Caracterização e análise de perfis de usuário em

Sistemas de compartilhamento de vídeo online

Fabrício Benevenuto[1], Adriano Pereira[2], Tiago Rodrigues[1],
Virgílio Almeida[1], Jussara Almeida[1], Marcos Gonçalves[1]

[1] Universidade Federal de Minas Gerais (UFMG)
Departamento de Ciência da Computação (DCC)
Belo Horizonte, MG, Brasil
{fabricio, tiagorm, virgilio, jussara, mgoncalv}@dcc.ufmg.br

[2] Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)
Departamento de Engenharia da Computação (DECOM)
Belo Horizonte, MG, Brasil
adriano@decom.cefetmg.br

Resumo. Recentemente, tem-se observado uma crescente popularização dos ambientes de compartilhamento de vídeo. Parte desse sucesso se deve à mudança na perspectiva do usuário de consumidor para criador de conteúdo, princípio básico da Web 2.0. Assim, os provedores de serviços de vídeo estão lidando com diferentes desafios, como armazenamento de conteúdo, desempenho e escalabilidade de servidores, personalização e diferenciação de serviço. Nesse contexto, é fundamental entender as características das solicitações que chegam a esses servidores e os padrões de navegação do usuário nesses sistemas interativos. Este trabalho aborda esses aspectos. Por meio da análise da carga de trabalho do serviço de vídeo do UOL, maior provedor de conteúdo da América Latina, apresentamos uma caracterização completa das sessões dos usuários, suas solicitações ao servidor e seu perfil de navegação. Tais análises são importantes não apenas para gerar carga de trabalho sintética, mas também para projetar e criar novas infra-estruturas para sistemas de compartilhamento de vídeo. Nossos resultados mostram que existem diferentes perfis de usuários e também fornecem uma melhor compreensão do padrão de acesso do usuário em sistemas de compartilhamento de vídeo.

Categorias e descritores de assuntos: H.3.5 [Serviços de informação online]: Serviços baseados na web

Palavras-chave: redes sociais, sistemas de compartilhamento de vídeo, servidor de vídeo, Web 2.0

1. INTRODUÇÃO

Recentemente, os sistemas de compartilhamento de vídeo online têm aumentado e ganhado popularidade rapidamente. Assistir e publicar vídeos na Internet está se tornando uma rotina no dia a dia dos internautas. Segundo a comScore, em maio de 2008, 74% do público norte-americano da Internet assistia a vídeos online, o que corresponde a 12 bilhões de vídeos transmitidos apenas naquele mês [comScore 2008].

Parte desse sucesso está associado à mudança na perspectiva do usuário, de simples espectador a um criador ativo de conteúdo. Além disso, esses ambientes permitem diversos tipos de interação entre usuários e vídeos, como relações de amizade, avaliação de vídeos e publicação de comentários. Associado a esta nova perspectiva da Web, também conhecida como Web 2.0, existem vários desafios que os fornecedores destes serviços têm de enfrentar, tais como armazenamento de conteúdos, desempenho e escalabilidade do servidor, personalização e diferenciação de serviços, detecção de conteúdos ilegais, etc. Assim, entender as características das solicitações e os padrões de acesso dos usuários, bem como os aspectos da navegação do usuário quando se conectam a esses sites, é importante por dois motivos principais. Primeiro, estudos

de navegação do usuário permitem avaliar o desempenho dos sistemas existentes e levar a um melhor design do site [Wilson et al. 2009; Burke et al. 2009] e políticas de posicionamento de anúncios [B. Williamson]. Em segundo lugar, entender como a carga de trabalho da mídia social está remodelando o tráfego da Internet é valioso para projetar a infraestrutura da Internet de próxima geração e os sistemas de distribuição de conteúdo [Rodriguez; Krishnamurthy 2009]. Apesar de alguns esforços que caracterizam as cargas de trabalho geradas pelo usuário [Cha et al. 2007; Gill et al. 2007; Benevenuto et al. 2009], não há um trabalho que forneça uma caracterização da navegação do usuário do ponto de vista de um servidor de vídeo.

Este trabalho dá o primeiro passo nessa direção. Por meio de um grande conjunto de dados obtido do serviço de vídeo do Universo OnLine (UOL)[1], o maior provedor de conteúdo da América Latina, apresentamos uma caracterização em profundidade da carga de trabalho de sessões e solicitações no servidor de vídeo. Nosso estudo usa traços, como dados de sequência de cliques, que capturam *tudo* atividades dos usuários [Chatterjee et al. 2003; Benevenuto et al. 2009]. Obtivemos um conjunto de dados de sequência de cliques, que descreve resumos no nível da sessão de mais de 3,6 milhões de solicitações HTTP de mais de 1 milhão de IPs diferentes durante um período de 26 dias.

Usando os dados de clickstream, conduzimos dois conjuntos de análises. Primeiro, caracterizamos o tráfego e os padrões de sessão da carga de trabalho. Examinamos a frequência com que as pessoas se conectam ao servidor de vídeo e por quanto tempo. Com base nos dados, fornecemos modelos de melhor ajuste de tempos entre chegadas de sessões e distribuições de duração de sessões. Em segundo lugar, fornecemos uma definição de sessão do usuário em nosso sistema e caracterizamos a navegação do usuário dentro das sessões. Nossa análise revela as atividades dominantes do usuário e as taxas de transição entre as atividades. Nosso estudo fornece muitas descobertas interessantes, incluindo:

- Uma sessão típica de usuário de sistemas de compartilhamento de vídeo online dura cerca de 40 minutos, um valor alto em comparação com os sistemas da Web tradicionais.
- As distribuições de popularidade de acessos de objetos (vídeos e tags) seguem longas caudas.
- As classificações da atividade do usuário em termos do número de solicitações enviadas e sessões criadas seguem distribuições de cauda longa e exponencial, respectivamente.
- A taxa de solicitação de chegadas no sistema apresenta um padrão periódico com maior intensidade durante o dia e menor intensidade durante a noite.
- As distribuições de tempo entre solicitações e tempo entre sessões podem ser modeladas por distribuições exponenciais.
- Para sessões mais longas, os usuários passam mais tempo vendo vídeos do que em sessões curtas.
- Nossa análise revela diferentes perfis de usuários que acessam o sistema, que podem ser usados pelos administradores do sistema para personalizar serviços.

O restante deste trabalho está organizado da seguinte forma. A próxima seção descreve o trabalho relacionado. A seção 3 apresenta estatísticas sobre a carga de trabalho do serviço de vídeo do UOL. A seção 4 discute a caracterização de solicitações e sessões. Na seção 5, apresentamos uma análise do perfil dos usuários que navegam no sistema. Finalmente, a Seção 6 conclui o artigo e apresenta orientações para trabalhos futuros.

2. TRABALHO RELACIONADO

A caracterização da carga de trabalho é fundamental para o entendimento e aprimoramento dos sistemas web. Existem vários estudos que apresentam caracterizações de carga de trabalho de diferentes tipos, como servidores Web [Arlitt e Williamson 1996], e-commerce [Menascé e Almeida 2000], blogs [Duarte et al. 2007], vídeo sob demanda [Costa et al. 2004] e vídeo ao vivo [Veloso et al. 2006]. Dentre as várias contribuições desses trabalhos, destacamos a criação de modelos valiosos capazes de descrever a carga de trabalho que chega a esses servidores, essenciais para a geração de cargas de trabalho sintéticas, que permitem a experimentação e simulação com base em cargas de trabalho realistas. Particularmente Costa *et al.* [Costa et al. 2004] analisou solicitações de dois servidores de vídeo em contexto educacional. Eles mostram que o tempo entre solicitações

[1] http://videolog.uol.com.br

segue uma distribuição de Pareto e a popularidade do objeto pode ser modelada pela concatenação de distribuições Zipflike. Diferentemente, em nosso trabalho, apresentamos uma caracterização da carga de trabalho de um servidor de conteúdo de vídeo gerado pelo usuário. Não temos conhecimento de nenhum outro trabalho que realize este tipo de caracterização do ponto de vista do servidor.

