CARACTERIZAÇÃO DE PADRÕES DE ACESSO A VÍDEOS EM PORTAIS DE MÍDIA ESPECIALIZADA

LUCAS CUNHA DE OLIVEIRA MIRANDA

CARACTERIZAÇÃO DE PADRÕES DE ACESSO A VÍDEOS EM PORTAIS DE MÍDIA ESPECIALIZADA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Alberto H. F. Laender
Coorientador: Rodrygo Luis Teodoro Santos

Belo Horizonte
Agosto de 2013

Miranda, Lucas Cunha de Oliveira

M672c Caracterização de Padrões de Acesso a Vídeos em Portais de Mídia Especializada / Lucas Cunha de Oliveira Miranda. — Belo Horizonte, 2013
xx, 80 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Alberto H. F. Laender

1. Computação - Teses. 2. Mineração de Dados - Teses. I. Orientador. II. Título.

CDU 519.6*72(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Caracterização de Padrões de Acesso a Vídeos em Portais de Mídia Especializada

LUCAS CUNHA DE OLIVEIRA MIRANDA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. ALBERTO HENRIQUE FRADE LAENDER - Orientador
Departamento de Ciência da Computação - UFMG

PROF. RODRYGO LUIS TEODORO SANTOS - Coorientador
Departamento de Ciência da Computação - UFMG

PROF. ARTUR ZIVIANI
Laboratório Nacional de Computação Científica - CNPq

PROF. DORGIVAL OLAVO GUEDES NETO
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 23 de agosto de 2013.

Dedico este trabalho a todos aqueles que, de alguma forma, colaboraram para sua realização.

"Success is stumbling from failure to failure with no loss of enthusiasm."
(Winston Churchill)

Resumo

Assistir a vídeos online é parte da rotina diária de uma fração considerável de usuários da Internet. Temos presenciado uma expansão de aplicações mais interativas, sociais e colaborativas na Web. Com isso, a produção e o consumo de conteúdos multimídia aumentou significativamente nos últimos anos. Esse aumento vem ocorrendo, principalmente, no Brasil, que hoje figura entre os dez maiores mercados de vídeos online do mundo. Nesse contexto, compreender os padrões de consumo de vídeos na Web é fundamental para a melhoria do planejamento de capacidade para os provedores de vídeos, da taxa de conversão para os anunciantes e da experiência de uso dos serviços e dos conteúdos providos ao usuário final. Enquanto muita pesquisa tem sido conduzida no sentido de analisar os padrões de acesso de vídeos de conteúdo gerado por usuário (CGU), pouco se sabe sobre como esses padrões se manifestam em sites de mídia especializada (ME). Neste trabalho, realizamos a primeira análise em larga escala de padrões de acesso de vídeos em sites de ME. Como estudo de caso, investigamos os registros de interação de usuários assistindo a vídeos em um total de 38 sites brasileiros de mídia especializada, incluindo seis dos maiores portais do país, durante um período de oito semanas. Nossas análises revelaram padrões de acesso estáticos, temporais e transacionais interessantes, que foram comparados e contrastados com padrões relatados para sites de CGU. No geral, os resultados obtidos fornecem discernimento para uma melhor compreensão do consumo de vídeos na Internet, além do contexto de sites de conteúdo gerado por usuário.

Palavras-chave: Web, Caracterização, Vídeos Online, Padrões de Acesso, Conteúdo Gerado por Usuário, Mídia Especializada.

Abstract

Watching online videos is part of the daily routine of a considerable fraction of Internet users nowadays. We have seen an expansion of more interactive, social and collaborative applications on the Web. Therefore, production and consumption of multimedia content increased significantly in recent years. This increase has occurred mainly in Brazil, which today ranks among the top ten online video markets worldwide. In this context, understanding the patterns of video consumption on the Web is paramount for improving the capacity planning for video providers, the conversion rate for advertisers, and the relevance of the whole online video watching experience for end users. While much research has been conducted to analyze video access patterns in user-generated content (UGC), little is known of how such patterns manifest in mainstream media (MSM) sites. In this work, we perform the first large-scale analysis of video access patterns in MSM portals. As a case study, we analyze interaction logs across a total of 38 Brazilian MSM Websites, including six of the largest portals in the country, over a period of eight weeks. Our analysis revealed interesting static, temporal and transactional video access patterns, which we compare and contrast to the access patterns reported for UGC Websites. Overall, our analysis provides several insights for an improved understanding of video access on the Internet beyond UGC Websites.

Keywords: Web, Characterization, Online Video, Access Patterns, User-Generated Content, Mainstream Media.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Vídeos *online* estão mudando o modo como as pessoas interagem e colaboram na Web. Atualmente, usuários podem compartilhar uma gravação em alta definição, de um dispositivo móvel conectado à Internet, com milhões de internautas. O uso de mídias ricas em aplicações diversas na Internet possibilita que usuários experimentem novas formas de comunicação, interação e troca de conhecimento.

O aprimoramento e a disponibilidade dos meios de transmissão em banda larga permitiu a expansão de aplicações multimídia na Web, tais como comunicação em tempo real (*VoIP*), transmissões de mídias *online* (*Streaming Media*), jogos em rede, dentre outras. Além disso, temos presenciado uma mudança importante no papel dos usuários da Web. Os internautas, que antes eram apenas consumidores de informação, estão se tornando colaboradores ativos de conteúdo. Essa mudança de paradigmas é o que representa o conceito de Web 2.0 [o'Reilly, 2005].

A Web tem se tornado mais interativa, social, dinâmica e colaborativa. Neste cenário moderno, o vídeo passou a ser um meio valioso de transmissão de conteúdo. São muitas as aplicações na Internet que adotam vídeos *online*: promoção de serviços e produtos, campanhas publicitárias, cursos a distância, vídeo-conferências, etc. Em termos de negócios, vídeos *online* representam uma oportunidade valiosa de tornar o compartilhamento de informações mais atrativo e acessível à audiência desejada.

Notoriamente, a publicação e o consumo de vídeos *online* aumentou significativamente nos últimos anos. Segundo uma pesquisa publicada em Junho de 2010 [Purcell, 2010], 69% de todos os internautas assistem vídeos na Web, enquanto 14% já fizeram algum compartilhamento. A Figura 1.1 traz um resumo de estatísticas mais recentes (2012) do consumo de vídeos *online* [Samba-Tech, 2013]. Números como esses (456,6M de vídeos foram assistidos em 2012 e 56% do tráfego da Web é referente ao consumo de vídeos *online*) ressaltam a relevância desse tipo de mídia na Web atual.

Figura 1.1. Estatísticas sobre o consumo de vídeos *online* em 2012.

O aumento no consumo de vídeos *online* se deu acompanhado também por uma mudança comportamental dos internautas. Em particular, a Figura 1.2 reporta, em números, a evolução no consumo de vídeos por telespectadores entre os anos de 2004 e 2012 [Samba-Tech, 2013]. Os usuários estão deixando de acompanhar programas exibidos na TV ao vivo (redução de 82%, em 2004, para 64%, em 2012) e assistindo mais vídeos transmitidos *online* (a categoria *streaming* de vídeo surge com 7% em 2012).

O aumento no consumo de vídeos *online* é uma tendência mundial. O Brasil, especificamente, figura entre os dez maiores mercados de vídeos *online* do mundo. Segundo uma pesquisa recente, 43 milhões de brasileiros consumiram vídeos *online* em dezembro de 2012 [ComScore, 2013]. Além disso, os usuários que acessam vídeos *online* no Brasil constituem 82% dos internautas do país. A Tabela 1.1 detalha a classificação dos países com maior representatividade em termos de consumo de vídeos *online*. O Brasil ocupa a sétima colocação.

Vídeos *online* vêm permeando diversas aplicações. Essa tendência é reforçada pela expansão dos dispositivos móveis (*smartphones*, *tablets*, etc) e pela fusão de serviços de TV e Internet (*Google TV*, *Apple TV*, *Netflix*, etc.). Os vídeos *online*, independente do contexto de aplicação, podem ser classificados considerando como critério o provedor de conteúdo [Cha et al., 2007]. Assim, duas possíveis categorias são:

Figura 1.2. Evolução do comportamento dos usuários no consumo de vídeos *online*.

	Usuários únicos	**% de internautas**
Mundo	1.279.264	83,8%
China	289.890	84,3%
EUA	188.130	84,9%
Japão	60.939	82,8%
Rússia	55.591	90,6%
Índia	51.718	73,1%
Alemanha	47.617	82,9%
Brasil	**42,998**	**82,2%**
França	40.662	84,6%
Inglaterra	37.477	83,6%
Itália	23.857	83,0%

Tabela 1.1. Os dez maiores mercados de vídeos *online*.

- "Mídia Especializada" (ME) ou *Mainstream Media* é a categoria de vídeos que foram publicados por provedores especializados de conteúdo (como portais de notícia e canais de entretenimento) em seus próprios *sites*. Em geral, os vídeos tratam de assuntos específicos ou fazem parte de canais (categorias) bem

definidos. Além disso, a produção das mídias segue, em geral, procedimentos profissionais ou atendem a um padrão de qualidade. São exemplos de provedores de vídeos do tipo ME: UOL[1] e CNN[2].

- "Conteúdo Gerado por Usuário" (CGU) ou *User-Generated Content* inclui vídeos compartilhados por usuários autônomos da Web, usualmente por meio de um *site* de compartilhamento ou através de redes sociais. Os vídeos, em geral, apresentam temas diversos e padrões de qualidade variados. Além disso, grande parte dos vídeos é produzida utilizando câmeras de dispositivos móveis. São exemplos de *sites* para compartilhamento de mídias do tipo CGU: YouTube[3] e Vimeo[4]. Redes sociais, como o Facebook[5], também permitem que usuários compartilhem seus vídeos.

A divisão proposta é apenas uma forma de classificação, dentre as muitas possibilidades de categorizar os vídeos *online*. As categorias sugeridas não devem ser tomadas como únicas e restritas.

Para que mídias do tipo ME ou CGU possam ser compartilhadas na Web e fiquem disponíveis para serem acessadas por um número grande de usuários, são necessários mecanismos especializados que permitam o funcionamento de todas as fases do processo, desde a captura do vídeo até a distribuição para o usuário final. Esses mecanismos envolvem sistemas complexos, infra-estruturas robustas e escaláveis e técnicas modernas para lidar com os vários aspectos envolvidos, tais como distribuição, armazenamento, disponibilidade e segurança.

Existem empresas especializadas em lidar com toda essa demanda exigida pelo processo de logística dos vídeos *online*. É muito comum que produtores de conteúdo e administradores de portais de mídias do tipo ME recorram a essas empresas. No Brasil, por exemplo, a empresa líder no segmento de plataformas de vídeos *online* é a Samba Tech[6] [Startupi, 2012; The Next Web, 2011].

Diante da complexidade do processo de gestão e distribuição de mídias, torna-se fundamental conhecer os padrões de consumo dessas mídias. Esse conhecimento afeta diretamente aspectos de infra-estrutura, tais como escalabilidade, disponibilidade e robustez, e do funcionamento dos procedimentos de logística. Além disso, não são apenas as empresas especializadas em plataformas de vídeos *online* que devem se interessar

[1] `http://www.uol.com.br`
[2] `http://www.cnn.com`
[3] `http://www.youtube.com`
[4] `http://www.vimeo.com`
[5] `http://www.facebook.com`
[6] `http://www.sambatech.com`

pelo comportamento dos usuários. Informações do perfil de consumo dos usuários podem ser valiosas para produtores de mídias que desejam melhorar a qualidade do conteúdo oferecido e prover uma experiência mais satisfatória em seus *sites*.

Apesar de existirem alguns esforços no sentido de investigar o consumo de vídeos *online*, principalmente sobre *sites* de CGU, ainda pouco se conhece sobre o assunto. Considerando a necessidade de uma maior investigação na área, a proposta desta dissertação é apresentar uma caracterização detalhada do consumo de vídeos *online* no mercado brasileiro de portais de mídia especializada.

1.1 Motivação

Como já foi mencionado, pretende-se, nesta dissertação, investigar o consumo de vídeos *online*, publicados por portais de mídias do tipo ME, no contexto do mercado brasileiro. As principais motivações para o presente trabalho podem ser encontradas no contexto atual da Web, na tendência de expansão do consumo de vídeos *online* e na necessidade de mais pesquisas sobre o assunto. São elas:

- *A importância dos vídeos online, que estão permeando as mais diversas aplicações.* A Web está cada vez mais colaborativa e sociável. Portanto, a tendência esperada é de crescimento do número de compartilhamentos e acessos de mídias ricas na Internet.

- *A necessidade de maiores estudos sobre o comportamento de consumo de vídeos online.* Por ser um fenômeno recente, pouco se sabe sobre como vídeos *online* são consumidos. Além disso, grande parte dos esforços de pesquisa estão focados em analisar o acesso de vídeos do tipo CGU [Benevenuto et al., 2008; Cha et al., 2007; Cheng et al., 2007; Szabo & Huberman, 2010]. Não há muitos trabalhos que investigam portais de mídias do tipo ME, principalmente, pela indisponibilidade de coleções com registros detalhados de interações de usuários em *sites* de mídias especializadas. Em geral, poucos portais de ME têm interesse em tornar públicos os *logs* de acessos de seus usuários. Pela dimensão do mercado de vídeos *online*, melhorias simples no processo de publicação e logística de conteúdo podem ter um impacto final bastante significativo.

- *A posição do Brasil como um dos principais consumidores de vídeos online.* O Brasil é o sétimo maior mercado de vídeos *online* do mundo e o quinto maior mercado mundial de Internet. E a tendência é de expansão. A previsão é de que sejam investidos 9 bilhões de dólares pelo Governo Federal em banda larga nos

próximos 4 anos. Em 2012, 12 bilhões de dólares foram gastos em movimentações *online* pelos brasileiros. Em 5 anos, prevê-se que os gastos na Web por internautos do Brasil cheguem a 25 bilhões de dólares ao ano [Samba-Tech, 2013]. Portanto, é fundamental conhecer o mercado de vídeos *online* brasileiro e seus consumidores.

1.2 Objetivos

Tendo em vista a carência de trabalhos na área, o principal objetivo desta dissertação é contribuir para um melhor entendimento do comportamento de acessos de usuários a vídeos *online*, específicamente, no contexto de portais brasileiros de mídias especializadas.

Pretende-se fazer uma caracterização detalhada de como ocorrem as visualizações em diferentes contextos de aplicação (vários portais). As análises devem considerar vários aspectos dos registros de acessos. Primeiramente, os dados devem ser explorados de maneira agregada, levando em conta os atributos estáticos (número total de visualizações, quantidade de mídias, duração dos vídeos, categorias, distribuições de acessos, etc). Outro aspecto considerado é o temporal, ou seja, como ocorre a evolução das visualizações ao longo do tempo e quais são os padrões recorrentes de acesso. Também é relevante investigar padrões de fluxos de navegação de usuários entre diferentes vídeos.

Outro objetivo importante da dissertação é apresentar um comparativo dos resultados das análises realizadas, sobre acessos a vídeos de portais de mídia especializada, com resultados previamente reportados sobre o contexto de *sites* do tipo CGU.

1.3 Contribuições

Os resultados desta dissertação permitem entender melhor o perfil de consumo de vídeos *online*. Além disso, as investigações propostas tornam possível estabelecer um paralelo entre os hábitos de acesso a vídeos produzidos por *sites* de conteúdo especializado e mídias compartilhadas por usuários autônomos na Internet. Assim, as principais contribuições desta dissertação são:

- Caracterização detalhada do comportamento de usuários visualizando vídeos *online* em *sites* provedores de mídias especializadas. As análises devem revelar padrões de acessos interessantes, considerando aspectos estáticos, temporais e de fluxos de navegação.

- Análise comparativa entre padrões de acesso em *sites* provedores de mídias especializadas e *sites* para compartilhamento de conteúdos gerados por usuários.

- Identificação de informações relevantes sobre o comportamento de consumidores de vídeos *online*, que possam ter aplicações práticas diversas, tais como:

  - Melhoria da experiência dos usuários pela implementação de serviços de recomendação e personalização da entrega de conteúdo.
  - Oferta de publicidade mais direcionada e relevante para usuários.
  - Aprimoramento dos serviços de distribuição de vídeos pela sofisticação das políticas de logística (considerando aspectos de localidade, padrões recorrentes de acesso, *cache*, etc.).

- Investigação, como estudo de caso, de um cenário importante, mas ainda pouco explorado, que é o de portais brasileiros provedores de mídias especializadas.

Parte dos resultados descritos nesta dissertação foram apresentados no *Third Temporal Web Analytics Workshop (TempWeb03) 2013*, realizado conjuntamente com a *WWW 2013 Conference* [Miranda et al., 2013].

1.4 Trabalhos Relacionados

A presente dissertação está relacionada com um amplo espectro de linhas de pesquisa sobre compartilhamento de vídeos *online*. As áreas correlatas podem variar desde "caracterização de tráfego de rede" [Gill et al., 2007; Gürsun et al., 2011; Saxena et al., 2008; Comarela et al., 2012; Zink et al., 2009], à "análise de redes sociais" [Benevenuto et al., 2008, 2009a; Cheng et al., 2008; Cheng & Liu, 2009; Paolillo, 2008; Szabo & Huberman, 2010; Maia et al., 2008]. Especificamente, esta dissertação pode ser situada entre esses dois extremos, concentrando-se na caracterização dos padrões de acesso de vídeos *online* no nível de aplicação (em oposição ao nível mais específico de rede).

Existem diversos estudos sobre caracterização de requisições e carga de trabalho (*workloads*) em sistemas Web variados, tais como *e-commerce* [Menascé et al., 1999; Vallamsetty et al., 2003], vídeo sob demanda (*video on demand - VoD*) [Costa et al., 2004], redes sociais [Benevenuto et al., 2009b; Maia et al., 2008] e servidores Web [Arlitt & Williamson, 1996]. [Acharya et al., 1999] está entre os trabalhos pioneiros na análise de padrões de acessos a vídeos na Web. Na época de sua publicação,

vídeos *online* não eram tão difundidos na Internet como ocorre atualmente. Em particular, os autores realizaram um experimento de pequenas proporções em uma universidade da Suécia, envolvendo 139 vídeos, cujas visualizações foram monitoradas por 6 meses, entre 1997 e 1998. Como resultado dessa análise, foi observada a existência de padrões de acesso cíclicos, sendo que o número de visualizações era reduzido nos fins de semana, em comparação com os demais dias. Também observou-se um padrão temporal de localidade nos acessos e uma baixa retenção para 45% de todos os vídeos, sendo que a maioria desses eram interrompidos tendo sido assistida apenas 5% de sua duração. Em nosso trabalho, conduzimos análises temporais similares, porém em uma escala muito maior, envolvendo milhões de vídeos publicados por grandes *sites* e portais brasileiros de mídia especializada.

Recentemente, com a emergência das mídias sociais na Web, vários estudos foram conduzidos objetivando analisar os padrões de acesso a vídeos publicados em *sites* para compartilhamento de mídias por usuários autônomos. Grande parte desses estudos envolve experimentos com a maior plataforma de vídeo sob demanda e publicação de conteúdos gerados por usuários, o YouTube [Cha et al., 2007; Cheng et al., 2007; Gill et al., 2008; Baluja et al., 2008]. Por exemplo, [Cheng et al., 2007] investigaram 3 meses de registros do YouTube datados de antes de 2007, compreendendo um total de 2,6 milhões de vídeos. Dos vídeos analisados, 22,9% pertenciam à categoria *"Music"*, a mais popular das 12 categorias pré-definidas pelo YouTube. Nesse estudo, foi observado que a maioria dos compartilhamentos no *site* eram de vídeos de tamanho relativamente reduzido, sendo que 97,8% de todas as mídias tinham menos de cinco minutos de duração. Em relação à distribuição de visualizações por vídeo, a análise mostrou que seguia uma lei de potência (*power-law*) truncada ao invés de uma distribuição *Zipf* padrão [**?**], indicando uma quantidade menor do que a esperada de vídeos pouco populares (fenômeno da cauda longa). Além disso, monitorando um total de 43 mil vídeos por um período de sete semanas, observou-se que 70% desses vídeos tinham um crescimento de popularidade decadente com o passar do tempo, o que denota, em geral, um tempo de vida curto.

Gill et al. [2008] também conduziram uma investigação sobre o YouTube. Nesse trabalho, os autores propuseram uma caracterização do comportamento de acesso a vídeos do *site* de compartilhamento por usuários. Para isso, foi adotado o conceito de "sessão de usuário", como sendo uma série de requisições subsequentes de um usuário, considerando uma única visita ao site [Menascé et al., 1999]. Além de algumas análises estáticas relativas ao usuário (número de acessos, quantidade de dados transferidos, etc.), foram investigados vários aspectos das sessões definidas, tais como duração, tempo entre sessões e tipo de conteúdo visualizado. Os resultados foram contrastados

com outras análises aplicando o conceito tradicional de sessões de usuários da Web. As diferenças identificadas têm implicações importantes para administradores de sistemas responsáveis pela arquitetura e pelo planejamento de infraestrutura e distribuição de conteúdo. Foi identificado, por exemplo, que usuários do YouTube, em geral, transferem maior quantidade de dados e apresentam intervalos mais longos entre requisições (*think times*) do que usuários de outros *sites* da Web. Em nosso trabalho, também realizamos análises de sessão de usuário e adotamos procedimentos similares para definir as sessões, mas no contexto de *sites* de mídia especializada. Além disso, enquanto as investigações apresentadas no artigo são mais voltadas para os aspectos de rede, priorizamos uma análise mais voltada à aplicação.

Outro trabalho importante sobre *sites* de compartilhamento de conteúdo gerado por usuário, tendo como estudo de caso o YouTube e o Daum (principal *site* coreano de CGU), é apresentado por [Cha et al., 2007]. Nesse trabalho, os autores executaram uma análise extensa do ciclo de vida da popularidade dos vídeos. Foram revelados aspectos interessantes em relação à distribuição de requisições sobre os vídeos, a evolução do foco dos usuários e as mudanças de popularidade. Ao analisar a evolução temporal das visualizações sobre vídeos, constatou-se, por exemplo, que a distribuição de popularidade de conteúdos compartilhados por usuários segue uma lei de potência truncada. Os resultados permitem a elaboração de sistemas que apliquem mecanismos de distribuição de dados mais eficientes (*caching* e *peer-to-peer*). Como investigação complementar, os autores mediram a ocorrência de duplicações e de carga de conteúdo ilegal. Em nosso trabalho, também realizamos algumas análises temporais e investigamos a evolução das visualizações de vídeos publicados em *sites* de mídia especializada.

Além da análise da evolução de popularidade dos vídeos, Cha et al. [2007] também estabeleceram uma comparação entre *sites* para compartilhamento de conteúdo gerado por usuários e *sites* de publicação de mídia especializada. Os autores compararam o YouTube e o Daum, com os portais de ME Netflix, LOVEFilM e Yahoo! Movies. Suas análises revelaram uma taxa de crescimento substancialmente maior de produção de conteúdos por usuários autônomos, se comparada a *sites* de mídia especializada. Por outro lado, a distribuição de vídeos por provedor entre os dois cenários mostrou um comportamento de lei de potência similar, com a ressalva de que a produção média de um produtor de mídia especializada (por exemplo, um diretor de cinema) é naturalmente limitada por restrições temporais e financeiras impostas pela indústria. A análise também mostrou que a duração média de um vídeo do tipo CGU é duas ordens de grandeza menor do que a de um vídeo de ME. Nessa dissertação, também procuramos realizar uma análise comparativa similar.

Goncalves et al. [2011] propuseram uma nova metodologia hierárquica de caracte-

rização de serviços Web multimídia. A metodologia propõe a segmentação das análises e entidades em quatro diferentes camadas: requisição, objeto, conteúdo e conhecimento. Essa metodologia foi aplicada em um estudo de caso utilizando registros de *logs* da Samba Tech. O objetivo do trabalho foi caracterizar e analisar de forma mais metódica os atributos das requisições, as propriedades dos objetos, o conteúdo associado através de seus metadados e os padrões implícitos derivados de informações associadas ao serviço multimídia. Os resultados obtidos podem ser aplicados no aprimoramento de serviços Web e na melhoria de serviços de entrega e distribuição de conteúdo. Em nosso trabalho, também desenvolvemos estudos similares empregando dados da Samba Tech, mas com enfoque maior em análises na camada de aplicação. Além disso, em vez de utilizar os *logs* de requisições, coletamos registros de interações de usuários com vídeos em diferentes *sites* de mídia especializada (maiores detalhes no Capítulo 2).

A maior parte dos esforços em pesquisas na área de vídeos *online* está voltada para a categoria de *sites* para compartilhamento de vídeos produzidos por usuários autônomos (CGU). Apesar disso, existem alguns trabalhos importantes no contexto de *sites* de mídia especializada. Benevenuto et al. [2010], por exemplo, apresentaram uma análise dos acessos a vídeos do portal *UOL*, maior provedor de conteúdo *online* da América Latina. Os autores realizaram uma caracterização em termos de sessões de usuários e de suas requisições ao servidor, similar àquela proposta por Gill et al. [2008] para o YouTube. Além disso, como contribuição complementar, foram identificados grupos de perfis de navegação no *site*. Algumas das conclusões importantes desse trabalho foram: uma sessão típica de usuário dura cerca de 40 minutos, um valor alto se comparado com sistemas Web tradicionais; as distribuições de popularidade de acessos de objetos (vídeos e *tags*) seguem o padrão de curvas com "cauda longa"; as distribuições de tempo entre requisições e de tempo entre sessões podem ser modeladas por distribuições exponenciais. Os resultados desse trabalho podem ser empregados para geração de dados sintéticos e na melhoria dos serviços de publicação de conteúdo (por meio de estratégias como personalização da experiência de uso e recomendação).

