Vídeos Gerados por Usuários: Caracterização de Tráfego*

Marcelo A. Maia[1], **Virgílio A. F. Almeida**[1], **Jussara M. Almeida**[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
31270-010 – Belo Horizonte – MG – Brasil

{mmaia,virgilio,jussara}@dcc.ufmg.br

Abstract. *The growth of Internet traffic derived from the user generated videos motivated this work. We seek to understand characteristics of users and videos from four online video sharing systems. We focus on traffic origin and temporal evolution of both recently added and already popular videos. The most recent videos experience on average a number of requests on their first days comparable to some of the most popular videos, indicating that these recent videos are potential candidates to be served from cache. We propose a two-level caching strategy that exploits both popularity and recency to reduce request traffic.*

Resumo. *O crescimento do tráfego na Internet derivado dos vídeos gerados por usuários motivou este trabalho. Busca-se entender características dos usuários e dos vídeos de quatro sistemas de compartilhamento de vídeos. O foco é voltado para a origem do tráfego e para a evolução temporal de vídeos tanto recentemente adicionados quanto já populares. Os vídeos mais recentes possuem em média um número de requisições nos primeiros dias comparável ao de alguns dos mais populares, indicando que esses vídeos recentes são candidatos em potencial para serem servidos da* cache. *É feita a proposta de uma estratégia de* caching *hierárquica que explora tanto a popularidade quanto a recência dos vídeos para reduzir o tráfego das requisições.*

1. Introdução

A maioria dos serviços atualmente disponíveis na Internet exibem uma sólida tendência a incentivar a interatividade dos usuários através da formação de diferentes tipos de comunidades sociais. Freqüentemente referenciados como Web 2.0, Orkut[1], Flickr[1], YouTube[1], DailyMotion[1], Veoh Networks[1] e Videolog[2] são apenas alguns exemplos. Em particular os sistemas de compartilhamento de vídeos *online* permitem que usuários utilizem os vídeos como veículos para se expressarem. Devido a natureza intrínseca dos seres humanos, espera-se que esses vídeos ofereçam ampla gama de assuntos, tópicos, qualidade e, principalmente, velocidade e quantidade produzida.

A multifatoriedade do conteúdo multimídia gerado e seu crescente tráfego na rede [Cha et al. 2007, Gill et al. 2007] configuram um desafio interessante: como gerenciar e

*O presente trabalho foi realizado com o apoio do UOL (www.uol.com.br), através do Programa UOL Bolsa Pesquisa, processo número 20060520221328a.

[1]http://www.{orkut,flickr,youtube,dailymotion,veoh}.com (2008)

[2]http://videolog.uol.com.br (2008)

usar eficientemente os recursos dos provedores de rede e serviços. Este trabalho busca entender, além dos aspectos do comportamento dos usuários produtores dos vídeos, aspectos relacionados ao tráfego associado a esse conteúdo.

Este trabalho está dividido em duas partes. A primeira discorre sobre o maior sistema existente, o YouTube [Alexa 2008, BIL 2007], e busca entender características da participação dos usuários e do conteúdo gerado por eles, bem como o impacto em possíveis publicidades ao permitir visualizar vídeos a partir de outras páginas. A segunda parte utiliza dados de quatro sistemas: YouTube, DailyMotion, Veoh Networks e Videolog, este último sediado no Brasil. Ela busca entender as variações temporais na popularidade de um vídeo e o potencial de *caching* para reduzir o tráfego na rede.

Dentre as contribuições deste trabalho, destacam-se:

- Caracterização do comportamento do usuário gerador de conteúdo multimídia.
- Identificação de que, apesar do YouTube permitir que vídeos sejam vistos a partir de outras páginas, possíveis publicidades dentro do YouTube teriam boa visibilidade, pois mais de 90% dos vídeos possuem menos de 20% de acessos externos.
- Contraste das características dos vídeos de quatro sistemas de compartilhamento.
- Estudo da evolução temporal do tráfego dos vídeos de quatro sistemas indicando que os mais recentes recebem a maior quantidade de acessos nos primeiros dias.
- Proposição de uma estratégia de *caching* para os vídeos menos populares que se baseia na recência dos mesmos para reduzir o tráfego na rede.

O restante do trabalho está organizado como descrito a seguir. A seção 2 apresenta os trabalhos relacionados. As seções 3 e 4 mostram, respectivamente, a primeira e a segunda parte do trabalho. As conclusões e futuras direções estão expostas na seção 5.

2. Trabalhos Relacionados

Até o presente momento apenas três trabalhos fizeram estudos do tráfego de vídeos gerados na Web 2.0. Os autores em [Duarte et al. 2007] analisaram características dos vídeos e usuários de diferentes regiões geográficas, em particular a América Latina, mas eles não avaliaram a popularidade no tempo e nem sistemas de *cache*.

Em [Cha et al. 2007] os autores coletaram um total de pouco menos do que 2 milhões de vídeos de 2 categorias do YouTube. Uma das categorias foi coletada uma única vez e as demais repedidas diariamente por 6 vezes. Eles também coletaram dados do sistema coreano Daum. Os autores avaliaram estratégias de cache que agregam os vídeos mais populares e os mais recentes. O número de acertos (*cache hit*) dos mais populares sobrepuja o dos mais recentes, encobrindo os benefícios da recência.

Os autores em [Gill et al. 2007] analisaram o tráfego do YouTube sob duas perspectivas: local, na Universidade de Calgary no Canadá, e global, a partir da lista dos 100 vídeos mais vistos disponibilizada no YouTube. Entretanto eles não avaliaram a popularidade no tempo e nem sistemas de *cache*.

Este trabalho difere fundamentalmente das referências citadas. Foram avaliados, além do YouTube, outros 3 sistemas que ainda não foram estudados. Foi mostrado que possíveis anúncios exibidos dentro do YouTube teriam boa visibilidade. Além disso, foi feita uma proposta de uma estratégia de *caching* que explora tanto a popularidade quanto a recência dos vídeos e avaliamos quantitativamente mostrando o potencial dos vídeos recentes que não ficou claro em [Cha et al. 2007].