Como complemento ao nosso esforço, existem vários trabalhos que caracterizam diferentes aspectos dos sistemas de compartilhamento de vídeo online, especialmente o YouTube. Em [Cha et al. 2007], os autores analisam a distribuição de popularidade, evolução e características dos vídeos do YouTube, além de avaliar diferentes abordagens para a distribuição de vídeos, como caches e compartilhamento P2P. Complementarmente, Duarte*et al.* [Duarte et al. 2007] caracteriza aspectos geográficos das interações dos usuários do YouTube. Rodrigues *et al.* [Rodrigues et al. 2010] estudou diferenças nas estatísticas de uso e metadados de vídeos duplicados. Gill*et al.* [Gill et al. 2007] apresentam uma caracterização da carga de trabalho do YouTube do ponto de vista de uma universidade, comparando suas propriedades com o tráfego da Web e outros servidores de vídeo. Em [Gill et al. 2008], os autores analisam as características das sessões de usuários no YouTube, por meio da análise de solicitações em proxy de uma universidade. No entanto, os autores avaliam apenas aspectos como a duração da sessão e a criação da sessão, diferentemente de nós, que investigamos as diferentes ações dos usuários em uma sessão. zinco*et al.* [Zink et al. 2008] realizam simulações para mostrar que o cache de vídeo, no cliente ou em um proxy, e a distribuição P2P podem reduzir o tráfego na rede e permitir um acesso rápido ao vídeo em sistemas de compartilhamento de vídeo online.

Recentemente, apresentamos uma caracterização abrangente das propriedades do YouTube *rede de resposta de vídeo*, ou seja, a rede que emerge das interações do usuário baseadas em vídeo [Benevenuto et al. 2009]. Em [Benevenuto et al. 2009], caracterizamos ainda o comportamento de três classes de usuários, a saber, usuários legítimos, spammers e promotores de conteúdo. Usando um algoritmo de aprendizado de máquina e explorando vários atributos dos perfis dos usuários, do comportamento social dos usuários no sistema (ou seja, as relações estabelecidas entre eles) e dos vídeos do usuário, fomos capazes de detectar a grande maioria dos promotores e spammers . Finalmente, a referência [Benevenuto et al. 2010] fornece uma visão geral abrangente de diferentes tipos de atividades maliciosas em sistemas de compartilhamento de vídeo, bem como suas implicações para usuários e sistemas.

Diferentemente desses esforços, nosso trabalho aqui visa não apenas caracterizar e compreender as solicitações que chegam ao servidor de conteúdo de vídeo gerado pelo usuário, mas também investigar e identificar o perfil dos usuários que acessam esses sistemas. Complementarmente, Benevenuto*et al.* [Benevenuto et al. 2009] usaram dados de clickstream para caracterizar a navegação do usuário e as interações sociais em redes sociais online, como Orkut, Hi5, MySpace e LinkedIn.

3. DESCRIÇÃO DA CARGA DE TRABALHO

Em nosso estudo, analisamos a carga de trabalho do serviço de vídeo do UOL, importante provedor de conteúdo no Brasil e na América Latina. Os dados clickstream obtidos correspondem a um período de quase um mês, de 12/12/2007 a 01/07/2008, contabilizando um total de 3.681.232 solicitações, de mais de 1.127.537 IPs diferentes.

Cada registro na carga de trabalho representa uma solicitação enviada por um usuário ao serviço de vídeo. As seguintes informações estão disponíveis para cada solicitação:*IP, hora, solicitação, status, tamanho, referenciador e agente*. O campo *IP* contém o endereço IP anônimo que gerou a solicitação. O campo *Tempo* corresponde ao momento, incluindo data e hora em segundos, em que a solicitação foi recebida pelo servidor. O campo *solicitar* contém não apenas a URL solicitada, mas também o método e o protocolo usados. O campo *status* mostra o código de resposta do protocolo HTTP à solicitação. O campo *Tamanho* indica o tamanho da solicitação em bytes. O campo *referenciador* mostra o URL de onde o pedido de visita foi originado. Por exemplo, se um usuário em uma página da Web A visitar um link que o redireciona para um vídeo B, o campo *solicitar* contém o URL B e o campo *referenciador* contém a página da Web A. O último campo, *agente*, identifica o navegador e o sistema operacional usados.

Nome do grupo	tipo de solicitação	Número de Pedidos	Percentagem
1: Ver	Veja um vídeo	2.758.883	74,94%
2: usuário	Lista de vídeos de um usuário	218.335	5,93%
	Lista de vídeos de um usuário com uma determinada tag	75.583	2,05%
3: listas	Lista dos principais vídeos	55.307	1,50%
	Lista de vídeos relacionados de um vídeo	32.838	0,89%
4: Interações	Avaliação de vídeo	22.038	0,60%
	Comentário de vídeo	14.131	0,38%
	Vídeo Favorito	10.774	0,29%
5: Pesquisa	Procurar	1.625	0,04%
	Lista de vídeos com uma determinada tag	421.700	11,46%
6: Outros	Página principal	2.679	0,07%
	Solicitações de erro ou registro não formatado	67.339	1,82%

Tabela I. Solicitar grupos

Os campos *referenciador* e *agente* pode estar faltando em alguns registros, uma vez que os usuários podem removê-los para aumentar a privacidade. Além disso, o campo *referenciador* não pode ocorrer quando o usuário digita a URL diretamente no navegador.

Em nossa carga de trabalho, existem vários tipos de solicitações, que organizamos em seis grupos, conforme mostrado na Tabela I. As solicitações do grupo 1 correspondem às visualizações do vídeo. No grupo 2, temos solicitações relacionadas a listar os vídeos do usuário e listar os vídeos de um usuário que contenham uma determinada tag. O terceiro grupo reúne solicitações de usuários para listas de vídeos relacionados e listas dos vídeos principais. No grupo 4, temos todas as solicitações relacionadas à avaliação de um vídeo (atribuir uma classificação de cinco estrelas) e as interações do usuário relacionadas à inclusão de adicionar um vídeo como favorito e postar um comentário em um vídeo. O grupo 5 corresponde a solicitações de busca de conteúdo por meio do mecanismo de busca de vídeos ou por meio de acessos a nuvem de tags. As solicitações de erro são identificadas por meio do campo status, de acordo com as definições apresentadas em [Fielding et al. 1999]. Para a análise das próximas seções, essas solicitações não são consideradas. Exceto pelo grupo 6, todos os grupos apresentados na Tabela I são usados na análise do perfil de navegação do usuário apresentada na Seção 5.

3.1 Limitações

Embora nossos dados nos dêem uma oportunidade única de estudar as atividades do usuário em sistemas de compartilhamento de vídeo, os registros têm várias limitações. Primeiro, não podemos identificar os IDs de usuário no sistema. Em segundo lugar, embora tenhamos informações sobre os endereços IP dos usuários, essas informações são anônimas. Assim, não podemos associar um endereço IP distinto a cada usuário. Isso ocorre porque muitos ISPs residenciais usam DHCP para atribuir dinamicamente um endereço IP a cada host quando ele se conecta à Internet. Quando um nó se desconecta, ele libera seu endereço IP atribuído, que pode então ser atribuído a um cliente residencial diferente. Portanto, não podemos agrupar várias sessões em eventos de um único usuário. Em segundo lugar, os campos *referenciador* e *agente* pode estar faltando em alguns registros do log, uma vez que os usuários podem removê-los para preservar a privacidade. Finalmente, o campo *referenciador* também pode estar ausente quando o usuário digita o URL diretamente no navegador.

4. CARACTERIZAÇÃO DA CARGA DE TRABALHO

Nesta seção, apresentamos uma caracterização da carga de trabalho do serviço de vídeo do UOL sob diferentes perspectivas, modelando diversos aspectos e distribuições.