O presente trabalho também constitui uma proposta de caracterização de padrões de acesso em *sites* produtores de mídia especializada (não produzida por usuários autônomos). O estudo dos trabalhos relacionados foi fundamental para a definição da metodologia e execução de análises comparativas (especialmente ao confrontarmos padrões de acessos em *sites* do tipo CGU com padrões de visualizações de vídeos em *sites* de ME). Foram conduzidas análises estáticas considerando diversos atributos dos dados (quantidade de visualizações, categorias, duração dos vídeos, etc), análises temporais (evolução de acessos, ciclo de vida das mídias, dentre outras) e investigações em termos de sessões de usuário (tais como tamanho de sessão e fluxo de visualizações entre ca-

tegorias e entre clientes). Contudo, não encontramos na literatura análises envolvendo portais produtores de mídia especializada de propósito geral. Os poucos trabalhos relacionados com esse tipo de *site* tratam de um contexto de aplicação bastante específico (análises sobre características próprias de um único *site*) [Benevenuto et al., 2010; Gill et al., 2008]. Além disso, investigamos tanto a produção quanto o consumo de conteúdo em portais de mídia especializada. Em nosso trabalho, as análises propostas foram aplicadas em uma coleção com registros de interações de usuários em diversos *sites*, envolvendo milhões de vídeos e centenas de milhares de usuários. Não temos conhecimento de análises similares sobre uma coleção tão abrangente.

1.5 Organização

A dissertação a seguir está organizada em cinco capítulos.

O Capítulo 2 descreve os mecanismos de coleta e tratamento da coleção e os procedimentos de análise experimental. Nesse capítulo também é descrito o estudo de caso desenvolvido. As análises experimentais realizadas e os resultados obtidos são detalhados nos três capítulos seguintes.

O Capítulo 3 discute as investigações sobre atributos estáticos da coleção.

No Capítulo 4 são detalhados os procedimentos experimentais que consideram a evolução dos acessos e propriedades dinâmicas dos registros.

No Capítulo 5 são apresentadas as análises sobre sessões de usuários e sobre fluxos de acessos.

Por fim, o Capítulo 6 apresenta as conclusões e um prospecto dos trabalhos futuros.

Capítulo 2

Metodologia

A difusão dos serviços e aplicações que adotam vídeos *online* na Web é um fenômeno recente que ainda demanda maiores estudos. Essa demanda é ainda maior no contexto de *sites* de mídias especializadas. Para investigar o consumo de vídeos em portais do tipo ME, é necessário fazer um acompanhamento detalhado das interações de usuários acessando *sites* provedores desse tipo de conteúdo. A análise de um conjunto de registros de acessos de usuários em portais de ME pode revelar padrões interessantes de comportamento de consumo de vídeos *online*. Os padrões identificados podem, então, ser aplicados para melhorar a experiência dos serviços e os conteúdos providos ao usuário.

Para que os estudos de registros de acessos sejam válidos e os padrões identificados possam ter aplicação em outros contextos, é necessário atentar para alguns aspectos dos dados utilizados nas análises experimentais:

- É necessário que os dados compreendam uma quantidade razoável de usuários, já que perfis específicos de usuários devem ter padrões próprios de comportamento.
- É importante que a coleção contenha vídeos suficientes para evitar que fenômenos excepcionais causados por vídeos específicos sejam tomados como padrões comuns.
- Os dados devem compreender um intervalo de tempo de acompanhamento dos acessos suficiente para observar padrões recorrentes. Além disso, é interessante poder inferir padrões temporais, que ocorrem ao longo do tempo, tais como a progressão do número de visualizações de um vídeo recém-publicado.
- A análise deve envolver diferentes contextos de aplicação para evitar que aspectos específicos de um *site* ou serviço possam influenciar nas conclusões gerais. O

objetivo do trabalho é ter uma visão geral do mercado de vídeos *online* brasileiro, portanto, as análises não devem se limitar à uma única amostra de todo o universo.

Nesse tipo de análise, deve-se ter cautela para que não se assuma como genéricos os aspectos específicos observados no cenário analisado. É necessário assegurar que todo padrão encontrado seja recorrente e aplicável em outros contextos compatíveis com o estudado. Atentando para os pontos levantados, para caracterizar o consumo de mídias especializadas, como um estudo de caso, analisamos registros de interações de internautas com vídeos hospeados por portais brasileiros de ME, conforme descrito a seguir.

2.1 Estudo de Caso

Já foi mencionado que o principal motivo para o número reduzido de pesquisas sobre o consumo de mídias especializadas é, provavelmente, a indisponibilidade de coleções com registros de acessos em portais de ME. Para realizar um estudo consistente, seria necessário analisar interações de vários usuários sobre um conjunto diversificado de vídeos em diferentes domínios. Porém, não é fácil encontrar *sites* de ME dispostos a publicar *logs* detalhados de seus acessos. Diante da dificuldade em obter dados junto aos portais de ME, recorremos a uma empresa que fornece soluções de gestão e distribuição de vídeos *online* para portais brasileiros de mídias especializadas. Os dados para o estudo de caso foram concedidos pela Samba Tech[1].

A Samba Tech é a empresa líder no setor de plataformas comerciais de vídeos *online* da América Latina. São clientes da Samba Tech grandes grupos de mídia, tais como SBT, GloboSat, Grupo Abril, Bandeirantes, Portal R7, dentre outros. Oito dos dez maiores grupos de mídias *online* da Internet brasileira adotam seus serviços. Além disso, a Samba Tech é responsável por 20 petabytes de tráfego por ano e recebe, aproximadamente, 300 milhões de *video views* por mês.

O principal produto da Samba Tech é uma Plataforma de Vídeos *online* (*Online Video Platform* - OVP). A *Liquid Platform*[2] provê armazenamento, gestão e distribuição profissional de mídias para portais de ME. Dentre as funcionalidades comuns de uma OVP, estão: *upload* de mídias; possibilidade de inserção e edição de metadados (título, descrição, *tags*, categoria, etc); armazenamento de mídias; logística de distribuição; codificação de vídeos, permitindo a execução em diferentes dispositivos;

[1] `http://www.sambatech.com`.
[2] `http://www.sambatech.com/liquid`.

servidores de processamento, redes e bandas para tráfego; mecanismos de segurança, tais como gestão de direitos digitais (*DRM*) e acesso restrito por regiões geográficas (*Geoblocking*) e mecanismos de análises, estatísticas e geração de relatórios.

Outro produto desenvolvido pela Samba Tech é um *player* de vídeos, que pode ser adotado pelos clientes da *Liquid Platform* em seus *sites*. Além da Plataforma de Vídeos *online* e do *player*, a Samba Tech oferece outras soluções em vídeos *online*. Porém, a descrição dos demais produtos foge do escopo deste trabalho.

2.2 Análise Experimental

O processo de experimentação se inicia na coleta de dados. Tendo organizado os dados na forma de uma coleção bem estruturada, devem ser formuladas questões de pesquisa relativas aos propósitos do trabalho. A busca de respostas para as questões levantadas serve de referência para a aplicação das análises experimentais.

2.2.1 Coleta de Dados

A coleta dos dados para o estudo de caso foi realizada em associação com a Samba Tech. A Samba Tech, por meio de sua plataforma de vídeos *online*, *Liquid Platform*, e seu *player*, participa de todas as etapas do processo de distribuição de mídias especializadas. Os vídeos produzidos por diversos portais brasileiros são enviados para a OVP da Samba Tech e, no fim do processo de logística, são entregues ao usuário final pelo *player*. Assim, é possível registrar dados consistentes do que é produzido e assistido em termos de conteúdos de mídias especializadas.

A coleta de dados foi realizada pelo registro de todas as interações de usuários no *player* da Samba Tech. A Figura 2.1 ilustra o processo geral de distribuição de vídeos e coleta de dados. Os passos envolvidos foram:

1. Vídeos são produzidos por provedores de conteúdo do tipo ME. Após a produção, o vídeo é enviado para a plataforma (OVP) da Samba Tech. A plataforma permite que o usuário faça a gestão de seu conteúdo, editando metadados, configurando permissões de acesso, removendo mídias, etc.

2. Vídeos armazenados na plataforma da Samba Tech são distribuídos. A distribuição ocorre pela publicação de URLs de mídias servidas por uma "Rede de Fornecimento de Conteúdo" (*Content Delivery Network* - CDN). A CDN é um sistema de computadores em rede que cooperam de modo transparente para entregar conteúdo.

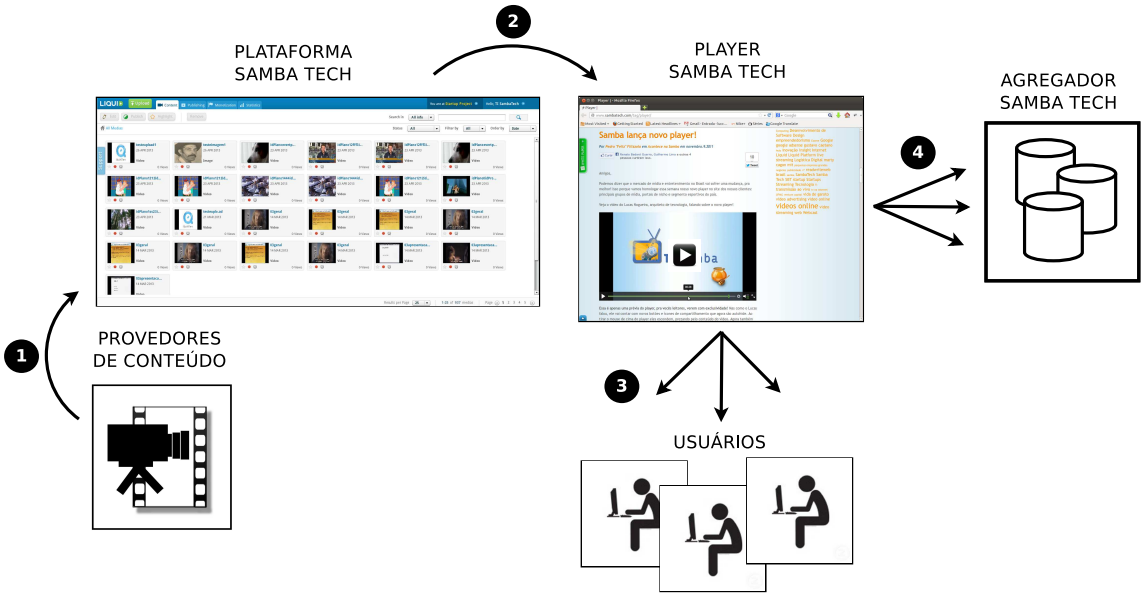

Figura 2.1. Fluxo de coleta de dados: 1. Vídeo é produzido e enviado para plataforma; 2. Vídeo é distribuído; 3. Usuários acessam o vídeo, interagindo com o player; 4. Eventos do player são registrados.

3. O usuário consome o vídeo acessando a página de um *site* de um cliente da Samba Tech. Durante a exibição do vídeo, o usuário interage com o *player* da Samba Tech.

4. Todos os eventos gerados por interações de usuários no *player* da Samba Tech são enviados para um servidor agregador de dados. O envio dos dados é feito por *scripts* (*JavaScript*) que compõem o *player*, utilizando uma API de arquitetura REST.

Os dados coletados possuem uma característica, em particular, bastante interessante: os registros incluem acessos de usuários em diversos *sites*. Com isso, é possível ter uma perspectiva privilegiada de como o internauta navega e consome conteúdo entre diferentes domínios. Para permitir esse tipo de análise, porém, é necessário relacionar acessos de um mesmo usuário em momentos (e *sites*) diferentes.

Como não há nenhum tipo de autenticação de usuários nos *sites* acompanhados (qualquer pessoa pode acessar os *sites* sem necessidade de registro), a identificação se deu por meio de *HTTP cookies*. *Cookies* são dados que podem ser armazenados e consultados no navegador (*browser*) do usuário. Um *site* pode adicionar um *cookie*, que fica armazenado no navegador associado ao domínio, e consultá-lo posteriormente. Assim, quando um internauta interage com o *player* da Samba Tech, esse atribui um identificador para o usuário e o salva em *cookie*. Nos acessos futuros, o *player* apenas

verifica que o *cookie* específico já existe no navegador e usa esse identificador para registrar as interações daquele usuário.

2.2.2 Descrição da Coleção

Os dados coletados foram armazenados em um banco de dados do tipo MongoDB[3]. Bancos de dados MongoDB são aplicações de código aberto, de alta performance, sem esquemas e orientados a documentos. Os dados armazenados em bancos de dados desse tipo são organizados como conjuntos de documentos no formato JSON (*JavaScript Object Notation*).

A Figura 2.2 apresenta um exemplo de registro no formato JSON, armazenado no processo de coleta. Cada registro possui informações detalhadas do acesso, incluindo:

- Conjunto de ações do usuário ao interagir com o *player*. Essas ações são registradas como eventos: *play_load*, *play*, *stop*, *resume*, *seek*, *progress* (ativado em cada quartil do tempo de execução do vídeo). Cada evento tem associado o tempo e informações específicas (como, por exemplo, *seekbartime*).

- Informações do contexto de acesso, incluindo título do *site*, URL do *site*, *referrer*, informações de *timezone* e *locale*, etc.

- Informações da mídia acessada. Essas informações são fornecidas pela plataforma no momento da distribuição do vídeo e incluem título, categoria, canal, duração, qualidade, descrição, *tags*, dentre outras. Todos os metadados de vídeos são preenchidos por produtores de conteúdo. Portanto, espera-se que sejam informações confiáveis e consistentes.

- Identificação do usuário (campo *user_id*), usado para associar acessos do mesmo usuário em diferentes *sites* em momentos distintos (informação previamente armazenada em *cookie*).

Além do conjunto de registros armazenados, a Samba Tech possui informações específicas de cada mídia no banco de dados de sua plataforma de vídeos *online*. Essas informações incluem a data de postagem (momento em que o vídeo foi disponibilizado na plataforma) e a data de publicação (momento em que o vídeo foi publicado para acesso de usuários em um *site*), dentre outros metadados.

No total, foram armazenados, aproximadamente, 280GB de dados, no formato de documentos JSON, como o do exemplo apresentado. O processo de coleta envolveu

[3] `http://www.mongodb.org/`.

```
{
    "_id": ObjectId("513843e227fa60c86c6c1e09"),
    "api_key": "123",
    "cookie_enabled": "true",
    "creation_time": NumberLong("1344816000382"),
    "domain": "sambatech.37.310",
    "events": {
        "play": NumberLong("1344816007576"),
        "player_load": NumberLong("1344816000382"),
        "resume": {
            "seekbartime": "129000",
            "time": NumberLong("1344816007617")
        },
        "seek": {
            "lastseektime": "129000",
            "time": NumberLong("1344816172451")
        }
    },
    "ip": "186.242.72.151",
    "last_update_time": NumberLong("1344816172451"),
    "player_session": "462b4df8-2f2f-d8e3-dd45-929e3d1a1687",
    "player_type": "flash",
    "referrer": "http://www.exemplo.com.br/videos/",
    "site_title": "Site exemplo",
    "site_url": "http://www.exemplo.com.br/videos/ultimos",
    "sub_domain": null,
    "user_agent": "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1;
        Trident/4.0; GTB7.3; BTRS127568; .NET CLR 2.0.50727;
        .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729;
        OfficeLiveConnector.1.3; OfficeLivePatch.0.0; InfoPath.1)",
    "user_id": "9839f7b6-1859-48f3-82f2-18c36b47b942",
    "user_time_zone": "180",
    "video_auto_play": "false",
    "video_channel": "Canal de Exemplo",
    "video_description": "Descrição detalhada do vídeo.",
    "video_duration": "152000",
    "video_genre": "ENTERTAINMENT",
    "video_id": "391dc09b0143cdf893de364a4b2cc914",
    "video_quality": "360p",
    "video_short_description": "Descrição curta do vídeo.",
    "video_tags": "video,exemplo,tags",
    "video_title": "Título do vídeo."
}
```


Figura 2.2. Exemplo de um registro armazenado no processo de coleta de dados.

oito semanas e 38 *sites* brasileiros de mídias especializadas, incluindo seis dos maiores portais brasileiros do gênero.

A Tabela 2.1 contém informações estatísticas de toda a coleção. Durante oito semanas, foram coletados mais de 110 milhões de registros, envolvendo mais de 43 milhões de usuários únicos. Foram registrados, aproximadamente, 127 mil vídeos, com

Data inicial	24 de Junho, 2012 (Dom)
Data final	18 de Agosto, 2012 (Sáb)
Registros	110.626.789
Usuários únicos	43.217.621
Vídeos	127.068
Duração dos vídeos (média)	433,5s
Duração dos vídeos (desvio padrão)	174,7s

Tabela 2.1. Estatísticas gerais da coleção.

duração média por volta de sete minutos. Essa é, certamente, uma coleção muito abrangente de dados extraídos de um ambiente real.

É importante ressaltar que não foi realizado nenhum tipo de amostragem durante o processo de coleta. Ou seja, no intervalo considerado de oito semanas, foram registradas todas as interações de usuários com vídeos publicados nos 38 *sites*.

A coleção obtida contém dados valiosos, já que envolve um grande volume de registros de acessos de usuários em diferentes *sites* de mídia especializada. Esta coleção é adequada para nossos propósitos de pesquisa segundo os critérios levantados (no início do capítulo). A coleção contém um número substancial de vídeos e usuários únicos registrados. Além disso, foram armazenados eventos de interações de usuários em diferentes contextos de aplicação (diversos *sites* e portais de ME), coletados durante um período relativamente longo. Esses atributos minimizam o efeito de aspectos específicos de parte dos dados sobre padrões identificados para toda a coleção e permitem a identificação de fenômenos recorrentes ao longo do tempo.

2.2.3 Questões de Pesquisa

Tendo coletado e pré-processado os dados, nosso objetivo é estudar o comportamento de usuários ao acessar vídeos de *sites* provedores de mídias especializadas. Para tanto, devemos submeter a coleção a um processo detalhado de caracterização.

A análise experimental dos dados coletados deve ser conduzida de forma a responder três questões de pesquisa:

Q1. Quais padrões de acesso emergem da análise de um perfil estático da coleção agregada?

Q2. Quais padrões temporais podem ser inferidos da análise das interações de usuários com vídeos ao longo do tempo?

Q3. Quais relações recorrentes podem ser identificadas de vídeos assistidos, frequentemente, em sequência?

Os três capítulos que seguem compreendem a análise experimental do estudo de caso. Cada capítulo aborda a pesquisa desenvolvida, orientada por uma das perguntas propostas, na ordem em que foram levantadas.

O Capítulo 3 trata da primeira pergunta e, portanto, aborda as análises estáticas. Nesse capítulo, a coleção é analisada considerando todo o período de oito semanas, de forma agregada.

A segunda pergunta proposta é investigada no Capítulo 4, que detalha as análises temporais. As investigações temporais consideram a evolução dos acessos ao longo das oito semanas de dados coletados.

Para responder à última pergunta, o Capítulo 5 apresenta estudos do relacionamento de vídeos assistidos conjuntamente. Nesse capítulo, o conceito de "sessão de usuário" é definido e aplicado à análise do fluxo de visualização de vídeos por usuários.

Em cada capítulo, são detalhados os procedimentos experimentais envolvidos e os respectivos resultados.

Capítulo 3

Análise Estática

No contexto de vídeos *online*, em geral, cada mídia possui um conjunto de atributos associados (metadados), que a descreve. Estudar quais padrões de visualizações ocorrem em decorrência desses atributos é fundamental para entender o mercado de vídeos *online* e seus consumidores. A primeira das investigações realizadas nesse sentido deve ser orientada pela questão de pesquisa número um (levantada na Seção 2.2.3): "quais padrões de acessos emergem da análise de um perfil estático da coleção agregada?".

Para realizar a análise estática, é necessário considerar todos os registros coletados durante as oito semanas de maneira agregada (como um único *snapshot* da coleção). Ou seja, os dados devem ser agrupados sem discriminação por data.

Além das informações contidas nos registros, a coleção usada no estudo de caso inclui metadados sobre todos os seus vídeos. Esses metadados foram adicionados pelo produtor de conteúdo por meio da plataforma (OVP) da Samba Tech. A atribuição dos metadados é opcional, mas é utilizada pela maioria dos produtores de conteúdo para gestão das mídias.

O objetivo da análise estática é obter padrões gerais sobre propriedades atemporais de usuários e vídeos, tais como número total de visualizações, categoria e duração. Além disso, os padrões revelados devem ser comparados com resultados reportados na literatura sobre usuários interagindo com *sites* de compartilhamento de conteúdo do tipo CGU.

3.1 Categoria

A categoria dos vídeos foi o primeiro atributo considerado nas análises estáticas. Categorizar vídeos é uma forma de organizar mídias agrupando-as pela similaridade de seus temas. A categoria de um vídeo é, certamente, uma propriedade muito discrimina-

tiva. Sendo assim, espera-se que muitos padrões de visualizações estejam relacionados diretamente com esse atributo.

Na plataforma de vídeos da Samba Tech, existem as seguintes categorias pré-definidas:

- Comédia (*Comedy*)
- Entretenimento (*Entertainment*)
- Filme (*Film*)
- Música (*Music*)
- Pessoas (*People*)
- Animais (*Pets*)
- Política (*Politics*)
- Ciência (*Science*)
- Esporte (*Sports*)

A seleção da categoria de um vídeo é opcional e única (não são permitidas múltiplas atribuições de categoria para um único vídeo). Os vídeos para os quais não foram especificadas alguma das nove categorias pré-definidas ficam registrados (em nosso estudo) como pertencentes à classe "Desconhecida" (*Unknown*).

A Figura 3.1 apresenta a distribuição do número de vídeos por categoria. Observa-se, imediatamente, a predominância de vídeos sem categoria. Em particular, para 53.6% de todos os vídeos (68.098 de 127.068) não houve atribuição de qualquer categoria (vídeos agrupados em *Unknown*).

As categorias com mais vídeos são, na ordem, *Politics* (17.1%), *Entertainment* (14.3%), *Sports* (11.1%), e *Science* (2.3%). Essas categorias também são muito representativas (em termos de quantidade de vídeos) no contexto de *sites* de compartilhamento de conteúdos gerados por usuário (CGU), tais como o YouTube. No YouTube, porém, enquanto *Music* e *Comedy* estão entre as categorias mais populares (22.9% e 12.1%, respectivamente) [Cheng et al., 2008], poucos vídeos foram atribuídos a essas categorias em nossa coleção (menos de 1%). Essa discrepância pode ser justificada pela ausência de provedores de música em nosso conjunto de dados e pela preferência presumível de usuários autônomos de compartilharem músicas e conteúdos de comédia em *sites* de CGU.

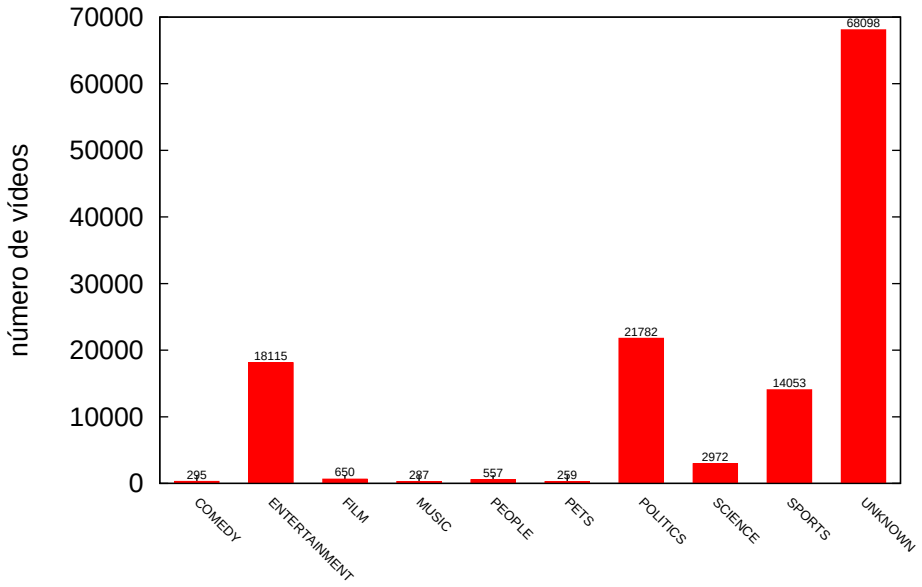

Figura 3.1. Distribuição do número de vídeos por categoria.

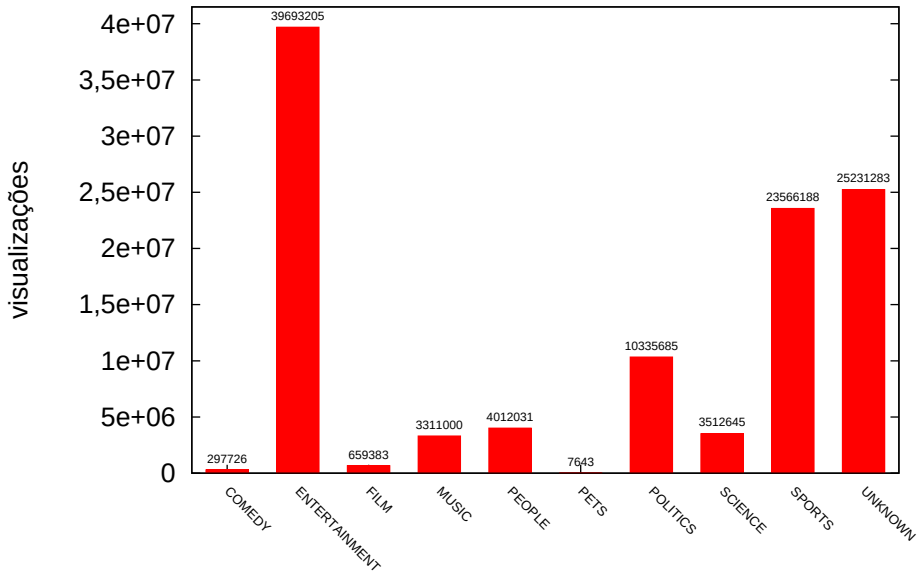

Figura 3.2. Distribuição de visualizações por categoria.