3. Entendendo Melhor o Conteúdo Gerado pelos Usuários no YouTube

O YouTube é um dos vários sistemas sociais *online* existentes para o compartilhamento de vídeos na Web 2.0. Criado em fevereiro de 2005 ele foi crescendo aos poucos até se tornar o maior sistema da atualidade [Alexa 2008, BIL 2007]. Devido à sua dimensão ele foi escolhido para as análises da primeira parte deste trabalho. Dentre as diversas características e possibilidades de interação social no YouTube destacam-se as 12 categorias em que os vídeos são classificados, os 9 diferentes tipos de ordenação (*rankings*), a possibilidade de adicionar outros usuários como amigos ou assinar os seus vídeos, selecionar um vídeo como favorito, além de permitir comentar e votar (de 1 a 5 estrelas) nos vídeos.

A próxima seção apresenta a metodologia utilizada na coleta dos dados e as seções seguintes mostram as principais características dos usuários e dos vídeos coletados.

3.1. Metodologia da coleta e sumário dos vídeos

A rede de amizades no YouTube é formada pelos usuários cadastrados (*nós da rede*) que adicionam uns aos outros em suas respectivas listas de amigos (*arestas da rede*). Um coletor (*crawler*) que realiza uma busca em largura nessa rede foi projetado. Essa estratégia, também conhecida como *snowball sampling* [Lee et al. 2006], parte de um usuário, ou de uma lista de usuários, e coleta todos os seus amigos colocando-os no fim da lista. Eventualmente os amigos desses amigos serão coletados e inseridos também no fim da lista. Esse ciclo continua até que o coletor seja interrompido ou seja esgotado o componente da rede. No caso deste trabalho a lista inicial foi composta pelos 100 usuários mais populares informados pelo YouTube. Para cada usuário descoberto na varredura da coleta foram obtidas informações sobre todos os seus vídeos enviados para o YouTube.

O YouTube não informa a quantidade de vídeos total no sistema, assim não é possível saber a fração exata coletada[3]. Entretanto, a amostra conseguida possui um número expressivo de vídeos sendo superior àquela dos trabalhos anteriores [Cha et al. 2007, Gill et al. 2007]. A tabela 1 apresenta um resumo da coleta.

Tabela 1. Resumo da coleta da rede de amizade no sistema YouTube.

Período da coleta	16/Out - 30/Out
Número total de usuários coletados	490.473
Número total de vídeos coletados	4.769.487
Número total de visualizações dos vídeos	34,19 bilhões
Média de vídeos por usuário (Coeficiente de Variação)	11,84 (3,84)
Média de visualizações por usuário (Coeficiente de Variação)	1.232 (3,22)
Estimativa de bytes transmitidos [4]	350 TB
Tempo total de exibição dos vídeos	35,95 anos

3.2. Participação dos usuários do YouTube

A tabela 2 apresenta um sumário da distribuição dos usuários com relação ao gênero, idade e localização geográfica, sendo esta última uma indicação da origem do tráfego.

[3]Uma busca de vídeo no YouTube por "*" retorna 54 milhões de vídeos (30/Out/2007). Se considerarmos esse valor como um limite inferior, a coleta realizada corresponderia a 8,83% do total de vídeos.

[4]Considerando uma média de 5 minutos por vídeo e uma taxa de codificação de 300kbps (seção 4.2), 34,19 bilhões de visualizações representam 350 TB de dados transmitidos.

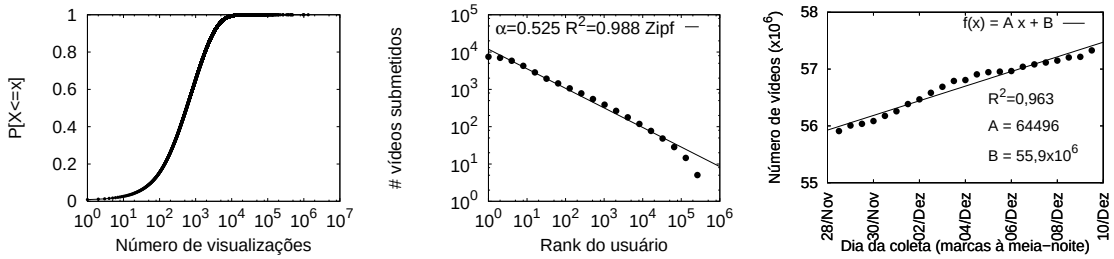
Tabela 2. Resumo dos usuários do YouTube.

Distribuição por Gênero		Distribuição por Idade		Distribuição Geográfica	
Masculino	55,12%	$<$ **20**	41,62%	**América do Norte**	48,16%
Feminino	34,00%	**[20 – 30)**	40,87%	**Europa**	17,70%
		[30 – 40)	9,43%	**Brasil**	1,16%
		$\geq$ **40**	7,42%	**Demais localidades**	10,50%

No YouTube a maioria é do sexo masculino e predominantemente jovem. Como era de se esperar a maioria dos usuários vem da América do Norte. Se comparado com o restante do mundo, o Brasil representa uma fração pequena dos usuários, entretanto o YouTube é o quinto[5] *site* mais popular da Internet brasileira.

A figura 1(a) mostra a participação dos usuários através da função de densidade acumulada (CDF) do número de visualizações de vídeos. Pode-se ver que os usuários são bastante ativos. Aproximadamente 85% deles assistiram mais do que 100 vídeos. A figura 1(b) mostra a distribuição do número de vídeos submetidos por cada usuário (*uploads*) em ordem dos mais ativos. Ela também mostra que uma distribuição que segue a lei de Zipf ($P[x] = kx^{-\alpha}$ [Zipf 1949]), com um $\alpha = 0,525$ e um corte em 10^5, representa um bom modelo (R^2=0,988). Uma reta com uma inclinação como esta indica que alguns usuários são muito mais ativos do que outros. Mais de 40% deles não submeteram nenhum vídeo e apenas quiseram assistir. Cerca de 95% dos usuários enviaram menos de 50 vídeos e apenas 2% enviaram mais do que 100 vídeos.

Apesar do YouTube não informar o total de vídeos existentes e nem a quantidade diária de novos vídeos, uma *estimativa* de ambos valores pode ser feita através de pesquisas periódicas usando o "*" como argumento na busca por vídeos. A figura 1(c) mostra os resultados obtidos diariamente entre 28/Nov/2007 e 9/Dez/2007. A figura também mostra uma regressão linear dos dados (R^2=0,963). A partir da inclinação da reta, pode-se estimar que o YouTube recebe aproximadamente 65 mil novos vídeos por dia. Essa estimativa está em consonância com o valor divulgado em [Reuters 2006].