4.1 Definição de Sessão

Antes de apresentar nossas análises, precisamos definir uma duração de sessão apropriada para as sessões em nossos dados. Uma sessão de usuário é definida como uma série de solicitações realizadas por um usuário a um site durante um determinado período de tempo [Menascé et al. 1999; Arlitt 2000]. Nas redes sociais de vídeo online, uma sessão típica inclui uma lista de vídeos por assunto, pesquisa, streaming de vídeo, interação com outros usuários por meio da publicação de comentários e avaliação dos vídeos. Essas solicitações são muito diferentes das solicitações em sessões de usuário dos sites tradicionais, que não fornecem o mesmo nível de interação entre usuários e objetos que ocorre nos sistemas Web 2.0.

Para determinar o início e o término de uma sessão no serviço de vídeo UOL é necessário analisar o tempo entre as solicitações de um mesmo usuário para medir o período de inatividade daquele usuário, uma vez que as sessões não apresentam registro de login e logout. Assim, é necessário realizar uma análise para identificar o limite de tempo entre as solicitações para considerá-las como pertencentes à mesma sessão. Consideramos duas solicitações consecutivas como pertencentes à mesma sessão se o tempo entre elas for menor que este limite, ou seja *sessão expirada*.

É importante escolher um tempo limite de sessão adequado para evitar a geração de sessões que não representem a utilização do serviço pelos usuários, evitando juntar diferentes momentos de utilização do serviço ou fragmentar a navegação do usuário. Seguindo a metodologia proposta em [Menascé et al. 1999], avaliamos o tempo limite de sessão adequado para nossa aplicação.

A Figura 1 (à esquerda) apresenta o número total de sessões para diferentes valores de tempo limite de sessão. Um valor extremamente pequeno (por exemplo, 1 minuto) pode resultar em um grande volume de sessões. À medida que o valor do tempo limite da sessão aumenta, o número de sessões é reduzido continuamente até se estabilizar. A estabilidade ocorre em torno de 40 minutos, indicando este valor como um tempo limite de sessão adequado.

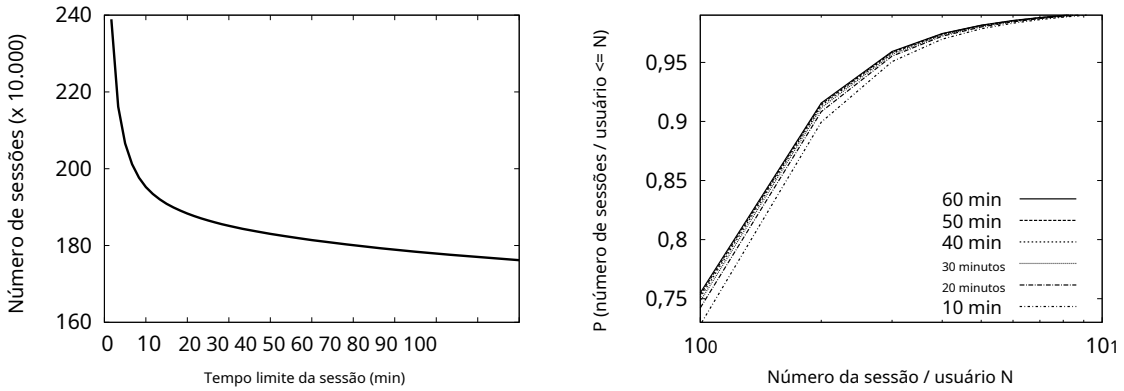

Figura 1. Tempo limite de sessão (esquerda) e CDF do número de sessões por usuário (direita)

Além dessa análise para escolher um tempo limite de sessão adequado, também geramos a função de distribuição cumulativa (CDF) do número de sessões por usuário para vários valores de tempo limite de sessão, conforme ilustrado na Figura 1 (à direita). A diferença entre as distribuições para diferentes valores de timeout da sessão é maior para valores menores, tornando-se muito pequena para valores maiores que 40 minutos. Assim, adotamos 40 minutos como tempo limite de sessão para nossas análises. No total, nossa carga de trabalho contém 1.127.537 sessões de usuário.

É interessante observar que nossa escolha é coerente com a análise feita em [Gill et al. 2008]. Em comparação com os esforços anteriores, que caracterizam as sessões em sites da Web tradicionais [Arlitt 2000; Oke e Bunt 2002], os valores de tempo limite mais longos obtidos são muito mais longos em comparação com os 10 minutos

geralmente observada. Os motivos mais intuitivos para esse comportamento são o tempo mais longo que os usuários levam para assistir a um vídeo e as ferramentas interativas, que podem fazer com que os usuários passem mais tempo no site.

4.2 Popularidade do Objeto

Inicialmente, avaliamos a popularidade de objetos com o objetivo de verificar se a popularidade de visualizações de vídeos e tags pesquisadas obedecem a uma lei de potência. As leis de potência estabelecem a seguinte relação $P(E_n) \propto n^{-\alpha}$, Onde $P(E_n)$ é a probabilidade de referência ao n^{o} elemento mais popular. A fim de verificar a precisão dos modelos propostos, medimos o fator R^2 da regressão linear [Trivedi 2002] para cada distribuição analisada. Em todos os modelos apresentados, os valores de R^2 são superiores a 0,97. Um valor de R^2 igual a 1 significa que não há diferenças entre o modelo e a carga de trabalho real.

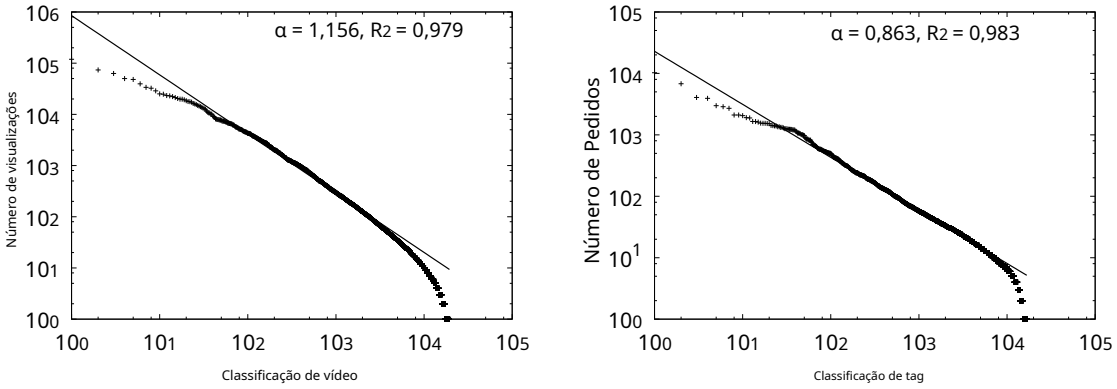

Fig. 2. Vídeos classificados pelo número de visualizações (à esquerda) e tags classificados pelo número de acessos (à direita)

Em seguida, analisamos se a distribuição de popularidade de vídeos e tags segue uma lei de potência. A Figura 2 (à esquerda) mostra a classificação dos vídeos classificados pelo número de visualizações. Podemos notar que um pequeno número de vídeos tem um alto número de visualizações e que um grande número de vídeos tem apenas algumas visualizações. Essa observação é importante, pois sugere uma oportunidade para armazenamento em cache de vídeo. Na verdade, a distribuição é modelada por uma função que segue uma lei de potência, com $\alpha = 1.156$ e $R^2 = 0.979$.

Da mesma forma, a Figura 2 (direita) mostra a classificação de acesso às tags (por exemplo, listas de todos os vídeos com uma determinada tag). Podemos notar que algumas tags concentram um grande número de acessos. A título de exemplo, a primeira tag do ranking possui 10.266 acessos. Esta classificação pode ser modelada por uma distribuição de lei de potência, com $\alpha = 0.983$ e $R^2 = 0.983$.

