ROCK: UMA METODOLOGIA PARA A CARACTERIZAÇÃO DE SERVIÇOS WEB MULTIMÍDIA BASEADA NUMA HIERARQUIA INFORMACIONAL

CHARLES FERREIRA GONÇALVES

ROCK: UMA METODOLOGIA PARA A CARACTERIZAÇÃO DE SERVIÇOS WEB MULTIMÍDIA BASEADA NUMA HIERARQUIA INFORMACIONAL

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: Wagner Meira Júnior
Co-orientador: Adriano C. M. Pereira

Belo Horizonte

Agosto de 2011

Gonçalves, Charles Ferreira
G635r *ROCK*: Uma Metodologia para a Caracterização de Serviços Web Multimídia baseada numa Hierarquia Informacional / Charles Ferreira Gonçalves. — Belo Horizonte, 2011
xxvi, 162 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais- Departamento de Ciência da Computação
Orientador: Wagner Meira Júnior
Co-orientador: Adriano C. M. Pereira

1. Computação - Teses. 2. Recuperação de informação - Teses. 3. World Wide Web (Sistema de recuperação da informação). I. Orientador.
II. Coorientador. III. Título.

CDU 519.6*73(403)

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

FOLHA DE APROVAÇÃO

ROCK: Uma metodologia para a caracterização de serviços web multimídia baseada numa hierarquia informacional

CHARLES FERREIRA GONCALVES

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. WAGNER MEIRA JÚNIOR - Orientador
Departamento de Ciência da Computação - UFMG

PROF. ADRIANO CÉSAR MACHADO PEREIRA - Co-orientador
CEFET-MG

PROF. FABRÍCIO BENEVENUTO DE SOUZA
Departamento de Computação - UFOP

PROF. ADRIANO ALONSO VELOSO
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 05 de agosto de 2011.

"É uma cilada Bino !!"
(Pedro)

Resumo

Apesar da crescente popularização do uso de conteúdos multimídia na Web Brasileira, pouco se sabe a respeito do consumo desses conteúdos. A maioria das metodologias de caracterização existente foca suas análises no acesso aos dados e nos conceitos de sessões e usuários, ignorando informações potencialmente relevantes associadas ao conteúdo dessas mídias. Neste trabalho apresentamos uma nova metodologia de caracterização de serviços Web multimídia organizada de forma hierárquica em quatro partes: Requisição (R), Objeto (O), Conteúdo (C) e Conhecimento (K) - ROCK. A metodologia propõe uma segmentação das análises em diferentes camadas visando a extração de informações existentes nos conteúdos. Aplicamos nossa metodologia aos dados reais de um distribuidor de conteúdos multimídia corporativo, demonstrando sua utilidade e aplicabilidade.

Palavras-chave: Multimídia, Web, Metodologia de Caracterização, Vídeos Online.

Abstract

Despite the increasing popularity of multimedia content in the Brazilian Web, we still do not understand much about the dynamics of how such content is consumed. The majority of previous characterization strategies focuses on data access, employing dimensions such as users and sessions and ignoring content-related information that may be potentially relevant. In this work we present ROCK, a novel characterization methodology of multimedia web services that has four parts: Request, Object, Content, and Knowledge. Besides quantitative characterization criteria, the methodology employs a segmentation-based analysis to grasp the semantic characteristics of the delivered content. We apply our methodology to real data from a distributor of corporate's multimedia content, demonstrating its applicability and usefulness.

Keywords: Multimedia, Web, Characterization Methodology, Online Videos.

Lista de Figuras

Lista de Tabelas

Lista de Siglas

API Application Programming Interface. Interface de Programação de Aplicativos é um conjunto de rotinas e padrões estabelecidos por um software para a utilização das suas funcionalidades por aplicativos que não pretendem envolver-se em detalhes da implementação do software, mas apenas usar seus serviços.

IAT *Inter Arrival Time*, (tempo entre chegadas) é definido como o tempo decorrido entre duas requisições consecutivas em um dado servidor, ou seja, $iat(j) = t(j+1) - t(j)$.

CDF *Cumulative Distribution Function* Função distribuição acumulada, descreve a probabilidade da variável avaliada assumir um valor menor ou igual a x

CCDF *Complementary Cumulative Distribution Function* Função distribuição acumulada complementar, tem a mesma característica da CDF porém com uma visão complementar. Avalia a probabilidade de uma variável assumir um valor maior a x.

CDN *Content Delivery Network*. Rede de distribuição de conteúdo. Sistema distribuído que otimiza a entrega de objetos na Internet.

codec Padrões tecnológicos utilizados para codificar e decodificar arquivos de mídia, favorecendo compactação para armazenagem e descompactação para visualização.

DRM Digital Rights Management. Gerenciador de direitos digitais. Protocolo e software responsáveis por assegurar a confidencialidade do conteúdo digital.

HDFS *Hadoop File System*, uma das abstrações do Hadoop, é um sistema de arquivos distribuído desenvolvido para aplicações de processamento em

lote que necessitam acesso contínuo a arquivos muito grandes que podem estar distribuídos em várias máquinas

IAT Inter Arrival Time. Tempo entre requisições também conhecido como tempo entre chegadas. Definimos como o tempo decorrido entre duas requisições consecutivas.

IO *Input-Output*. Entrada e saída. Termo utilizado para se referir a operação de escrita e leitura em processos computacionais

OVP Online Video Platform. Plataforma de vídeos online, ferramentas que facilitam o trabalho com vídeos na Web.

PaaS Platform as a Service. Plataforma como serviço é uma modalidade de prestação de serviço possibilitada pela tecnologia de computação em nuvens onde todas ferramentas como infraestrutura, bibliotecas, serviços, etc são fornecidas pela Internet. Com essa modalidade o consumidor não precisa se preocupar com infraestrutura nem programas para ter sua aplicação/serviço funcionando

SaaS Software as a Service. Programas como serviço é a habilidade que um consumidor possui de utilizar um programa sobre demanda através de um cliente leve como um navegador da Internet.

SLA Service Level Aggremment. Acordo de nível de serviço é um acordo firmado entre a área de TI e seu cliente interno, que descreve o serviço de TI, suas metas de nível de serviço, além dos papéis e responsabilidades das partes envolvidas no acordo.

UGC *User Generated Content*. Conteúdo gerado pelo usuário. Modalidade de produção de conteúdos digitais onde não há a participação de um produtor de conteúdo profissional, mas sim os próprios usuários de um sistema que se encarregam por gerar os principais conteúdos.

User Agent É um identificador textual que define qual dispositivo está acessando um dado recurso na Web.

VOD Video On Demand. Vídeo sobre demanda. Modalidade de entrega de vídeos onde o cliente solicita o conteúdo no momento desejado, diferentemente de *streamming* que funciona como a TV aberta, o conteúdo é transmitido a revelia da audiência.

Sumário

Capítulo 1

Introdução

1.1 Contexto

Nos últimos anos temos presenciado uma enorme expansão no uso de mídias ricas em aplicações e portais na *Web*, possibilitando aos usuários novas formas de interação, comunicação e aprendizagem. A maior parte dessas aplicações utiliza recursos multimídia para enriquecer a experiência do usuário, enquanto outras baseiam suas funcionalidades principais nestes recursos multimídia. Aplicações de rádios online e áudio, como o *Grooveshark*[1], fotos, como o *Flicker*[2] e *Picasa*[3], e vídeos , como Youtube [4], Hulu [5] e NetFlix[6], já fazem parte do cotidiano das pessoas e continuam se tornando cada vez mais populares.