(a) (b) (c)

Figura 1. Participação dos usuários do YouTube

3.3. Relação entre características dos vídeos e o tráfego gerado por eles

A tabela 3 mostra a distribuição dos vídeos coletados na seção 3.1 por categoria. Para cada uma, são mostradas as frações correspondentes do total de vídeos coletados e do total de

[5]Busca pelos *top sites* do Brasil mostra Orkut em 1º, UOL em 4ºe YouTube em 5º[Alexa 2008].

visualizações. O coeficiente de correlação (c) entre essas frações é de 0,93, indicando que o tráfego gerado pelos vídeos também é proporcional à sua distribuição por categoria. Distribuição dos vídeos similar também foi encontrada em [Gill et al. 2007].

Tabela 3. Distribuição da classificação por categoria dos vídeos coletados (% do total do número de vídeos - % do total do número de visualizações).

Autos & Vehicles	1,72% - 1,70%	**Music**	22,43% - 34,44%
Comedy	11,91% - 12,89%	**News & Politics**	3,80% - 2,59%
Entertainment	19,56% - 19,34%	**People & Blogs**	13,07% - 6,60%
Film & Animation	8,97% - 7,90%	**Pets & Animals**	1,74% - 1,29%
Gadgets & Games	7,93% - 5,74%	**Sports**	5,06% - 4,90%
Howto & DIY	1,65% - 1,81%	**Travel & Places**	2,16% - 0,80%

A figura 2(a) apresenta a CDF do tempo de exibição dos vídeos coletados. Cerca de 25% dos vídeos possuem duração inferior a um minuto e existem menos de 3% com duração superior a 10 minutos. Considerando um *bitrate* médio de 300kbps (fig. 6(b)), o conjunto total de vídeos da amostra coletada ocuparia um espaço da ordem de 40 TB.

A figura 2(b) mostra a distribuição do número de visualizações dos vídeos coletados em ordem dos mais vistos. Essa figura também mostra que uma distribuição que segue a lei de Zipf ($\alpha = 0,665$ e corte em 10^6) é um bom modelo. Essa lei de potência indica que alguns vídeos geram muito mais tráfego que outros, sendo eles os melhores candidatos a serem colocados em uma rede de distribuição de conteúdo (CDN). De fato, apenas 10% (20%) dos vídeos são responsáveis por 85% (93%) de todo o tráfego gerado. Dessa forma, o restante da seção analisa o que leva um vídeo a ser muito mais popular do que outro e quais as funcionalidades do sistema indicam ou contribuem para a popularidade.

A figura 3(a) apresenta a CDF do número de votos recebidos pelos vídeos coletados. Com o sistema de votação os usuários autenticados podem avaliar os vídeos através de voto numa escala de 1 a 5 estrelas que indica seu nível de satisfação. Cerca de 23% dos vídeos não possuem voto algum e apenas 3% possuem mais de 100. Um coeficiente de correlação de 0,79 entre o número de visualizações e o número de votos recebidos, pode indicar que quanto mais votos o vídeo tem (independente da quantidade de estrelas atribuída) maior será o tráfego gerado por ele. Entretanto, o número médio de estrelas recebido e o número de visualizações possuem um coeficiente de correlação de apenas 0,08, indicando que os usuários podem ser pouco exigentes com relação à qualidade teo-

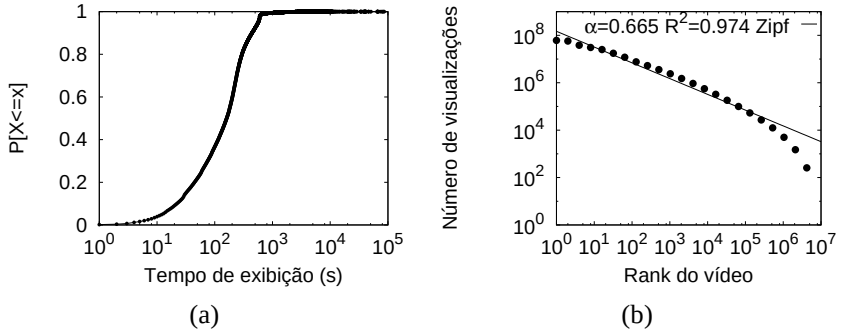
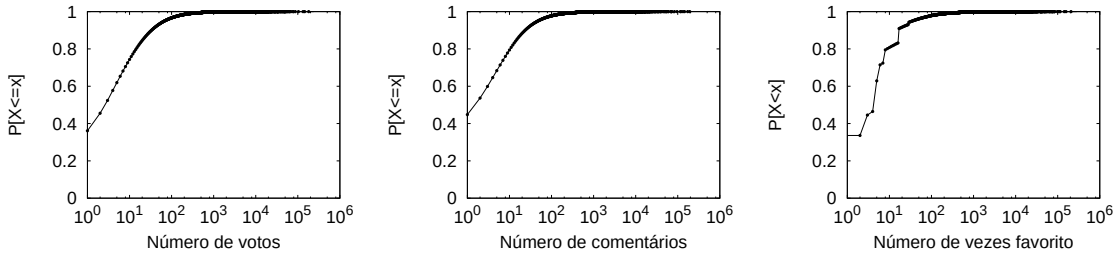


Figura 2. Tempo de exibição e distribuição do num. de visualizações dos vídeos.



(a) c = 0,79 (b) c = 0,61 (c) c = 0,48

Figura 3. Distribuição do número de votos, comentários e vezes que o vídeo foi selecionado como favorito e suas correlações com o número de visualizações.

ricamente representada por um número médio maior de estrelas.

A figura 3(b) mostra a CDF do número de comentários recebidos pelos vídeos coletados. Cerca de 32% dos vídeos não foram comentados e apenas 2% possuem mais do que 100 comentários. Um coeficiente de correlação de 0,61 entre o número de visualizações e o número de comentários pode indicar que os vídeos mais polêmicos têm maiores chances de gerar tráfego. Esse coeficiente pode ser ainda maior, pois a confiabilidade dos usuários nesses comentários pode ter sido comprometida devido à dificuldade de se reconhecer *spam* em sistemas sociais como o YouTube [Joshi et al. 2007].