4.3 Atividade do usuário

A seguir, analisamos o nível de atividade dos usuários no sistema. Sabemos que o usuário pode acessar o serviço de vídeo do UOL várias vezes na mesma sessão ou em sessões diferentes. Assim, de forma a modelar o nível de atividade dos utilizadores no sistema, caracterizamos a classificação dos utilizadores em termos de pedidos enviados e em termos de número de sessões criadas. Por usuário, queremos dizer cada IP anônimo de nossa carga de trabalho.

A Figura 3 (à esquerda) mostra a classificação dos usuários de acordo com o número de solicitações enviadas ao servidor. Podemos notar que há um pequeno número de usuários que geram grande quantidade de solicitações ao servidor e um grande número de usuários que enviam poucas solicitações. Na verdade, a distribuição é bem modelada por uma lei de potência do tipo $f(x) = bx^{-a}$, com $\alpha = 0.745$, e $R^2 = 0.987$.

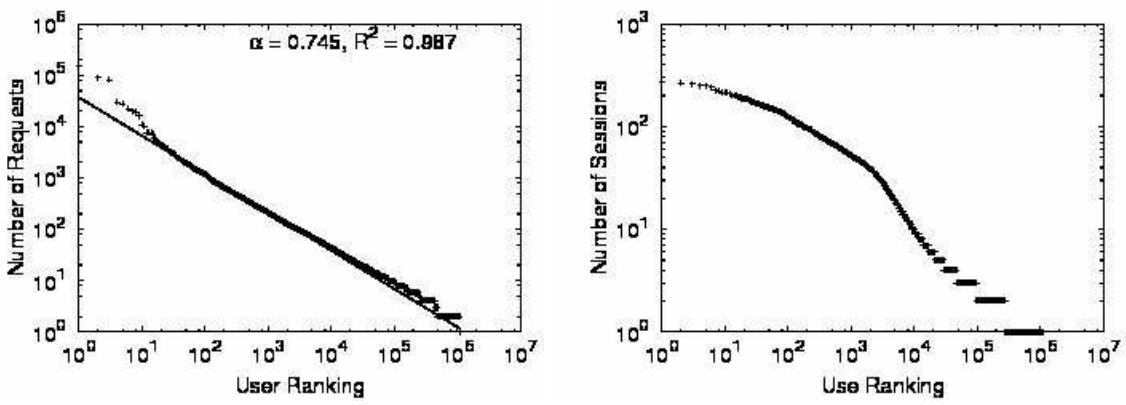

Fig. 3. Classificação da atividade do usuário em termos de solicitações (à esquerda) e sessões (à direita)

Em termos de sessões criadas no servidor, a análise mostra que uma distribuição exponencial é a função que melhor modela os dados. Assim, a classificação das sessões pode ser modelada por uma distribuição exponencial do tipo $f(x) = \alpha e^{\beta x}$, com $\alpha = 175.2$ e $\beta = -0.002681$. Esse resultado enfatiza o comportamento de que poucos usuários podem gerar grandes quantidades de sessões, enquanto a maioria dos usuários gera apenas algumas sessões.

4.4 Padrões Temporais

Esta seção analisa o número de requisições que chegam ao servidor em função do tempo. As solicitações sobre streaming de vídeo não são registradas na carga de trabalho. Registramos apenas solicitações de HTML que fornecem acesso ao vídeo. Assim, não podemos quantificar o tráfego em bytes transferidos pelo serviço de vídeo do UOL com nossa carga de trabalho.

A Figura 4 (canto superior esquerdo) mostra o número de solicitações que chegam ao servidor em intervalos de tempo de uma hora. A curva apresenta um padrão periódico, com maior intensidade de solicitações durante o dia e pequena intensidade durante a noite, semelhante a outros servidores Web tradicionais [Veloso et al. 2006; Arlitt e Williamson 1996]. Observe que existem alguns pontos em que podemos ver 50.000 solicitações em 1 hora. Esses pontos representam links para vídeos disponíveis em páginas populares do portal UOL.

Para analisar a participação dos usuários que visitam o sistema, caracterizamos o tempo entre requisições e entre sessões. Apresentamos nas Figuras 4 (canto superior direito) e (abaixo) a função de distribuição cumulativa complementar para essas duas métricas. Podemos notar que a probabilidade do tempo entre requisições ser maior que 5 segundos é menor que 1%, enquanto 57% das requisições que chegam ao servidor possuem intervalos de tempo menores que 1 segundo. Da mesma forma, cerca de 96% dos intervalos entre as sessões são menores que 5 segundos.

Ambas as distribuições são melhor modeladas por uma função exponencial do tipo $f(x) = \alpha e^{\beta x}$. Para a distribuição dos tempos entre solicitações, obtivemos um $\alpha = 0.424$ e $\beta = -1.298$ com $R^2 = 0.996$, e para a distribuição dos tempos entre as sessões, encontramos um $\alpha = 0.5518$ e $\beta = -0.7309$ com $R^2 = 0.989$.

4.5 Referenciador de solicitações e sessões

A seguir, analisamos o referenciador de solicitações e sessões de usuários que acessam o sistema. Para analisar como os usuários começam a navegar no sistema de vídeo, analisamos o referenciador da primeira solicitação de cada sessão. Cerca de 50% das sessões não possuem o referenciador de campo na primeira solicitação, sendo, portanto, descartadas. Da mesma forma, cerca de 40% das solicitações não possuem esse campo e, portanto, não foram utilizadas. A Tabela II mostra o referenciador das sessões e das solicitações de acesso ao sistema. Notamos que a maior parte das sessões e

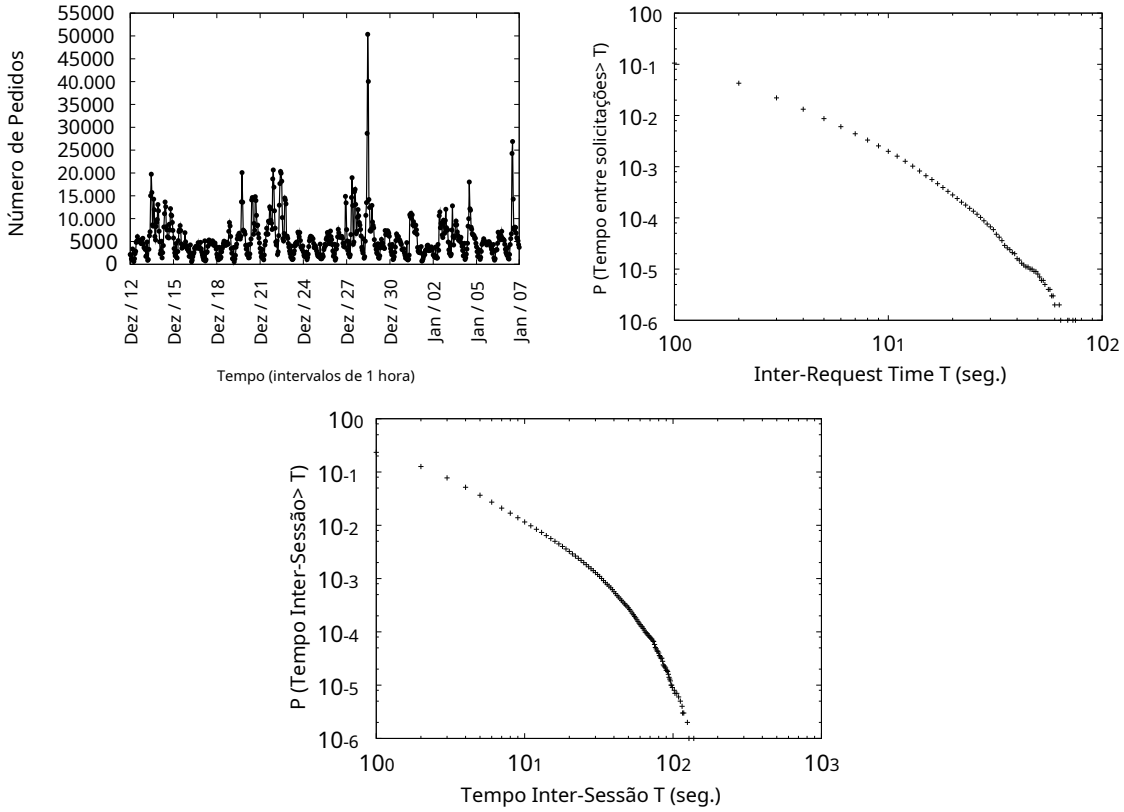

Fig. 4. Número de solicitações em intervalos de uma hora (canto superior esquerdo). CCDF para o tempo entre solicitações (canto superior direito) e tempo entre sessões (para baixo)