Todas essas aplicações foram possíveis devido ao aumento da disponibilidade de tecnologia de banda larga que consequentemente também promoveu o crescimento do uso de outras aplicações como videoconferência, televisão interativa, jogos online, aplicações de transmissão de vídeos pela rede e comunicação em tempo real (*Instant Messengers* e VoIP). Essa mudança de tecnologia deu origem ao conceito de Web 2.0, que caracteriza-se por [Boll, 2007]:

- Aplicações dinâmicas, onde os usuário as usam tanto na interação com o sistema como na participação do mesmo na geração de conteúdo;
- A utilização de mídias ricas como imagens, sons e vídeos para a comunicação;

[1] `http://listen.grooveshark.com/`
[2] `http://www.flickr.com/`
[3] `http://picasaweb.google.com/`
[4] `http://www.youtube.com/`
[5] `http://www.hulu.com`
[6] `http://www.netflix.com/`

- O estabelecimento de plataformas distribuidoras de conteúdo;

Notoriamente, o vídeo é a mídia que mais ganhou força nos últimos anos com o fortalecimento do aspecto colaborativo das aplicações *Web*. Atualmente, assistir filmes, noticiários, vídeo-aulas e realizar videoconferências pela Internet já são tarefas rotineiras. Estima-se que em 2010 [Pingdom, 2010] mais de 2 milhões de vídeos foram assistidos diariamente no Youtube. Segundo um estudo da Limelight [Limelight, 2010], o tráfego de vídeos nos EUA já representa 51% de todo o tráfego da Internet, seguido por 23% de tráfego *peer-to-peer* que, em grande parte, se deve a *downloads bit-torrent*, como podemos ver na Figura 1.1 retirada do estudo.

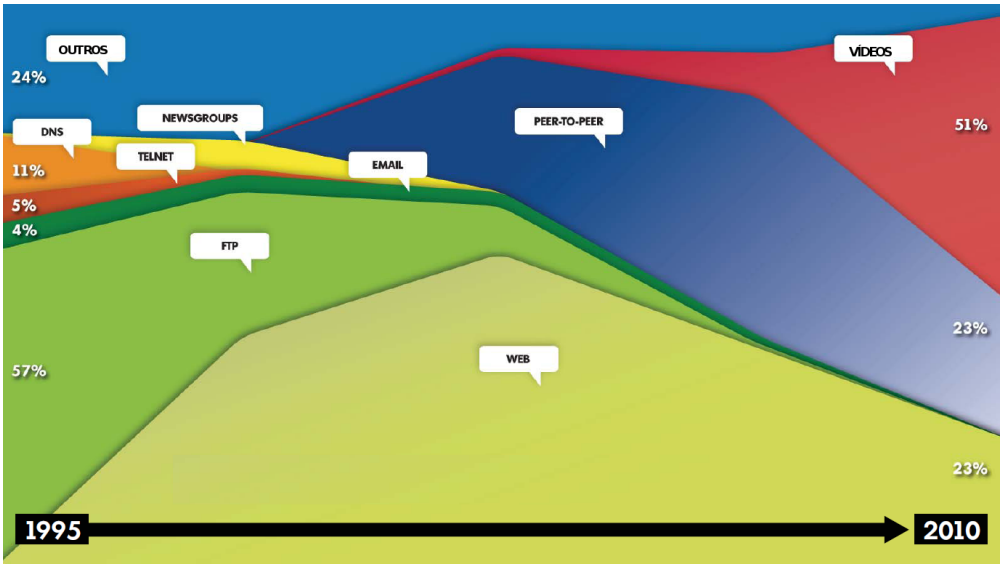


Figura 1.1. Proporção do tráfego da Internet americana de 1995 a 2010 (Limelight)

Entretanto, como mostra o levantamento realizado por [Sar, 2011] (apresentado na Figura 1.2), a maior parte dos conteúdos trafegados nas redes *peer-to-peer* também são relacionadas a vídeos, o que nos leva a estimar que a grande maioria do tráfego de vídeos na Internet americana está, de alguma forma, relacionada a vídeos.

Com a expansão de dispositivos móveis, como celulares do tipo *Smartphone*, *Tablets* e *Netbooks*, a tendência é que vídeos sejam utilizados cada vez mais e para os mais diversos fins. Além desses dispositivos, uma nova tendência é a fusão entre os televisores e Internet ganhando força recentemente com a divulgação do Google Tv (`http://www.google.com/tv/`) e Apple TV (`http://www.apple.com/appletv/`) que, caso se estabeleçam, transformarão por completo a forma como os *telespectadores* utilizam seus televisores.

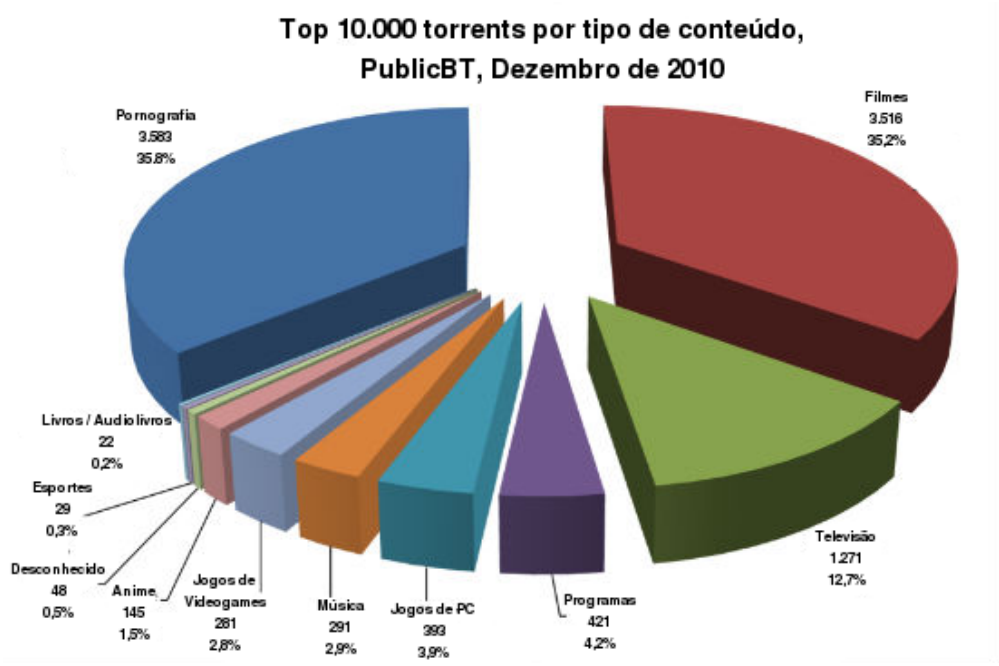


Figura 1.2. Resumo dos 10.000 torrents mais populares por tipo de conteúdo

Para que todo esse universo multimídia de entretenimento, notícias e negócios seja possível é necessário conhecimentos técnicos específicos e avançados, infra-estrutura robusta e escalável, além de ferramentas que suportem todos os processos existentes desde a captura do vídeo até a entrega ao usuário final. Tal fluxo, representa um processo muito complexo como podemos ver na Figura 1.3.