A figura 3(c) mostra a CDF do número de vezes que o vídeo coletado foi selecionado como favorito. Cerca de 29% dos vídeos nunca foram selecionados e cerca de 2% apenas o foram mais de 100 vezes. Ao inserir um vídeo de sua preferência na lista de favoritos, o usuário facilita o seu acesso a esse vídeo. Um coeficiente de correlação de 0,48 entre o número de visualizações e o número de vezes que o vídeo foi selecionado como favorito pode indicar que as listas de favoritos podem ser usadas como uma referência rápida para inúmeros acessos futuros e conseqüentemente gerar mais tráfego.

Todas as três interações mencionadas anteriormente somente são possíveis mediante a autenticação dos usuários no sistema. Um usuário que entra na página do YouTube, assiste a alguns vídeos e vai embora, fica impossibilitado de interagir. Outro fator que limita a participação dos usuários é o fato do YouTube disponibilizar um código HTML que permite a visualização de vídeos a partir de outras páginas mas não permite que, por exemplo, um comentário seja adicionado. Dessa forma, os três coeficientes de correlação citados poderiam ser ainda maiores se não houvesse a necessidade da autenticação.

3.4. Tráfego dos vídeos gerado pelos acessos que vêm de fora do YouTube

Esta seção analisa o tráfego gerado pelos vídeos embutidos e visualizados em outras páginas da Internet. São os chamados *embedded links*. Uma vez que os vídeos podem ser vistos a partir de outras páginas populares como MySpace ou Orkut grande parte do número de visualizações dos vídeos pode vir dessas comunidades e não ter relação com as funcionalidades oferecidas pelo YouTube. Um vídeo pode inclusive receber visibilidade exclusivamente porque foi embutido em uma página popular e externa ao YouTube.

O YouTube mostra atualmente para cada vídeo que possui algum acesso externo apenas as 5 maiores origens de requisições. Para esses vídeos são disponibilizadas URL's que representam a origem do acesso. Essas URL's foram usadas para fazer um casamento

(*match*) com expressões como “myspace.com” ou “orkut.com” e assim contabilizar a quantidade de visualizações que cada vídeo recebeu vinda de cada origem.

Os resultados indicam que 44,29% de todos os vídeos coletados possuem pelo menos 1 acesso vindo de fora do YouTube. De todas as visualizações recebidas pelos vídeos coletados, os acessos externos representam 2,48% do total. Esses valores estão em consonância com os encontrados em [Cha et al. 2007]. A tabela 4 apresenta os resultados encontrados para algumas das principais origens de acesso.

Tabela 4. Resumo das principais origens de acessos externos do YouTube.

***Site* de origem**	**% do total de visualizações**	**% do total das origens**
MySpace	0,422%	17,052%
Blogspot	0,053%	2,129%
Friendster	0,026%	1,042%
Orkut	0,012%	0,498%
Total das comunidades[6]	0,655%	26,426%
Google	0,050%	2,003%
Yahoo	0,009%	0,365%

Apesar dos acessos externos representarem uma fração pequena (2,48%) do total de visualizações, ela não é desprezível. O gráfico mais externo da figura 4 mostra a fração dos acessos externos para todos os vídeos. Lembrando que para cada vídeo o YouTube disponibiliza apenas as 5 maiores fontes de requisições, ou seja, essa curva representa um limite inferior, uma vez que podem haver mais acessos. Mais de 90% dos vídeos possuem menos de 20% dos acessos externos. Isso indica que a maioria realmente é descoberta e visualizada de dentro do YouTube, o que oferece boa visibilidade para possíveis publicidades a serem exibidas dentro dele. Entretanto, existem alguns poucos vídeos (0,4%) que a fração de acessos externos é superior a 50%, ou seja, mais da metade das visualizações são provenientes de outros ambientes. Para essa pequena fração dos vídeos, esforços de publicidade dentro do YouTube não seriam eficientes.

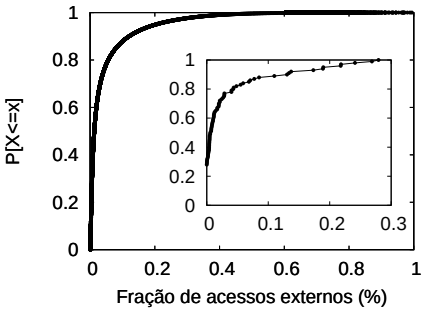


Figura 4. Fração de acessos externos de todos os vídeos e dos mais populares.

O gráfico mais interno da figura 4 mostra a mesma fração dos acessos externos apenas para os 100 vídeos mais vistos. Embutir o vídeo em uma página popular fora do

[6]Incluídas as comunidades: MySpace, Blogspot, Friendster, Orkut, Hi5, Multiply, Flixter, Facebook, SkyBlog, SkyRock, Twitter, Netlog, Stumbleupon, Tagged, Bebo, Livejournal, Dabble, GaiaOnline e Ilike.

YouTube pode significar algumas visualizações a mais e, apesar da maioria dos acessos vir de dentro do YouTube, alguns vídeos tiraram proveito dos acessos externos para ajudá-los a se tornar populares. Dos 100 vídeos mais vistos, 28 não têm acesso externo, mas 20 vídeos possuem de 4,3% até 27,9% das visualizações vindas de fora.

4. A Evolução Temporal do Tráfego dos Vídeos

Nesta seção busca-se entender a evolução temporal do tráfego dos vídeos gerados pelos usuários da Web 2.0, aqui representada por quatro diferentes sistemas, e mostrar o potencial de *caching* para minimizar o tráfego na rede. A próxima seção apresenta a metodologia utilizada na monitoração dos vídeos e as seções seguintes, os resultados.

4.1. Metodologia de monitoração e sumário dos vídeos monitorados

Para realizar o estudo temporal do tráfego dos vídeos foram utilizados, além do YouTube, outros 3 sistemas *online* de compartilhamento de vídeo da Web 2.0: DailyMotion, bastante popular na Europa [Alexa 2008], Veoh Networks, com presença asiática marcante [Alexa 2008], e Videolog, o sistema sediado no Brasil mais acessado [Alexa 2008]. Todos apresentam funcionalidades similares às descritas para o YouTube e usam na transmissão do vídeo o Adobe Flash Video (FLV) com a tecnologia de *download progressivo*[7].

