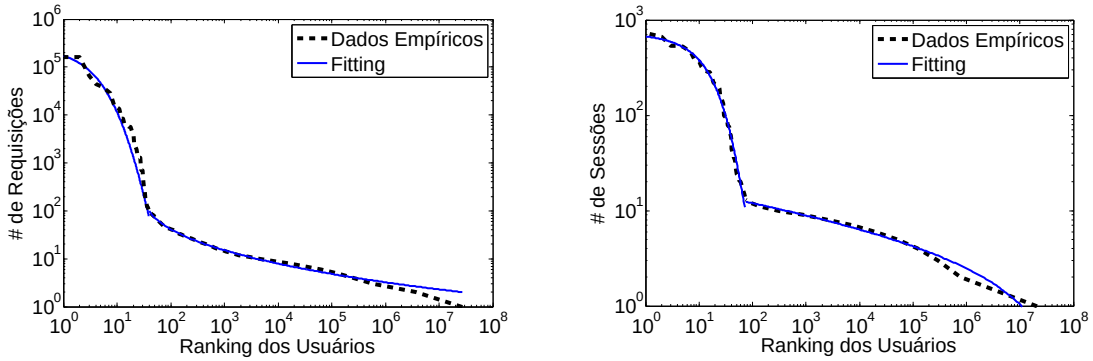
É interessante observar que esse resultado é similar as análises realizadas no trabalho [Arlitt 2000, Oke and Bunt 2002], que mostram resultados que caracterizam sessões em *sites* Web tradicionais, o valor de tempo de expiração da sessão aqui obtido também é de 10 minutos, como foi tipicamente observado nesses estudos. Esse tamanho curto de sessão, semelhante ao da Web 1.0, pode indicar que os usuários do Apontador permanecem por períodos curtos de tempo navegando no Apontador. De fato, outros sistemas Web como o YouTube e o UOL Mais possuem tempos de expiração de sessão de cerca de 40 minutos [Benevenuto et al. 2010, Gill et al. 2008].

4.3. Nível de Atividade dos Usuários

A seguir analisamos o nível de atividade dos usuários. Sabemos que usuários podem acessar o serviço de busca local repetidas vezes dentro da mesma sessão ou retornar ao sistema constantemente, gerando um grande número de sessões. Sendo assim, para modelarmos

o nível de atividade dos usuários, caracterizamos o *ranking* dos usuários em termos do número de requisições enviadas e em termos do número de sessões criadas no sistema. Chamamos de usuário cada endereço *IP* anonimizado da carga de trabalho.

A figura 3(a) mostra o *ranking* dos usuários ordenados de acordo com o número de requisições enviadas ao servidor. Podemos notar que existe uma pequena quantidade de usuários que fazem muitas requisições ao servidor e uma grande quantidade de usuários que fazem poucas requisições. Mas essa queda tende a estabilizar quando o número de requisições é menor que 100, com isso foram necessárias duas funções para obtermos uma modelagem que represente bem a distribuição. Sendo a primeira função que vai até 100 requisições segue a lei de potência do tipo $f(x) = \alpha(x+b)^n$ com $\alpha = 2,5109$ x 10^{14}, $b = 18,39$, $n = -7,115$ e $R^2 = 0,962$. E a segunda função com os valores menores que 100 requisições segue uma distribuição logarítmica do tipo $g(x) = \alpha(\log(x+b))^c$ com $\alpha = 1074$, $b = -21,024$, $c = -2,211$ e $R^2 = 0,988$.