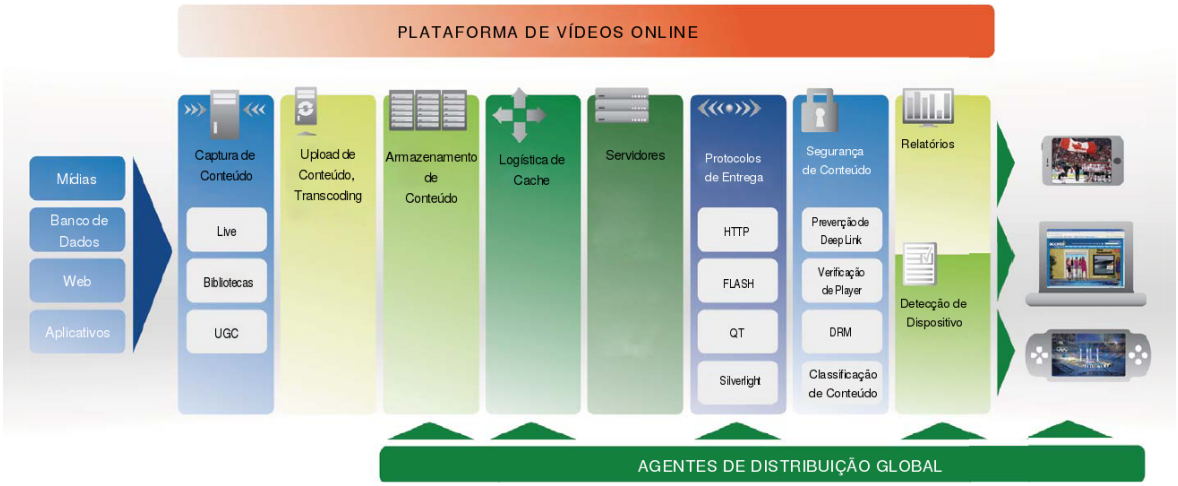


Figura 1.3. Complexidade do processo de distribuição de conteúdos online

Por exemplo, no caso de distribuição de vídeos, as plataformas de vídeos online

(*OVP*, Online Video Platform) geralmente realizam as seguintes tarefas:

- Processos de aquisição do material a ser distribuído como servidores de streaming para transmissões ao vivo, ou ferramentas que permitam a coleta ou gravação dos vídeos por parte dos usuários.
- Inserção de metadados, *upload* de material;
- Armazenamento;
- Logística de distribuição;
- Servidores de processamento, redes capacitadas, bandas para tráfego;
- Softwares de distribuição em diversos protocolos;
- Mecanismos de segurança como gestão de direitos digitais (*DRM*), acesso restrito a regiões geográficas (*geoblocking*);
- Sistemas de análises e estatísticas;

Para suprir essa demanda, algumas empresas estão se especializando em prover as ferramentas para que empresas focadas em produção de conteúdos possam se dedicar exclusivamente ao seu trabalho especializado e não se preocupar com detalhes que não fazem parte da sua atividade fim.

Nos Estados Unidos várias empresas como Brightcove, Kaltura, Ooyala, dentre outras, já se consolidaram nesse mercado. Em [StreamingMedia, 2011] temos uma lista das maiores plataformas de vídeos online daquele país. No Brasil, a empresa líder no setor de plataformas comerciais de vídeos online é a Samba Tech (http://www.sambatech.com/), detentora de cerca de 70% do mercado de vídeos corporativo do Brasil [Dalmazo, 2011]. Atualmente esta empresa trabalha com grandes grupos de mídias, como Portal R7, GloboSat, SBT, Grupo Abril, IG, Bandeirantes, RBS, dentre outros grandes grupos possuindo dentre seus clientes 8 dos 10 maiores grupos de mídias online da Internet brasileira.

1.2 Motivação

Como podemos observar, com o fortalecimento do aspecto colaborativo da Web, o tráfego e utilização de vídeos é dominante na Internet e tudo indica que seu uso será cada vez maior. Recentemente, o Youtube anunciou que a cada minuto 24 horas de conteúdo de vídeo são inseridos em sua plataforma [Blog, 2010]. Adicionalmente os

números mostram uma tendência crescente na publicação de conteúdo. É importante notar que toda essa mudança traz novos desafios e consequências que devem ser estudados e avaliados. Porém, como tratar tais desafios se não se conhece a fundo o perfil de uso e as características intrínsecas a tais recursos multimídia? É preciso conhecer em detalhes a dinâmica de tais serviços e é isso que se pretende com essa dissertação.

Poucos estudos apresentam caracterizações suficientemente amplas sobre o consumo de vídeos na Web brasileira que ofereçam uma boa compreensão acerca do perfil de uso e do impacto da distribuição de vídeos online. A maioria dos trabalhos [Gill et al., 2007, Cha et al., 2007, Gill et al., 2008] explora a hierarquia de acesso aos objetos em termos de requisições, sessões e usuários, ignorando a natureza do conteúdo multimídia e as informações inerentes a ele, diferenciando do presente estudo nos seguintes aspectos:

1. O foco é em conteúdos gerados pelo usuário (UGC). Esta dissertação foca em conteúdos profissionais (PC *professional content*) gerados por empresas especializadas em notícias, esportes, música, entretenimento dentre outros gêneros. Seus conteúdos são criados por profissionais especializados e pretendemos avaliar as principais diferenças deste perfil de conteúdo comparado aos UGC.

2. Em sua grande maioria, os estudos são realizados com informações coletadas na Internet [Cha et al., 2007] ou através de *proxies* em redes controladas [Gill et al., 2007], raramente se encontra na literatura trabalhos que possui acesso aos dados (logs, metadados, etc) de forma privilegiada como temos nesse caso, o que nos permite realizar um estudo mais amplo e detalhado.

3. A grande maioria dos usuários finais dos estudos existentes, estão localizados em um contexto diferente do brasileiro. Os dados que são analisados nesta dissertação possuem em sua maioria usuários brasileiros que possuem culturas e hábitos diferentes que podem influenciar na forma de utilização e consumo dos serviços. Além disso, no Brasil a disponibilidade e custo de rede e infraestrutura é completamente diferente de outros países, outra motivação que justifica tal estudo específico.

Observa-se a carência de métodos que possibilitem avaliar os serviços multimídia também sob o ponto de vista da informação que oferecem e como esse ponto de vista pode ser útil em diferentes contextos como (i) organizar melhor a infraestrutura de acesso e distribuição das informações; (ii) avaliar o impacto de diferentes tipos de mídias; (iii) avaliar a popularidade de seus conteúdos; e (iv) extrair padrões que possam inferir informações chave relacionadas ao negócio em questão.

Assim sendo, a relevância do presente trabalho se mostra pelo aspecto diferencial de outros trabalhos da literatura e também por estar em consonância com alguns dos "Desafios de Pesquisa em Computação 2006-2016" [SBC, 2006], definidos em evento da Sociedade Brasileira de Computação (SBC), realizado em maio de 2006. Os desafios referidos são:

- *Gestão da Informação em grandes volumes de dados multimídia distribuídos*: o presente trabalho aborda o estudo de uma plataforma para a gestão de conteúdos multimídias de alcance nacional que atualmente faz a gestão de grandes grupos de mídias Brasileiros e nos permitirá entender melhor como funciona a dinâmica da distribuição de vídeos no Brasil e assim prover *insights* de como melhorar a gestão dessas informações.

- *Desenvolvimento tecnológico de qualidade - Sistemas disponíveis, corretos, seguros, escaláveis, persistentes e ubíquos.* Para o desenvolvimento tecnológico de qualidade é necessário conhecimentos sólidos e pesquisa. É impraticável o avanço na área de distribuição de conteúdo multimídia sem conhecer os detalhes de uso e as características do ambiente em que as aplicações serão distribuídas e como aplicar tais requisitos em novos dispositivos que permitem a ubiquidade de tais aplicações.