A Figura 3.2 apresenta a distribuição do número de visualizações por categoria. Comparando essa distribuição com a da Figura 3.1, observamos que o número de visualizações por categoria não corresponde, necessariamente, ao número de vídeos por categoria. Ou seja, a disponibilidade de muitas mídias de uma categoria não implica, impreterivelmente, em uma maior quantidade de acessos que outras categorias com número inferior de vídeos. Por exemplo, a categoria com mais vídeos, *Politics*, rece-

beu menos visualizações que outras categorias menos representativas em quantidade de vídeos, tais como *Entertainment* e *Sports*.

Entertainment foi a categoria com a maior quantidade de visualizações (35.9%), provavelmente pelo fato dos vídeos dessa categoria, geralmente, terem um forte "efeito viral", tendendo a serem mais compartilhados.

Considerando-se a proporção de visualizações por quantidade de vídeos, as categorias *Music* e *People* se destacam, com 11.536,59 e 7.202,93 acessos por vídeo, respectivamente. Enquanto, *Pets* e *Politics* apresentam as piores relações: 29,51 e 474,51 acessos por vídeo, respectivamente.

3.2 Duração

A duração dos vídeos pode ter influência direta sobre o comportamento de acesso dos usuários. Dependendo de sua duração, um vídeo pode ser mais ou menos atrativo, pode ter baixa ou alta taxa de retenção (porcentagem do vídeo assistida pelo usuário) e, eventualmente, pode afetar o interesse do espectador.

Estatísticas reportadas para o YouTube têm mostrado que o *site* é formado, em geral, por vídeos mais curtos (20,6% dos vídeos do YouTube têm menos que um minuto de duração e 17,1% têm entre 3 e 4 minutos apenas) [Cheng et al., 2008]. Em contraste, a duração média dos vídeos em nossa coleção é de 433,5 segundos, pouco mais que sete minutos (Tabela 2.1).

A duração média dos vídeos está diretamente relacionada com as categorias, como mostrado na Figura 3.3. Na figura, a duração média é expressa em segundos, com barras de erro denotando um intervalo de confiança de 95% sobre a média. Podemos observar que vídeos da categoria *Pets* e aqueles não categorizados possuem a maior duração (mais de 600 segundos) em relação aos demais vídeos. *Entertainment* e *Politics* também apresentaram durações médias elevadas. Esse conjunto de vídeos, em geral, envolve curtos documentários, reportagens e partes de programas de televisão. Em contraste, *Comedy*, *Music* e *Sports* contêm vídeos mais curtos, em média, compreendendo notícias rapídas, anúncios publicitários e clipes de música. A predominância de vídeos de curta duração nessas categorias também se assemelha ao cenário observado nos estudos com o YouTube, onde essas categorias estão entre as mais populares [Cheng et al., 2008].

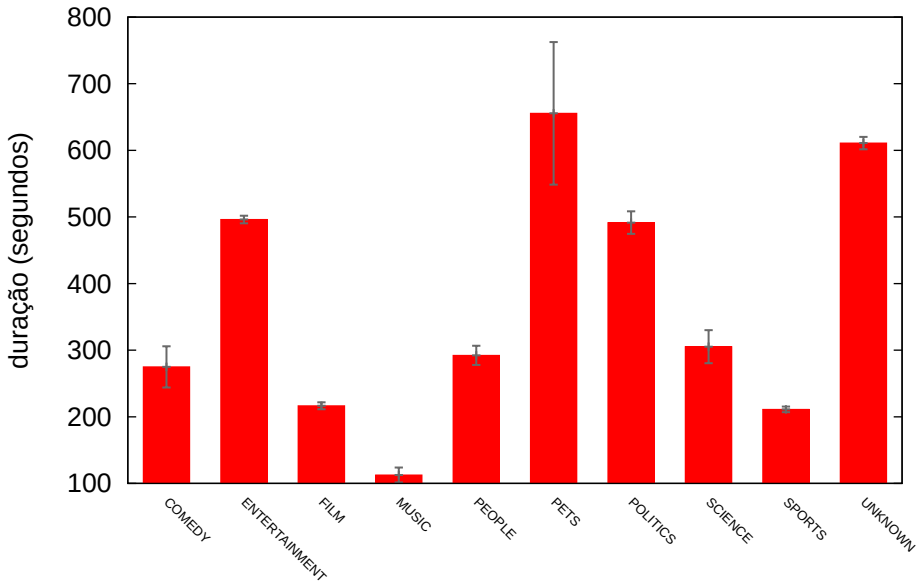

Figura 3.3. Duração média dos vídeos por categoria (barras de erro indicam um intervalo de confiança de 95%).

3.3 *Sites* Provedores de Conteúdo

A coleção desse estudo de caso envolve registros de 38 *sites* provedores de mídias especializadas da Internet brasileira. As identidades dos *sites* foram preservadas por questões de privacidade de conteúdo. Porém, é importante caracterizá-los para termos uma visão geral do mercado de *sites* de ME no Brasil.

A Figura 3.4 exibe a distribuição de vídeos por *site*, com cada coluna do gráfico correspondendo a um *site* da coleção (totalizando 38). Esse experimento evidencia como é a distribuição das publicações dos vídeos, assistidos no período de coleta (oito semanas), entre os produtores de conteúdo. Vídeos publicados antes da data inicial da coleta, mas assistidos no período coletado, foram, portanto, considerados.

A distribuição da Figura 3.4 é bastante heterogênea, incluindo poucos provedores com uma grande quantidade de vídeos e muitos *sites* que publicaram, relativamente, poucos vídeos (menos de dois mil). O portal com maior número de publicações, no período de coleta, disponibilizou 32.964 vídeos, enquanto que a menor quantidade de vídeos produzidos por um *site* foi 124.

A Figura 3.5 apresenta as distribuições de vídeos publicados por categoria para os seis *sites* com maior quantidade de visualizações no período de coleta. Podemos observar, pela figura, que a representatividade de cada categoria varia muito de um *site* para outro. Em alguns portais, onde o conteúdo é variado, existem vídeos de diferentes categorias. O quinto *site*, por exemplo, só não possui publicações de vídeos em

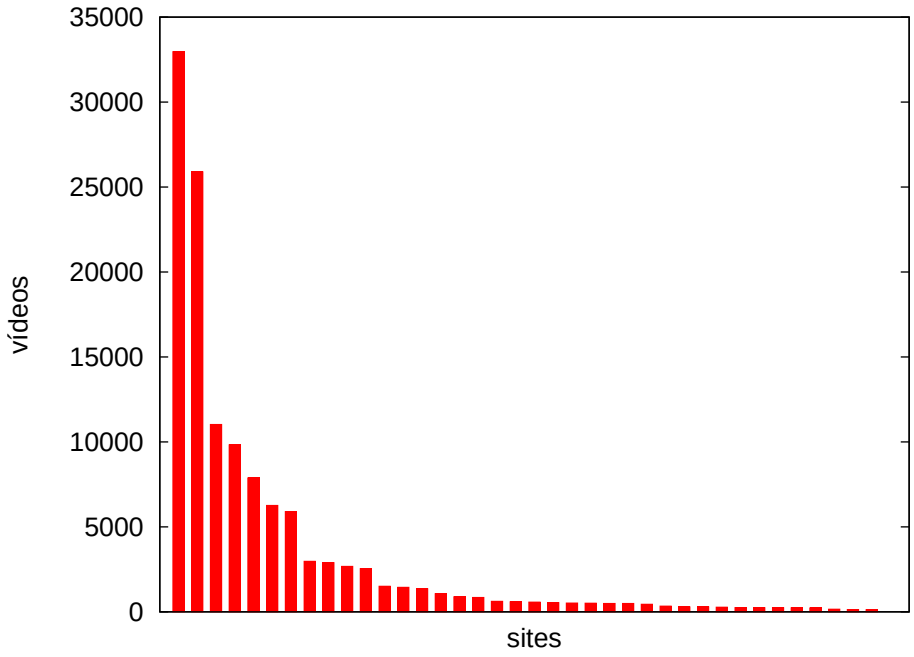

Figura 3.4. Distribuição de vídeos por *site*.

Comedy. Por outro lado, alguns *sites* são provedores de conteúdos mais especializados, privilegiando uma ou algumas poucas categorias. O sexto *site*, por exemplo, tem praticamente todos seus vídeos classificados como *Sports*. Um aspecto comum da maioria dos *sites* é a ocorrência de muitas mídias não categorizadas (em *Unknown*). Alguns provedores de conteúdo, como é o caso do quarto *site* com maior número de vídeos, não atribuem categoria a nenhum de seus vídeos. As distribuições de vídeos publicados por categoria para todos os *sites* da coleção podem ser visualizadas no Apêndice A.

A Figura 3.6 mostra a distribuição de visualizações por *site*, considerando todo o período de coleta (oito semanas). Cada coluna do gráfico corresponde a um *site* e a disposição das colunas segue a mesma ordem do gráfico da Figura 3.4. Ou seja, as colunas foram ordenadas pelo número de vídeos do *site* correspondente. Analisando os gráficos das Figuras 3.4 e 3.6, verificamos que a distribuição de acessos de vídeos de um *site* não é proporcional à quantidade de vídeos publicados nesse *site*. De fato, o *site* com maior quantidade de visualizações (mais de 30 milhões em oito semanas) é o quinto *site* com maior quantidade de vídeos. Enquanto que o portal com mais vídeos ficou em quarto lugar em número de acessos (cerca de 14 milhões em oito semanas). O segundo *site* com mais acessos (aproximadamente 17 milhões em todo período de coleta) é apenas o décimo quinto *site* com mais vídeos publicados.

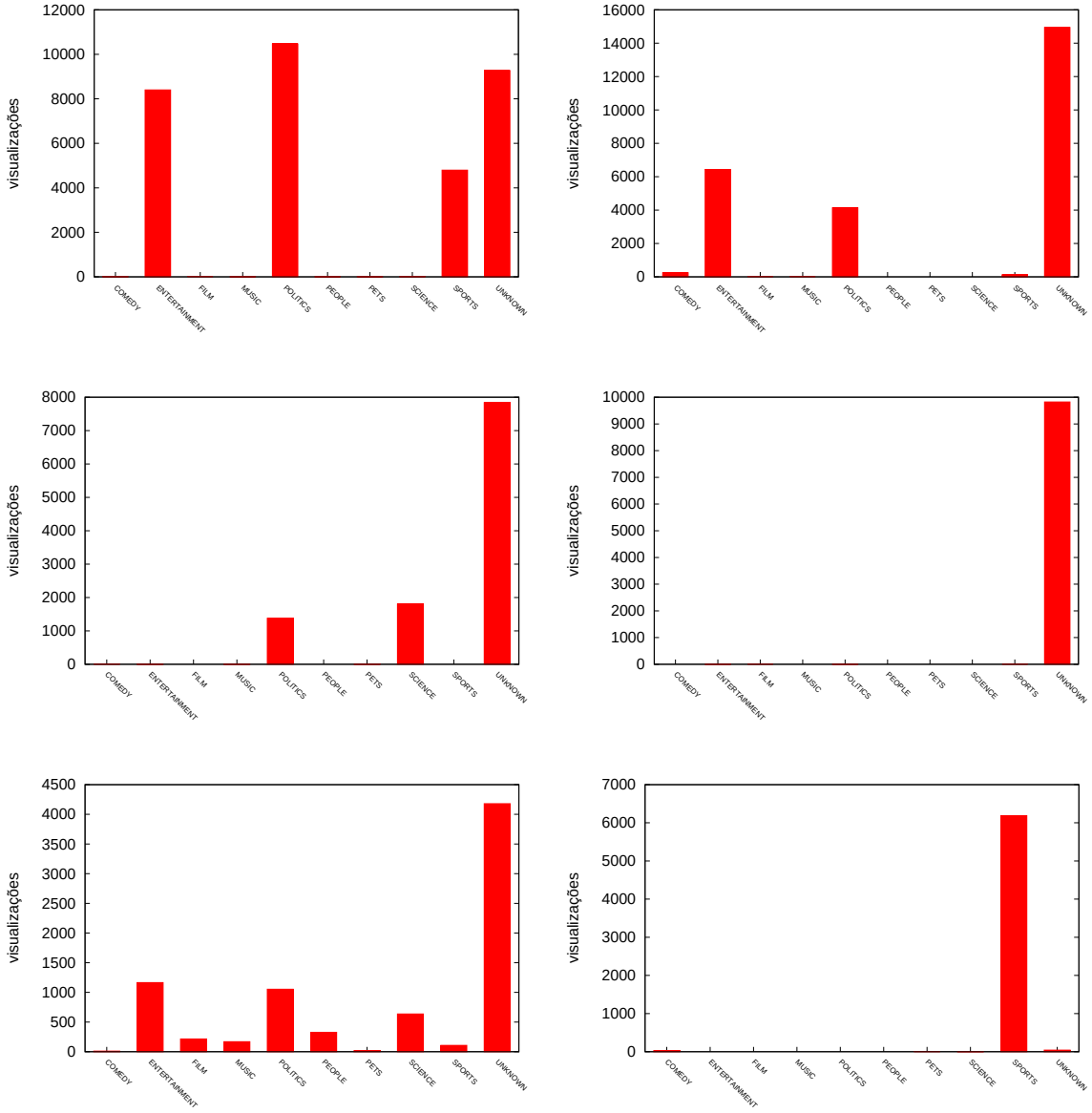

Figura 3.5. Distribuições de vídeos publicados por categoria para os seis *sites* com maior número de visualizações.

3.4 Visualizações

O número de visualizações de um vídeo é uma medida bastante relevante, que pode ser tomada como uma estimativa de popularidade ou até mesmo como uma métrica do quão interessante é o vídeo para o público espectador. Deve ser levado em conta, porém, que existem muitos outros fatores que afetam essa medida, tais como posicionamento do vídeo no *site*, tempo de disponibilidade, momento e contexto de publicação.

Os acessos aos vídeos pelos portais de mídias especializadas analisados, em ge-

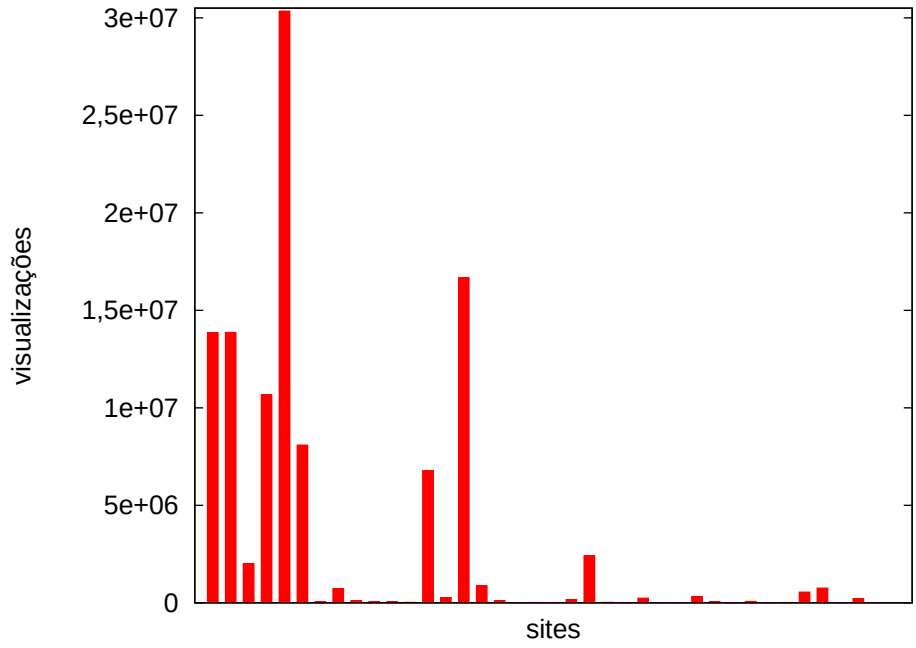

Figura 3.6. Distribuição de visualização por *site* (*sites* ordenados por quantidade de vídeos publicados), durante oito semanas.

ral, não dependem de nenhum procedimento de autenticação. No entanto, é possível acompanhar quais vídeos foram visualizados por um usuário, em diferentes contextos, durante um período de tempo, utilizando o recurso de *cookies* (como foi detalhado na Seção 2.2.1). Com base no acompanhamento dos acessos, foram computados, na etapa de análise estática, o total de visualizações para cada usuário e vídeo da coleção.

3.4.1 Visualizações por Usuário

A Figura 3.7 apresenta a distribuição acumulada complementar (*Complementary Cumulative Distribution Function - CCDF*) do número de visualizações por usuário de todo o conjunto de dados. Interpretando o gráfico, pode-se dizer que alguns poucos usuários chegaram a fazer mais de 100 mil acessos (provavelmente por meio de algum *script*).

A distribuição acumulada complementar revela uma curva com o comportamento de uma "cauda longa" (*long tail*) [Bingham & Spradlin, 2011]: muitos usuários assistiram uma quantidade reduzida de vídeos, enquanto uma pequena fração dos usuários visualizaram substancialmente mais vídeos. De fato, menos de 10% dos espectadores assistiram pelo menos dez vídeos e muito menos de 1% assistiu pelo menos uma centena de vídeos. A curva característica da cauda longa pode ser melhor observada ao se gerar

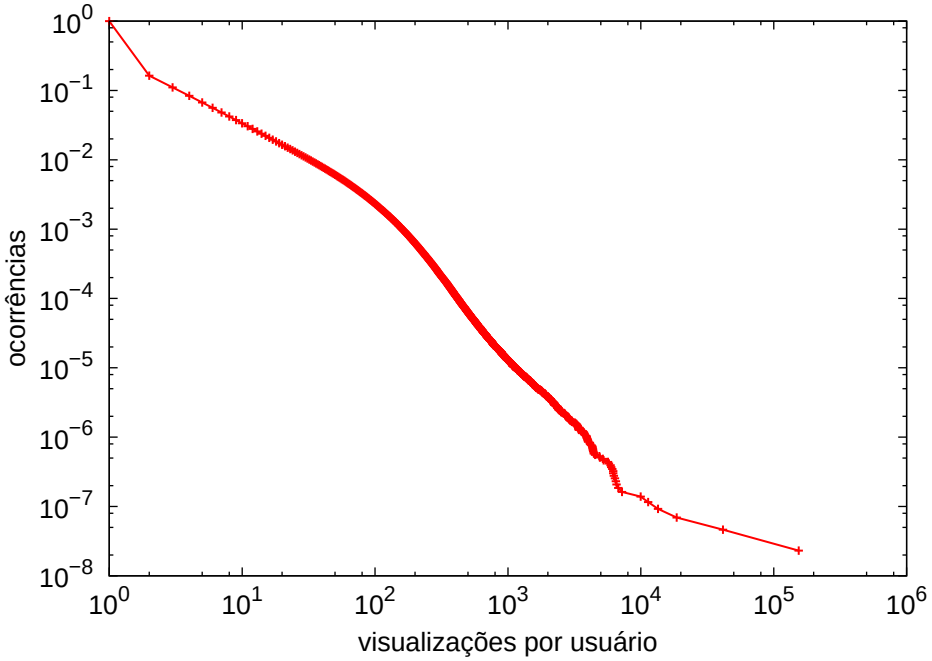

Figura 3.7. Distribuição acumulada complementar de visualizações por usuário (*CCDF*).

o gráfico da distribuição de acessos por usuário aplicando a escala logarítimica em seus eixos.

O mesmo comportamento identificado também é observado em estudos no contexto de *sites* de conteúdo gerado por usuários. Em [Gill et al., 2008], por exemplo, é mostrado que as distribuições de popularidade de acessos de objetos (vídeos e *tags*) no YouTube também seguem o fenômeno da cauda longa. Outras semelhanças foram encontradas no estudo das visualizações por vídeo, descrito a seguir.

3.4.2 Visualizações por Vídeo

A Figura 3.8 exibe a distribuição acumulada complementar (*CCDF*) do número de visualizações por vídeo de todo o conjunto de dados.

Pela observação do gráfico, é possível notar uma distribuição com curva menos íngreme, onde mais de 10% de todos os vídeos receberam pelo menos uma centena de acessos. Por outro lado, uma parcela muito menor de vídeos recebeu mais de mil visualizações. Esse comportamento, típico de curvas com cauda longa, pode ser melhor visualizado em um gráfico com eixos em escala logarítimica.

Apesar de apresentar uma cauda longa, a distribuição de visualizações por vídeo não segue a "Lei de *Zipf*" [Manning & Schütze, 1999], uma lei de potências sobre a

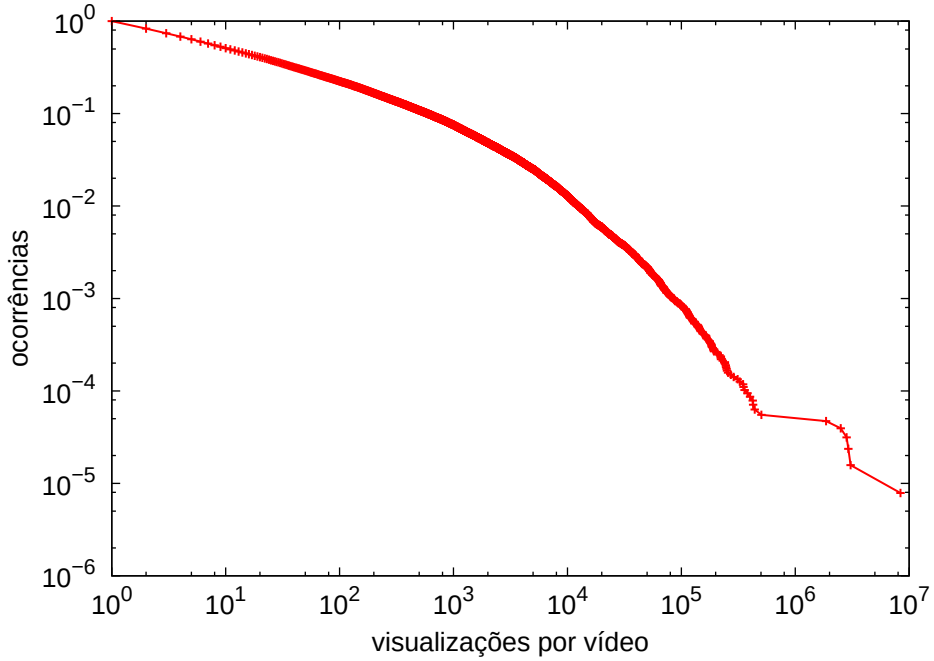

Figura 3.8. Distribuição acumulada complementar de visualizações por vídeo (*CCDF*).

distribuição de valores de uma lista, de acordo com a ordem de seus itens. A Lei de *Zipf* formula que, em uma lista, a frequência de qualquer item é inversamente proporcional à sua posição. Assim, o item mais frequente ocorrerá, aproximadamente, duas vezes mais do que o segundo item mais frequente, três vezes mais do que o terceiro mais frequente e assim sucessivamente [Manning & Schütze, 1999]. Visualmente, a curva da distribuição deveria se aproximar de uma reta descendente para se enquadrar na Lei de *Zipf*.

As análises da distribuição de acessos por vídeo estão de acordo com estudos reportados sobre *sites* de CGU. Em particular, resultados similares foram apresentados em trabalhos sobre o YouTube [Acharya et al., 1999; Cheng et al., 2008].

Uma possível justificativa para o comportamento observado na Figura 3.8 é o fato de os produtores de vídeos tenderem a acessar recorrentemente os vídeos recém publicados para validá-los (verificar se não ocorreram erros no processo de publicação) e para compartilhá-los. Esse procedimento comum pode afetar a distribuição de acessos dos vídeos, fazendo com que a curva não corresponda exatamente à Lei de *Zipf*, apesar de apresentar uma cauda longa.

3.5 Conclusões da Análise Estática

Considerando a primeira questão de pesquisa levantada (Seção 2.2.3), sobre os padrões de acesso derivados de uma visão estática das oito semanas de dados coletados, a análise apresentada neste capítulo revela um interessante paralelo com estudos anteriores sobre padrões de visualização de vídeos em *sites* de CGU (principalmente no YouTube).

Em particular, sobre a popularidade de categorias (em termos de número de vídeos) identificamos que as duas categorias que prevalecem no YouTube, *Music* e *Comedy*, têm pouca representatividade em nossa coleção de *sites* de mídias especializadas. Por outro lado, o restante da distribuição de vídeos por categoria é, em geral, semelhante entre portais de ME e *sites* de CGU [Cheng et al., 2008].

Também foi observado que a duração média de vídeos nos portais de mídias especializadas é bastante dependente da categoria dos vídeos, enquanto que a maioria dos vídeos do YouTube apresenta durações menores, independentemente da categoria atribuída [Cheng et al., 2008].

Investigando a distribuição de vídeos e de acessos entre os *sites* da coleção, foi possível verificar que o cenário dos provedores brasileiros de mídia especializada é bastante heterogêneo. A diversidade ocorre no tamanho dos portais (número de vídeos e de acessos) e também em relação ao conteúdo publicado.

Por fim, em linha com pesquisas anteriores no contexto de *sites* de CGU [Acharya et al., 1999; Cheng et al., 2008], observamos o comportamento de cauda longa para as distribuições de acessos por usuário e por vídeo. No caso da distribuição de visualizações por vídeo, também foi identificado que a curva não segue estritamente a Lei de *Zipf*, com vídeos na cauda longa apresentando substancialmente mais visualizações que o esperado.