No dia 20 de outubro de 2007 foram obtidas 8 listas de vídeos, os mais populares e os mais recentemente inseridos em cada um dos 4 sistemas. Os mais recentes foram usados para representar parte dos menos populares. Cada lista foi composta de cerca de 1000 vídeos, exceto os mais populares do YouTube. No caso específico do YouTube, ele apenas disponibiliza os 100 mais populares e 100 mais recentes. Entretanto, uma lista maior dos mais recentes pode ser obtida fazendo consultas sucessivas em curtos intervalos de tempo, suficientes apenas para que o sistema seja atualizado com novos vídeos.

Foram monitorados, para cada vídeo em cada lista, o número de visualizações, comentário, votos e vezes em que ele foi considerado favorito por algum usuário. A monitoração foi realizada durante 24 dias. Alguns vídeos, geralmente os mais recentes, não puderam ser acompanhados todos os dias, e portanto foram removidos de sua lista. Eles tiveram suas informações bloqueadas ou não disponibilizadas devido a problemas de violação dos termos de uso do sistema ou remoção por parte do próprio dono. A tabela 5 apresenta o número de vídeos monitorados com sucesso em cada lista, para cada sistema.

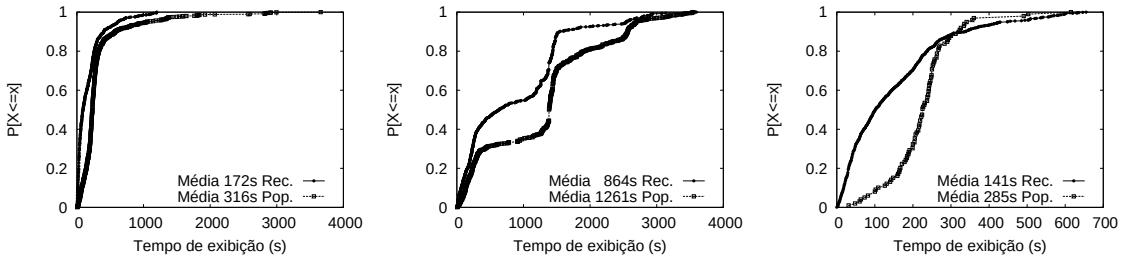
Tabela 5. Número de vídeos recentes e populares monitorados com sucesso.

Sistema	DailyMotion		Veoh		Videolog		YouTube	
	Rec.	Pop.	Rec.	Pop.	Rec.	Pop.	Rec.	Pop.
# Vídeos coletados	969	970	818	865	898	953	791	97

4.2. Características dos vídeos monitorados

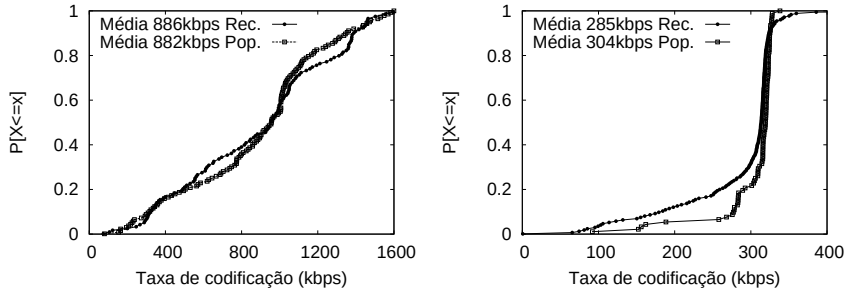
Para cada sistema analisado, as figuras 5(a-c) mostram a CDF do tempo de exibição dos vídeos monitorados. Essa informação não está disponível no Videolog, pois a duração do vídeo está embutida no arquivo FLV usado. As figuras 6(a-b) mostram a CDF da taxa de codificação média (*bitrate*) dos vídeos mais recentes e mais populares monitorados

[7] Adobe Developer Center - http://www.adobe.com/devnet/flash/articles/flv_download.html (2008)



(a) DailyMotion (b) Veoh (c) YouTube

Figura 5. Tempo de exibição dos vídeos monitorados.



(a) Veoh (b) YouTube

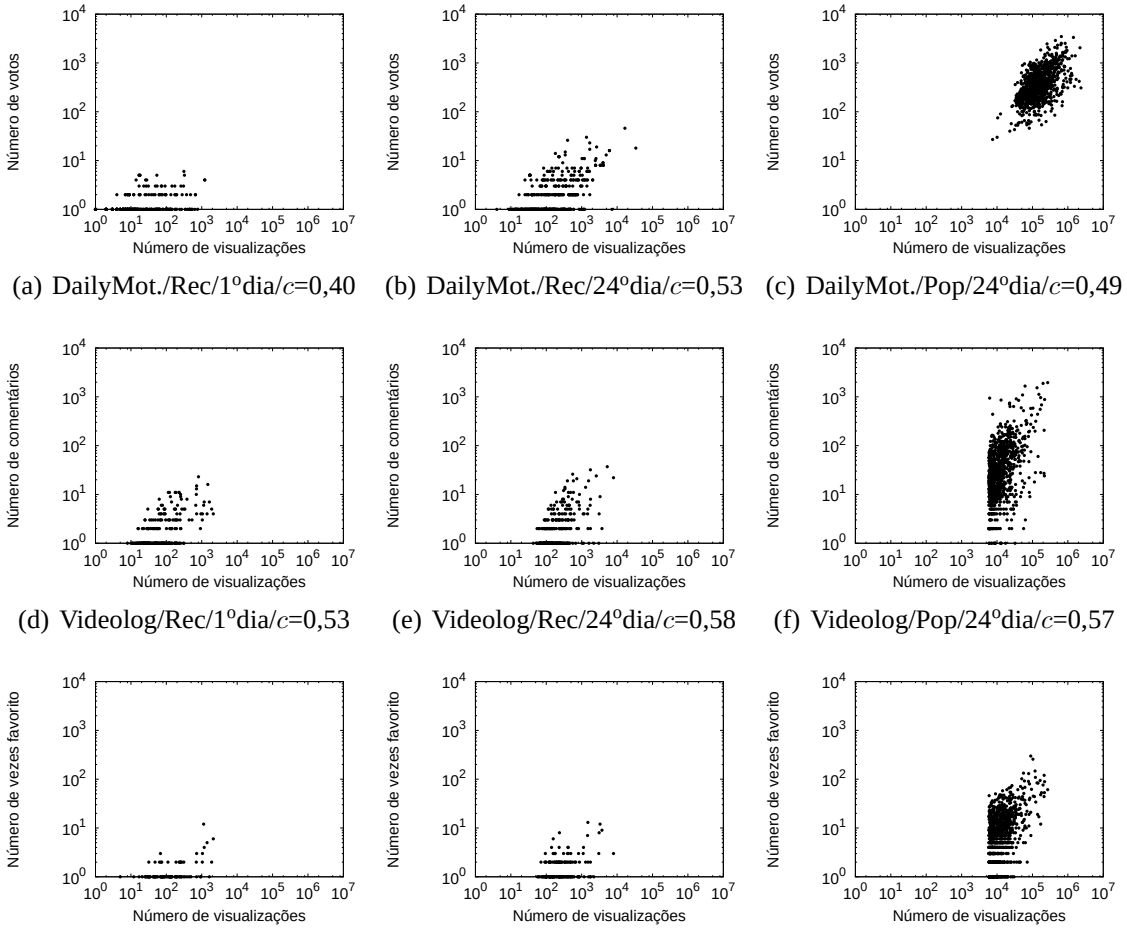
Figura 6. Taxa de codificação dos vídeos monitorados.