Domínio	% de acessos	% de sessões
uol.com.br	50,46%	75,71%
videos.uol.com.br	39,58%	12,10%
. br	7,26%	7,89%
Outros	2,69%	4,30%

Tabela II. Referenciador de solicitações e sessões do usuário

as solicitações vêm de outros serviços do UOL. Porém, uma fração significativa (cerca de 40%) das solicitações tem origem no serviço de vídeo, correspondendo, assim, a usuários que assistem a outros vídeos ou interagem com outros usuários do sistema. Apenas uma pequena parte das solicitações e sessões vem de outros sites.

4.6 Probabilidade de atividade ao longo do tempo

Em seguida, investigamos se existe alguma correlação entre a ocorrência de um tipo específico de atividade (por exemplo, Pesquisar, visualizar um vídeo, etc.) e a duração da sessão. Para verificar essa correlação, categorizamos as sessões de usuário em quatro classes não sobrepostas com base em suas durações de sessão: (*uma*) com menos de 1 minuto de duração, (*b*) entre 1 e 10 minutos de duração, (*c*) entre 10 e 20 minutos de duração, e (*d*) mais de 20 minutos. Para as sessões pertencentes a cada um desses intervalos, examinamos a proporção média da duração total da sessão que um usuário gastou em cada atividade.

A Figura 5 mostra a fração de tempo gasto em cada tipo de atividade em função da duração da sessão.

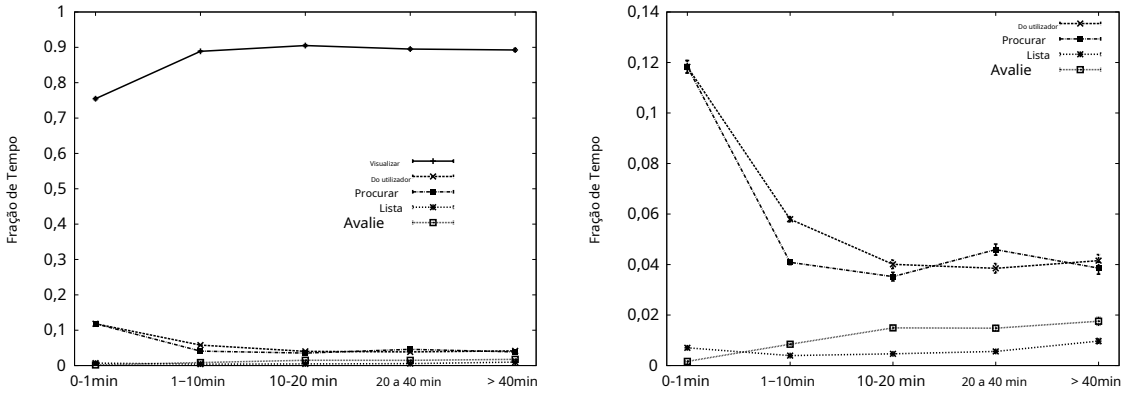

Fig. 5. Probabilidade de diferentes tipos de atividades do usuário como funções da duração da sessão (barras de erro indicam intervalo de confiança de 95%)

Os resultados são mostrados em dois gráficos separados, um contendo todos os grupos de atividades e o outro removendo o grupo Visualização para enfatizar as tendências para as atividades menos populares. Encontramos dois padrões principais. Primeiro, independentemente da duração da sessão, os usuários passam a maior parte do tempo vendo vídeos. Em sessões muito curtas (ou seja, com menos de 1 minuto de duração), os usuários gastam 76% do seu tempo nessas atividades. Para sessões mais longas (ou seja, 20 minutos ou mais), os usuários passam mais de 89% do tempo de suas sessões vendo vídeos. Em segundo lugar, as categorias restantes de atividades tornam-se menos prevalentes em sessões mais longas. A exceção é o tempo gasto na avaliação do conteúdo, que aumenta em um fator de 9 quando comparamos as sessões com duração inferior a 1 minuto com aquelas com duração superior a 20 minutos.

5. MODELAGEM DOS PERFIS DE NAVEGAÇÃO DO USUÁRIO

Esta seção modela os perfis de navegação dos usuários do serviço de vídeo do UOL. Na Seção 5.1, apresentamos a estratégia de modelagem básica, aplicando-a para construir um modelo geral de todos os usuários do sistema. Na Seção 5.2, categorizamos os usuários em grupos separados e analisamos os diferentes perfis de navegação do usuário.

5.1 Perfil geral de navegação do usuário

Para entender os perfis de navegação dos usuários durante suas sessões no sistema, construímos um gráfico direto probabilístico, onde os nós representam os possíveis tipos de solicitações do usuário (por exemplo, pesquisa, visualização, etc.) e os arcos representam a navegação entre um tipo de solicitação para outro em uma única sessão. Além disso, os pesos representam as probabilidades de ocorrência do padrão de navegação. Chamamos este gráfico de UBMG (User Behavior Model Graph). O UBMG é baseado no *Gráfico do modelo de comportamento do cliente* - CBMG [Menascé e Almeida 2000], uma metodologia para representar a navegação do usuário em serviços de e-commerce. Os nós UBMG correspondem aos grupos de pedidos definidos na Tabela I. O *inicial* e *fifinal* são apresentados para representar a primeira e a última solicitações das sessões do usuário, respectivamente.

A Figura 6 ilustra um UBMG típico, considerando todas as sessões de usuário em nossa carga de trabalho. Podemos notar que a maioria dos usuários inicia suas sessões visualizando vídeos (86%), enquanto os demais visitam perfis de usuários ou realizam pesquisas. Em contraste, apenas uma pequena fração dos usuários inicia suas sessões navegando por listas de vídeos ou avaliando vídeos. Depois de ver um vídeo, a maioria dos usuários tende a continuar vendo os vídeos ou até mesmo a terminar a sessão, embora as transições para os outros estados possam ocorrer com probabilidade não desprezível. Curiosamente, encontramos auto-loops fortes em quase todos os estados. Por exemplo, a pesquisa é seguida por outra pesquisa com uma probabilidade de 0,562. Da mesma forma, há uma alta probabilidade (0,545) de um usuário continuar navegando pelos perfis de usuário repetidamente. A repetição também é evidente para navegar nas listas de vídeos (probabilidade de 0,475). o