Capítulo 4

Análise Temporal

No Capítulo 3, todo o conjunto de dados do estudo de caso foi analisado de forma agregada, considerando apenas os atributos estáticos da coleção. A análise estática nos forneceu uma visão geral do mercado de vídeos *online*, mais especificamente do mercado brasileiro de mídia especializada, em termos de aspectos como volume de acesso, *sites* provedores, consumo de vídeos e características do conteúdo publicado.

Além dos padrões estáticos, existem comportamentos recorrentes que se dão ao longo do tempo e que não são perceptíveis a partir de uma análise agregada dos registros. Para identificar esses padrões, é necessário acompanhar como ocorrem a distribuição e o consumo de vídeos no decorrer de um período de tempo.

Os padrões de acesso temporais podem ser bastante relevantes, pois servem, muitas vezes, como uma previsão do impacto da escolha do conteúdo e da estratégia de entrega para o usuário final. Porém, esses são padrões difíceis de serem obtidos, por requererem um acompanhamento minucioso da evolução dos acessos, em diferentes contextos e envolvendo grande quantidade de usuários e vídeos.

Como detalhado na Seção 2.2.2, cada registro de interação de um usuário com um vídeo, através do *player*, pode conter diversos eventos, tais como *play*, *stop* e *progress*. Cada evento tem associado o tempo de ocorrência. Nesse capítulo, essas informações serão aplicadas na investigação dos aspectos temporais do consumo de vídeos *online*. As análises serão conduzidas com base na questão de pesquisa Q2, levantada na Seção 2.2.3: "quais padrões temporais podem ser inferidos da análise das interações de usuários com vídeos ao longo do tempo?".

Em particular, a análise temporal deve abranger quatro aspectos principais: os padrões de acesso ao longo do tempo, a taxa de retenção dos vídeos, a publicação de conteúdo e a expectativa de vida (*lifespan*) dos vídeos.

4.1 Padrões de Acesso

O comportamento de usuários ao consumirem vídeos *online* pode variar de acordo com as particularidades da rotina diária de cada um. Além disso, o tipo de conteúdo assistido, os momentos e a frequência de acesso também dependem muito das preferências individuais. Apesar das especificidades, é possível identificar alguns padrões de acesso comuns pela análise de interações de usuários com vídeos, em *sites* de mídia especializada, ao longo do tempo.

A Figura 4.1, por exemplo, exibe o número de visualizações (acessos) de vídeos por dia, ao longo de todo o intervalo da coleta. O primeiro dia das oito semanas (56 dias), 24 de Junho de 2012, é um domingo, enquanto o último dia, 18 de Agosto de 2012, é um sábado. O número médio de visualizações por dia, no período, é de 1.848.972,21 (quase dois milhões), com valores extremos de 1.094.646 e 2.416.967.

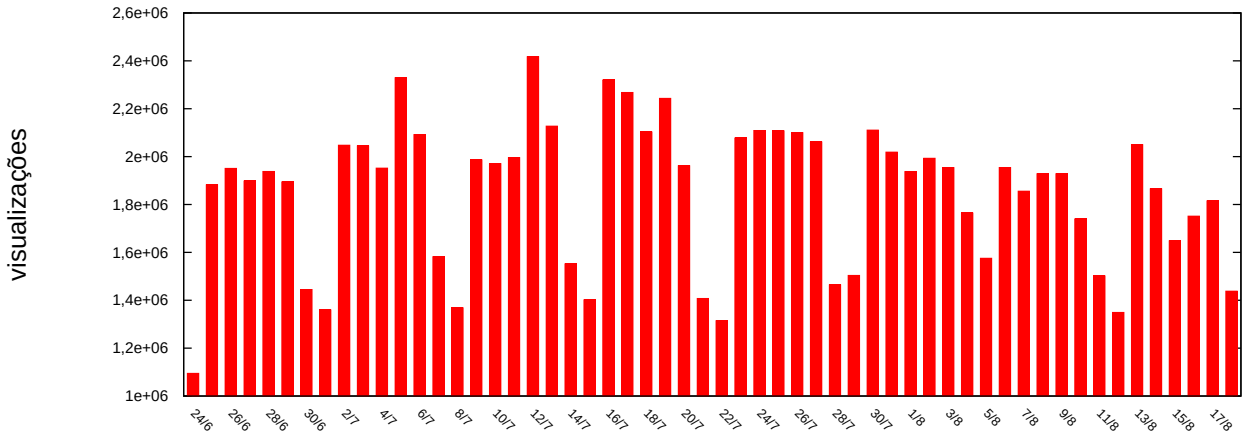

Figura 4.1. Número de visualizações por dia ao longo de oito semanas.

A partir da Figura 4.1, observamos, de imediato, que existe um padrão cíclico de acesso aos vídeos do tipo ME. Em todo o período analisado, o número de visualizações em um dia de semana (dia útil) é quase duas vezes maior que a quantidade de acessos no sábado ou domingo. Também deve ser observado que, com exceção do fim de semana dos dias 28 e 29 de julho, mais vídeos foram assistidos no sábado que aos domingos. Na verdade, o domingo é o dia com o menor número médio de acessos da semana, 1.370.790,87, enquanto que quinta-feira é o dia com a maior média, 2.087.578,37. O primeiro dia do gráfico (24 de Junho) aparece com um número relativamente menor de visualizações pelo fato da coleta ter sido iniciada na metade desse dia.

Para entender melhor os padrões temporais de acesso, analisamos a distribuição de visualizações ao longo do tempo em uma granularidade menor. A Figura 4.2 mostra o número de acessos por hora do dia. Para facilitar a visualização, os resultados foram

apresentados para uma única semana: do dia 15, um domingo, até o dia 21 de Julho, um sábado. Essa é uma semana típica da coleção, já que os mesmos padrões se repetem nas demais semanas do conjunto de dados.

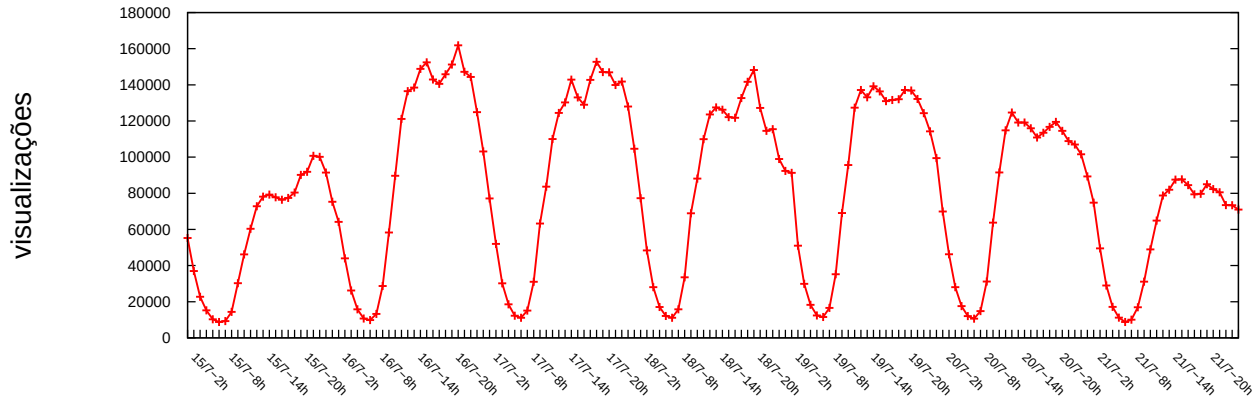

Figura 4.2. Número de visualizações por hora do dia ao longo de sete dias.

A partir da análise da Figura 4.2, notamos que a distribuição de acessos por hora do dia segue um padrão esperado e bem definido. Mais especificamente, há um aumento acelerado do número de visualizações entre 7:00 e 12:00 horas. Então, os acessos continuam em crescimento, porém desacelerando. O pico do número de visualizações ocorre, geralmente, entre 19:00 e 20:00 horas. Depois disso, a quantidade de visualizações começa a declinar rapidamente, até cerca das 6:00 horas, quando, normalmente, o valor mais baixo é atingido. Esse padrão de acesso é observado, invariavelmente, em todos dias. Mesmo nos fins de semana, quando o total de visualizações é menor, o mesmo comportamento é verificado, porém, em escala reduzida.

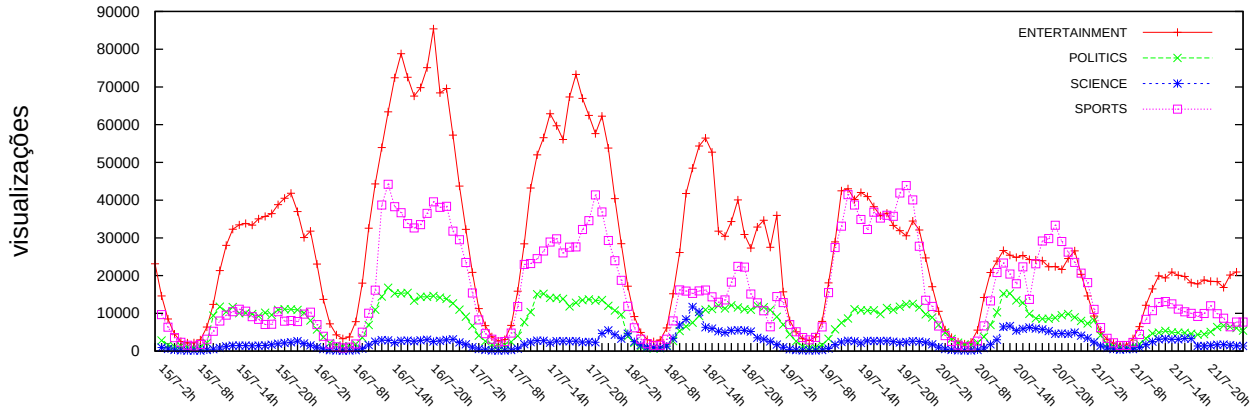

Figura 4.3. Número de visualizações por hora do dia, para as quatro categorias mais representativas, ao longo de sete dias.

A Figura 4.3 exibe as distribuições de visualizações por hora do dia para as quatro categorias mais representativas (em quantidade de vídeos), em uma semana. Podemos observar que o padrão comum identificado na Figura 4.2 também é notado nas distribuições individuais de cada categoria. Entretanto, é possível inferir alguns padrões específicos. Por exemplo, para a categoria *Sports*, o número de acessos é maior na segunda-feira e na quinta-feira. Esse fenômeno pode ser explicado pelo fato dos principais jogos de futebol no Brasil ocorrerem no domingo e na quarta-feira, o que resulta em mais visualizações de vídeos sobre o assunto nos dias subsequentes.

4.2 Taxa de Retenção

O número de visualizações de um vídeo pode ser tomado como um indicativo de sua popularidade, mas não, necessariamente, como uma medida do quanto os usuários gostaram do vídeo. Para servir de estimativa da qualificação de um vídeo por um determinado usuário, adotamos como métrica a "taxa de retenção". A taxa de retenção de um vídeo é dada pelo tempo despendido por um usuário assistindo o vídeo, dividido por sua duração.

Nos experimentos deste estudo de caso, adotou-se a diferença de tempo entre o último e o primeiro evento de um registro como uma aproximação para o tempo em que um usuário ficou assistindo um vídeo. Existe um evento para cada interação possível do usuário com o *player* (*play*, *pause*, *resume*, etc.). Além disso, é registrado um evento de progresso para cada quartil assistido do vídeo (incluindo um evento para o fim do vídeo). O primeiro evento de um registro é sempre um *play*, enquanto o último evento representa a última ação do usuário no *player*, podendo ser um evento de progresso, *pause* ou de finalização do vídeo.

A Figura 4.4 apresenta a distribuição acumulada complementar ($CCDF$) dos valores da taxa da retenção para todos os registros da coleção (coletados em oito semanas). A partir da análise do gráfico, verificamos que somente, aproximadamente, 25% dos registros (das interações entre um usuário e um vídeo) possuem uma taxa de retenção igual ou superior a 0,1. Isso quer dizer, que a maioria dos usuários assistem, em geral, menos que 10% do conteúdo dos vídeos.

A Figura 4.5 exibe a distribuição acumulada complementar ($CCDF$) da taxa de retenção para todos os registros, discriminando por categoria. Para favorecer a visualização, somente as distribuições para as quatro categorias mais representativas foram incluídas no gráfico. Analisando a figura, nota-se que todas as categorias apresentam curvas similares, seguindo o mesmo comportamento identificado na Figura 4.4. A ten-

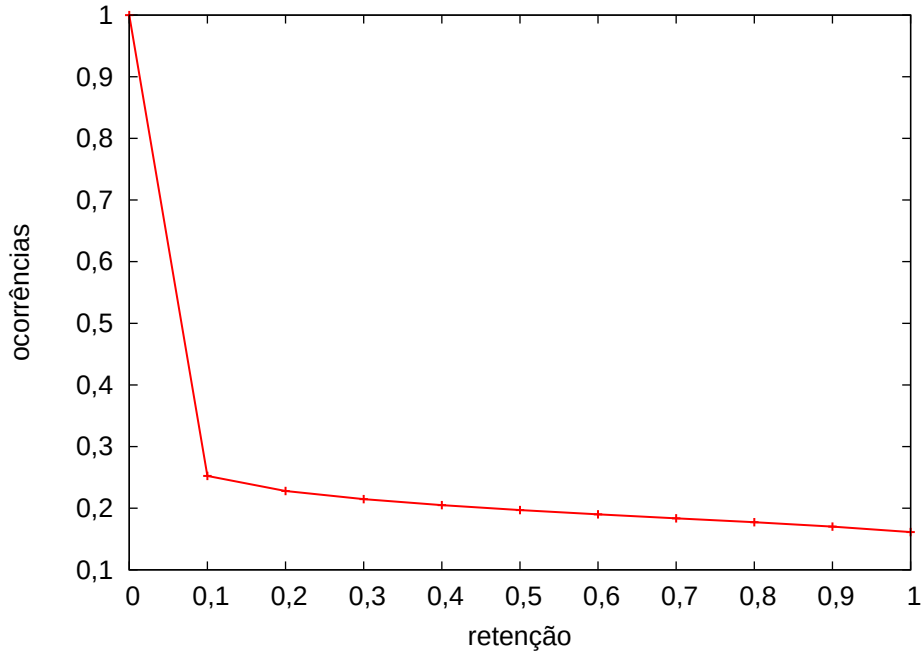

Figura 4.4. Distribuição acumulada complementar da taxa de retenção (*CCDF*).

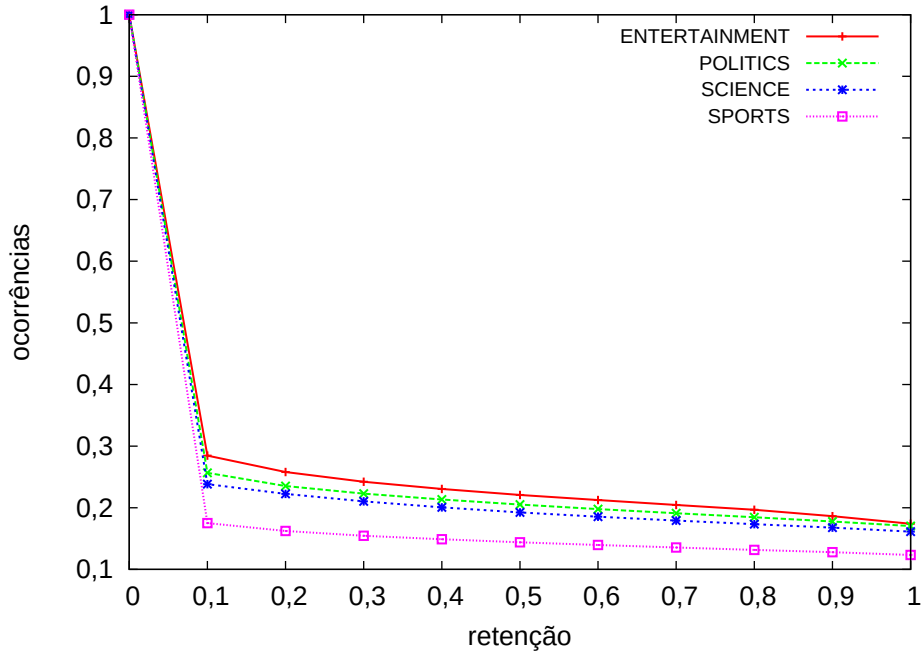

Figura 4.5. Distribuição acumulada complementar da taxa de retenção, para as quatro categorias mais representativas (*CCDF*).

dência, independente de categoria, é que quanto maior a taxa de retenção esperada, menor seja o número de ocorrências de registros que atinjam essa taxa de retenção.

Porém, as distribuições diferem mais substancialmente no percentual de ocorrências de registros que satisfaçam um valor mínimo de taxa de retenção baixo. As diferenças são mais aparentes, entre categorias, na quantidade de registros com pelo menos 10% de taxa de retenção. A categoria *Sports*, por exemplo, tem menos de 20% de seus registros com taxa de retenção de pelo menos 0,1 (10%), enquanto que, na categoria *Entertainment*, quase 30% das interações de usuários com vídeos resultam em pelo menos 0,1 (10%) de taxa de retenção.

4.3 Publicações

As análises anteriores contemplam o consumo de vídeos em *sites* de mídia especializada. Porém, para entender melhor os padrões de acesso, também é pertinente investigar um pouco da perspectiva dos provedores de conteúdo. Nesse sentido, a Figura 4.6 apresenta a distribuição do número de vídeos publicados por dia, ao longo das oito semanas da coleção. O primeiro dia, 24 de Junho de 2012, é um domingo, enquanto o último dia, 18 de Agosto de 2012, é um sábado.

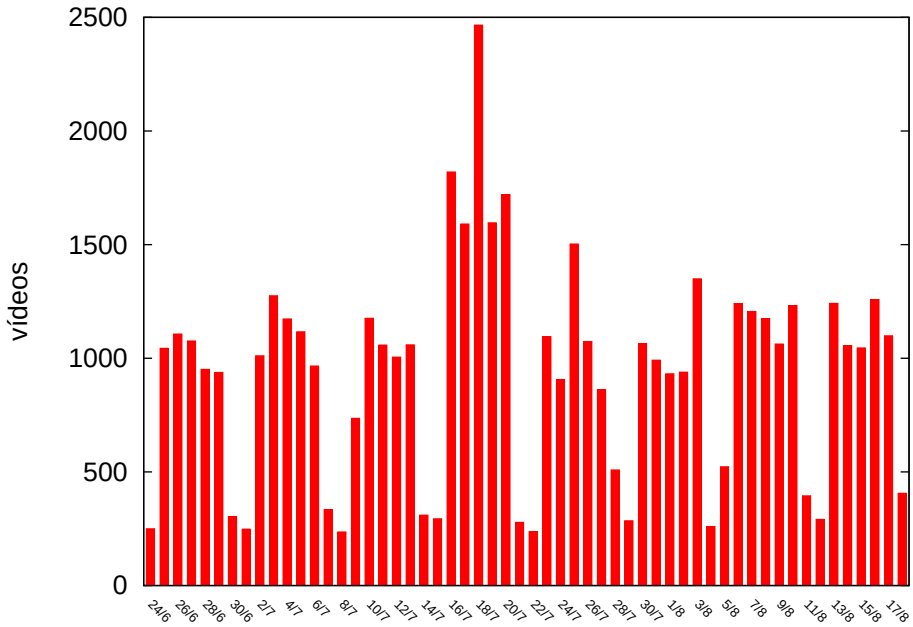

Figura 4.6. Distribuição de vídeos publicados por dia, ao longo de oito semanas.

A coleção de nosso estudo de caso contém registros de acessos envolvendo 127.068 vídeos (Tabela 2.1), que podem ser divididos em dois grupos:

- Vídeos publicados durante o período de coleta: 52.381;

- Vídeos publicados antes do período de coleta, mas que receberam acessos no período coletado: 74.687.

A distribuição da Figura 4.6 engloba somente os vídeos publicados durante o período de coleta (52.381). Podemos observar a ocorrência de um padrão cíclico de publicações, similar aquele identificado nas análises da distribuição de visualizações ao longo do tempo. Durante os dias de semana ocorrem mais publicações. Em particular, aproximadamente mil vídeos são disponibilizados diariamente, de segunda a sexta-feira, enquanto que, nos fins de semana, o número de publicações é reduzido, não passando de 300, em geral.

Os dias entre 16 e 20 de Julho, especialmente o dia 18, experimentaram um número excepcionalmente superior de publicações. Comparando a Figura 4.6 com a distribuição de visualizações por dia (Figura 4.1), verificamos que o aumento das publicações não implicou, diretamente, em um aumento da quantidade de acessos. Ou seja, aparentemente, não há uma relação de proporcionalidade entre a quantidade de vídeos publicados e o número de visualizações. O aumento no número de publicações foi ocasionado, principalmente, por dois portais de emissoras de televisão. Os conteúdos publicados pelos portais são, em geral, bastante variados, incluindo temas jornalísticos, notícias e vídeos de entretenimento. Na semana de pico, não houve um assunto específico predominante, mas alguns dos temas mais recorrentes foram: política (a semana em questão precedeu as eleições em dois meses e marcou o início da exibição do horário eleitoral na TV), Olimpíadas (os jogos olímpicos iniciaram na semana seguinte) e programação das emissoras (vídeos sobre novelas, jornais e programas diversos exibidos na televisão pelos canais).

4.4 Evolução das Visualizações

Na Seção 3.4, analisamos a distribuição de visualizações agregadas em todo o intervalo de coleta (oito semanas). Nessa análise, apenas o total de acessos recebidos por cada vídeo e o total de visualizações de cada usuário foram considerados. A análise agregada fornece uma visão geral da quantidade de acessos, de como esses estão distribuídos e de quais são os padrões de visualização de todo um intervalo de tempo. Porém, essa análise não mostra a evolução dos acessos. Nesta seção, investigamos como as visualizações ocorreram ao longo do tempo.

Para compreender a evolução dos acessos, é necessário acompanhar todos os acessos recebidos por vídeos, desde o momento a publicação. Para esse fim, consideramos apenas os vídeos que foram publicados na primeira semana de nossa coleção (entre os

dias 24 e 30 de Junho de 2012). Todos os acessos a esses vídeos foram registrados, diariamente, a partir do dia de publicação, até o último dia de coleta (18 de Agosto de 2012). No total, foram considerados 5.670 vídeos nesse experimento (Figura 4.6).

A Figura 4.7 apresenta a distribuição acumulada (*Cumulative Distribution Function - CDF*) das visualizações dos vídeos, fornecida pelo número de acessos diários desde o momento da publicação de cada vídeo. Analisando essa figura, verificamos que, aproximadamente, 37% das visualizações recebidas por um vídeo ocorreram no mesmo dia de sua publicação (representado pelo valor zero no eixo x). Além disso, 67% ocorreram no dia seguinte da data de publicação, 85% dos acessos aconteceram antes do quinto dia e 90% ocorreram antes de passados nove dias pós-publicação.

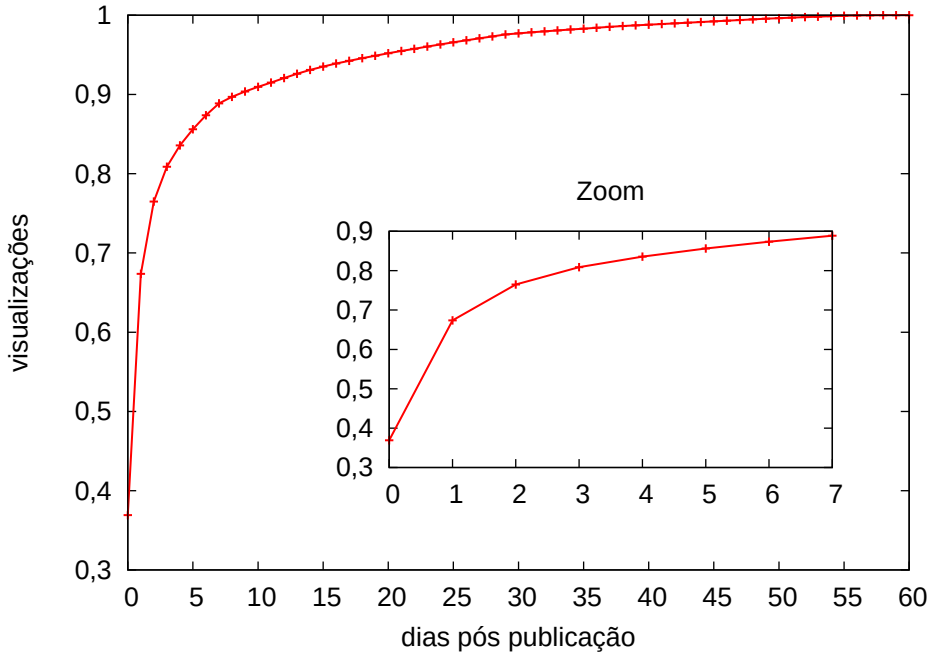

Figura 4.7. Evolução de visualizações (*CDF*).

Claramente, a evolução das visualizações, expressa pelo gráfico da Figura 4.7, segue uma função logarítmica. A maior parte dos acessos ocorreram logo que um vídeo foi publicado, enquanto poucas visualizações aconteceram nos dias subsequentes ao longo de poucas semanas. Além disso, poucos vídeos continuaram recebendo acessos após um mês da data de publicação, o que indica que o tempo de vida de vídeos do tipo ME é, em geral, muito curto.

A Figura 4.8 apresenta a distribuição acumulada (*CDF*) das visualizações diárias dos vídeos (como exibido na Figura 4.7), discriminando pelas quatro categorias mais representativas. Para facilitar a visualização dos padrões de acesso, a Figura 4.9 mostra a mesma distribuição somente para a primeira semana do intervalo.