(a) *Ranking* Usuários x Número de Requisições

(b) *Ranking* Usuários x Número de Sessões

Figura 3. Nível de Atividade dos Usuários

Em termos das sessões criadas no servidor, também foram utilizadas duas funções para modelar a distribuição sendo que a análise dos dados com o número de sessões maiores que 13 mostrou que a primeira função de distribuição exponencial do tipo $f(x) = \alpha e(bx)+c$ com $\alpha = 709,7$, $b = -0,063$, $c = 3,393$ e $R^2 = 0,988$ é a que melhor modela esses dados. A segunda função do *ranking* de sessões é melhor aproximado (com $R^2 = 0,9620$) por uma distribuição que segue a lei de potência do tipo $g(x) = \alpha x^n + c$ com $\alpha = 26,052$, $c = -5,802$ e $n = -0,083$. Esse resultado enfatiza o comportamento de que poucos usuários possuem muitas sessões, enquanto muitos possuem poucas sessões.

4.4. Padrões Temporais do Acesso

Nesta seção analisamos o número de requisições que chegam ao servidor ao longo do tempo. A figura 4 mostra o número de requisições que chega ao servidor em intervalos de uma hora. A curva apresenta um padrão periódico, com maior intensidade de acessos durante o dia e menor intensidade durante a noite, similar ao descrito em estudos sobre servidores tradicionais da *Web* [Veloso et al. 2006, Arlitt and Williamson 1996]. Podemos notar que durante os finais de semanas e nos feriados, como por exemplo, o feriado de 7 de setembro ocorrem quedas de acesso ao sistema. Como analisado os picos que normalmente passam de 250.000 requisições em 1 hora, em dias de semana, nos finais de

semana e feriados ficam em torno de 100.000 requisições em 1 hora, uma queda de mais de 50%.

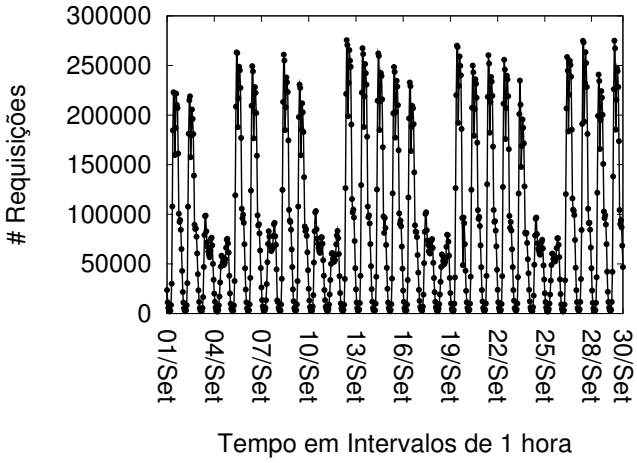
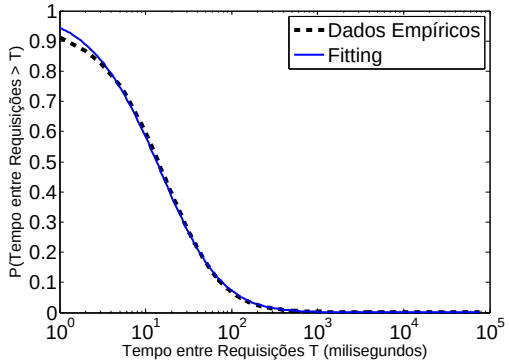
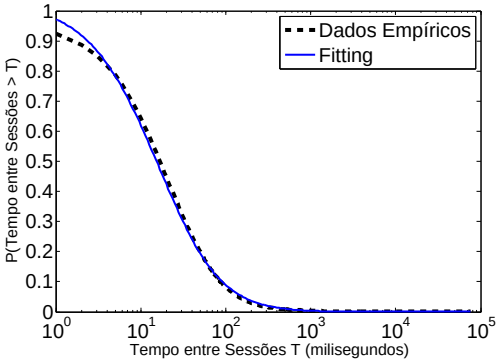


Figura 4. Número de Requisições em intervalos de 1h

Para analisarmos a participação dos usuários visitantes do sistema, caracterizamos o intervalo de tempo entre chegadas de requisições e sessões ao sistema. Apresentamos nas figuras 5(a) e 5(b) a probabilidade acumulada complementar (CCDF) para os intervalos de tempo entre requisições e sessões, respectivamente. Podemos notar que a probabilidade do intervalo de tempo entre requisições ser maior do que 500 milisegundos é menor do que 1%, sendo que 62% das requisições chegam ao servidor com intervalos menores do que 100 milisegundo. Da mesma forma, cerca de 99% dos intervalos entre sessões são menores do que 1 segundo.