Até onde sabemos não há, na literatura, até o momento, um estudo de caracterização de vídeos da Web brasileira tão amplo como este, capaz de produzir um painel do uso de vídeos online do Brasil.

1.3 Objetivos

Neste trabalho apresentamos uma nova metodologia hierárquica de caracterização de serviços Web multimídia. A metodologia propõe uma segmentação das análises em quatro diferentes camadas: Requisição (R), Objeto (O), Conteúdo (C) e Conhecimento (K, do inglês *knowledge*), dando enfoque na extração de conhecimento a partir das informações contidas no conteúdo dos arquivos multimídia.

Essa metodologia tem por objetivo principal o foco nas informações existentes nos conteúdos multimídia, de forma a complementar a análise geral das requisições realizadas pelos usuários.

Adicionalmente, a metodologia foi aplicada em um estudo de caso de um cenário real utilizando um arcabouço experimental distribuído. O estudo de caso utilizou informações da maior distribuidora de conteúdos multimídia corporativos da América

Latina, a Samba Tech, uma empresa provedora de infraestrutura, gerenciamento e distribuição de vídeos em canal digital. Tivemos acesso aos conteúdos (vídeos, imagens, áudios, etc), metadados e registros de 17 meses de publicação de vídeos.

Os dados foram gerados pela principal plataforma de distribuição da empresa, sendo essa utilizada pelos maiores produtores de conteúdo multimídia da Internet brasileira. Os registros totalizam cerca de 3,5 Petabytes de tráfego e com os metadados referentes a cada conteúdo garante-se as informações semânticas dos conteúdos sem esforço de extração e categorização dos mesmos.

1.4 Contribuições

Os resultados deste trabalho visam ilustrar o potencial da metodologia proposta que, quando aplicada de forma extensiva e direcionada, pode produzir conhecimentos relevantes para o aprimoramento de serviços Web, como a personalização da experiência do usuário, construção de serviços de recomendação, aperfeiçoamento da tarifação de publicidade em vídeos online ou mesmo no direcionamento da produção de novo conteúdo. Além disso, serviços de entrega também podem se utilizar dessas informações para otimizar as políticas de distribuição, explorando o perfil de uso e localidade geográfica ou mesmo sofisticando os mecanismos de replicação e *cache*.

Além dos exemplos citados no parágrafo anterior podemos enumerar as principais contribuições desse trabalho como:

1. A proposta de uma nova metodologia de caracterização e análise de serviços Web multimídia que possibilita a melhor compreensão da natureza dos conteúdos multimídia trafegados pelas plataformas distribuidoras de vídeos e seus *Produtores de Conteúdo*.

2. O uso do arcabouço experimental de computação distribuída, cuja arquitetura pode ser adotada em diferentes trabalhos de processamento, caracterização e análise de dados.

3. A aplicação de tal metodologia em um cenário real através das informações de um empresa de distribuição de conteúdos multimídia corporativos da América Latina, que engloba parte significativa do tráfego dos vídeos online no Brasil.

De forma complementar, a metodologia proposta pode ser utilizada na realização de estudos semelhantes por outros grupos de pesquisa com posse de dados similares de conteúdo multimídia. Uma vez que a metodologia utiliza ferramentas escaláveis, não há

restrições de quantidades de dados a serem analisados o que pode permitir até mesmo realizar tais análises em tempo de distribuição (*online*) e não num pós-processamento (*offline*), o que agilizaria a tomada de decisões por parte dos editores e gestores dos *Produtores de Conteúdo*.

A metodologia aqui proposta foi aceita para ser apresentada no Simpósio Brasileiro de Sistemas Multimídia e Web em outubro deste ano [Gonçalves et al., 2011a]. Complementarmente, o trabalho de avaliação do arcabouço distribuído utilizado neste trabalho foi apresentado no SCECR [Gonçalves et al., 2011b] (*Seventh Symposium on Statistical Challenges in Electronic Commerce Research*).

1.5 Organização

O presente trabalho se divide da seguinte maneira: no Capítulo 2, temos uma discussão a respeito de alguns trabalhos que se relacionam com a presente dissertação. Em seguida, no Capítulo 3, apresentamos a metodologia *ROCK*. Explicamos o conceito hierárquico e as entidades das camadas *ROC* e mostramos como podemos aplicar métodos de descoberta do conhecimento em banco de dados (*KDD*) em análises transversais às camadas anteriores para extrapolar a compreensão básica, possibilitada pelas análises intrínsecas a cada nível da hierarquia, e inferir conhecimentos (*K*) sobre o conteúdo alvo das requisições.

A explicação e algumas análises sobre o arcabouço utilizado para a viabilização do estudo e processamento do grande volume de dados é mostrado no Capítulo 4. A metodologia proposta é aplicada nos registros da maior plataforma de distribuição de vídeos corporativos da América Latina no Capítulo 5, compondo nosso estudo de caso. Finalizamos esta dissertação apresentando as conclusões e trabalhos futuros no Capítulo 6.

Capítulo 2

Trabalhos Relacionados

Nesta seção serão discutidos os trabalhos de maior relevância, disponíveis na literatura, relacionados ao trabalho realizado nesta dissertação.

Caracterização de carga é fundamental para o entendimento e aprimoramento de sistemas *Web*. Há vários estudos que apresentam caracterizações de carga de trabalho de diferentes tipos, tais como servidores *Web* [Arlitt & Williamson, 1996], de comércio eletrônico [Menasce & Almeida, 2000], de blogs [Duarte et al., 2007b], de vídeo sob demanda [Costa et al., 2004] e de vídeo ao vivo [Veloso et al., 2006b]. Dentre as várias contribuições desses trabalhos, destacamos a criação de modelos capazes de descrever a carga que chega nesses servidores, essenciais para a geração de carga sintética que, por sua vez, possibilita a realização de experimentação e simulação baseadas em distribuições realistas.

Em [Veloso et al., 2006b], os autores estudaram a carga de trabalho de um servidor comercial de fluxos de vídeo ao vivo localizado no Brasil. O foco do estudo foi a caracterização do processo de chegada de sessões de usuários, bem como dos tempos destas sessões com o intuito de utilizar os resultados em um gerador de cargas sintéticas realistas. Os autores consideraram um modelo composto por períodos de atividade (quando o usuário recebe mídia) intercalados por períodos de inatividade, disparados por alguma ação do usuário (como pausa). Alguns dos principais resultados obtidos são: (1) o processo de chegada dos usuários segue uma distribuição Poisson; (2) os períodos de atividade são bem modelados por uma distribuição lognormal; (3) os períodos de inatividade são bem modelados por uma distribuição exponencial. A metodologia hierárquica proposta em [Veloso et al., 2006b] serviu de *baseline* para que pudéssemos organizar nossa hierarquia de informação, que será explicada no Capítulo 3.

Em [Sripanidkulchai et al., 2004], foi caracterizada uma carga de trabalho de fluxos de vídeo e áudio ao vivo de uma grande CDN, o que possibilitou a análise

de uma ampla diversidade de conteúdos. Mais de 90% do conteúdo analisado foi apenas áudio. Observou-se que a popularidade dos conteúdos segue uma distribuição Zipf de duas partes. Foi também observado que os clientes entram no sistema de acordo com uma distribuição exponencial e que a duração das sessões dos usuários apresenta cauda pesada.