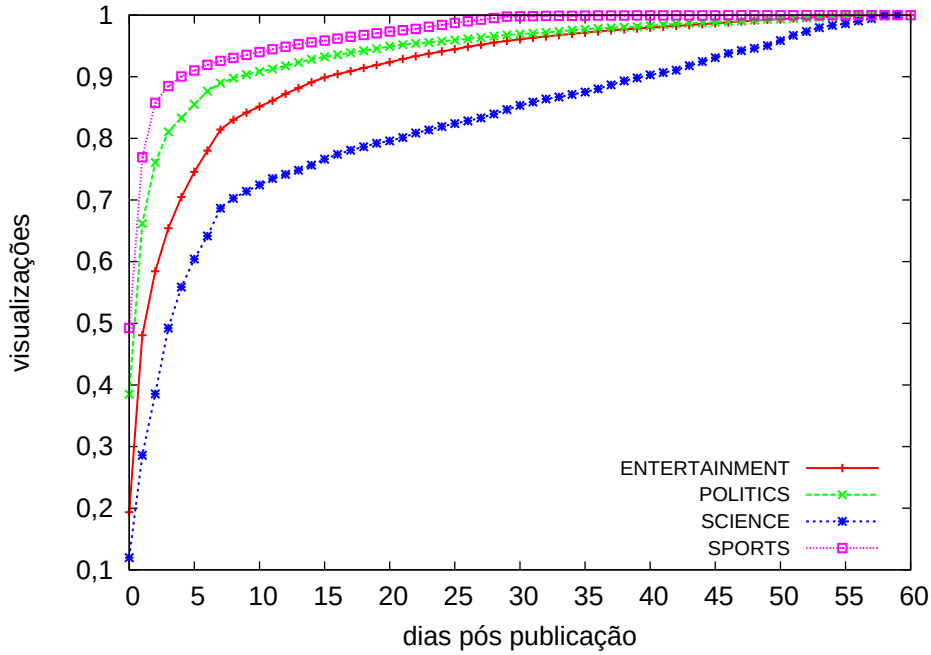

Figura 4.8. Evolução de visualizações por categoria (CDF).

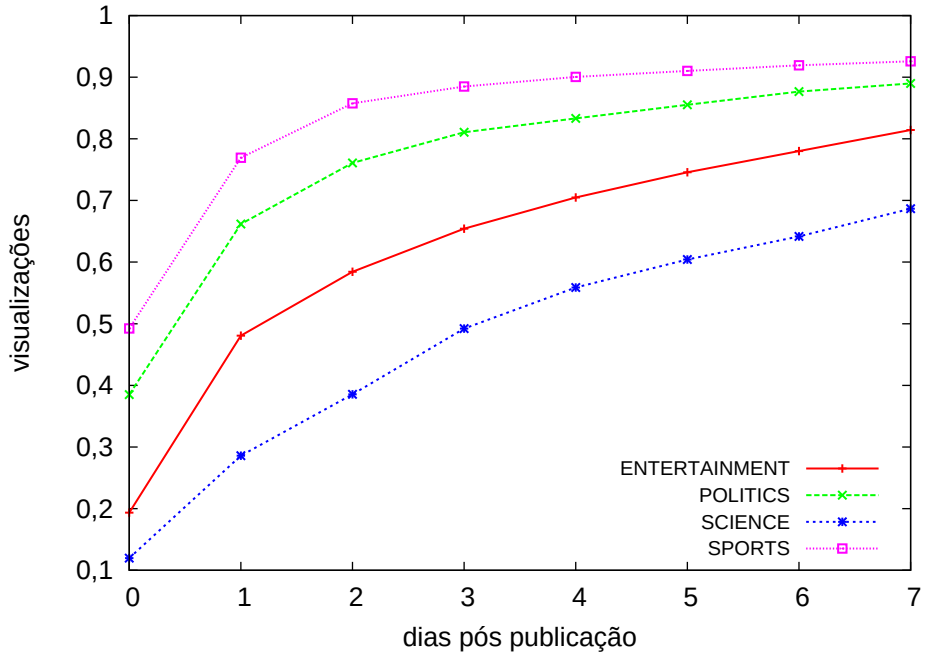

Figura 4.9. Evolução de visualizações por categoria (CDF) na primeira semana.

A partir da análise das Figuras 4.8 e 4.9, identificamos o mesmo padrão logarítmico observado na Figura 4.7. Porém, a diferença mais distinta entre as distribuições

de cada categoria é a taxa de convergência. Em particular, cerca de 77% dos acessos aos vídeos da categoria *Sports* ocorreram no primeiro dia de publicação, enquanto vídeos da categoria *Science* atingiram apenas cerca de 29% do total de acessos no mesmo período. Na verdade, para que ocorressem 77% das visualizações de vídeos da categoria *Science*, foram necessários quase 16 dias.

4.5 Conclusões da Análise Temporal

Tendo como referência a questão de pesquisa Q2, levantada na Seção 2.2.3, identificamos, no presente capítulo, diversos padrões temporais de visualização de vídeos, no período coberto pelo nosso conjunto de dados.

Ao analisar as visualizações de mídias especializadas por dia, observou-se a existência de um padrão cíclico, em que o número de acessos durante os dias úteis da semana é muito maior do que a quantidade de acessos nos fins de semana. Um padrão recorrente também ocorre na distribuição de acessos diários. Com isso, é possível estimar a quantidade de visualizações em função do dia e horário. Esse tipo de análise é bastante relevante para os provedores de conteúdo estimarem o impacto de suas publicações e para os responsáveis pela distribuição dos vídeos administrarem os recursos necessários (infraestrutura de redes e servidores, por exemplo) com base nos padrões de acesso identificados.

Também notamos que os valores de retenção dos vídeos variam ligeiramente entre as categorias, mas, em geral, são baixos para a maioria dos usuários. A retenção pode ser tomada como uma medida do interesse de usuários pelos vídeos. É interessante, portanto, para os produtores de conteúdo, terem como objetivo o aumento das taxas gerais de retenção. Verificamos, nas análises estáticas, que a duração dos vídeos do tipo ME é, geralmente, grande. A duração elevada dos vídeos é uma possível justificativa para os valores baixos de retenção.

Analisando o lado dos produtores de conteúdo, identificamos um padrão cíclico e esperado na distribuição de vídeos publicados por dia. Esse comportamento é similar ao padrão da distribuição de visualizações por dia, mas não foi verificada uma correspondência direta entre o número de vídeos publicados e o número de visualizações.

Com a análise da evolução dos acessos ao longo do tempo, podemos concluir, como já relatado por [Cheng et al., 2007] para vídeos do YouTube, que a expectativa de vida dos vídeos (*lifespan*) é, em geral, muito baixa. Vídeos de algumas categorias recebem a maior parte de suas visualizações no primeiro dia pós-publicação, enquanto vídeos de outras categorias demoram, usualmente, um pouco mais de tempo para

atingir a mesma porcentagem de acessos. Por fim, os provedores de conteúdo podem usar os padrões temporais identificados para estimar com que frequência deve ser feita a atualização (ou reposição) dos vídeos publicados e como meio de predição de estatísticas de acesso. Também é importante considerar as diferenças entre as categorias em relação aos padrões de evolução das visualizações de mídias especializadas.

Capítulo 5

Análise Transacional

Nos Capítulos 3 e 4, analisamos o consumo de vídeos em *sites* brasileiros de mídia especializada, considerando aspectos estáticos e temporais. Com isso, foi possível revelar padrões de acesso gerais, entender melhor o mercado de vídeos do tipo ME e identificar comportamentos recorrentes de usuários assistindo vídeos ao longo de um período de tempo.

Nas análises estáticas, todas as visualizações foram agrupadas em um único conjunto, enquanto que, nas análises temporais, os acessos foram agregados por data. Porém, ainda falta analisar os acessos agrupando-os de uma forma próxima de como esses ocorrem na prática. Em geral, um usuário comum assiste frequentemente mais de um vídeo em sequência, criando uma relação implícita entre eles. Neste capítulo, pretendemos analisar os vídeos e suas visualizações, agrupadando-os por relações desse tipo. Além disso, queremos entender os padrões de transação dos usuários entre diferentes conteúdos.

As análises serão conduzidas tendo como referência a questão de pesquisa Q3, levantada na Seção 2.2.3: "Quais relações recorrentes podem ser identificadas de vídeos assistidos, frequentemente, em sequência?".

Esperamos que os padrões identificados possam ser relevantes para entender o comportamento geral dos usuários ao navegarem entre vídeos, possibilitanto, assim, o aprimoramento dos serviços e conteúdos oferecidos.

5.1 Modelagem

Vídeos assistidos em sequência, frequentemente, têm uma relação (implícita) entre si, mesmo que não sejam similares em termos de conteúdo. Identificar esse tipo de relação pode ser muito relevante para o conhecimento dos padrões de consumo e, consequente-

mente, para a melhoria da experiência dos usuários. Porém, muitas vezes, as relações entre vídeos não são muito óbvias, o que dificulta a identificação dos padrões.

A coleção usada em nosso estudo de caso contém registros de usuários acessando diversos vídeos, em diferentes *sites*. Esses dados constituem uma visão privilegiada do mercado brasileiro de mídia especializada. A abrangência dos dados nos permite acompanhar o fluxo de navegação dos usuários assistindo vídeos de diferentes provedores de conteúdo (em diferentes *sites*).

Para essa análise, foi selecionada uma semana típica, do dia 15 (domingo) ao dia 21 de Julho (sábado). A escolha de uma semana de dados é apropriada, nesse caso, pois consideraremos, em nossas análises, dados agrupados em conjuntos reduzidos (acessos feitos por um mesmo usuário entre intervalos de minutos). O uso de toda a coleção iria implicar em uma quantidade muito grande de conjuntos, o que inviabilizaria os experimentos.

Data inicial	15 de Julho, 2012 (Dom)
Data final	21 de Julho, 2012 (Sáb)
Registros	13.703.842
Usuários únicos	6.179.640
Vídeos únicos	59.319
Duração dos vídeos (média)	398,0s
Duração dos vídeos (desvio padrão)	131,9s

Tabela 5.1. Estatísticas gerais dos dados da semana selecionada para as análises transacionais.

A Tabela 5.1 contém estatísticas gerais da semana selecionada para análise. Apesar de termos restringido os dados a uma semana, são quase 14 milhões de registros, mais de seis milhões de usuários e quase 60 mil vídeos considerados.

Pretendemos, nas análises seguintes, agrupar as visualizações considerando vídeos acessados em sequência. Para tanto, adotamos o conceito de "sessão". Uma sessão é definida como sendo uma série de requisições realizadas por um usuário, em um *site*, durante um determinado período de tempo [Menascé et al., 1999]. O conceito de sessão, nesse caso, não pode ser confundido com o conceito tradicional de "sessão de *websites*". Sessões de *sites* são usadas para identificação de um usuário em um *site* específico e, em geral, exigem algum mecanismo de autenticação. Em nosso estudo de caso, não há autenticação explícita, os acessos de um usuário em diferentes *sites* são associados por meio de *cookies* (como foi detalhado na Seção 2.2.1).

O que define, na prática, quais acessos pertencem a uma sessão é a sua duração, que é dada em função do "tempo de expiração da sessão". O tempo de expiração da

sessão é uma medida do tempo de inatividade do usuário e serve para delimitá-la. Uma sessão inclui todos os vídeos assistidos em sequência pelo mesmo usuário desde que o intervalo entre cada vídeo subsequente (na ordem em que foram assitidos) não ultrapasse o tempo de expiração estipulado. Ou seja, para agrupar vídeos em sessões, basta considerar a lista dos vídeos assistidos por cada usuário em ordem cronológica e ir adicionando, a partir do primeiro, os vídeos em uma sessão. Esse procedimento deve ser repetido até que o intervalo entre o próximo vídeo da sequência e o último adicionado seja maior que o tempo de expiração, quando uma nova sessão deve ser iniciada.

Em nosso estudo de caso, cada registro de acesso corresponde a um conjunto de ações de um usuário interagindo com um vídeo. Para definição da sessão, o intervalo de tempo entre vídeos foi considerado como sendo a diferença de tempo entre o momento do último evento de um registro e o momento do primeiro evento do registro de acesso seguinte.

A partir do conceito de sessão, podemos dividir os registros em entidades que representam acessos subsequentes. Esses acessos definem relações implícitas entre vídeos. Pretendemos identificar padrões relevantes a partir dessas relações. Para tanto, propomos uma modelagem em rede dos vídeos e de seus relacionamentos.

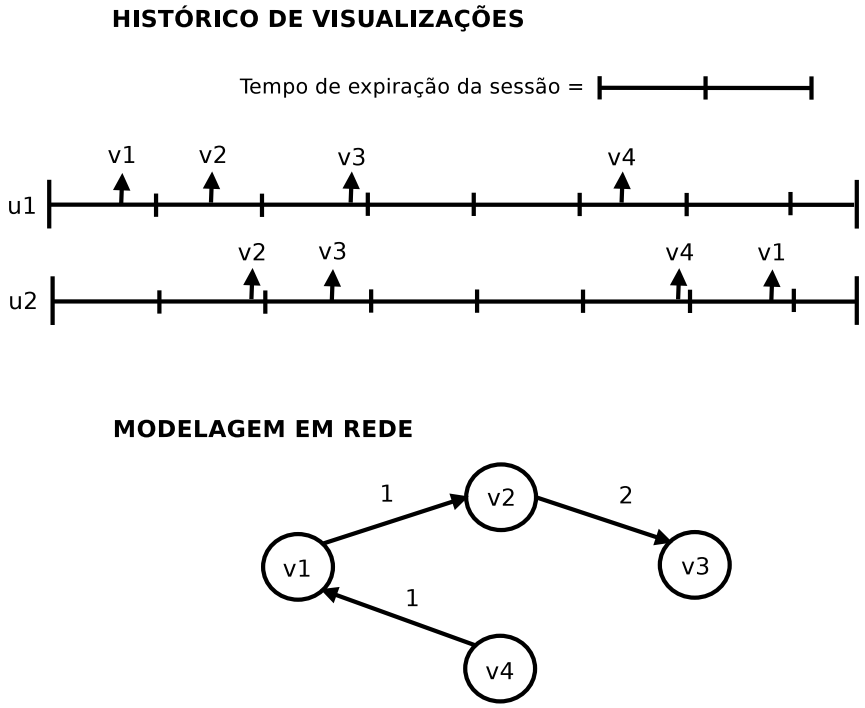

Figura 5.1. Modelagem em rede do histórico de vídeos assistidos por usuários.

A Figura 5.1 ilustra a modelagem em rede a partir do conceito de sessão. Na figura são mostrados os esquemas dos históricos de acesso de dois usuários, u_1 e u_2. Cada parte da linha do tempo de um usuário corresponde a uma unidade de tempo. Assim sendo, o tempo de expiração da sessão foi definido, para efeito de ilustração, como sendo equivalente a duas unidades de tempo. O usuário u_1 assistiu aos vídeos v_1, v_2, v_3 e v_4 (em ordem de ocorrência), enquanto o usuário u_2 assistiu aos vídeos v_2, v_3, v_4 e v_1 (nessa ordem). O primeiro passo da modelagem consiste em dividir as visualizações em sessões. No caso do esquema da Figura 5.1, uma sessão é formada por vídeos assistidos por um mesmo usuário, desde que o intervalo entre vídeos visualizados em sequência não seja maior que duas unidades de tempo. Portanto, as sessões identificadas no exemplo são:

- Sessões do usuário u_1: $\{v_1, v_2, v_3\}$ e $\{v_4\}$;
- Sessões do usuário u_2: $\{v_2, v_3\}$ e $\{v_4, v_1\}$.

Os relacionamentos entre vídeos de uma mesma sessão podem ser modelados por meio de uma "rede complexa". Uma rede complexa modela entidades de um sistema, no qual existe uma regra que estabelece conexões entre essas entidades [Newman, 2003]. Para a representação de redes complexas, normalmente, são usados "grafos". Grafo é uma forma de representação de um conjunto de objetos, denominados vértices, em que alguns pares são conectados por ligações (*links*), denominados arestas [Bondy & Murty, 1976].

Em nossa modelagem, as entidades de interesse são os vídeos e a regra que estabelece a existência de uma conexão entre um par de vídeos é um vídeo suceder outro em alguma sessão de usuário. Os relacionamentos entre vídeos são representados por "grafos ponderados direcionados", em que os vértices do grafo são vídeos, as arestas relacionam vídeos assistidos em sequência em uma sessão, sendo direcionados de um vídeo para seu sucessor, e os pesos das arestas representam o número de ocorrências daquela sucessão de vídeos nas sessões.

A Figura 5.1 também apresenta a representação em grafo do exemplo descrito. O grafo contém quatro vértices, representando os vídeos v_1, v_2, v_3 e v_4, e três arestas ponderadas e direcionadas de acordo com as sessões identificadas. Por exemplo, a partir da sessão $\{v_1, v_2, v_3\}$, do usuário u_1, foram geradas as arestas $e_1 = \{v_1, v_2\}$ e $e_2 = \{v_2, v_3\}$. O peso da segunda aresta, de valor igual a dois ($p(e_2)$=2), ocorre por causa da sucessão dos vídeos v_2 e v_3, que também ocorre na sessão $\{v_2, v_3\}$, do usuário u_2.

5.2 Definição do Tempo de Expiração da Sessão

O principal parâmetro do modelo proposto para as análises transacionais (análise das transações de usuários) é o tempo de expiração da sessão. Como já foi mencionado, o tempo de expiração da sessão mede o período de inatividade do usuário e determina o tamanho da sessão. Assim, vídeos assistidos consecutivamente por um usuário pertencem à mesma sessão se o intervalo entre eles for menor que o tempo de expiração definido.

A escolha do tempo de expiração da sessão deve ser cuidadosa. Um tempo elevado faz com que atividades desassociadas sejam tratadas como sendo consecutivas, relacionando, assim, vídeos que não têm conexão. Por outro lado, um tempo de expiração baixo faz com que vídeos assistidos em sequência sejam considerados separadamente, omitindo, assim, várias possíveis relações.

O tempo de expiração da sessão determina a quantidade de vídeos englobados por cada sessão e, portanto, o número total de sessões para um intervalo de tempo. O tempo de expiração da sessão é definido, empiricamente, pela análise do número de sessões para diferentes valores de tempo de expiração da sessão [Benevenuto et al., 2010].

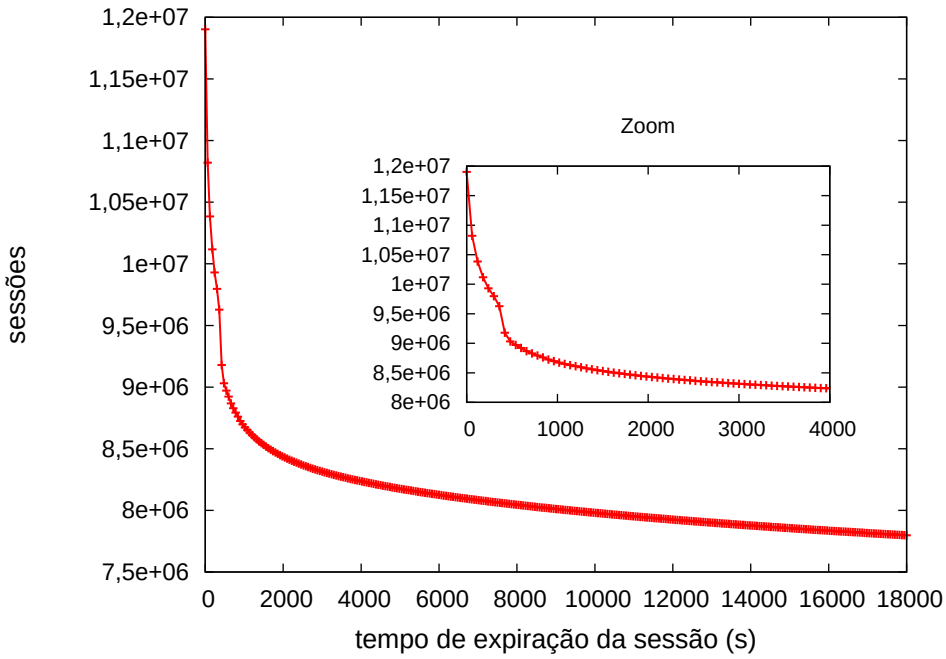

Figura 5.2. Número de sessões por tempo de expiração da sessão.

A Figura 5.2 apresenta o número de sessões para diferentes valores de tempo de expiração da sessão. O tempo de expiração foi variado, aumentando-se de um em um

minuto, a partir de 0, em que uma sessão é sempre formada por um único acesso, até 18000 segundos (cinco horas). O aumento no tempo de expiração faz com que mais vídeos sejam incluídos em cada sessão e, consequentemente, menor seja o número total de sessões.

Analisando a Figura 5.2, verificamos que a curva da distribuição do número de sessões por tempo de expiração decai exponencialmente. Ou seja, para valores baixos do tempo de expiração da sessão, um pequeno aumento desse valor faz com que o número de sessões reduza drasticamente, enquanto que, para valores mais altos, um aumento tem pouco efeito sobre o número total. A curva de convergência do número de sessões acontece por volta de 1800 segundos (30 minutos), quando o número total estabiliza próximo de 8,5 milhões. Portanto, 1800 segundos (30 minutos) parece uma boa escolha para o tempo de expiração da sessão.

Para nos certificarmos da coerência da escolha do tempo de expiração, analisaremos o número de sessões por usuário.

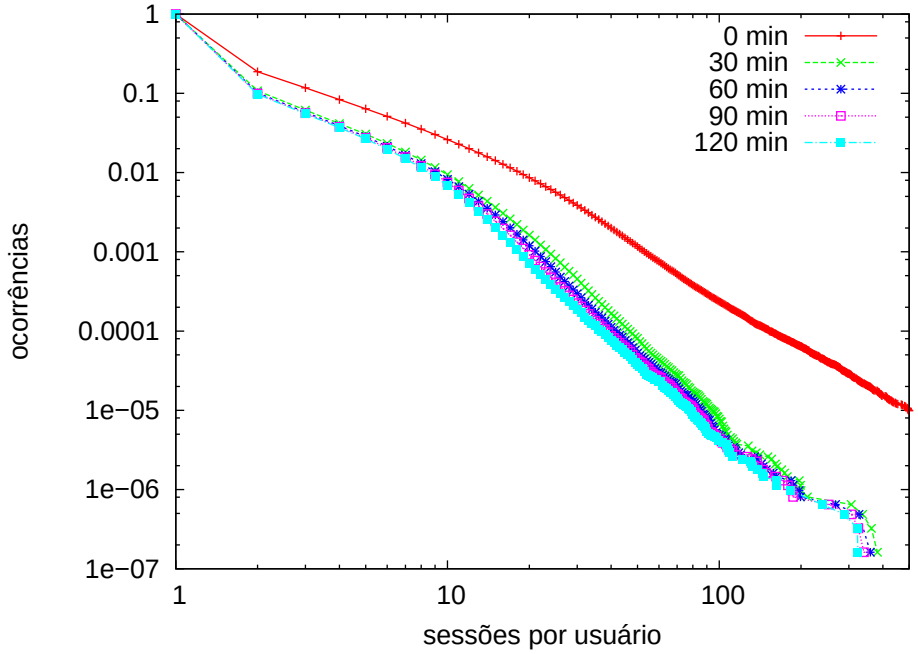

Figura 5.3. Distribuição acumulada complementar do número de sessões por usuário para diferentes valores de tempo de expiração da sessão (*CCDF*).

A Figura 5.3 exibe a distribuição *CCDF* da relação do número de sessões por usuário para diferentes valores do tempo de expiração. O tempo de expiração foi variado, aumentando-se de 30 em 30 minutos, a partir de 0, em que uma sessão é sempre formada por um único acesso, até 120 minutos. Em geral, a distribuição do número de sessões por usuário segue o comportamento de uma *long tail* (que pode

ser visualizada em uma plotagem com os dois eixos em escala logarítmica). Ou seja, poucos usuários possuem um número grande de sessões, enquanto a grande maioria dos usuários detém algumas poucas sessões. No caso, apenas cerca de 1% dos usuários têm mais que 10 sessões. Porém, um número pequeno de usuários detém mais de 100 sessões.

Pela análise do gráfico da Figura 5.3, nota-se que, as curvas das distribuições são, praticamente, coincidentes, para todos os valores do tempo de expiração da sessão a partir de 30 minutos. Essa constatação reforça a escolha de 30 minutos como um tempo de expiração apropriado.

Outros trabalhos existentes na literatura apresentam uma análise da sessão de usuários similar à desenvolvida em nosso estudo de caso. Verificamos que os tempos de expiração da sessão adotados nesses trabalhos estão de acordo com o que foi obtido em nossas análises. Benevenuto et al. [2010] propuseram uma caracterização de sessões de usuários do portal *UOL*, maior provedor de mídia especializada da América Latina, enquanto Gill et al. [2008] realizaram uma análise similar sobre o YouTube. Os valores do tempo de expiração da sessão adotados nos dois estudos foram 30 e 40 minutos, respectivamente.

A partir das análises apresentadas, estipulamos, em nosso estudo de caso, o tempo de expiração da sessão como sendo de 1800 segundos (30 minutos).