do Veoh e do YouTube (DailyMotion e Videolog não disponibilizam essa informação). Em média, os vídeos mais recentes do YouTube possuem uma duração e uma taxa de codificação menores do que os mais populares, sugerindo que um vídeo mais recente gera menos tráfego na rede do que um mais popular. Mesma observação pode ser feita com relação ao Veoh, exceto pelas taxas de codificação médias que são equivalentes. Para a visualização dos vídeos, os requisitos de largura de banda exigidos pelo YouTube são menos rígidos do que os exigidos pelo Veoh.

4.3. Evolução temporal da interatividade dos usuários através dos vídeos

Um usuário pode interagir com outros utilizando os vídeos como veículo. Quando o usuário adiciona um vídeo como favorito ele demonstra seu interesse, assim como quando deixa um comentário na página de um vídeo. Os usuários ainda podem interagir com o sistema e expressar suas opiniões ao votar nos vídeos. Esta seção busca caracterizar o nível de interatividade dos usuários através dos vídeos à medida que o tempo passa.

As figuras 7(a-i) mostram para dois diferentes sistemas e para os vídeos mais recentes e populares a correlação entre o número de visualizações dos vídeos e o número de votos, comentários e vezes em que o vídeo foi selecionado por um usuário como favorito. Para os vídeos mais recentes são mostradas as correlações no primeiro e no último dia monitorado. Já para os vídeos mais populares, são mostrados os gráficos apenas relativos ao último dia, pois quantitativamente as correlações permaneceram similares durante todo o período monitorado. À medida que os vídeos mais recentes ganham popularidade, qualitativamente as correlações tendem a se aproximar dos valores dos mais populares.



(a) DailyMot./Rec/1ºdia/c=0,40 (b) DailyMot./Rec/24ºdia/c=0,53 (c) DailyMot./Pop/24ºdia/c=0,49

(d) Videolog/Rec/1ºdia/c=0,53 (e) Videolog/Rec/24ºdia/c=0,58 (f) Videolog/Pop/24ºdia/c=0,57

(g) Videolog/Rec/1ºdia/c=0,58 (h) Videolog/Rec/24ºdia/c=0,46 (i) Videolog/Pop/24ºdia/c=0,50

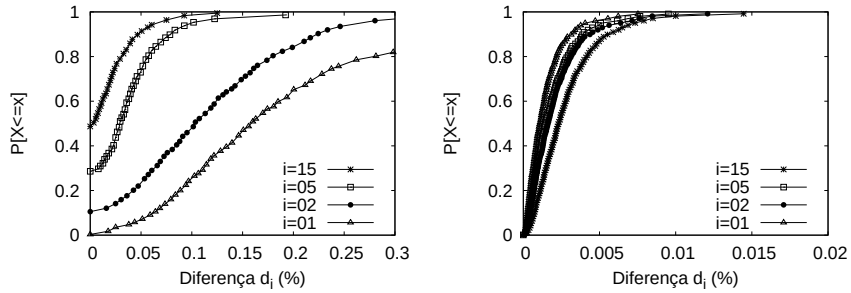
Figura 7. Correlação entre o número de visualizações e diferentes atributos dos vídeos (sistema/recentes ou populares/dia da coleta/coeficiente de correlação).

4.4. Avaliação temporal do número de visualizações dos vídeos

Esta seção mostra um estudo da evolução temporal do tráfego gerado pelos vídeos. Como principal resultado tem-se que os vídeos mais recentes possuem maior número de acessos nos primeiros dias, indicando um potencial para *caching* avaliado na próxima seção.

Considere n o número total de dias monitorados, v_i como sendo o número de visualizações de um determinado vídeo no dia i. O valor $d_i = \frac{v_{i+1}-v_i}{v_n}, (1 \leq i < n)$ representa o percentual, com relação ao último dia monitorado, da diferença do número de visualizações de um vídeo em dois dias subseqüentes.