Em [Cha et al., 2008], os autores apresentam o que eles clamam ser a primeira análise do maior sistema de IPTV do mundo. O trabalho foca na análise do padrão de visualização dos conteúdos do sistema pelos usuários. Foram consideradas 250.000 casas (cada casa pode ter mais de 1 usuário), por um período de 6 meses. Dentre os principais resultados, os autores observaram a ocorrência de um elevado número de sessões de curta duração (aproximadamente 60% até 10 segundos), que eles consideram ser provenientes de mudanças de canais (período em que o usuário está surfando entre canais com a finalidade de escolha de algum desses canais). Os autores identificaram que a maioria das mudanças de canais não leva o usuário a efetivar a escolha da visualização desse novo canal, procedimento que pode sobrecarregar a rede com sessões de mudanças de canais supérfluas. Além disso, os autores analisaram os processos de chegada e partida dos usuários no sistema, identificando um comportamento bastante dinâmico, que por sua vez impõe sérios desafios na construção de um sistema de IPTV baseado em P2P.

Dois servidores de conteúdo educacional de renomadas universidades americanas, tiveram suas cargas de trabalho analisados no trabalho de [Almeida et al., 2001]. Durante períodos aproximadamente estacionários de taxa de chegada de usuários, o processo de chegada de sessões foi observado ser aproximadamente Poisson e o tempo entre requisições de interatividade seguiam uma distribuição de Pareto. Foi observado também que a popularidade dos objetos analisados era melhor modelada por uma distribuição Zipf de duas partes.

Em [Costa et al., 2004], os autores realizaram uma extensão do trabalho apresentado em [Almeida et al., 2001] analisando também dois servidores de vídeos em um contexto educacional, entretanto, o foco foi principalmente no comportamento de interatividade dos usuários em sistemas para transmissão de mídia pré-armazenada. As cargas de trabalho analisadas representam três diferentes tipos de conteúdo, educacional, entretenimento baseado em áudio e entretenimento baseado em vídeo. Foram encontrados inúmeros resultados, dentre eles o que mostra que a probabilidade de um cliente pausar, avançar ou voltar um conteúdo depende fortemente da interação realizada imediatamente antes, dentro da mesma sessão, e não no número de requisições efetuadas desde o início da sessão. Além disso, o trabalho mostra que o tempo entre requisições segue uma distribuição Pareto e que a popularidade de objetos multimídia

pode ser modelada pela concatenação de duas distribuições do tipo Zipf.

Especificamente com relação ao estudo da popularidade do conteúdo analisado, podemos citar [Acharya & Smith, 2000] que realizou suas análises baseadas em acessos de usuários a vídeos transmitidos na WEB, e [Chesire et al., 2001], que analisou uma carga de trabalho de um servidor de fluxo de mídia de uma grande empresa. O primeiro trabalho identificou que a popularidade do conteúdo não segue uma distribuição Zipf, já o segundo trabalho observou que a popularidade do conteúdo analisado segue uma distribuição Zipf. Representam também essa classe de trabalhos [Yu et al., 2006] e [Cherkasova & Gupta, 2002]

Complementar ao nosso esforço, existem vários trabalhos que caracterizam diferentes aspectos de sistemas de compartilhamento de vídeos gerados pelos usuários, especialmente do *YouTube*. A geração de conteúdo por parte dos usuários, em oposição à geração de conteúdo profissional de grupos de mídias ou empresas, é a chave para o sucesso de vários serviços da Web 2.0 atual. Devido ao significativo crescimento na produção de vídeos amadores e o consequente aumento do tráfego gerado por eles, vários trabalhos buscam entender melhor esse novo fenômeno, por exemplo, estudando o comportamento dos usuários. Nessa classe de trabalhos também existem vários representantes, no entanto discutiremos apenas os mais relevantes

Em [Cha et al., 2007], os autores analisam a distribuição de popularidade, evolução, e características dos vídeos do *YouTube*. Segundos os autores, este foi o primeiro grande trabalho para um melhor entendimento de sistemas que permitem a participação dos usuários na geração de conteúdo. Baseado no estudo do comportamento dos usuários foram identificados elementos chave que auxiliam em uma melhor compreensão, por exemplo, da popularidade do conteúdo. Os autores também realizaram análises em relação ao uso de estratégias P2P e de um melhor uso de caches, com o intuito de tornar esse tipo de sistema mais eficiente.

Gill *et al.* [Gill et al., 2007] apresentam uma caracterização do tráfego do *YouTube* sob duas perspectivas: local, na Universidade de Calgary no Canadá, e global, a partir da lista dos 100 vídeos mais vistos que é disponibilizada no website do mesmo. Os autores mostram que a utilização de cache traz benefícios para os usuários e provedores de conteúdo.

Com relação ao comportamento dos usuários nesses sistemas de compartilhamento de vídeos, o trabalho de [Benevenuto et al., 2009b] mostra a existência de usuários maliciosos e oportunistas, enquanto [Benevenuto et al., 2009a] aborda o problema de identificar tais usuários.

No trabalho de Crane e Sornete [Crane & Sornette, 2008] é feito um estudo da

resposta de relaxamento [1] de um sistema social após rajadas de atividades utilizando a série temporal de visualizações diárias de quase 5 milhões de vídeos no Youtube. Os autores mostram que tal atividade pode ser precisamente descrita como um processo de Poisson. No estudo, quase 90% dos vídeos apresentavam tal distribuição. Esses são os vídeos que não registravam muita atividade. Contudo, foram encontrados outras centenas de milhares de exemplos, aproximadamente 10% das amostras, no qual a atividade em rajada é seguida por um processo power-law de relaxamento que descreve o intervalo das visualizações.

Crane e Sornete [Crane & Sornette, 2008] acreditam que esses expoentes de relaxamento se agrupam em três classes distintas e sugerem a classificação em:

Vídeos virais: são aqueles com uma divulgação boca a boca prévia que resulta em uma propagação epidêmica através das redes sociais

Vídeos de qualidades: são similares aos vídeos virais, contudo experimentam uma súbita explosão de atividade ao invés de um crescimento *bottom-up*. Devido à qualidade do seu conteúdo estes vídeo provocam um subsequente cascateamento epidêmico através das redes sociais.

Vídeos sem valor: são aqueles que experimentam uma explosão de atividades por alguma razão (spam, acaso, etc.) porém não se propagam através das redes sociais.

Os autores afirmam que embora alguns possam argumentar que essa classificação seja inerentemente subjetiva, elas refletem um medida objetiva contida nas respostas coletivas a eventos e informações. Além disso eles acreditam que esse padrão é consistente com um modelo epidemiológico sobre uma rede social que contém dois ingredientes: uma distribuição power-law de tempo de espera entre causa e ação e uma cascata epidemiológica de ações se tornando a causa de ações futuras. Este modelo é uma extensão conceitual do teorema dissipação-flutuação para sistemas sociais e provê um arcabouço único para a investigação de intervalos de tempo em sistemas complexos.