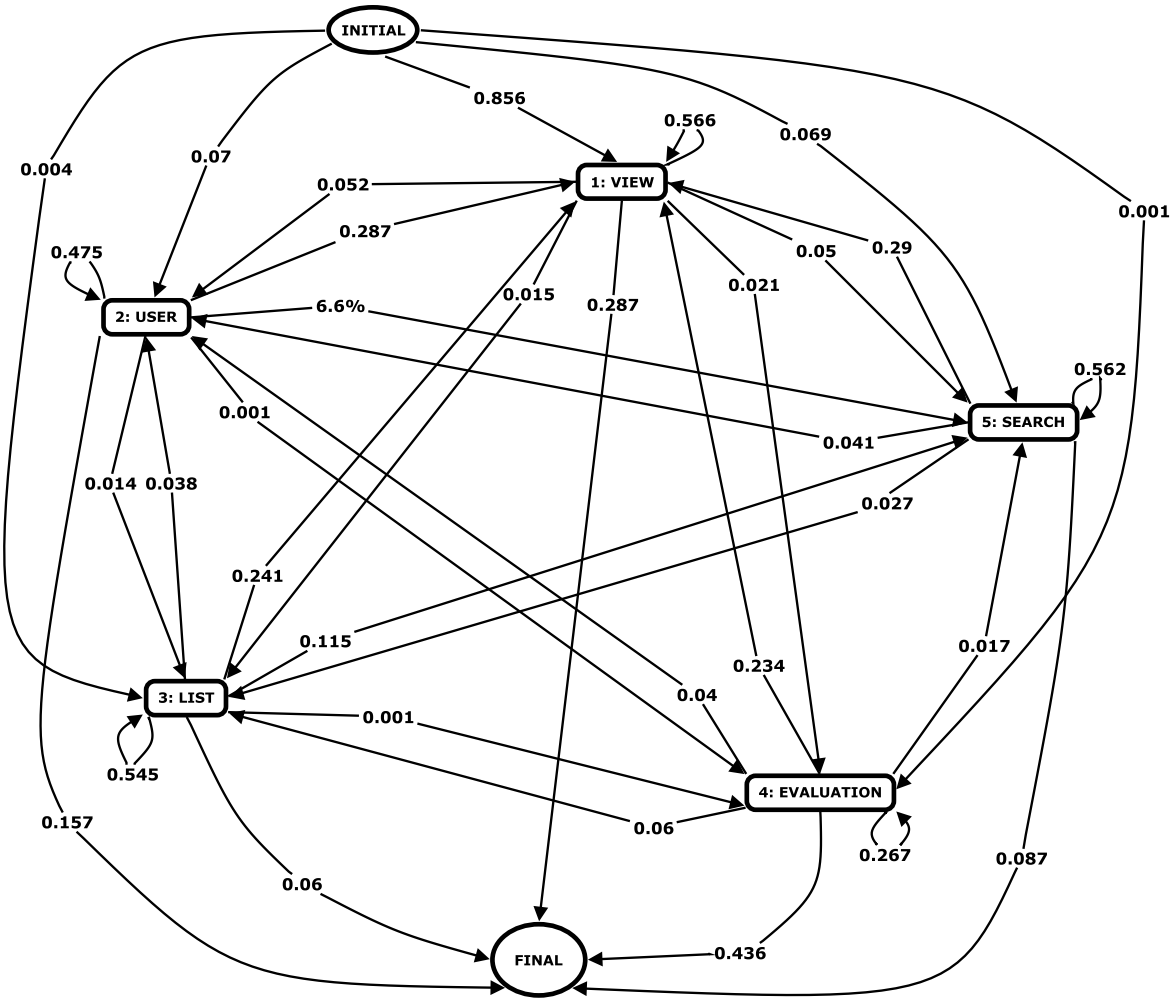

Fig. 6. UBMG típico: comportamento geral do usuário

a ocorrência de tais probabilidades altas de atividades repetidas do mesmo tipo, particularmente navegação, pode ajudar a conduzir o projeto de mecanismos de pré-busca.

A seguir, propomos um método para categorizar os usuários em grupos de acordo com os diferentes perfis de navegação. Essa diferenciação é importante para apoiar o design de serviços customizados.

5.2 Grupos de perfis de navegação do usuário

Até agora, examinamos os padrões gerais de navegação dos usuários em todas as sessões. Embora o UBMG mostrado na Figura 6 seja útil para descobrir o padrão de navegação típico no sistema, ele fornece apenas uma imagem geral e, portanto, não mostra a heterogeneidade entre os usuários em termos de seus perfis individuais. A seguir, propomos um método para categorizar os usuários em grupos de acordo com seus perfis de navegação.

Começamos calculando o UBMG de cada usuário individual, considerando todas as suas sessões. A seguir, aplicamos um algoritmo de agrupamento [Bock 2002], a fim de identificar grupos com características semelhantes a partir de um vetor de atributos. Mais especificamente, definimos cada sessão como um vetor unidimensional, onde cada posição no vetor contém a probabilidade de um usuário navegar de uma categoria de atividade para

outro. Para cada usuário, calculamos seu UBMG individual em todas as suas sessões e, em seguida, usamos as probabilidades dos 35 arcos UBMG possíveis como atributos do usuário para o algoritmo de agrupamento.

Usamos o algoritmo de agrupamento X-means [Pelleg e Moore 2000], que estende o popular Kmeans [Jain et al. 1999] algoritmo. Uma vantagem principal do X-means sobre o K-means é que o algoritmo não apenas fornece os clusters, mas também estima o melhor número possível de clusters. Portanto, não temos que decidir a priori o número de perfis. O algoritmo X-means encontra clusters minimizando a soma das distâncias quadradas entre cada vetor e o centróide do cluster, um vetor que representa as propriedades médias de cada grupo. Consideramos a distância euclidiana entre dois vetores, que é calculada da seguinte forma:

$$D = \frac{1}{n} \sum_{eu=1}^{n} (x_{eu} - y_{eu})^2 \quad (1)$$

Onde n é o tamanho de qualquer vetor, e x e y são os dois vetores.

Usamos a implementação de X-means disponível na ferramenta Weka [Witten e Frank 2005] e definimos o número máximo de grupos como 20. O algoritmo X-means indicou que 15 grupos distintos era a melhor escolha para ajustar nosso conjunto de dados. As sessões com apenas uma solicitação foram descartadas, pois não agregam valor à nossa análise (ou seja, suas representações UBMG possuem apenas arcos que incluem o estado inicial ou final). No total, descartamos 779.384 sessões, concentrando nossa análise a seguir nas 348.153 sessões restantes e em 345.152 usuários.

A Tabela III apresenta os grupos identificados de usuários identificados, o número de usuários e a frequência de ocorrência de cada grupo. Ele também mostra o predominante*Inicial* e *Final* transições de cada grupo, apresentando o estado *para* e *a partir de* que um usuário de cada grupo normalmente navega.

Grupo	Transição Predominante		Número de Comercial	Frequência (%)
	Inicial *declarar*	Final *do estado*		
0	1	1	195.028	55,64
1	2	2 e 1	15.102	4,38
2	1	1	11.424	3,31
3	3	1 e 3	1.352	0,39
4	4	1 e 4	273	0,08
5	5	1	13.211	3,83
6	1	2 e 4	28.562	8,28
7	1	5	8.427	2,44
8	5	5	9.296	2,69
9	5	2 e 1	803	0,23
10	2 e 1	3 e 1	366	0,11
11	1	1	33.137	9,60
12	1	1 e outros	3.726	1.08
13	1 e 5	1 e 5	6.722	1,95
14	1 e 2	1	20.723	6,00
		Total	345.152	100,00

Tabela III. Perfil de navegação do usuário - Grupos

Agora voltamos nossa atenção para os UBMGs de todos os 15 grupos, que são mostrados na Figura 7. Começamos discutindo os perfis dos usuários que visualizam predominantemente vídeos, aqui referidos como *Visualizadores*. Esses perfis correspondem aos grupos 0, 2, 6, 7, 11, 12, 13 e 14. Observamos que omitimos arcos com probabilidades menores que 0,03 por uma questão de clareza. Também optamos por excluir estados que possuem apenas arcos com baixas probabilidades.