As figuras 8(a-b) mostram a CDF da diferença d_i em 4 dias diferentes para os vídeos mais recentes e mais populares no sistema DailyMotion. Possíveis interpretações da figura 8(a) seriam que com o passar dos dias a fração de vídeos que mantêm o *mesmo* número de visualizações do dia anterior ($d_i = 0$) aumenta (0%, 10%, 29% e 49% para os dias 1, 2, 5 e 15, respectivamente), ou seja, alguns vídeos *param* de gerar tráfego. Para outros valores de d_i, como 10% ou 20%, o número de vídeos que apresentam essas variações aumenta, indicando que os vídeos geram *menos* tráfego com o passar dos dias. No caso dos vídeos mais populares a *diferença* d_i de um dia para outro, se comparada

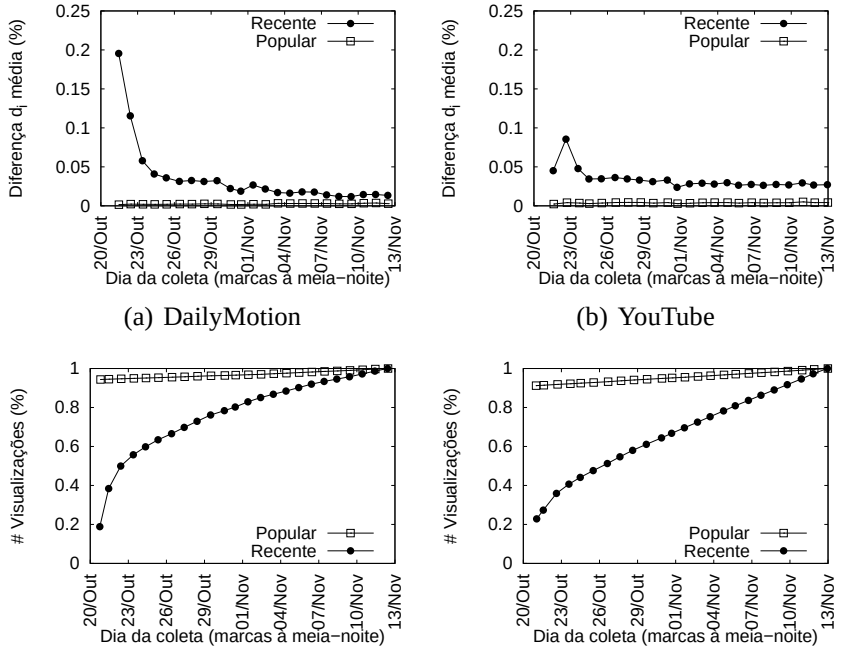


(a) Vídeos mais recentes (b) Vídeos mais populares

Figura 8. Variação de d_i em 4 dias de monitoração diferentes no sistema DailyMotion.

com o grande total de visualizações já recebido, é pequena e constante (Note que se trata da diferença diária e não do número absoluto de visualizações). Um comportamento qualitativamente semelhante também foi observado para os vídeos mais recentes e mais populares nos demais sistemas analisados tanto para o número de visualizações quanto para o número de comentários, votos e vezes em que o vídeo foi considerado favorito.

As figuras 9(a-b) mostram a média diária da diferença d_i dos vídeos dos sistemas DailyMotion e YouTube, respectivamente. Pode-se ver que a diferença é maior nos primeiros dias e após o quarto ou quinto dia tende a ter um decaimento linear. No caso do YouTube houve um aumento da diferença do primeiro para o segundo dia, sugerindo que os vídeos podem gastar algum tempo para ganhar visibilidade antes de atingirem o auge. Para os vídeos mais populares a *diferença* (e não o número absoluto) é constante e baixa.



(a) DailyMotion (b) YouTube

(c) DailyMotion (d) YouTube

Figura 9. Evolução da diferença d_i média de todos os vídeos e da porcentagem do número de visualizações médio durante todo o período monitorado.

As figuras 9(c-d) apresentam os mesmos dados, porém mostrados como a porcentagem acumulada do número de visualizações em relação ao total no último dia monitorado. Os sete primeiros dias representam mais de 50% do total de visualizações obtidas durante os 24 dias monitorados e após essa primeira semana o crescimento segue linear.

Os resultados encontrados estão de acordo com [Cha et al. 2007], onde foi mostrado que é improvável que um vídeo antigo e não popular receba uma repentina rajada de visualizações. Além disso, se a visibilidade recebida nos primeiros dias não foi suficiente para que ele aparecesse na lista dos mais populares é provável que ele nunca o faça. Motivado por esses resultados, a próxima seção mostra o potencial de economia de banda de rede se esses vídeos recentes forem mantidos em *cache* durante os primeiros dias.

4.5. Estratégia hierárquica de *caching* para minimizar o tráfego de requisições

Na seção 3.3 foi mostrado que a distribuição do número de visualizações dos vídeos coletados no YouTube segue uma lei de potência (fig. 2(b)). Apenas 10% (20%) dos vídeos são responsáveis por 85% (93%) de todo o tráfego gerado. Esses vídeos mais populares são os melhores candidatos a serem distribuídos através de uma CDN e, de fato, é o que o YouTube faz [GSCS 2008]. O restante dos vídeos, os menos populares e que correspondem à cauda da distribuição, são atendidos diretamente pelos servidores do YouTube. Segundo o engenheiro chefe do YouTube [GSCS 2008], este é um dos principais problemas enfrentados por eles no planejamento do *cache* dos servidores. As requisições para esses vídeos menos populares são de caráter aleatório e o custo para recuperar os vários vídeos dos discos é muito alto. Sendo assim, o uso eficiente de recursos dos servidores como memória e banda de rede e disco depende fundamentalmente da estratégia de escolha dos vídeos que serão colocados e servidos a partir da *cache*.

A tabela 6 mostra o número médio de visualizações recebidas pelos vídeos mais recentes de cada sistema em cada um dos 4 primeiros dias. Esses correspondem aos dias com o maior número de requisições (fig. 9). A tabela também mostra para cada sistema o número médio de visualizações recebidas por dia durante o mesmo período de 4 dias dos vídeos mais populares que ocupam as 10 primeiras e as 10 últimas posições do *ranking* formado pelos vídeos monitorados (tab. 5). Foi feita uma média do número de requisições do período para os vídeos mais populares porque este valor não varia significativamente (fig. 9). Em três sistemas o número médio de requisições recebidas nos primeiros dias dos vídeos mais recentes é da mesma ordem de grandeza (e às vezes até maior) que alguns dos vídeos dentre os mais populares monitorados (aqueles que ocupavam as últimas posições do *ranking*). Assumindo que os vídeos mais populares monitorados seriam distribuídos por uma CDN, uma fração significativa das requisições que chegam aos servidores dos

Tabela 6. Número médio de visualizações recebidas nos 4 primeiros dias.