Em [Benevenuto et al., 2009b] os autores exploram a rede de relacionamento criada através do recurso de *Video Response* do *YouTube* e possui como principal objetivo entender o comportamento da rede social criada pelo recurso citado. Um vídeo que possuir outro vídeo como resposta é denominado *responded video*. Neste cenário emerge dois tipos de usuários: (1) *reponded user* (um de seus vídeos é um responded video) e o (2) *responsive user* (que posta vídeos em resposta a outros). É feita uma análise do

[1] relaxation response

grafo gerado pelos relacionamentos criados entre os vídeos de uma amostra coletada de forma consistente e significativa dos vídeos do *YouTube*. As interações dos usuários possuem uma característica similar ao grafo da web com usuários que atuam como *hubs* e *authorities*. As observações são que usuários que atuam como *authorities* são geralmente empresas de mídia que submetem conteúdo profissional como esportes, entretenimento e seriados de televisão. Um fato interessante ressaltado é que o grafo não possui características de rede sociais (assortative mixing, significante grau de simetria). Outra análise feita é de comportamentos maliciosos e anti-sociais. É analisado como alguns usuários tentam realizar a auto promoção de alguns de seus vídeos, o que pode ser caracterizado como spam. O problema no caso de vídeos, é que ao contrário de material textual, ainda não existem técnicas para predizer o conteúdo malicioso, e é preciso iniciar o streaming do vídeo para detectar tal conteúdo, o que consome recursos computacionais e de rede. Para detectar tal comportamento é utilizada a métrica *IRD* (inter-reference distance) e o algoritmo *PageRank*.

Em [Gill et al., 2008], os autores analisam as características das sessões dos usuários do *YouTube*, analisando as requisições que partem de uma universidade. Entretanto, os autores avaliam apenas aspectos como o tamanho e chegada de sessões. Em [Zink et al., 2008], os autores realizam simulações que mostram que *cache* de vídeos, tanto no cliente quanto no *proxy*, e distribuição P2P podem reduzir tráfego de rede e permitir acesso mais rápido a vídeos em sistemas de compartilhamento de vídeos.

O trabalho de [Benevenuto et al., 2010] propões mecanismos para diferenciar tipos de usuários a partir de seus padrões de navegação em um servidor de vídeos do UOL, um dos maiores provedores de conteúdo da América Latina. Nesse trabalho é apresentado uma completa caracterização de sessões de usuários, suas requisições ao servidor além do padrão de navegação desses usuários. Entre os achados do trabalho se encontram:

- Uma sessão típica de um usuário de um sistema de compartilhamento de vídeos online dura cerca de 40 minutos, um valor alto em comparação com sistemas Web tradicionais.

- As distribuições de popularidade dos acessos dos objetos (vídeos e etiquetas [2]) seguem uma distribuição em cauda longa.

- O ordem de atividades dos usuários em termos de requisições enviadas e as sessões criadas seguem, respectivamente, distribuições de cauda longa e exponencial.

[2] tags

- A taxa de chegada de requisições no sistema apresenta um padrão periódico com intensidade alta durante o dia e menor intensidade durante a noite.

- As distribuições do tempo entre requisições e dos tempos entre sessões podem ser modeladas por uma distribuição exponencial.

- Os usuários gastam mais tempo tempo vendo vídeos em sessões longas do que em sessões curtas.

- Descoberta de diferentes perfis de usuários que acessam o sistema que podem ser utilizados pelos administradores para personalizar os serviços.

O estudo do crescimento de popularidade dos vídeos foi abordado no trabalho [Figueiredo et al., 2011] utilizando análises de dados recém disponibilizados pelas ferramentas de estatísticas do *Youtube*. O estudo é conduzido sobre três tipos diferente de vídeos. Os vídeos mais acessados, os vídeos removidos por violarem direitos autorais e vídeos selecionados aleatoriamente.

Os resultados mostram que o padrão de crescimento de acesso varia de acordo com o conjunto de vídeos. Em particular, os vídeos protegidos por copyright tendem a obter a maior parte das suas visualizações muito cedo, frequentemente exibindo um crescimento de popularidade caracterizada por uma processo de propagação semelhante a uma epidemia viral. Em contrapartida, os vídeos mais populares tendem a experimentar uma repentina explosão de popularidade, sendo que uma grande dessas visualizações ocorrem em um único dia (ou semana) atípico.

Ao caracterizarem as fontes que levavam ao acesso dos vídeos, os autores alegam abordar um aspecto fundamental que teria sido ignorado por trabalhos anteriores. Fazem isso na tentativa de elucidar os mecanismos que conduzem os usuários aos vídeos. Finalmente, como contribuições adicionais, Figueiredo [Figueiredo et al., 2011] mostra que não somente a busca mas também outros mecanismos internos do *YouTube* assumem papeis importantes para atrair usuários aos vídeos de todos os conjunto de dados analisados.

Outro trabalho para a compreensão da popularidade de vídeos é realizado por Chatzopoulou em [Chatzopoulou et al., 2010]. São analisados aproximadamente 37 milhões de vídeos, o que corresponde aproximadamente 25% da base do *YouTube*. A análise é feita de maneira compreensiva através de investigações das propriedades e padrões no tempo além de várias outras métricas de popularidade. As relações entre as métricas de popularidade são estudadas e conclui-se que quatro destas são altamente correlacionadas (número de visualizações, número de comentários, número

de qualificações[3], número de favoritos) enquanto que a quinta, média das qualificações, exibe muito pouca correlação com as outras métricas. Também é relatado um "número mágico" no comportamento mágico dos vídeos: a cada 400 visualizações que um vídeo recebe, tem-se uma de cada ação : escrita de um comentário, qualificação do vídeo e adição do mesmo aos favoritos. Os autores sugerem que outras análises a ser seguidas são : (a) a co-evolução das métricas de popularidade no tempo, ou seja, qual métrica cresce primeiro e qual métricas segue, (b) o comportamento dos usuários em termos de reação a um vídeos, isto é, se ele é o mesmo usuário que deixa um comentários e qualifica um vídeo, e (c) o efeito dos "standard feeds" (listas de vídeos fornecidas pelo *Youtube*) na popularidade de vídeos ao longo do tempo.

Em [Duarte et al., 2007a] foram analisadas as características dos vídeos e dos usuários de diferentes regiões geográficas. Focou-se em particular no estudo dos usuários localizados na América Latina. Os autores enfatizaram seus esforços na identificação de redes sociais formadas a partir de relacionamentos ocorridos no *YouTube*. Foram mostradas evidências de que a localidade dos usuários pode ser explorada para melhorar a infra-estrutura utilizada pelo provedor do serviço.

No trabalho [Maia & Virgilio Almeida, 2009] foram avaliados quatro sistemas para transmissão de fluxo de vídeo pré-armazenado, o *YouTube*, o *DailyMotion*, o *Veoh Networks* e o *Videolog* (sediado no Brasil). Foi proposta uma estratégia de *caching* que explora tanto o fato do vídeos serem recentes quanto a popularidade destes.

Em [Saxena et al., 2008] os autores apresentaram uma comparação entre o *YouTube*, o *Dailymotion* e o *Metacafe*. O foco das análises realizadas foi inferir sobre o sistema de transmissão de conteúdo utilizado por tais aplicações, como também identificar como as meta-informações associadas a um vídeo impactam no seu armazenamento e transmissão.

Os estudos discutidos nesta seção foram importantes fontes para definir as análises que propomos na metodologia. Contudo, não encontramos na literatura nenhum trabalho que abordasse o problema de caracterização de dados multimídia com o foco nas informações existentes nos conteúdos e objetos. Dessa forma, utilizamos como ponto de partida a estratégia hierárquica do trabalho [Veloso et al., 2006b] para modelar e organizar as informações relativas aos objetos e conteúdos multimídia. O resultado deste processo será apresentado no Capítulo 3.