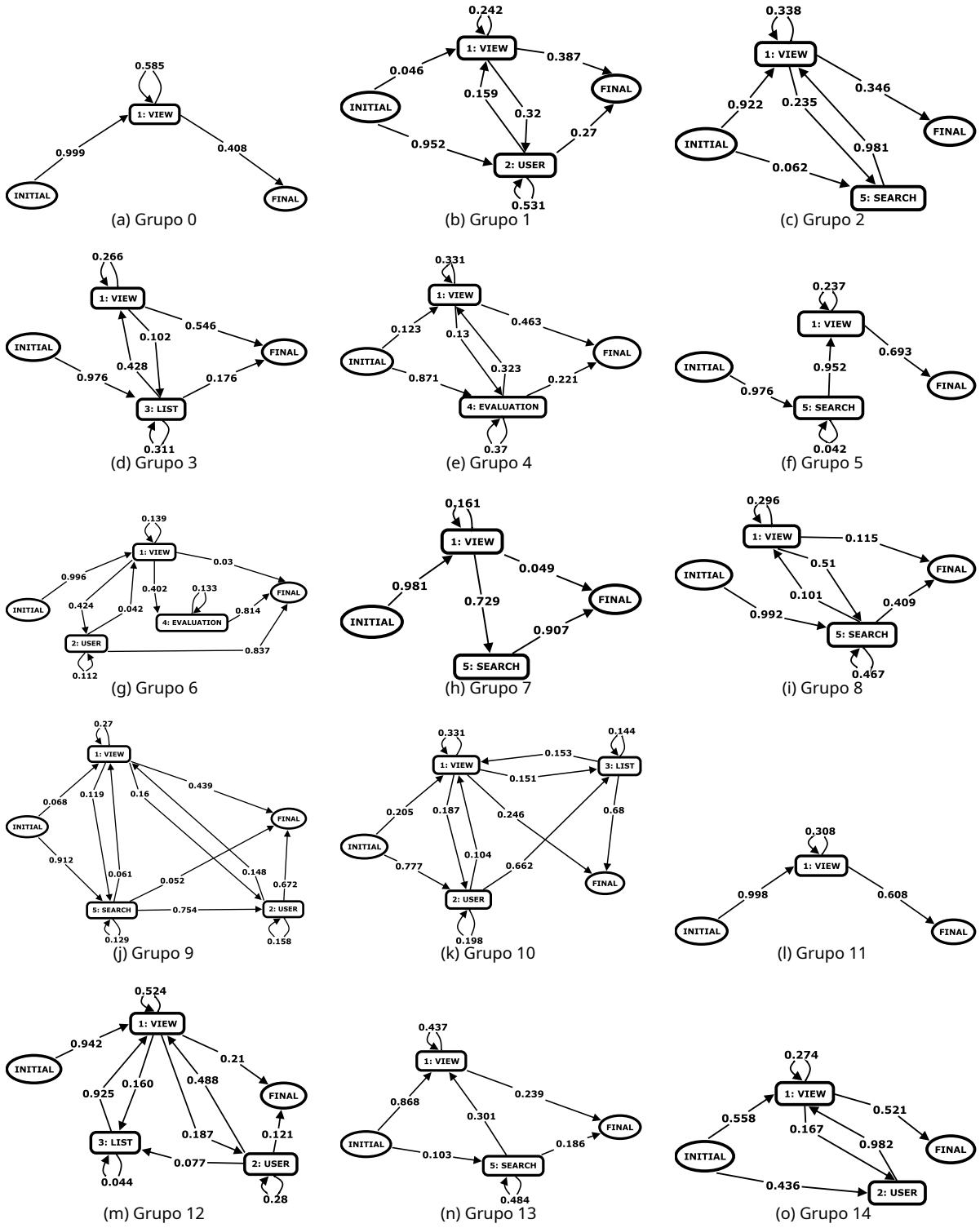

Fig. 7. Gráficos do modelo de comportamento do usuário para diferentes perfis de usuário

A Figura 7 (a) apresenta o gráfico do perfil de navegação do grupo 0, sendo o mais frequente responsável por 55,64% dos usuários. Esses usuários geralmente iniciam suas sessões assistindo a um vídeo. Em seguida, 58,5% deles assistem a outros vídeos, enquanto o restante sai do sistema. Observe que os grupos 0 e 11, representados na Figura 7 (l), são muito semelhantes. A diferença é que os usuários do grupo 11 têm pequena chance de realizar outras atividades (não representadas), como visualizar o perfil do usuário, avaliar conteúdo e listar vídeos. Juntos, esses dois grupos representam mais de 65% dos usuários.

A Figura 7 (g) mostra o perfil típico de navegação do usuário do grupo 6. Da mesma forma que os usuários dos grupos 0 e 11, esses usuários também iniciam suas sessões assistindo a um vídeo. No entanto, a maioria deles visita um perfil de usuário ou avalia o conteúdo posteriormente. Os grupos 2 e 7 (Figuras 7 (c) e 7 (h), respectivamente) também são um pouco semelhantes aos grupos 0 e 11, enquanto os usuários do grupo 2 geralmente assistem aos vídeos após a pesquisa (probabilidade de 0,981), os usuários do grupo 7 deixam o sistema após a pesquisa. Uma diferença importante é que, depois de assistir a pelo menos um vídeo, esses usuários também podem pesquisar um novo vídeo.

Os últimos três grupos de *Visualizadores* são os grupos 12, 13 e 14, representados nas Figuras 7 (m), 7 (n) e 7 (o), respectivamente. Depois de assistir a um vídeo, os usuários do grupo 12 geralmente acessam listas de vídeos e, em seguida, assistem a outro vídeo ou acessam um perfil de usuário, com uma grande chance (quase 49%) de assistir a um vídeo novamente. Alguns usuários do grupo 13 iniciam suas sessões realizando pesquisas (probabilidade de 0,103), enquanto uma grande fração dos usuários do grupo 14 inicia suas sessões navegando em um perfil de usuário (probabilidade de 0,436) e, em seguida, assistir a vídeos com probabilidade de 0,98.

A seguir, analisamos os perfis dos usuários que iniciam suas sessões predominantemente por meio da pesquisa de conteúdo. Esses usuários, chamados de*Buscadores*, correspondem aos grupos 5, 8 e 9. A Figura 7 (f) ilustra o perfil típico do grupo 5. Esses usuários iniciam suas sessões de busca (probabilidade de 0,976) antes de assistir a vídeos com probabilidade de 0,95. Da mesma forma, os usuários dos grupos 8 e 9 também iniciam suas sessões pesquisando (probabilidades de 0,99 e 0,91, respectivamente). No entanto, em vez de assistir a vídeos depois de pesquisar como a maioria dos usuários do grupo 5, os usuários do grupo 8 normalmente continuam pesquisando repetidamente, com probabilidade de 0,47 e saem do sistema com probabilidade de 0,41. Depois de realizar uma pesquisa de grupo, 9 usuários têm uma alta probabilidade (0,75) de navegar em um perfil de usuário.

Os grupos 1 e 10, ilustrados nas Figuras 7 (b) e 7 (k), correspondem aos usuários que iniciam suas sessões navegando nos perfis. De fato, os usuários desses grupos apresentam perfis muito semelhantes, com a diferença de que os usuários do grupo 1 também têm alguma chance (probabilidade de 0,53) de acessar listas de vídeos mantidas pelo sistema.

Apenas uma pequena fração (menos de 1%) dos usuários inicia suas sessões listando vídeos. Esses usuários são representados pelo grupo 3 (Figura 7 (d)). Basicamente, eles iniciam suas sessões listando vídeos e depois assistem a vídeos com probabilidade de 0,43 ou continuam navegando nas listas de vídeos.

Por fim, o grupo 4 (Figura 7 (e)) exibe um perfil suspeito: alguns usuários iniciam suas sessões por *avaliando* vídeos. Esse comportamento sugere algum tipo de ação mal-intencionada ou oportunista, pois seria de se esperar que uma avaliação apareça apenas depois que o usuário assistir a pelo menos um vídeo.

6. CONCLUSÕES E TRABALHO FUTURO

Neste trabalho utilizamos uma carga de trabalho real e representativa para caracterizar os padrões de acesso de uma rede social de compartilhamento de vídeo online e estudar os perfis de navegação do usuário deste sistema. Como resultados, fornecemos vários modelos estatísticos para várias características do sistema, como popularidade de vídeos, usuários e tags, distribuições de tempo entre solicitações e sessões, etc. Nossas análises fornecem insights novos e úteis sobre o usuário de sistemas de compartilhamento de vídeo online , o que pode auxiliar no desenho da futura geração de carga de trabalho sintética, bem como impulsionar o desenvolvimento de novas infra-estruturas para este tipo de serviço.