5.3 Análise da Rede

Tendo determinado o tempo de expiração da sessão, podemos, então, aplicar a modelagem ilustrada na Figura 5.1. Primeiramente, os vídeos assistidos por usuários foram divididos em sessões, considerando o tempo de expiração da sessão, estipulado em 30 minutos. Assim, para o conjunto de dados utilizado (uma semana), tem-se um total de 8.469.239 sessões. A média de sessões por usuário é 1,371 (com desvio padrão de 1,855 e variância de 3,442). Esse valor é baixo por existirem, na coleção, muitos usuários que assistiram poucos vídeos (Figura 3.7). Porém, um valor baixo do número de sessões por usuário não compromete nossas análises, já que nosso interesse é obter padrões recorrentes em sessões, independente do usuário. Além disso, uma quantidade menor de sessões por usuário (ou uma distribuição mais heterogênea dos usuários das sessões) reduz a interferência das preferências pessoais nos padrões obtidos.

O conjunto de sessões foi modelado como uma rede complexa, onde vídeos de uma mesma sessão, assistidos em sequência, estão relacionados.

5.3.1 Métricas da Rede

Como detalhado na Seção 5.1, o modelo proposto pode ser representado como um grafo direcionado ponderado. A Tabela 5.2 contém os valores para algumas medidas tradicionais de grafos, extraídos do grafo gerado a partir das relações dos vídeos em sessões.

Vértices	42.260
Arestas	301.110
Média de graus (de saída)	7,58
Média de graus ponderando (de saída)	57,687
Diâmetro	78
Caminho médio	5,65
Modularidade	0,831
Componentes conectados	881
Coeficiente de clusterização médio	0,089

Tabela 5.2. Medidas da rede de vídeos relacionados.

Dentre as medidas obtidas, é importante salientar que o número de vértices, mais de 42 mil e de arestas, mais de 300 mil, denotam que esse é um grafo grande. O diâmetro do grafo (maior distância entre qualquer par de vértices ou maior caminho mínimo), 78, também indica que estamos lidando com um grafo extenso. Além disso, os valores elevados de modularidade e de componentes conectados, sugerem que a rede está dividida em comunidades bastante distintas. Ou seja, existem vértices muito conectados entre si, formando comunidades, e pouco conectados com vértices externos à comunidade.

A Figura 5.4 apresenta a distribuição *CCDF* de graus da rede. Como o grafo é direcionado, foram consideradas somente as arestas de saída de cada vértice. A distribuição segue o comportamento de uma *long tail*, com mais de 70% dos vídeos da rede com grau inferior a 10. Pelos dados da Tabela 5.2, a média dos graus de saída é 7,58 e a média ponderada (pelos pesos das arestas de saída) é 57,687. Ou seja, cada vídeo antecede, em média, cerca de oito vídeos nas sessões do intervalo avaliado.

5.3.2 Visualização da Rede

As medidas de rede fornecem uma visão geral das relações entre vídeos estabelecidas pelas ocorrências em diversas sessões. Sabemos, por exemplo, que a rede contém muitos componentes conectados e uma divisão em comunidades bastante distintas. Porém, sabemos pouco sobre a organização da rede, sobre o que define a divisão em comunidades, por exemplo.

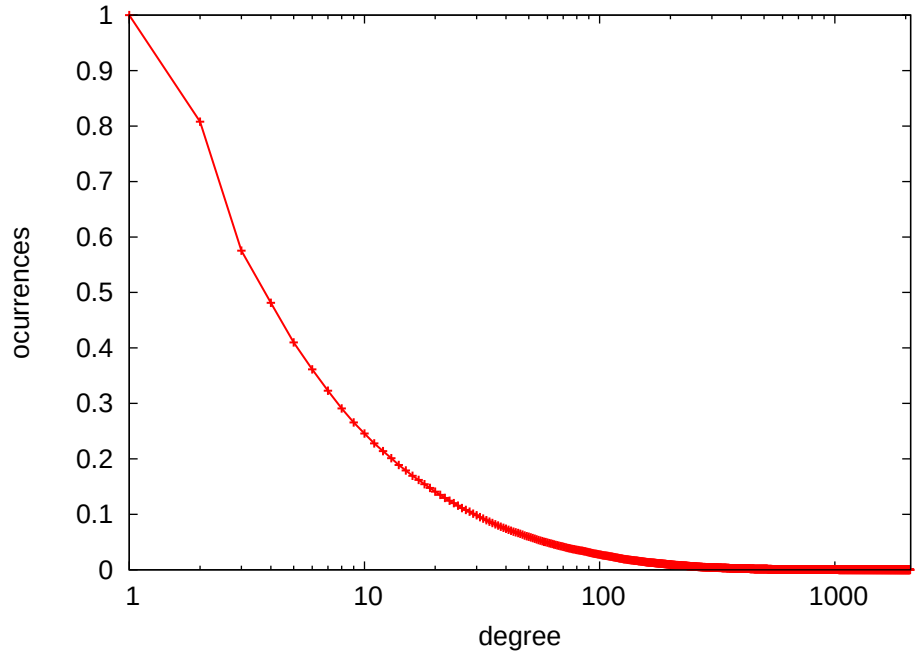

Figura 5.4. Distribuição acumulada complementar dos graus de saída da rede (*CCDF*).

Para entender melhor sobre como a rede está organizada e como os vídeos assistidos em sequência estão relacionados, geramos visualizações do grafo que representa a rede. Nas visualizações, os vértices (vídeos) são representados como pontos e as arestas (relações entre vídeos) como linhas ligando pontos. Foram geradas três visualizações, variando apenas o critério para coloração dos vértices. Primeiramente, gerou-se uma visualização com vértices coloridos segundo a categoria do vídeo. Em seguida, foi gerada uma visualização em que foi atribuída aos vértices uma cor para cada *site* provedor.

A Figura 5.5 exibe a visualização da rede de vídeos relacionados em sessões, considerando a categoria dos vídeos como critério para coloração dos vértices. Ou seja, foi associada uma cor para cada uma das 10 categorias (*Unknown* inclusive) e cada vértice foi colorido segundo a cor da categoria do vídeo correspondente.

Na visualização da Figura 5.5, é evidente a separação de vértices da mesma categoria em componentes, com muitas conexões entre si. Em alguns pontos, as cores se confundem, mas existe uma separação clara dos vértices pela categoria. Apesar dos vídeos não categorizados serem a maioria (a categoria *Unknown* representa 48,15% de toda a amostra), o maior grupo de vértices é o que representa vídeos da categoria *Entertainment*. Os vídeos não categorizados podem ser sobre assuntos diversos, relacionados com categorias diferentes, o que justifica a maior dispersão desses vértices.

As arestas ocorrem somente para vértices representando vídeos de uma mesma

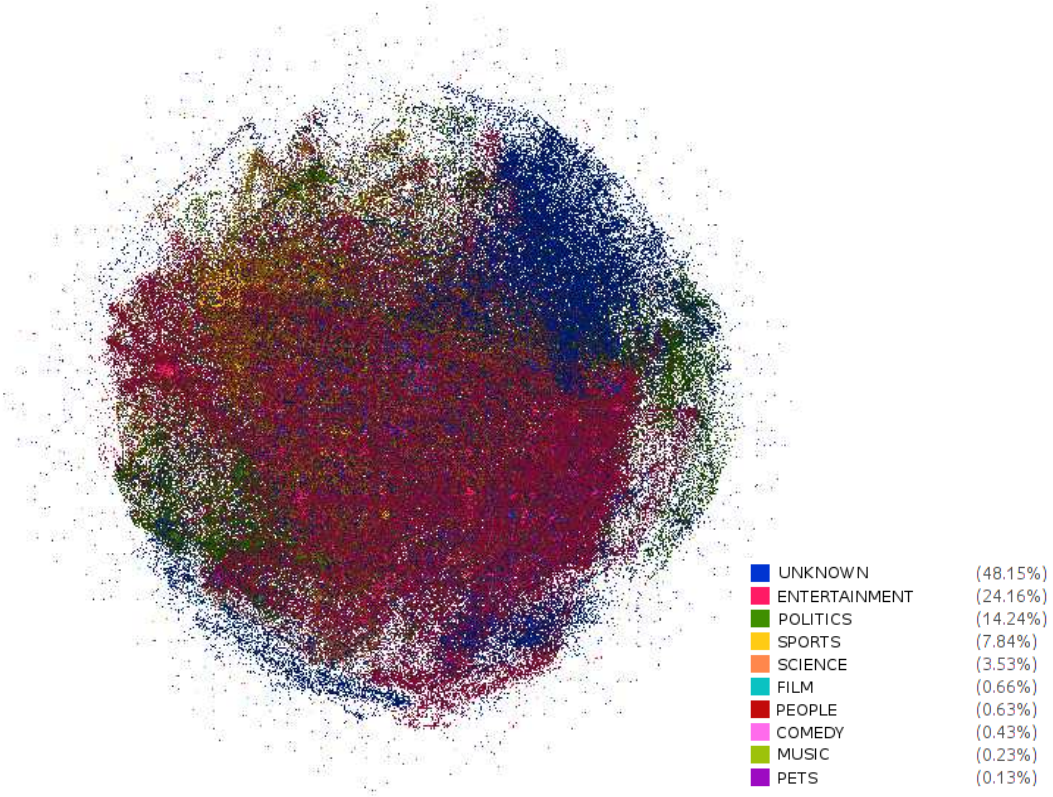

Figura 5.5. Representação da rede com categorias em destaque.

sessão, assistidos em sequência. Portanto, o padrão observado na visualização sugere que a categoria é um fator forte de relação entre vídeos. Ou seja, vídeos assistidos em sequência são, frequentemente, da mesma categoria.

A Figura 5.6 apresenta a visualização da rede de vídeos relacionados em sessões, considerando o *site* provedor dos vídeos como critério para coloração dos vértices. Ou seja, foi associada uma cor para cada um dos 38 provedores de conteúdo e cada vértice foi colorido segundo a cor do *site* de origem do vídeo correspondente.

Observamos uma distinção dos grupos de vértices de cada *site*. As cores dos provedores mais representativos são mais evidentes, mas, em geral, os grupos de vértices representando vídeos de cada *site* estão muito bem definidos. Ou seja, vídeos assistidos em sequência são, frequentemente, do mesmo *site*.

Comparando com a visualização da Figura 5.5, percebemos que a coloração por *site* parece gerar grupos mais distintos e coesos que a coloração por categoria. Ou seja, o *site* de origem é, provavelmente, um fator mais influente que a categoria para relacionar vídeos.

Outra observação relevante é que o grupo de vértices que representam os vídeos do *site* de cor roxa (na parte superior direita da visualização) é muito similar a um

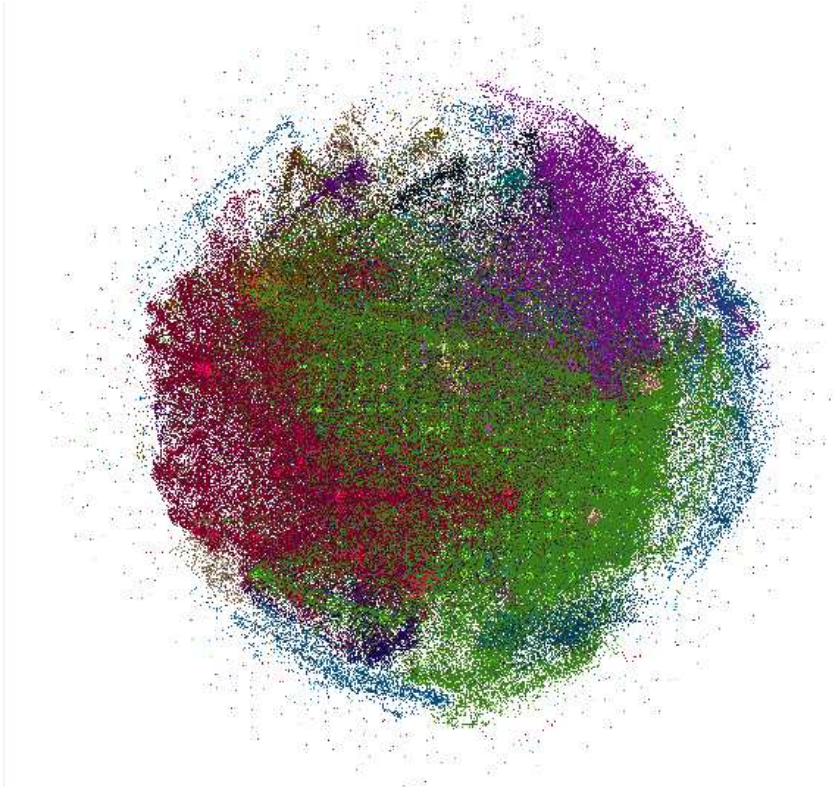

Figura 5.6. Representação da rede com *sites* provedores em destaque.

grupo bastante conectado de vídeos não categorizados (cor azul) da Figura 5.5. Na Figura 5.5, apesar dos vídeos sem categoria serem maioria, esse grupo é o único componente conectado formado por vídeos não categorizados (os demais vértices da categoria *Unknown* estão dispersos na visualização).

A Figura 5.7 apresenta a distribuição de vídeos por categoria para os quatro provedores mais representativos, na semana considerada. Observando o quarto provedor, verificamos que não houve atribuição de categoria para nenhum dos vídeos desse *site*. Esse provedor foi representado pela cor roxa na visualização da Figura 5.6. Portanto, provavelmente, o motivo da formação do grupo de vértices azul, na parte superior direita da visualização da Figura 5.5 (ou em roxo na Figura 5.6), é a origem comum dos vídeos.

Os padrões observados nas duas visualizações sugerem que os atributos considerados (categoria e *site* provedor) têm uma influência forte nos relacionamentos entre vídeos.

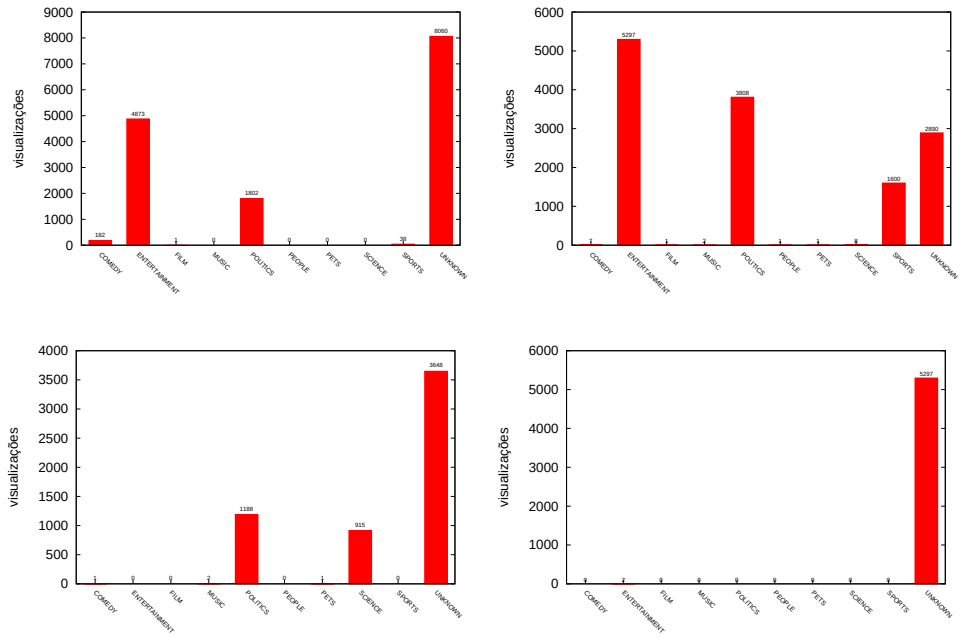

Figura 5.7. Distribuição dos vídeos por categoria, entre os dias 15 e 21 de Julho, para os quatro provedores mais representativos.

5.4 Associações entre Vídeos

As análises anteriores mostraram que a categoria e o *site* provedor são atributos que influenciam as associações entre vídeos. Pretendemos, então, entender melhor essa influência.

Para investigar as relações entre vídeos, vamos recorrer ao conceito de "Regras de Associação". Regras de associação são representações de padrões frequentes de relacionamentos entre itens de um determinado conjunto de dados [Agrawal et al., 1993]. Uma de suas típicas aplicações é a análise de transações, que são conjuntos de itens transacionados conjuntamente por um usuário.

O problema de mineração de regras de associação pode ser formalizado segundo a definição dada por Agrawal et al. [1993]. Seja $I = \{i_1, i_2, \ldots, i_n\}$ um conjunto de n itens. Seja $D = \{t_1, t_2, \ldots, t_m\}$ um conjunto de transações, em que, cada transação em D possui um único identificador e contém um subconjunto de itens em I. Uma regra é definida como sendo uma implicação da forma $X \Rightarrow Y$, onde $X, Y \subseteq I$ e $X \cap Y = \emptyset$. Os conjuntos de itens (*itemsets*) X e Y são chamados *antecedentes* e *consequentes* da regra, respectivamente.

O processo de mineração de regras de associação pode envolver diversas métricas para seleção de regras relevantes. As mais populares são "suporte", "confiança" e "*lift*":

- O suporte $sup(X)$ de um *itemset* X é definido como a proporção das transações consideradas que contêm X.
- A confiança de uma regra é definida como $conf(X \Rightarrow Y) = \frac{\sup(X \cup Y)}{\sup(X)}$, que pode ser interpretada como a proporção de ocorrências da regra nas transações em que seu antecedente ocorre.
- O *lift* de uma regra é definido como a razão do suporte observado, considerando X e Y independentes: $lift(X \Rightarrow Y) = \frac{\sup(X \cup Y)}{\sup(X)\sup(Y)} = \frac{\text{conf}(X \Rightarrow Y)}{\sup(Y)}$. Um valor da métrica *lift* maior do que 1 indica que o consequente é mais frequente quando o antecedente ocorre. Enquanto um valor menor que um informa que o consequente é mais frequente nas transações em que o antecedente está ausente. Para um valor de *lift* igual a 1, a frequência do consequente é igual independente do antecedente. Portanto, quanto maior o valor do *lift*, maior a relação entre os dois lados da regra.

Observando nosso modelo de dados, é fácil perceber a semelhança de sessões com transações de regras de associação. Ou seja, cada sessão pode ser visualizada como uma transação, em que usuários consomem vídeos associadamente. Assim, vídeos podem ser relacionados por serem, frequentemente, consumidos em conjunto. Portanto, modelamos os vídeos como sendo itens, cada sessão como sendo uma transação e desejamos encontrar regras que possam ser traduzidas em padrões de acesso frequentes.

Regras envolvendo vídeos específicos (do tipo $V_a \Rightarrow V_b$, sendo V_a e V_b vídeos da coleção), não são relevantes para nosso estudo, já que não podem ser generalizadas para outros contextos. Vamos obter, portanto, regras envolvendo atributos dos vídeos, mais especificamente, categoria e *site* provedor.

Em nossas análises, empregamos um dos algoritmos mais tradicionais para extração de regras de associação, o *Apriori* [Agrawal & Srikant, 1994].

5.4.1 Associações por Categoria

Primeiramente, investigamos as relações entre categorias de vídeos. Cada sessão foi modelada como uma transação contendo as categorias dos vídeos pertencentes à sessão. A partir do conjunto de transações, foram extraídas regras de associação relevantes.

As Tabelas 5.3 e 5.4 apresentam as 10 principais regras de associação obtidas considerando a categoria dos vídeos, ordenadas pela confiança e pelo *lift*, respectivamente.

A maioria das regras que ocorrem com maior confiança são bastante intuitivas, como, por exemplo, COMEDY → ENTERTAINMENT e MUSIC → ENTERTAINMENT. Porém, algumas das 10 regras mais relevantes surpreendem, tais como POLITICS → ENTERTAINMENT, SCIENCE → POLITICS e SPORTS → POLITICS. A

	Regras	**Confiança**
1	COMEDY → ENTERTAINMENT	0,73
2	MUSIC → ENTERTAINMENT	0,71
3	PEOPLE → POLITICS	0,59
4	SPORTS → ENTERTAINMENT	0,53
5	POLITICS → ENTERTAINMENT	0,47
6	SCIENCE → ENTERTAINMENT	0,38
7	SCIENCE → POLITICS	0,35
8	ENTERTAINMENT → POLITICS	0,34
9	ENTERTAINMENT → SPORTS	0,32
10	SPORTS → POLITICS	0,31

Tabela 5.3. Regras de associação relevantes, pela métrica *confiança*, considerando as categorias dos vídeos.

	Regras	***Lift***	**Confiança**
1	FILM → SCIENCE	3,22	0,59
2	SCIENCE → FILM	3,22	0,13
3	POLITICS → PEOPLE	1,27	0,2
4	PEOPLE → POLITICS	1,27	0,59
5	COMEDY → ENTERTAINMENT	1,14	0,73
6	ENTERTAINMENT → COMEDY	1,14	0,11
7	ENTERTAINMENT → MUSIC	1,11	0,1
8	MUSIC → ENTERTAINMENT	1,11	0,71
9	FILM → MUSIC	0,99	0,08
10	MUSIC → FILM	0,99	0,04

Tabela 5.4. Regras de associação relevantes, pela métrica *lift*, considerando as categorias dos vídeos.

categoria *Entertainment* está presente em quase todas as regras frequentes, isso por ser a maior categoria (desconsiderando o conjunto de vídeos não categorizados) e por ter uma abrangência maior (vídeos muito diversificados podem ser classificados em *Entertainment*).

Analisando as regras mais relevantes pela métrica *lift* (Tabela 5.4), verificamos que as oito primeiras regras possuem *lift* maior que 1. Isso significa que a relação do consequente com o antecedente é forte, já que, na maior parte das ocorrências do consequente, o antecedente da regra está presente. Por exemplo, na maioria das sessões em que um usuário acessa um vídeo da categoria *Science*, ele também acessou, anteriormente, um vídeo da categoria *Film*. O *lift* 3,22 para a regra FILM → SCIENCE pode ser interpretado como: usuários que assistem um vídeo da categoria *Film* têm uma probabilidade 3,22 vezes maior de também assistirem um vídeo da categoria *Science*.

Algumas regras estão presentes nas duas ordenações, como, por exemplo, CO-

MEDY → ENTERTAINMENT e MUSIC → ENTERTAINMENT e são, portanto, regras bastante fortes e relevantes.

5.4.2 Associações por *Site*

O mesmo procedimento anterior para extração de regras de associação foi aplicado, mas, dessa vez, considerando o *site* que publicou o vídeo como atributo na modelagem das transações. Assim, objetivamos encontrar *sites* relacionados, que são acessados, frequentemente, em sequência.

	Regras	**Confiança**
1	SITE6 → SITE2	0,59
2	SITE1 → SITE2	0,48
3	SITE5 → SITE2	0,42
4	SITE5 → SITE1	0,39
5	SITE2 → SITE1	0,37
6	SITE1 → SITE5	0,34
7	SITE8 → SITE5	0,34
8	SITE2 → SITE6	0,33
9	SITE8 → SITE2	0,31
10	SITE2 → SITE5	0,28

Tabela 5.5. Regras de associação relevantes, pela métrica *confiança*, considerando os *sites* dos vídeos.

	Regras	***Lift***	**Confiança**
1	SITE2 → SITE6	0,91	0,33
2	SITE6 → SITE2	0,91	0,59
3	SITE1 → SITE5	0,79	0,34
4	SITE5 → SITE1	0,79	0,39
5	SITE5 → SITE8	0,78	0,3
6	SITE8 → SITE5	0,78	0,34
7	SITE1 → SITE2	0,74	0,48
8	SITE2 → SITE1	0,74	0,37
9	SITE2 → SITE5	0,65	0,28
10	SITE5 → SITE2	0,65	0,42

Tabela 5.6. Regras de associação relevantes, pela métrica *lift*, considerando os *sites* dos vídeos.

As Tabelas 5.5 e 5.6 apresentam as 10 principais regras de associação obtidas considerando o *site* provedor dos vídeos, ordenadas pela confiança e pelo *lift*, respectivamente. Por questões de privacidade de conteúdo, os *sites* foram identificados por

números. Apesar disso, a interpretação das regras nos permite entender como os *sites* provedores de mídia especializada estão associados dentro do cenário geral do mercado brasileiro de vídeos *online*.

Observamos, primeiramente, que os valores de confiança e *lift* para as regras envolvendo relações entre *sites* são menores se comparados com os valores obtidos para regras relacionando categorias (Seção 5.4.1). Ou seja, a relação entre categorias é, em geral, mais forte do que a relação entre *sites*.

Analisando as Tabelas 5.5 e 5.6, verificamos que a regra SITE2 → SITE6 é a mais relevante segundo os dois critérios considerados (confiança e *lift*). Além dessa, existem outras regras que aparecem nas duas listagens, tais como SITE1 → SITE2 e SITE5 → SITE1. Nenhuma regra obtida teve um valor de *lift* superior a 1. Ou seja, para nenhuma regra, o antecedente ocorre na maioria das vezes que o consequente ocorre. Essa é mais uma evidência de que a relação entre *sites* é, em geral, fraca.

5.4.3 Associações por Categoria e *Site*

Por último, os dois atributos analisados foram combinados com intuito de encontrar relações mais específicas. Assim, o par "*site*-categoria" foi usado na modelagem das transações para obtenção de regras relevantes.

	Regras	**Confiança**
1	SITE6-COMEDY → SITE6-SPORTS	0,98
2	SITE1-COMEDY → SITE1-ENTERTAINMENT	0,93
3	SITE2-POLITICS → SITE2-ENTERTAINMENT	0,74
4	SITE1-POLITICS → SITE1-ENTERTAINMENT	0,69
5	SITE5-PEOPLE → SITE5-ENTERTAINMENT	0,61
6	SITE5-FILM → SITE5-SCIENCE	0,6
7	SITE2-SPORTS → SITE2-ENTERTAINMENT	0,52
8	SITE5-SCIENCE → SITE5-FILM	0,47
9	SITE5-MUSIC → SITE5-ENTERTAINMENT	0,37
10	SITE6-SPORTS → SITE2-ENTERTAINMENT	0,37

Tabela 5.7. Regras de associação relevantes, pela métrica *confiança*, considerando as categorias e os *sites* dos vídeos.