[3] *ratings*

Capítulo 3

Metodologia

Os padrões de acesso e as cargas de trabalho dos servidores Web já foram extensamente estudados por vários autores na literatura [Almeida et al., 2001, Gill et al., 2007, Cha et al., 2007]. Entretanto, somente nos últimos anos, os padrões de acesso a conteúdos diversificados como mídias ricas de diferentes tipos (p. ex., vídeos, comércio eletrônico, redes sociais, etc.) vêm sendo mais explorados, até mesmo devido à popularização de serviços desse tipo.

Vários são os estudos que abordam as análises de acesso e tráfego sobre aspectos como tipos de objetos existentes (p. ex., vídeos, áudios, imagens), suas propriedades (p. ex., formato, tamanho, duração) além de outras análises. Nesta seção descrevemos uma nova metodologia de caracterização de serviços Web multimídia, organizada de forma hierárquica e composta por quatro partes: Requisição (R), Objeto (O), Conteúdo (C) e Conhecimento (K, do inglês *knowledge*), - ROCK. Essa organização estratificada possibilita caracterizar e analisar de forma mais metódica os **atributos** das requisições, as **propriedades** dos objetos, o conteúdo associado a esses objetos através de seus **metadados**, bem como explorar e investigar as **relações e padrões implícitos** que podem ser derivados de todas as informações associadas ao serviço multimídia.

Essa metodologia tem por **objetivo principal**, estabelecer uma abordagem incremental que possibilite a análise dos conteúdos multimídias em dois níveis. O primeiro nível é hierárquico e utiliza as informações existentes nos objetos e seus conteúdos de forma a complementar a análise geral das requisições realizadas pelos usuários. Já o segundo nível, complementar ao primeiro, é marcado pela aplicação de um processo mais complexo de análise de dados visando a extração do conhecimento das informações avaliadas na camadas anteriores. Tais conceitos e objetivos serão detalhados nas seções a seguir.

Através do emprego da metodologia ROCK pretendemos compreender melhor

a forma com que os recursos multimídia são consumidos atualmente nos canais digitais da Internet. Pretende-se obter melhores informações a respeito das interações dos usuários, objetos e servidores a fim de descrever a forma com que os conteúdos multimídias são acessados em tais servidores.

3.1 Conceitos Básicos

Como dito anteriormente temos duas hierarquias, a de acesso e a de informação. A primeira hierarquia nos diz **quem** e de **que forma** os acessos são realizados. A segunda nos diz **o que** e **qual conteúdo** está sendo acessado, além de adicionalmente possibilitar a extração de **conhecimento** (*padrões e relações*, dentre outras possibilidades) destes dados através de técnicas de inteligência computacional. Esta última hierarquia é a contribuição principal deste trabalho.

A interação com os servidores se inicia com as requisições dos usuários. Tais usuários podem acessar os conteúdos em diferentes períodos de tempo, sendo que cada período em que o usuário permanece ativamente consumindo tais conteúdos é conhecido como sessão. Dessa forma, o usuário pode possuir sessões distintas dependendo da forma que as requisições estão distribuídas temporalmente. A relação *requisições* $\rightarrow$ *sessões* $\rightarrow$ *usuários* forma a **hierarquia de acesso**. A Figura 3.1 ilustra esta relação mostrando, de forma esquemática, como as diferentes visões se agregam na hierarquia. A hierarquia de acesso tem como foco o usuário e suas requisições, não a informação presente nas requisições e como esta se organiza. Por esse motivo e por já ter sido exaustivamente estudada na literatura [Veloso et al., 2006b], essa relação não será detalhada nesse trabalho.

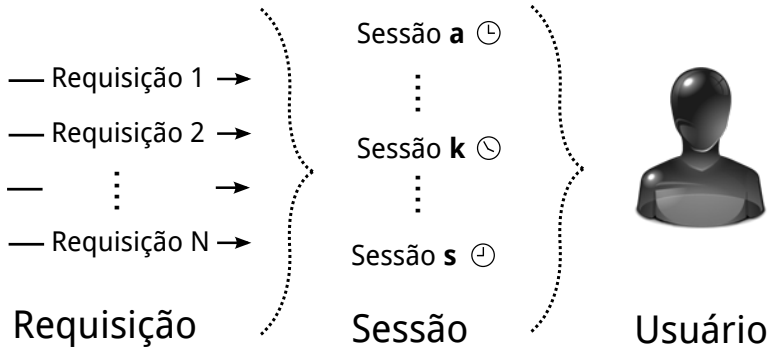


Figura 3.1. Ilustração de como as diferentes visões estão agregadas na *Hierarquia de Acesso*

As várias requisições que ocorrem estão distribuídas entre os diversos objetos existentes nos servidores. Tais objetos possuem tipos (vídeos, imagens, áudio, aplicativos, etc.) e propriedades diferentes (tamanho, duração, formato, etc.), além de estarem associados a algum conteúdo, como, por exemplo, uma determinada categoria (jornalismo, esportes ou entretenimento). Tal conteúdo pode ser identificado, dentre várias evidências, pelos metadados associados aos objetos. Por exemplo, título, descrição, *tags* e categorias ou até mesmo por outros metadados não textuais específicos de cada tipo de objeto. Além disso, após análise básica das informações anteriores é natural que se questione a respeito de correlações, associações dentre outras relações implícitas que podem existir nesses dados. Tais questionamentos podem ser respondidos por técnicas de análises de dados como algoritmos para descoberta de conhecimento em bases de dados (**KDD**), incluindo técnicas de mineração de dados que, se aplicados aos insumos das camadas anteriores, podem gerar um conhecimento efetivo.

A relação *requisição* → *objeto* → *conteúdo* constitui o que denominamos **hierarquia de informação** que é o primeiro nível da nossa metodologia. A aplicação dos processos de **KDD** nesta hierarquia implica na geração de conhecimento e constitui o segundo nível da metodologia. A representação esquemática da metodologia ROCK é ilustrada pela Figura 3.2.

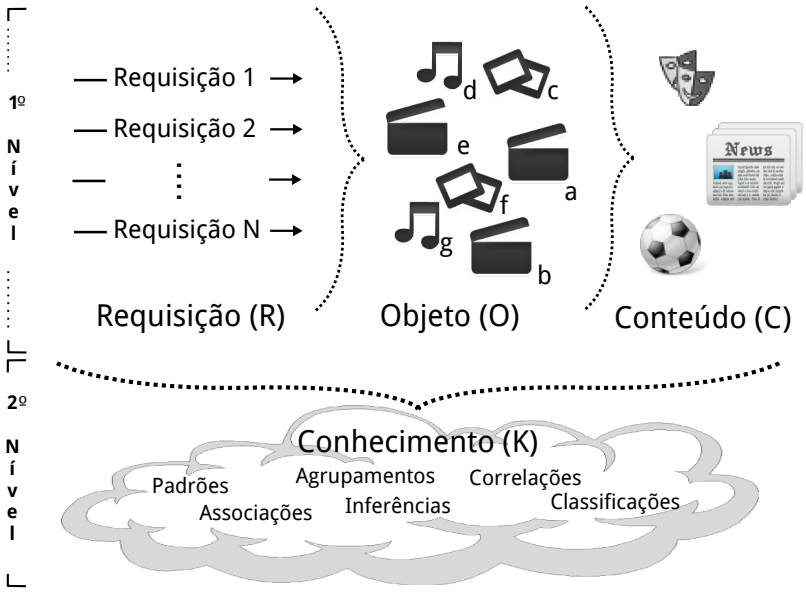


Figura 3.2. Hierarquia de informação - ROCK

A nomenclatura definida e adotada neste trabalho é explicada a seguir. Para as camadas da hierarquia da informação: **R** (*Requests*) para a camada de requisições, **O** (*Objects*) para a camada de objetos, **C** (*Content*) para a a camada de conteúdos

e **K** (*Knowledge*) para o processo de aplicação de técnicas de **KDD** às informações provindas da hierarquia da informação.