	Vídeos mais recentes				**Vídeos mais populares**	
Sistema	**1ºdia**	**2ºdia**	**3ºdia**	**4ºdia**	**Primeiras pos.**	**Últimas pos.**
DailyMotion	30,1	66,8	38,3	17,0	897,2	54,3
Veoh	31,7	48,4	38,8	36,1	3116,8	35,5
Videolog	7,9	10,6	8,2	7,6	84,1	5,7
YouTube	11,9	37,7	19,2	11,6	112542	25573

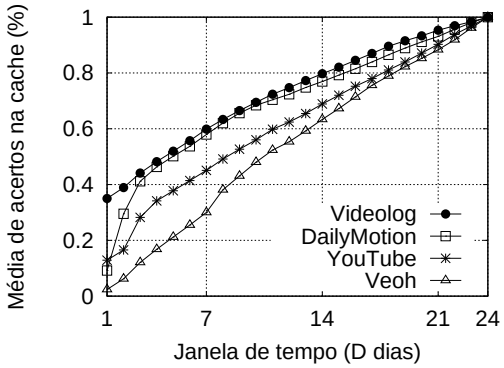


Figura 10. Média de acertos na *cache* do nível baseado na recência.

sistemas seriam para os vídeos mais recentes, motivando o desenvolvimento de uma estratégia de *caching* que leva em consideração essa grande quantidade de requisições.

Esta seção propõe uma estratégia hierárquica de *caching*. Um nível possui a mesma política de troca da CDN baseada na freqüência para capturar os mais populares. O outro nível baseia-se na recência e captura o grande número de visualizações dos primeiros dias. Após uma janela de tempo de D dias, os vídeos são retirados da *cache* ou, caso consigam um número de visualizações alto o suficiente, são promovidos para a *cache* baseada na freqüência. Os vídeos permanecem assim até que outros superem seu número de visualizações forçando sua retirada da *cache* ou ele receba um número expressivo de visualizações que justifique a sua promoção para o conjunto de vídeos da CDN.

Foi feita uma simulação do primeiro nível da *cache* com os 4 sistemas onde os vídeos mais recentes foram mantidos em uma *cache* infinita por uma janela de tempo de D dias, contados a partir da data de inclusão. Para cada sistema, a figura 10 mostra o percentual de acertos (*hits*) variando o tamanho da janela de tempo. Se os vídeos da última semana dos sistemas Veoh, YouTube, DailyMotion e Videolog fossem mantidos em *cache*, 30%, 45%, 58% e 60% das requisições para esses vídeos mais recentes, no período dos 24 dias monitorados, seriam atendidas da *cache*, respectivamente.

Considerando um tamanho médio de arquivo de 10MB e a adição diária de 65 mil novos vídeos (fig 1(c)), o maior sistema, o YouTube, necessitaria de menos de 5 TB para manter os vídeos dos últimos 7 dias em *cache*, ou seja, um investimento de baixo custo [Gray 2003] se comparado, por exemplo, ao potencial para economia de banda de rede.

5. Conclusões e Trabalhos Futuros

A interatividade por parte dos usuários, em oposição à geração de conteúdo profissional de algumas poucas grandes empresas, é a chave para o sucesso de vários serviços da Web 2.0 atual. Por conta do significativo crescimento na produção de vídeos amadores e conseqüente aumento do tráfego gerado por eles, esse trabalho buscou entender algumas características desse novo paradigma como a participação dos usuários, origem e evolução temporal do tráfego. Para isso foram utilizados dados de quatro sistemas: DailyMotion, Veoh Networks, Videolog (sediado no Brasil) e YouTube, o maior sistema existente.

O trabalho foi dividido em duas partes. A primeira mostrou características dos usuários do YouTube e dos vídeos gerados por eles. Foi mostrado que o número de

visualizações dos vídeos segue uma lei de potência e também quais as funcionalidades do sistema que contribuem para a popularidade do vídeo. Outro resultado relevante é que possíveis publicidades dentro do YouTube teriam boa visibilidade, pois, apesar da tecnologia permitir assistir aos vídeos fora do YouTube, a maior parte das visualizações ocorrem mesmo é dentro do próprio sistema.

A segunda parte utilizou dados dos quatro sistemas e buscou entender a evolução temporal do tráfego gerado pelos vídeos. Foi verificado que os primeiros dias dos vídeos recentemente incorporados ao sistema são responsáveis por uma quantidade de requisições da mesma ordem de grandeza de alguns dos vídeos dentre os mais populares coletados e servidos pela CDN. Esse resultado motivou o desenvolvimento de uma estratégia hierárquica de *caching* que levasse em consideração também a recência dos vídeos. Foi mostrado que manter os vídeos da última semana em *cache* é capaz de atender uma parcela significativa de requisições e representa um investimento de baixo custo, se comparado com a economia de banda de rede.

Informações das redes sociais formadas pelos usuários podem ser futuramente usadas em conjunto à freqüência e à recência para aprimorar a estratégia de *caching*.

Referências

Alexa (2008). Alexa Web Search. `http://www.alexa.com`.

BIL (2007). Business Intelligence Lowdown: Top 10 Largest Databases in the World.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., and Moon, S. (2007). I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *Proc. ACM Internet Measurement Conference (IMC)*, San Diego, CA, USA.

Duarte, F., Benevenuto, F., Almeida, V., and Almeida, J. (2007). Geographical Characterization of YouTube: A Latin American View. In *Proc. Latin American Web Conf.*, Washington, DC, USA.

Gill, P., Arlitt, M., Li, Z., and Mahanti, A. (2007). Youtube Traffic Characterization: A View From the Edge. In *Proc. ACM Internet Measurement Conference (IMC)*, San Diego, CA, USA.

Gray, J. (2003). Distributed Computing Economics. Technical Report MSR-TR-2003-24.

GSCS (2008). Google Seattle Conference on Scalability. `http://video.google.com/videoplay?docid=-6304964351441328559`.

Joshi, A., Finin, T., Java, A., Kale, A., and Kolari, P. (2007). Web (2.0) Mining: Analyzing Social Media. In *Proc. NSF Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation (NGDM)*, Baltimore, MD, USA.

Lee, S. H., Kim, P.-J., and Jeong, H. (2006). Statistical Properties of Sampled Networks. *Physical Review E*, 73:016102.

Reuters (2006). Usa Today: YouTube serves up 100 million videos a day online - San Francisco (Reuters). `http://www.usatoday.com`.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Reading MA (USA).