Tabelas 5.7 e 5.8 apresentam as 10 principais regras de associação obtidas considerando o *site* provedor e a categoria dos vídeos conjuntamente, ordenadas pela confiança e pelo *lift*, respectivamente.

Com a combinação dos atributos, conseguimos regras com valores bastante elevados de confiança e *lift*. A regra SITE6-COMEDY → SITE6-SPORTS, por exemplo, ocorre com confiança 0,98 e *lift* 4,48. Porém, com exceção da regra SITE6-SPORTS →

	Regras	*Lift*	Confiança
1	SITE5-FILM → SITE5-SCIENCE	8,85	0,6
2	SITE5-SCIENCE → SITE5-FILM	8,85	0,47
3	SITE6-COMEDY → SITE6-SPORTS	4,48	0,98
4	SITE6-SPORTS → SITE6-COMEDY	4,48	0,14
5	SITE5-ENTERTAINMENT → SITE5-PEOPLE	3,00	0,16
6	SITE5-PEOPLE → SITE5-ENTERTAINMENT	3,00	0,61
7	SITE1-COMEDY → SITE1-ENTERTAINMENT	2,75	0,93
8	SITE1-ENTERTAINMENT → SITE1-COMEDY	2,75	0,26
9	SITE1-ENTERTAINMENT → SITE1-POLITICS	2,04	0,14
10	SITE1-POLITICS → SITE1-ENTERTAINMENT	2,04	0,69

Tabela 5.8. Regras de associação relevantes, pela métrica *lift*, considerando as categorias e os *sites* dos vídeos.

SITE2-ENTERTAINMENT, todas as regras obtidas são de associações entre categorias diferentes em um mesmo *site*. Ou seja, poucas regras relevantes envolvem interações entre *sites* diferentes ou mesmo entre categorias diferentes em *sites* diferentes. Isso é mais uma evidência de que as relações entre categorias são mais fortes do que as relações entre *sites*. Esse é um padrão esperado, já que um usuário tende a navegar em vídeos de um mesmo *site* mais do que acessar, em sequência, vídeos de diferentes portais.

5.5 Transições

Na Seção 5.3.2 analisamos diferentes visualizações da rede, enquanto na Seção 5.4, investigamos as associações entre vídeos. Agora, vamos visualizar o fluxo de acessos para termos uma visão geral de como os usuários transitam entre categorias e entre *sites*.

Para cada sessão, registramos as transições entre categorias e *sites* (incluindo transições entre vídeos de mesma categoria ou mesmo *site*). Ou seja, se um usuário assistiu um vídeo da categoria C_i e do *site* S_i e, em seguida, acessou, em uma mesma sessão, outro vídeo da categoria C_j e do *site* S_j, ocorreu uma transição de C_i para C_j e de S_i para S_j. Com isso, é possível representar as transições por meio de redes complexas.

A Tabela 5.9 contém as ocorrências de transições entre cada par de categorias. As linhas são as origens das transições e as colunas são os destinos. Assim, por exemplo, o número de transições partindo da categoria *Music* para a categoria *Entertainment* é 1605, enquanto que, a quantidade de transições partindo de *Entertainment* para a

	COMEDY	ENTERT.	FILM	MUSIC	POLITICS	PEOPLE	PETS	SCIENCE	SPORTS	UNKNOWN
COMEDY	4208	5850	4	33	175	4	0	20	1525	633
ENTERT.	7063	1023220	1037	6665	17883	2528	1	6652	14736	35597
FILM	7	594	24724	365	410	23	0	2362	123	1414
MUSIC	1	1605	78	172	154	9	0	39	105	59184
POLITICS	183	16701	148	597	162111	2963	17	5033	9177	14080
PEOPLE	1	3138	107	360	10586	3455	6	991	2460	3063
PETS	0	4	0	0	27	5	49	9	3	10
SCIENCE	15	2598	2592	916	3852	584	2	24341	2027	3718
SPORTS	2401	14775	129	935	7775	400	9	1464	297094	84789
UNKNOWN	1010	25840	1988	2264	15131	1690	12	4306	49629	431073

Tabela 5.9. Transições entre categorias, considerando os acessos entre os dias 15 e 21 de Julho de 2012. As linhas são as origens das transições, enquanto as colunas são os destinos.

categoria *Music* é 6665.

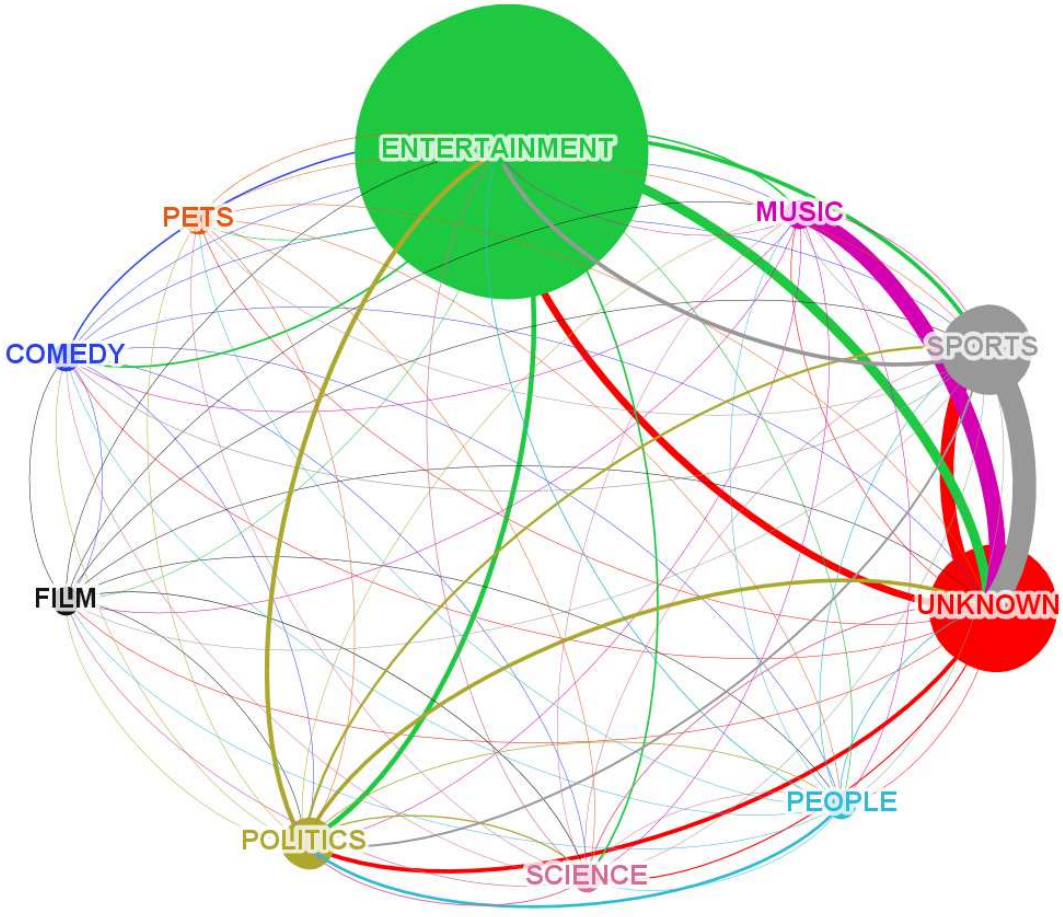

Figura 5.8. Rede de fluxo entre categorias, considerando os acessos entre os dias 15 e 21 de Julho de 2012.

A Figura 5.8 apresenta uma visualização da rede de fluxos entre categorias segundo os dados da Tabela 5.9. Nessa representação, atribuiu-se uma cor para cada

categoria. Os *links*, representando fluxos entre categorias, foram coloridos com a cor da categoria de origem. A espessura dos *links* indica a quantidade de transições entre as categorias conectadas (quanto mais espesso o *link*, maior é o fluxo). A dimensão dos círculos em volta dos nomes das categorias indicam a quantidade de transições entre vídeos da mesma categoria.

Verificamos a ocorrência de um fluxo grande entre vídeos de *Entertainment*, a maior categoria. Esse fluxo ocorre internamente (entre dois vídeos de *Entertainment*) e externamente, principalmente, com as categorias *Sports* e *Politics*.

Apesar de existirem fluxos entre todos os possíveis pares de categorias, algumas relações específicas são mais fortes, como, por exemplo, entre as categorias *Entertainment* e *Politics*. Também é interessante notar que algumas relações entre categorias são fortes somente em um sentido, como o que ocorre com o par *People* e *Politics*, no qual muitos usuários migram da categoria *People* para *Politics*, mas poucos usuários fazem o fluxo contrário.

A Figura 5.9 exibe a rede de fluxos entre *sites* provedores de conteúdo. Essa visualização é similar à representação de fluxos entre categorias, mas considera as transições entre *sites*. Apesar dos *sites* não serem identificados, a Figura 5.9 nos permite visualizar como ocorrem as interações de usuários no cenário geral, transitando entre diferentes *sites*.

Os *sites* foram numerados de acordo com sua representatividade (o *site* 1 é o mais representativo, sendo o mais acessado dentre os 38). É interessante notar que o terceiro portal em número de acessos possui um círculo em volta de si, na visualização, menor que os de outros *sites* com menos visualizações. Nesse caso, o número de transições entre vídeos do mesmo *site* não é tão grande, o que deve ser compensado pela quantidade de transições para outros *sites*.

Comparando as Figuras 5.8 e 5.9, verificamos que as interações entre categorias são, em geral, mais fortes do que as interações entre *sites*. Para alguns pares de *sites*, praticamente, não há ocorrências de transições. Porém, notamos uma relação forte entre dois *sites* (representados na figura pelos números 6 e 15). Nesse caso, o fluxo entre esses *sites* é justificado pelo fato de um deles, um portal de notícias, ter *links* diretos para o outro, um *site* de previsões metereológicas. O fluxo no sentido contrário provavelmente ocorre por que o usuário continua navegando no portal de notícias após acessar (ou mesmo enquanto acessa) as informações no *site* de previsões metereológicas.

As diferenças nas dimensões dos círculos indicam diferentes fluxos entre vídeos do mesmo *site*. Não houve normalização dos fluxos para os diversos *sites*, portanto, provedores com maior número de acessos tendem a ter um círculo maior, indicando uma quantidade grande de transições entre vídeos do próprio *site*. Assim, a Figura 5.9

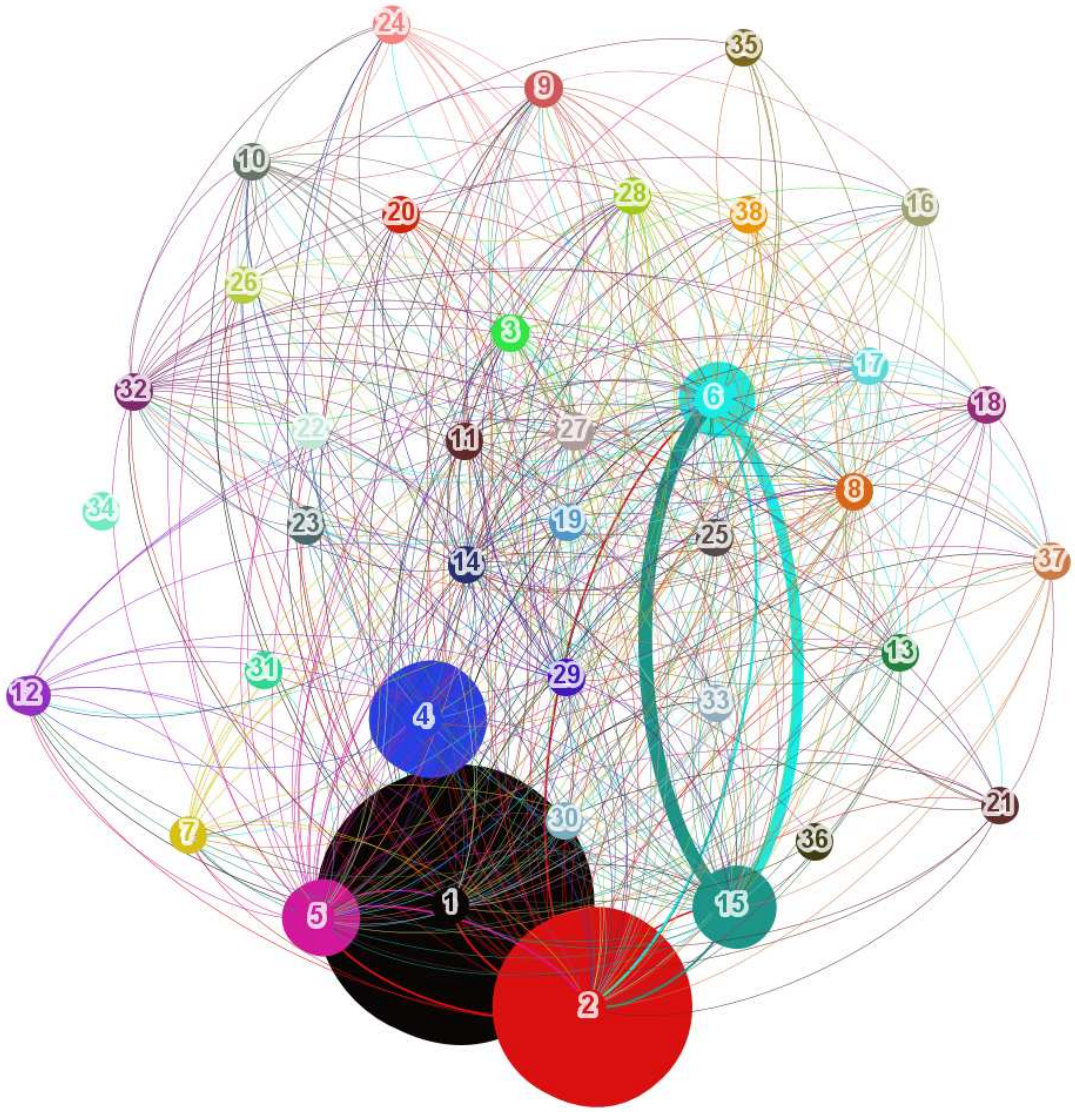

Figura 5.9. Rede de fluxo entre *sites*, considerando os acessos entre os dias 15 e 21 de Julho de 2012.

também nos fornece uma visão geral da representatividade de alguns dos principais provedores de mídia especializada do mercado brasileiro.

5.6 Conclusões da Análise de Transacional

As análises apresentadas neste capítulo têm como referência a terceira questão de pesquisa levantada na Seção 2.2.3. Foram identificados, portanto, padrões interessantes na investigação do fluxo de navegação de usuários assistindo vídeos.

Para identificar os padrões de transição entre vídeos, foi necessário modelar os dados empregando o conceito de sessão. Uma sessão constitui um conjunto de acessos realizados em um período de tempo por um usuário. O tempo de expiração da sessão foi determinado, empiricamente, como sendo de 30 minutos. Esse tempo é semelhante aos valores usados em outros estudos, inclusive em trabalhos sobre *sites* de conteúdo gerado por usuário [Gill et al., 2008; Benevenuto et al., 2010].

A partir da divisão dos acessos em sessões, foi possível aplicar uma modelagem de rede complexa, associando vídeos assistidos em sequência em sessões de usuários. A representação da rede como um grafo nos permitiu extrair métricas gerais e ter uma visão de como os vídeos estão relacionados. Constatamos, por exemplo, que existem grupos de vídeos muito conectados entre si, formando comunidades na rede. Pela visualização da rede, foi possível identificar a ocorrência de agrupamentos pela categoria e pelo *site* provedor dos vídeos.

Em nossas análises, percebemos que uma sessão também pode ser interpretada como uma transação realizada por um usuário. Assim, aplicamos o conceito de regras de associação para extrair padrões interessantes de associação entre vídeos. Com isso, foi possível identificar comunidades e *sites* que apresentam uma relação forte. Notamos, nas análises transacionais, que as relações entre categorias são, em geral, mais fortes que as relações entre *sites*.

Por fim, analisamos o fluxo entre pares de categorias e *sites*. Considerando todas as transições entre vídeos dentro de cada sessão, representamos como uma rede complexa os fluxos entre as categorias e entre os diversos *sites* provedores. A visualização das redes nos permitiu enxergar o fluxo geral de acessos aos vídeos sob uma perspectiva ampla do mercado brasileiro de mídia especializada.

As análises desenvolvidas neste capítulo têm aplicações práticas importantes. Para os provedores de conteúdo, por exemplo, é interessante entender como as comunidades estão relacionadas e quais são os comportamentos típicos de usuários ao assistirem diversos vídeos. Com esse conhecimento, é possível prover personalização de serviços e recomendação de conteúdo. A análise do fluxo de usuários entre *sites* também pode ser útil para que os provedores de conteúdo tentem atrair o interesse dos usuários para seus vídeos, evitando, assim, a migração desses para outras fontes de conteúdo. Outra possível aplicação dos padrões identificados é no planejamento de estratégias para a entrega de conteúdos. Os responsáveis por serviços de *CDN* (*Content Delivery Network*) podem se beneficiar das informações de padrões de transições e associações entre *sites* para aprimorarem a logística e evitarem imprevistos.

Capítulo 6

Conclusões e Trabalhos Futuros

6.1 Conclusões

Temos acompanhado o aumento da produção e do consumo de vídeos *online* em uma *Web* cada vez mais interativa, social e colaborativa. A expansão de aplicações multimídias na Internet trouxe a necessidade de mecanismos especializados capazes de garantir serviços de distribuição e entrega de conteúdo eficientes. Para tanto, é primordial conhecer o comportamento de usuários ao visualizarem vídeos *online*.

Nesta dissertação, apresentamos uma análise extensiva dos padrões de acesso em *sites* produtores de mídia especializada (ME). Devido à disponibilidade limitada de dados públicos, pouco se sabe sobre o consumo de vídeos em portais de ME. No entanto, com os dados coletados em associação com a *Samba Tech*, empresa líder no segmento de plataformas de vídeos *online*, foi possível ter uma perspectiva privilegiada das interações de usuários ao acessarem vídeos de alguns dos principais portais brasileiros de ME. A coleção, que compreende registros de milhões de usuários acessando milhares de vídeos, foi usada como estudo de caso na investigação de padrões de visualização em diferentes contextos de aplicação. Nossas análises consideraram, primeiramente, os dados de maneira agregada, levando em conta os atributos estáticos. Em seguida, os aspectos temporais foram investigados considerando como ocorre a evolução das visualizações ao longo do tempo e quais são os padrões recorrentes de acesso. Por fim, estudamos os padrões de transições de usuários entre diferentes vídeos e suas correlações.

As análises estáticas revelaram padrões gerais interessantes sobre o consumo de mídia especializada. Identificamos, por exemplo, pela observação das distribuições de vídeos e acessos entre *sites*, que o cenário dos provedores brasileiros de ME é bastante heterogêneo, incluindo alguns provedores com muita representatividade (tanto em termos de acesso quanto em número de vídeos publicados) e muitos *sites* pequenos. Foi

possível também, a partir dos resultados, comparar os padrões de acesso identificados com comportamentos recorrentes reportados no contexto de *sites* de conteúdo gerado por usuário (CGU). Assim, constatamos, por exemplo, que a duração média de vídeos nos portais de ME é, em geral, maior que a dos vídeos de *sites* de CGU. As investigações realizadas envolveram muitos aspectos gerais das mídias, tais como categorias, visualizações, número de publicações por *site* e duração. Análises desse tipo fornecem uma visão geral do mercado brasileiro de ME e de seu consumo, trazendo informações sobre a quantidade de acessos, a distribuição desses acessos e os padrões globais de visualização.

As análises temporais permitiram a identificação de padrões recorrentes de visualização de vídeos. Foi verificado, por exemplo, que padrões cíclicos e bem definidos de acesso ocorrem semanalmente e diariamente. Apesar desses padrões serem comuns entre as categorias, existem comportamentos de acesso próprios de cada categoria. Do lado dos produtores de conteúdo, também foi identificado um padrão cíclico na distribuição das publicações de vídeos. Além disso, analisamos a taxa de retenção do consumo de mídias especializadas, descobrimos que essa é, em geral, muito baixa, e investigamos a evolução das visualizações de vídeos do tipo ME, quando observamos que o seu "tempo de vida" é, geralmente, pequeno. Os resultados obtidos com as análises temporais são relevantes para a realização de estimativas do impacto das publicações e para a compreensão de como variam os acessos ao longo do tempo, o que pode ser aplicado na gestão dos recursos para garantir a qualidade da entrega dos vídeos em cada momento. Os produtores de mídia especializada também podem aproveitar os padrões obtidos para conhecerem o melhor momento para a substituição de vídeos publicados ou para a publicação de novos conteúdos.

Por fim, as análises transacionais foram úteis na identificação de relações implícitas entre vídeos assistidos frequentemente em sequência. Através do conceito de sessão e de uma modelagem dos dados usando redes complexas, foi possível descobrir associações entre vídeos. Concluímos, por exemplo, que os vídeos estão relacionados entre si pela categoria e pelo *site* provedor de conteúdo. O agrupamento dos acessos em sessões também nos permitiu modelar os dados como transações para extrairmos regras de associação. As regras mais relevantes nos indicaram relações entre categorias e entre *sites*. Além disso, foi possível mapear as transições de usuários entre categorias e *sites*. Essas últimas investigações revelam como os vídeos estão associados e quais são os padrões de comportamento de usuários ao assistirem vídeos em sequência, sendo úteis para tarefas como personalização de serviços, recomendação de conteúdo e aprimoramento de mecanismos para entrega de conteúdo.

Em geral, as análises desenvolvidas foram proveitosas por proverem uma visão de-

talhada do mercado brasileiro de mídia especializada, sobre o qual ainda pouco se sabe. O trabalho cumpriu com o objetivo principal de fornecer informações relevantes sobre um cenário carente de pesquisas. O cenário dos vídeos *online*, mais especificamente, de mídia especializada, é relativamente novo e pode se beneficiar muito de estudos sobre o consumo e a distribuição dos vídeos. A caracterização do mercado brasileiro de ME, a identificação dos padrões de acesso e a análise comparativa desses com resultados de pesquisas em sites de conteúdo gerado por usuário são as principais contribuições deste trabalho. Apesar dos procedimentos de caracterização terem sido executados sobre um estudo de caso envolvendo um cenário específico, os padrões de acesso obtidos têm aplicabilidade geral, assim como os procedimentos de análise e caracterização empregados. Esses padrões possuem aplicações práticas diversas, tais como melhoria da experiência dos usuários pela implementação de estratégias como recomendação e personalização de serviços, oferta de publicidade mais direcionada para o usuário final e aprimoramento dos serviços de distribuição e entrega de conteúdo.

6.2 Trabalhos Futuros

Nesta dissertação, foi apresentada uma caracterização detalhada do consumo de vídeos *online* no cenário de *sites* provedores de mídia especializada, considerando aspectos estáticos, temporais e transacionais. Apesar da abrangência de nossas análises, existem muitos outros aspectos promissores a serem investigados no contexto de ME. Nossos estudos podem ser expandidos, por exemplo, para considerar aspectos sociais. Seria interessante entender como usuários se relacionam no consumo dos vídeos e quais são os padrões de similaridade entre eles.

Além de explorar outras possibilidades de análise, pretendemos investir em alguma aplicação dos padrões de acesso identificados. Os resultados de nossas investigações podem ser utilizados, por exemplo, na composição de um sistema de recomendação de conteúdo. Os padrões de acesso identificados podem ser incorporados no recomendador para que as indicações de conteúdo considerem aspectos como categoria, duração do vídeo e o momento da recomendação (por exemplo, deve-se considerar que vídeos de esporte recebem muitos acessos segunda e quinta-feira).

Por fim, considerando que o principal motivo para a escassez de trabalhos no contexto de *sites* de ME é a indisponibilidade de dados, pretendemos contribuir com o fornecimento de coleções para novas pesquisas na área. Não estamos autorizados a publicar os dados de nosso estudo de caso, coletados em associação com a Samba Tech. Porém, esses dados podem ser usados como modelo para geração de coleções sintéticas

(*synthetic workload generation*) [Ganger, 1995]. As coleções sintéticas preservam os padrões dos dados originais e são vantajosas por permitirem a definição de propriedades gerais (como a dimensão da coleção) de acordo com o interesse de pesquisa.

Referências Bibliográficas

Acharya, S.; Smith, B. C. & Parnes, P. (1999). Characterizing user access to videos on the World Wide Web. In *Proceedings of the 7th SPIE Multimedia Computing and Networking Conference*, pp. 130 – 141.

Agrawal, R.; Imieliński, T. & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 22(2):207 – 216. ISSN 0163-5808.

Agrawal, R. & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of 20th International Conference on Very Large Data Bases*, volume 1215, pp. 487 – 499.

Arlitt, M. F. & Williamson, C. L. (1996). Web server workload characterization: The search for invariants. In *ACM SIGMETRICS Performance Evaluation Review*, volume 24, pp. 126 – 137.

Baluja, S.; Seth, R.; Sivakumar, D.; Jing, Y.; Yagnik, J.; Kumar, S.; Ravichandran, D. & Aly, M. (2008). Video suggestion and discovery for youtube: taking random walks through the view graph. In *Proceedings of the 17th International Conference on World Wide Web*, pp. 895 – 904.

Benevenuto, F.; Duarte, F.; Rodrigues, T.; Almeida, V. A.; Almeida, J. M. & Ross, K. W. (2008). Understanding Video Interactions in YouTube. In *Proceedings of the 16th ACM International Conference on Multimedia*, pp. 761 – 764.

Benevenuto, F.; Pereira, A.; Rodrigues, T.; Almeida, V.; Almeida, J. & Gonçalves, M. (2010). Characterization and analysis of user profiles in online video sharing systems. *Journal of Information and Data Management*, 1(2):261.