Modelamos os padrões de navegação das sessões do usuário usando o conceito de UBMG. Usando uma técnica de agrupamento, fornecemos uma análise de diferentes perfis de usuários que acessam o sistema. Nossos resultados podem ser usados para orientar as políticas de personalização de serviço, bem como a recomendação de conteúdo para os usuários.

Como trabalhos futuros, pretendemos caracterizar novas cargas de trabalho do serviço de vídeo do UOL, incluindo aspectos de criação de conteúdo e interações sociais. Mais importante, pretendemos estudar os aspectos que influenciam a popularidade dos vídeos, que são fundamentais para um mercado emergente, a associação de anúncios a vídeos.

Reconhecimentos

Este trabalho teve apoio do Instituto Nacional de Ciência e Tecnologia para a Web (bolsa CNPq nº 573871 / 2008-6), CNPq, CAPES, Finep e Fapemig. Agradecemos também ao Universo OnLine SA - UOL (www.uol.com.br) pelos dados que disponibilizaram, que tornaram esta pesquisa possível.

REFERÊNCIAS

Arlitt, M. Caracterizando sessões de usuários da web. *Análise de Avaliação de Desempenho SIGMETRICS* 28 (2): 50–63, 2000.

Arlitt, M. e Williamson, C. Caracterização da carga de trabalho do servidor Web: a busca por invariantes. *SIGMÉTRICOS Análise de avaliação de desempenho* 24 (1): 126-137, 1996.

B. Williamson. Marketing de rede social: gastos e uso de anúncios. *EMarketer Report*, 2007. http://tinyurl.com/ 2449xx. Acessado em março / 2010.

Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., e Gonçalves, M. Detecção de spammers e conteúdo promotores em redes sociais de vídeo online. No *Proceedings of the International ACM SIGIR Conference on Research & Development of Information Retrieval*. Boston, EUA, pp. 620-627, 2009.

Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., Gonçalves, M., e Ross, K. Poluição de vídeo em a teia. *Primeira segunda-feira* 15 (4): 1-14, 2010.

Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J. e Ross, K. Interações de vídeo em vídeo social online redes. *ACM Transactions on Multimedia Computing, Communications and Applications (TOMCCAP)* 5 (4): 1-25, 2009.

Benevenuto, F., Rodrigues, T., Cha, M. e Almeida, V. Caracterizar o comportamento do usuário nas redes sociais online. No *Procedimentos da Conferência de Medição da Internet ACM SIGCOMM*. Chicago, EUA, pp. 49–62, 2009.

Bock, H. *Tarefas e métodos de mineração de dados: Classificação: o objetivo da classificação*. Oxford University Press, Inc., Nova York, NY, EUA, 2002.

Burke, M., Marlow, C. e Lento, T. Alimente-me: Motivando a contribuição de recém-chegados em sites de redes sociais. No *Anais da Conferência ACM SIGCHI sobre Fatores Humanos em Sistemas de Computação*. Boston, EUA, pp. 945-954, 2009

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y. e Moon, S. I tube, you tube, todos tubos: Analyzing the maior sistema de vídeo de conteúdo gerado pelo usuário do mundo. No *Proceedings of the ACM SIGCOMM Conference on Internet Measurement*. San Diego, EUA, pp. 1-14, 2007.

Chatterjee, P., Hoffman, DL e Novak, TP Modelando o fluxo de cliques: implicações para anúncios baseados na web tising esforços. *Ciência de Marketing* 22 (4): 520-541, 2003.

comScore, R. Americanos viram 12 bilhão videos conectados no poderia 2008 http://www.comscore.com/press/release.asp?press=2324, 2008.

Costa, C., Cunha, I., Vieira, A., Ramos, C., Rocha, M., Almeida, J., e Ribeiro-Neto, B. Analisando cliente interatividade em streaming media. No *Anais da World Wide Web Conference*. Nova York, EUA, 2004.

Duarte, F., Benevenuto, F., Almeida, V., e Almeida, J. Caracterização geográfica do YouTube: um latino visão americana. No *Anais da Conferência Latino-americana da Web*. Santiago, Chile, 2007.

Duarte, F., Mattos, B., Bestavros, A., Almeida, V., e Almeida, J. Características de tráfego e comunicação padrões na blogosfera. No *Anais da Conferência sobre Weblogs e Redes Sociais*. Boulder, EUA, 2007.

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P. e Berners-Lee, T. *RFC 2616: Protocolo de transferência de hipertexto - HTTP / 1.1*. The Internet Society, 1999.

Gill, P., Arlitt, M., Li, Z. e Mahanti, A. Caracterização do tráfego do YouTube: uma visão da borda. No *Processos da Conferência ACM SIGCOMM sobre Medição da Internet*. San Diego, EUA, pp. 15-28, 2007.

Gill, P., Arlitt, M., Li, Z. e Mahanti, A. Caracterizando sessões de usuários no YouTube. No *Processos do IEEE Multimedia Computing and Networking*. San Jose, EUA, 2008.

Jain, A., Murty, M. e Flynn, P. Agrupamento de dados: uma revisão. *Pesquisas de computação ACM* 31 (3): 264–323, 1999.

Krishnamurthy, B. Uma medida de redes sociais online. No *Atas da Conferência sobre Comunicação Sistemas e Redes*. Bagalore, Índia, 2009.

Menascé, D. e Almeida, V. *Escalonamento para E Business: tecnologias, modelos, desempenho e planejamento de capacidade*. Prentice Hall PTR, 2000.

Menascé, D., Almeida, V., Fonseca, R., e Mendes, M. Uma metodologia para caracterização da carga de trabalho de e-sites de comércio. No*Anais da conferência ACM sobre comércio eletrônico*. Denver, EUA, pp. 119-128, 1999.

Oke, A. e Bunt, R. Caracterização hierárquica da carga de trabalho para um servidor web ocupado. No*Proceedings of the International Conferência sobre Avaliação de Desempenho de Computadores, Técnicas e Ferramentas de Modelagem*. Londres, Reino Unido, 2002.

Pelleg, D. e Moore, A. X-médias: Estendendo k-médias com estimativa eficiente do número de clusters. No *Anais da Conferência Internacional sobre Aprendizado de Máquina*. Stanford, EUA, pp. 727-734, 2000.

Rodrigues, T., Benevenuto, F., Almeida, V., Almeida, J., e Gonçalves, M. Igual, mas diferente: um contexto análise de vídeos duplicados no youtube. *Revista da Sociedade Brasileira de Computação* 16 (3): 201–214, 2010.

Rodriguez, P. Infraestrutura web para o século 21. *WWW'09 Keynote, 2009*. http://tinyurl.com/mmmaa7. Acessado em março / 2010.

Trivedi, KS *Probabilidade e estatísticas com aplicativos de confiabilidade, enfileiramento e ciência da computação*. John Wiley and Sons Ltd., Chichester, Reino Unido, 2002.

Veloso, E., Almeida, V., Jr., WM, Bestavros, A., e Jin, S. Uma caracterização hierárquica de uma transmissão ao vivo carga de trabalho de mídia. *Transações IEEE / ACM na rede (TON)* 14 (1): 217–230, 2006.

Wilson, C., Boe, B., Sala, A., Puttaswamy, KPN e Zhao, BY Interações do usuário em redes sociais e suas implicações. No*Procedimentos da ACM European Professional Society on Computer Systems*. Nuremberg, Alemanha, pp. 205–218, 2009.

Witten, I. e Frank, E. *Mineração de dados: ferramentas e técnicas práticas de aprendizado de máquina*. Morgan Kaufmann, 2005.

Zink, M., Suh, K., Gu, Y. e Kurose, J. Assista global, cache local: rastreamentos de rede do YouTube em uma rede de campus - medições e implicações. No*Procedimentos do IEEE Multimedia Computing and Networking*. San Jose, EUA, 2008.