(a) Intervalos de tempo entre requisições - CCDF



(b) Intervalos de tempo entre sessões - CCDF

Figura 5. Padrões Temporais do Acesso

As duas distribuições são melhores aproximadas por uma distribuição que segue a lei de potência do tipo $f(x) = \alpha(x + b)^n$. Para a distribuição do tempo entre requisições obtivemos um $\alpha = 202,52$, $b = 25,203$ e $n = -1,6436$ com $R^2 = 0,9992$, e para a distribuição do tempo entre sessões encontramos um $\alpha = 144,8$, $b = 25,335$ e $n = -1,5296$ com $R^2 = 0,9977$.

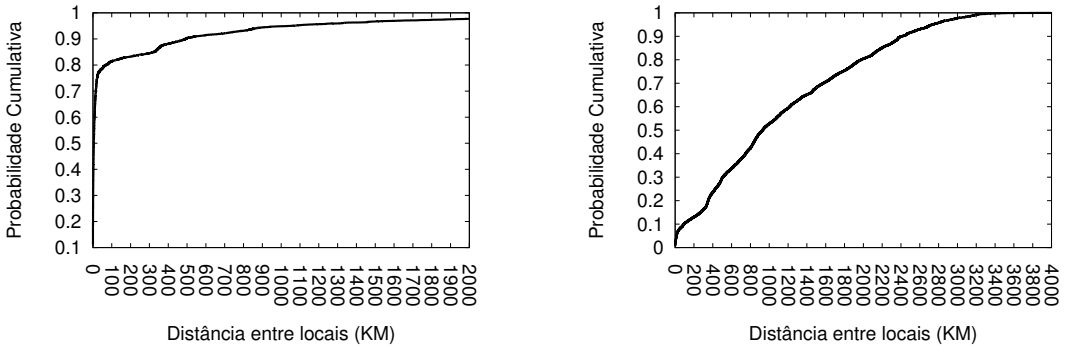
4.5. Distância entre Locais

A seguir, nós investigamos as distâncias entre locais que aparecem de forma consecutiva em requisições de usuários. Para medir a distância entre cada par de locais em uma sessão, reconstruímos as sessões dos usuários logados e dos não logados e consideramos apenas as sessões com mais de 2 requisições para locais diferentes. Em seguida, computamos a distância entre cada par de localização utilizando os dados da latitute e longitude desses locais. As Figuras 6 mostram as distribuições de probabilidade acumulada das distâncias entre locais dos usuários logados e dos não logados.

Podemos observar conforme a Figura 6(b) referente aos usuários não logados que cerca de 50% das distâncias entre os locais visitados durante uma sessão estão até 1000 KM e as demais 50% acima. Em contrapartida, a Figura 6(a) referente aos usuários logados possui as distâncias entre locais bem menores, cerca de 70% estão até 15 KM.

Este fato ocorre porque o usuário não logado é redirecionado para o *site* do Apontador através de outro buscador. Por exemplo, se ele está procurando um local chamado "Minerais Itaguaçu" , o buscador retorna vários *links* de locais no *site* do Apontador com este nome, porém um fica localizado no Estado de São Paulo, outro no Rio de Janeiro e assim por diante. E para saber qual é o local realmente desejado o usuário acaba acessando todos os lugares sugeridos.

Acreditamos que as distâncias com valores altos correspondem a robôs acessando o *site* do Apontador e coletando suas páginas. Como exemplo, algumas sessões possuem milhares de acessos para localizações de diferentes pontos do país, realizadas em poucos minutos. De fato, todas as localizações compartilhadas pelo Apontador podem ser livremente acessadas por robôs e indexadas por máquinas de busca, conforme pode ser percebido pelo arquivo robots.txt do servidor do Apontador[5].