Benevenuto, F.; Rodrigues, T.; Almeida, V. A. F.; Almeida, J. M. & Gonçalves, M. A. (2009a). Detecting spammers and content promoters in online video social networks. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 620 – 627.

Benevenuto, F.; Rodrigues, T.; Cha, M. & Almeida, V. (2009b). Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*, pp. 49 – 62.

Bingham, A. & Spradlin, D. (2011). *The Long Tail of Expertise*. Pearson Education.

Bondy, J. A. & Murty, U. S. R. (1976). *Graph theory with applications*, volume 290. Macmillan London.

Cha, M.; Kwak, H.; Rodriguez, P.; Ahn, Y.-Y. & Moon, S. (2007). I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pp. 1 – 14.

Cheng, X.; Dale, C. & Liu, J. (2007). Understanding the Characteristics of Internet Short Video Sharing: YouTube as Case Study. *CoRR*, abs/0707.3670.

Cheng, X.; Dale, C. & Liu, J. (2008). Statistics and Social Network of YouTube Videos. In *Proceedings of the 16th International Workshop on Quality of Service*, pp. 229 – 238.

Cheng, X. & Liu, J. (2009). NetTube: Exploring Social Networks for Peer-to-Peer Short Video Sharing. In *Proceedings of the 28th IEEE International Conference on Computer Communications*, pp. 1152 – 1160.

Comarela, G.; Crovella, M.; Almeida, V. & Benevenuto, F. (2012). Understanding factors that affect response rates in twitter. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, pp. 123 – 132.

ComScore (2013). Brazilian Online Video Audience Reaches 43 Million Unique Viewers in December 2012. http://www.comscore.com.

Costa, C. P.; Cunha, I. S.; Borges, A.; Ramos, C. V.; Rocha, M. M.; Almeida, J. M. & Ribeiro-Neto, B. (2004). Analyzing client interactivity in streaming media. In *Proceedings of the 13th International Conference on World Wide Web*, pp. 534 – 543.

Ganger, G. R. (1995). Generating representative synthetic workloads: An unsolved problem. In *in Proceedings of the Computer Measurement Group (CMG) Conference*.

Gill, P.; Arlitt, M.; Li, Z. & Mahanti, A. (2007). YouTube Traffic Characterization: A View from the Edge. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pp. 15 – 28.

Gill, P.; Arlitt, M.; Li, Z. & Mahanti, A. (2008). Characterizing user sessions on YouTube. In *Electronic Imaging 2008*.

Goncalves, C.; Totti, L.; Duarte, D.; Meira, W. & Pereira, A. (2011). ROCK: Uma Metodologia para a Caracterização de Serviços Web Multimídia Baseada em Hierarquia da Informação. *XVII Simpósio Brasileiro de Sistemas Multimídia e Web WebMedia*, pp. 174 – 181.

Gürsun, G.; Crovella, M. & Matta, I. (2011). Describing and Forecasting Video Access Patterns. In *Proceedings of the 30th IEEE International Conference on Computer Communications*, pp. 16 – 20.

Maia, M.; Almeida, J. & Almeida, V. (2008). Identifying user behavior in online social networks. In *Proceedings of the 1st Workshop on Social Network Systems*, pp. 1 – 6.

Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*, volume 999. MIT Press.

Menascé, D. A.; Almeida, V. A.; Fonseca, R. & Mendes, M. A. (1999). A methodology for workload characterization of e-commerce sites. In *Proceedings of the 1st ACM Conference on Electronic Commerce*, pp. 119 – 128.

Miranda, L. C.; Santos, R. L. & Laender, A. H. (2013). Characterizing video access patterns in mainstream media portals. In *Proceedings of the 22nd International Conference on World Wide Web*, pp. 1085 – 1092.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167 – 256.

o'Reilly, T. (2005). *What is web 2.0*. O'Reilly.

Paolillo, J. C. (2008). Structure and network in the YouTube core. In *Proceedings of the 41st Hawaii International Conference on System Sciences*, pp. 156 – 156.

Purcell, K. (2010). The state of online video. Relatório técnico, Pew Internet & American Life Project.

Samba-Tech (2013). Blog. http://www.sambatech.com/blog/.

Saxena, M.; Sharan, U. & Fahmy, S. (2008). Analyzing video services in Web 2.0: a global perspective. In *Proceedings of the 18th International Workshop on Network and Operating Systems Support for Digital Audio and Video*, pp. 39 – 44.

Startupi (2012). Rede de negøcios do Vale do Silício aponta Samba Tech como uma das empresas mais promissoras do mundo em inovação. http://startups.ig.com.br.

Szabo, G. & Huberman, B. A. (2010). Predicting the popularity of online content. *Communications of the ACM*, 53(8):80 – 88.

The Next Web (2011). Brazilian video service Samba Tech expands to Latin America, aiming for IPO in 2016. http://thenextweb.com.

Vallamsetty, U.; Kant, K. & Mohapatra, P. (2003). Characterization of e-commerce traffic. *Electronic Commerce Research*, 3(1-2):167 – 192.

Zink, M.; Suh, K.; Gu, Y. & Kurose, J. (2009). Characteristics of YouTube network traffic at a campus network: measurements, models, and implications. *Computer Networks*, 53(4):501 – 514.

Apêndice A

Distribuição de Vídeos por Site

As Figuras A.1 a A.6 exibem a distribuição dos vídeos por categoria para os 38 *sites* da coleção. Comparando os gráficos é possível perceber que os conteúdos são bastante variados entre os *sites*. Alguns portais são especializados em uma única categoria de vídeos (ou em poucas categorias), enquanto outros *sites* oferecem um conteúdo mais diversificado.

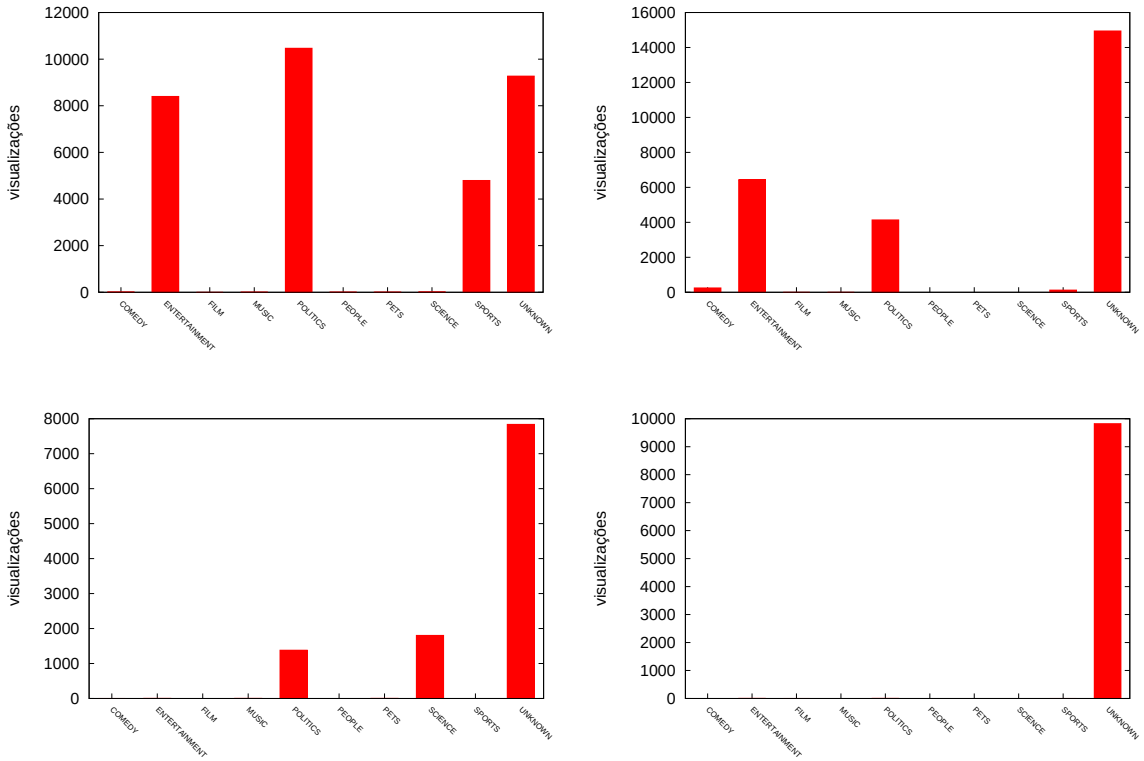

Figura A.1. Distribuição de vídeos por categoria para cada *site* (ordenados pelo número de vídeos total) - *sites* de 1 a 4.

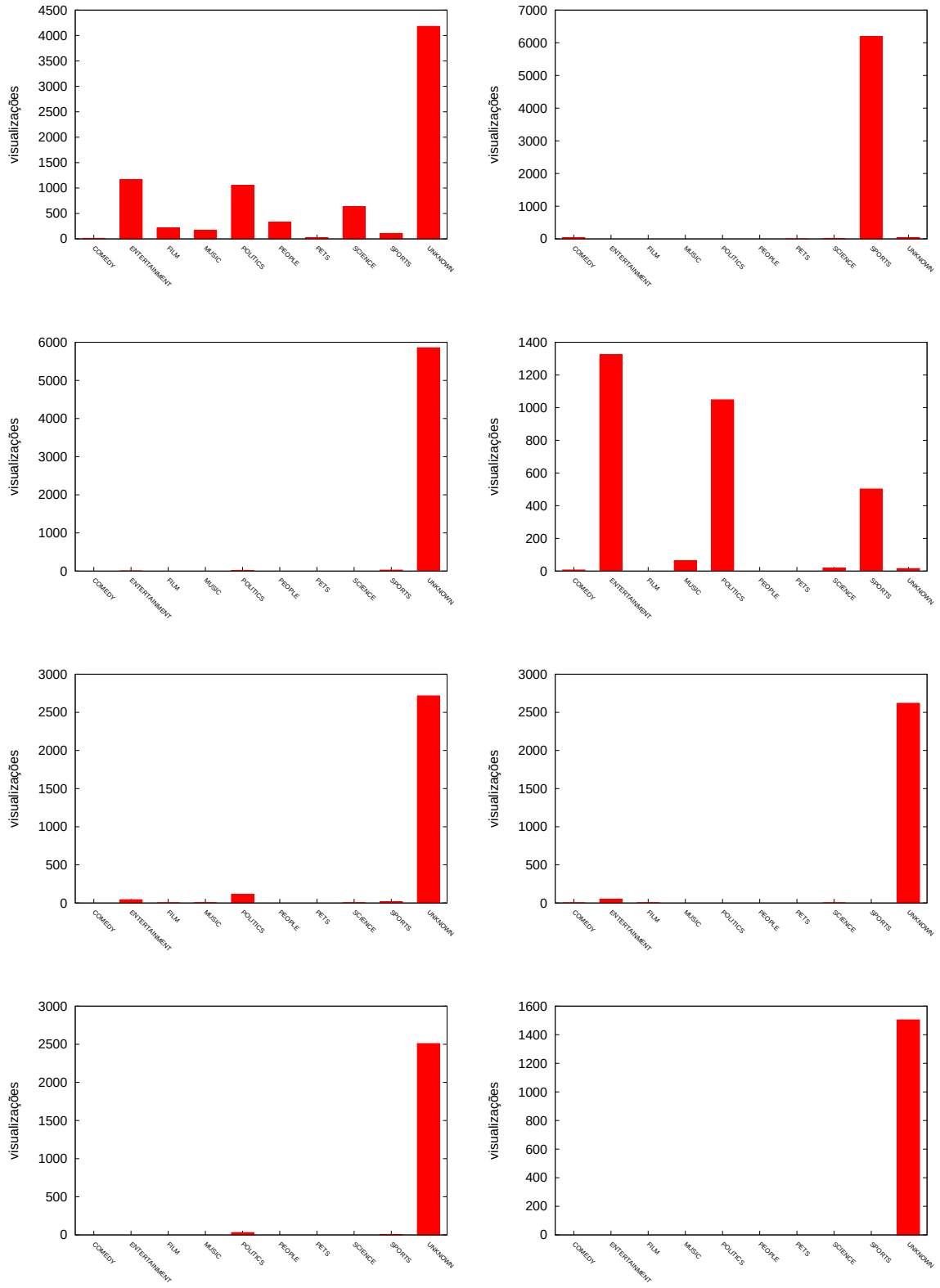

Figura A.2. Distribuição de vídeos por categoria para cada *site* (ordenados pelo número de vídeos total) - *sites* de 5 a 12.

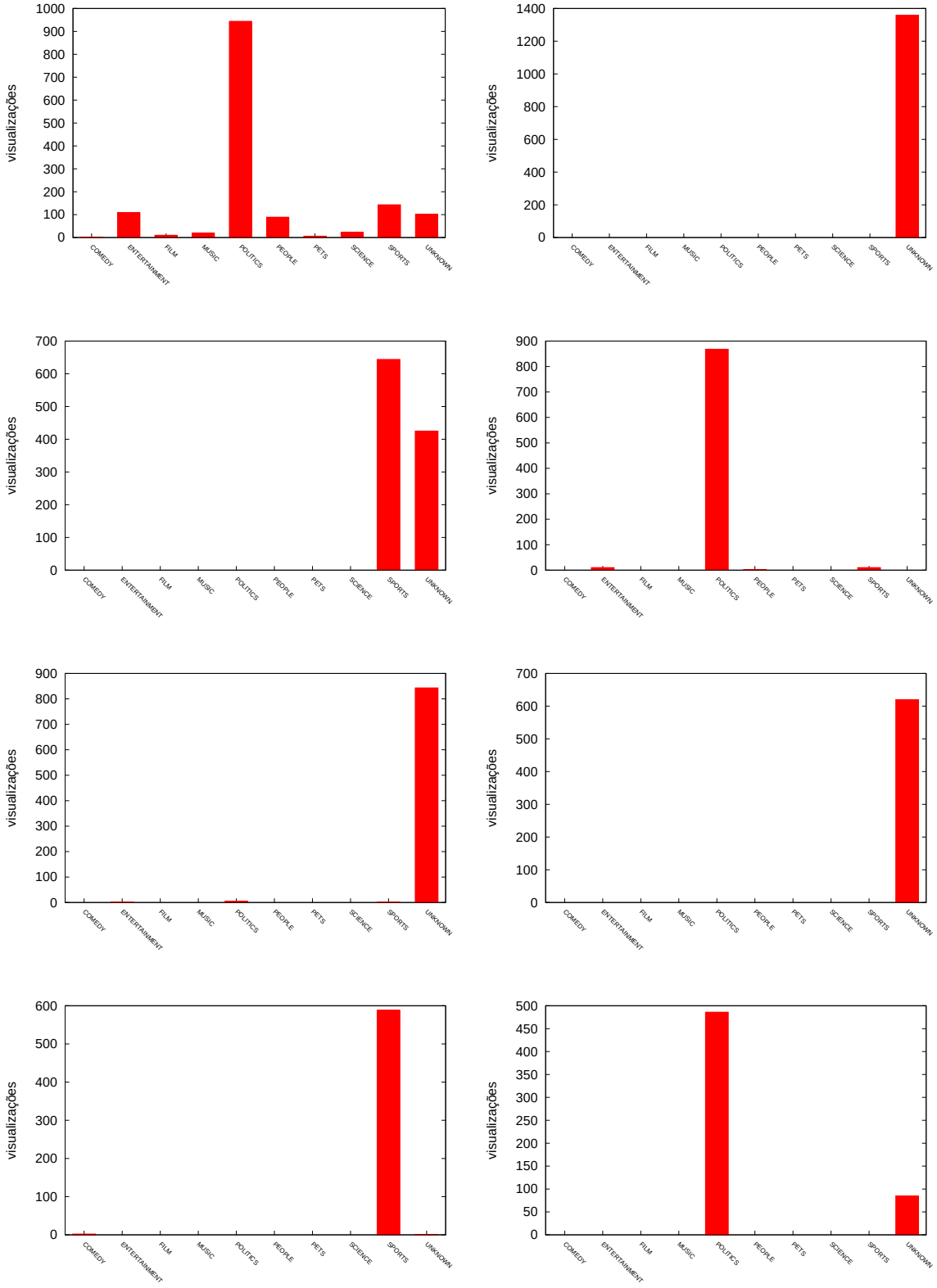

Figura A.3. Distribuição de vídeos por categoria para cada *site* (ordenados pelo número de vídeos total) - *sites* de 13 a 20.

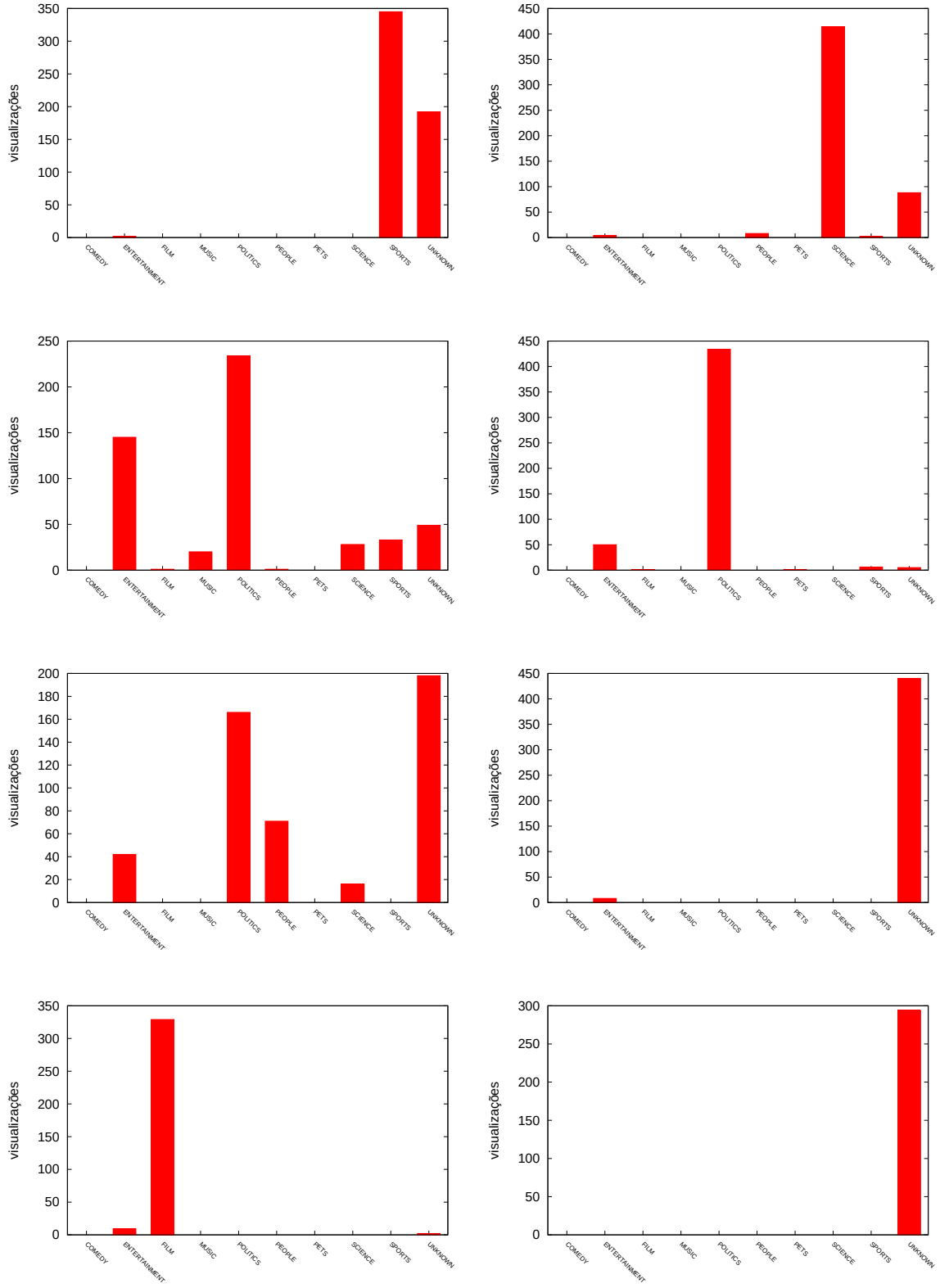

Figura A.4. Distribuição de vídeos por categoria para cada *site* (ordenados pelo número de vídeos total) - *sites* de 21 a 28.

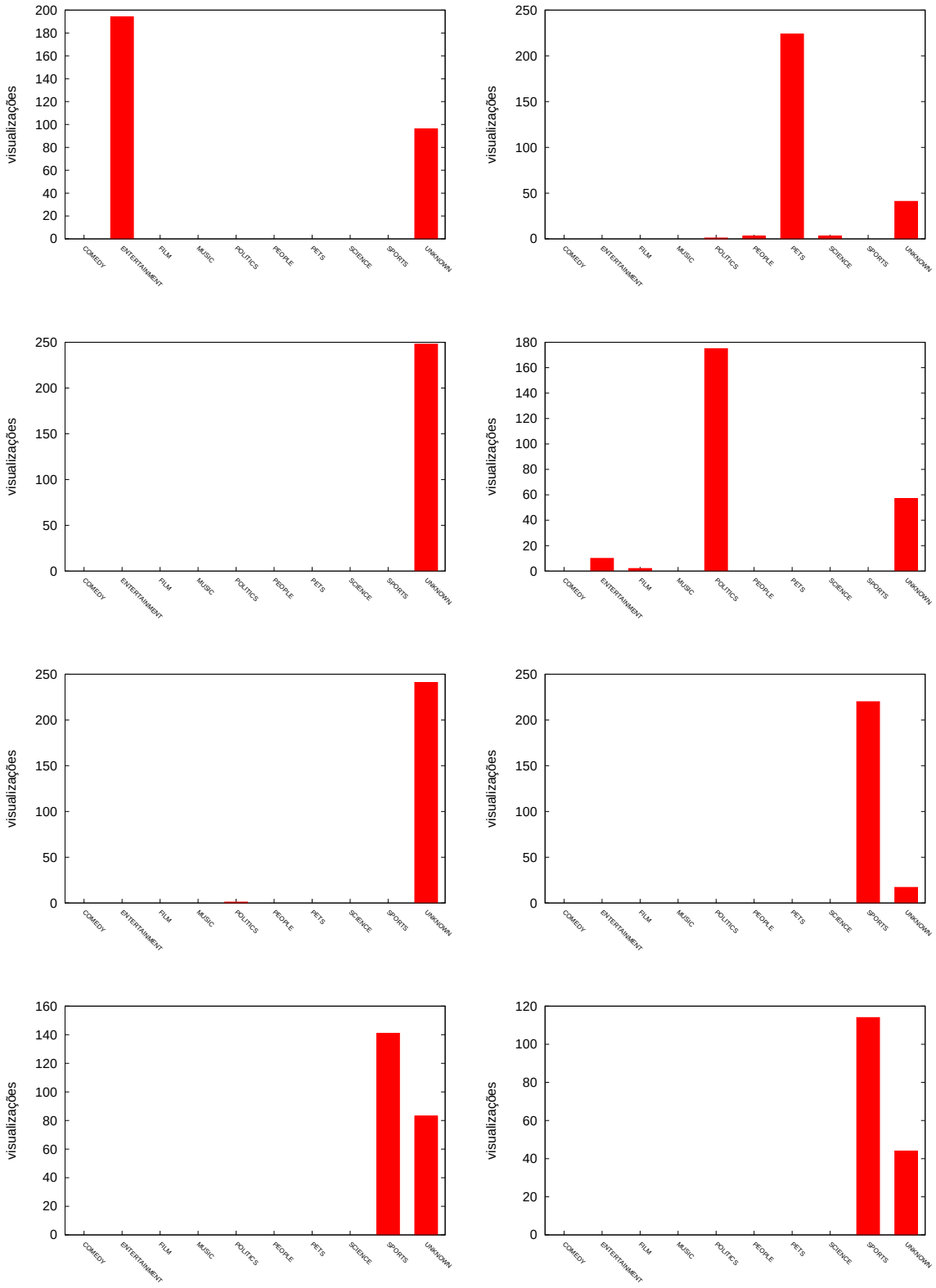

Figura A.5. Distribuição de vídeos por categoria para cada *site* (ordenados pelo número de vídeos total) - *sites* de 29 a 36.

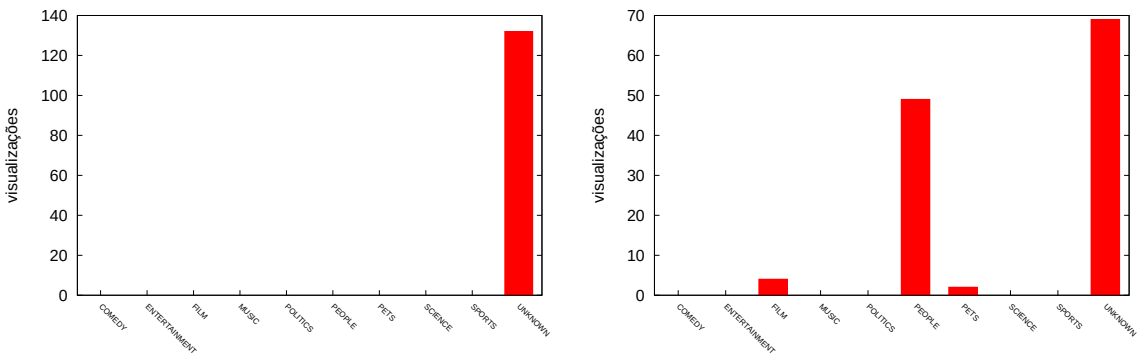

Figura A.6. Distribuição de vídeos por categoria para cada *site* (ordenados pelo número de vídeos total) - *sites* 37 e 38.