(a) Usuários logados (b) Usuários não logados

Figura 6. Distâncias entre os locais na sessão

Poucas requisições por sessão associada ao fato do tempo de expiração das sessões ser pequeno, aproximadamente 10 minutos, e atípico se comparado com o tempo de expiração de outras redes sociais, como o Youtube [Gill et al. 2008], sugerem que existe pouca atividade de navegação dos usuários. Além disso, a estratégia do Apontador de permitir que máquinas de busca indexem seu conteúdo atrai um grande número de acessos ao sistema, porém, não mantém os usuários navegando no sistema. Acreditamos

[5]www.apontador.com.br/robots.txt

que tal cenário sugere como direção de trabalho futuro a construção de um sistema de recomendação de locais.

Também comparamos as distâncias entre locais das requisições de usuários logados com relação à quantidade de categorias iguais e diferentes, conforme a Figura 7. Podemos notar que cerca de 35% do total de distâncias entre locais possuem categorias iguais, sendo que 80% dessas ocorrências possuem distâncias até 10 KM. Já para categorias diferentes, os outros 65% do total, 54% das ocorrências possuem distâncias até 10KM. Ou seja, quando o usuário está procurando por locais com categorias iguais, a probabilidade desses lugares serem muito próximos é maior do que quando está procurando por locais de categorias diferentes. Padrões como este podem ser úteis para um sistema de recomedação de locais.

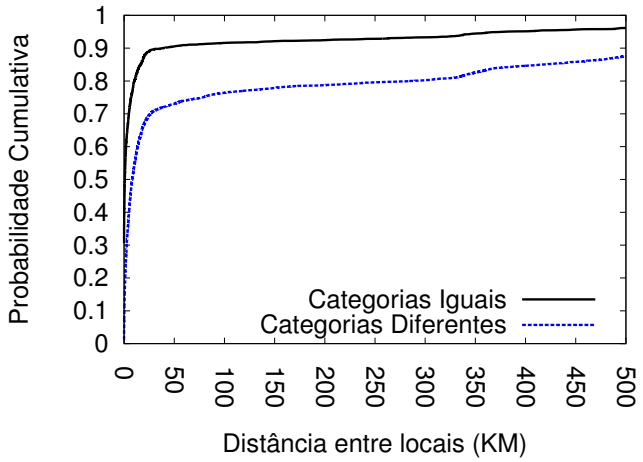


Figura 7. Distâncias entre os locais na sessão de usuários logados para categorias iguais e diferentes

5. Conclusão e Trabalhos Futuros

Neste trabalho utilizamos uma carga de trabalho real e representativa para caracterizar os padrões de acesso ao servidor do Apontador, de forma a caracterizar e modelar os padrões de acessos dos usuários a esses sistemas. Como resultados, fornecemos modelos estatísticos para várias características de acesso, como popularidade dos locais e dos usuários, tempo entre chegada de requisições e sessões, etc. O estudo apresentado é inovador por ser o primeiro a analisar uma rede social baseada em localização sob o ponto de vista do servidor. Os modelos apresentados são úteis não só para a geração de carga sintética, mas também para o projeto e criação de novas infra-estruturas para esse tipo de serviço.

Como trabalhos futuros, planejamos construir um gerador de carga sintética que possibilite realizar experimentação e simulação baseadas em distribuições realistas. Outra direção consiste em construir um sistema de recomendação que ajude os usuários na busca por locais.

Agradecimentos

Os autores gostariam de agradecer ao Apontador pelos dados fornecidos, que tornaram possível a realização desse trabalho.

Referências

Arlitt, M. (2000). Characterizing web user sessions. *SIGMETRICS Performance Evaluation Review*, 28(2):50–63.

Arlitt, M. and Williamson, C. (1996). Web server workload characterization: the search for invariants. *SIGMETRICS Performance Evaluation Review*, 24(1):126–137.

Benevenuto, F., Pereira, A., Rodrigues, T., Almeida, V., Almeida, J., and Gonçalves, M. (2010). Characterization and analysis of user profiles in online video sharing systems. *Journal of Information and Data Management*, 1(2):115–129.

Benevenuto, F., Rodrigues, T., Cha, M., and Almeida, V. (2009). Characterizing user behavior in online social networks. In *ACM SIGCOMM conference on Internet measurement conference (IMC)*, pages 49–62.

Benevenuto, F., Rodrigues, T., Cha, M., and Almeida, V. (2012). Characterizing user navigation and interactions in online social networks. *Information Sciences*, 195(15):1–24.

Costa, C., Cunha, I., Vieira, A., Ramos, C., Rocha, M., Almeida, J., and Ribeiro-Neto, B. (2004). Analyzing client interactivity in streaming media. In *World Wide Web Conference (WWW)*.

Duarte, F., Mattos, B., Bestavros, A., Almeida, V., and Almeida, J. (2007). Traffic characteristics and communication patterns in blogosphere. In *Proc. Int'l Conference on Weblogs and Social Media (ICWSM)*.

Gill, P., Arlitt, M., Li, Z., and Mahanti, A. (2008). Characterizing user sessions on youtube. In *IEEE Multimedia Computing and Networking (MMCN)*.

Lins, T., Benevenuto, F., Dores, W., and Barth, F. (2012). Object popularity distributions in online social networks. In *ACM SIGWEB Web Science Conference (WebSci)*.

Menasce, D. and Almeida, V. (2000). *Scaling for E Business: Technologies, Models, Performance, and Capacity Planning*. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Menascé, D., Almeida, V., Fonseca, R., and Mendes, M. (1999). A methodology for workload characterization of e-commerce sites. In *ACM Conf. on Electronic Commerce (EC)*.

Noulas, A., S. Scellato, C. M., and Pontil, M. (2011a). Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. *SMW 2011*.

Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011b). An empirical study of geographic user activity patterns in foursquare. In *Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 570–573.

Oke, A. and Bunt, R. (2002). Hierarchical workload characterization for a busy web server. In *Int'l Conf. on Computer Performance Evaluation, Modelling Techniques and Tools (TOOLS)*.

Pereira, A., Silva, L., and Meira, Jr., W. (2006). Evaluating the impact of reactive workloads on the performance of web applications. In *Proceedings of the 25th IEEE Interna-*

tional Performance, Computing, and Communications Conference (IPCCC), Phoenix, Arizona, USA. IEEE CS.

Pujol, J., Erramilli, V., Siganos, G., Yang, X., Laoutaris, N., Chhabra, P., and Rodriguez, P. (2010). The little engine(s) that could: scaling online social networks. In *ACM SIGCOMM Conferece*, pages 375–386.

Scellato, S. (2011). Beyond the social web: the geo-social revolution. *SIGWEB Newsletter*, pages 5:1–5:5.

Scellato, S., Noulas, A., Lambiotte, R., and Mascolo, C. (2011). Socio-spatial Properties of Online Location-based Social Networks. In *Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*.

Schneider, F., Feldmann, A., Krishnamurthy, B., and Willinger, W. (2009). Understanding online social network usage from a network perspective. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 35–48.

Trivedi, K. S. (2002). *Probability and statistics with reliability, queuing and computer science applications*. John Wiley and Sons Ltd.

Vasconcelos, M., Ricci, S., Almeida, J., Benevenuto, F., and Almeida, V. (2012a). Caracterizacão e influência do uso de tips e dones no foursquare. *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*.

Vasconcelos, M., Ricci, S., Almeida, J., Benevenuto, F., and Almeida, V. (2012b). Tips, dones and to-dos: Uncovering user profiles in foursquare. In *ACM International Conference of Web Search and Data Mining (WSDM)*.

Veloso, E., Almeida, V., Jr., W. M., Bestavros, A., and Jin, S. (2006). A hierarchical characterization of a live streaming media workload. *IEEE/ACM Transactions on Network*, 14(1).