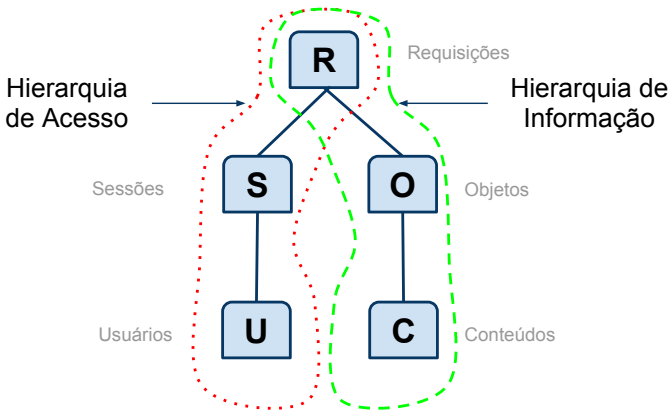


Figura 3.3. As duas hierarquias e o relacionamento de ambas.

Podemos ver a representação esquemática de ambas as hierarquias na Figura 3.3. Note que, em ambas hierarquias, a camada base é constituída das requisições realizadas ao servidor multimídia pelos usuários. É natural que a camada de requisições (R) seja comum a ambas hierarquias, pois esta é a forma de interação entre os servidores multimídia e seus usuários, sem as requisições não há interação entre tais agentes e não há o que se analisar.

3.1.1 Subdivisão da Hierarquia

Como podemos ver na Figura 3.2 existem dois níveis distintos na nossa metodologia. O primeiro, caracterizado pela hierarquia de informação, composta pelas camadas R,O e C, onde se realizam as análises básicas e um nível marcado por análises complementares que denominamos nível K.

Para cada uma das camadas do primeiro nível da hierarquia existem **unidades de análise** que são os dados associados a cada camada que fornecem informações para a caracterização e compreensão da respectiva camada. Cada camada possui um conjunto de dados que se associam primariamente a tal camada. Para a camada R são os atributos dos registros de acesso (*logs*), para a camada de objetos são as propriedades dos objetos e, finalmente, no caso dos conteúdos são os metadados associados aos objetos.

A hierarquia da informação e os relacionamentos entre suas entidades são descritos a seguir. As requisições se segmentam nos objetos que por sua vez estão associados aos conteúdos e, consequentemente, também se segmentam neste nível. Além disso, para

cada uma das camadas básicas existem unidades de análise inerentes a cada camada que possibilitam análises estatísticas básicas como histogramas, funções de distribuição, análises temporal dos dados, correlação direta entre as medidas, etc. Por exemplo, pode-se analisar o número de requisições por segundo que chega ao servidor de forma direta na Camada R ou a distribuição do *bitrate* dos diversos vídeos na Camada O.

Entretanto, a camada K de *conhecimento* não se agrega de forma linear à hierarquia composta por R,O e C. O nível K atua como uma componente ortogonal de análise que surge após o estudo das camadas básicas e pode se relacionar diretamente com qualquer uma das outras camadas (R,O ou C) de acordo com as avaliações que se pretende conduzir.

Para essa parte complementar não existem unidades de análise inerentes ou mesmo informações exclusivas que possam ser analisadas diretamente como no caso das outras camadas. Os insumos desse nível são os três grupos de informações básicas, atributos dos *logs*, propriedades dos objetos e metadados associados, que podem ser pré-processados, ou não, para compor as informações que alimentarão os algoritmos e técnicas de mineração de dados que revelarão os conhecimentos implícitos que caracterizam a Camada K.

Assim, a camada K atua como uma camada de extração de informações implícitas que permite análises mais complexas através de técnicas de extração de conhecimento em base de dados (**KDD**) [Fayyad et al., 1996] que normalmente são aplicados após um entendimento mais completo do cenário em estudo, que neste caso são as análises básicas.

Em síntese, o primeiro nível, composto das camadas R, O e C, possibilita um entendimento geral a partir das caracterizações elaboradas e suas análises, compondo a parte básica da hierarquia. Por sua vez, o nível K vem possibilitar um novo leque de análises que poderão enriquecer a caracterização geral agregando conhecimento às análises iniciais, sendo uma extensão complementar da hierarquia da informação.

Isso não quer dizer que outra abordagem que excluísse K e fizesse todas as análises conjuntas estaria errada, porém defende-se aqui que a metodologia proposta segue o processo natural de caracterização e avaliação do cenário em estudo e permite uma exploração mais adequada do processo de KDD após instanciação dos níveis R, O e C em determinado estudo de caso.

Nas próximas seções explicaremos cada parte da metodologia ROCK, suas respectivas *unidades de análises* e exemplificaremos algumas possíveis análises a serem realizadas em cada etapa.

3.2 (R) Requisições

Nesta Seção descrevemos a camada R de requisições. Esta é a camada de menor granulação e se concentra no fluxo de transferências individuais entre os usuários e servidores de conteúdo multimídia. Tal camada descreve de que forma os objetos são consumidos no sistema.

As **unidades de análise** desta camada são os atributos associados à representação da requisição. Geralmente uma requisição é representada por uma entrada nos registros de acesso dos servidores (*logs*). Tais entradas contêm os atributos que descrevem a requisição e que geralmente são: o horário de ocorrência, a informação de quantidade de Bytes trafegados e o tempo de *download* da informação, dentre outros atributos. O número de atributos e seus significados são variáveis de acordo com o servidor que registra *logs*. Exemplos de tais atributos podem ser vistos na Tabela 3.1.

Horário da requisição	Campo que representa o horário da requisição descrita pela entrada no *log*. Este campo está na grande maioria dos *logs* de acesso através de um *timestamp* representado os segundos desde o *epoch time* (00:00:00 UTC de 1o de Janeiro de 1970) ou uma representação mais legível como *29/Nov/2010 00:00:01*.
Duração da Requisição	Tempo que o servidor gasta escrevendo as informações no canal de transmissão para o cliente. Geralmente representado em milissegundos.
Método de Transferência	Método da requisição cliente-servidor. Representa a forma em que a interação entre o servidor e cliente foi realizada. Por exemplo: HTTP-GET, HTTP-POST, etc.
Código de Retorno	Código que representa o estado da resposta da requisição feita, pode indicar erro ou ações a serem tomadas.
Tamanho da Transferência	Este atributo representa o quanto, em bytes, foi trafegado entre o servidor e o cliente.
Identificador de Origem	Este atributo representa a origem do acesso. Pode ser, por exemplo, o endereço ip da máquina que originou a requisição ou um identificador de um usuário no sistema em estudo.
Identificador dos Agente de Origem	Atributo que representa a forma com que o acesso é feito. Geralmente é utilizado o *User Agent* da requisição HTTP se existente. Permite caracterizar a plataforma de acesso e o sistema utilizado.

Tabela 3.1. Exemplos de atributos que podem ser utilizados nas análises em **R**

3.2.1 Peculiaridades da Camada R

As análises da camada **R** permitem ver as informações do acesso com a menor granulação possível do sistema. O objetivo é estudar a carga de trabalho sobre os servidores

e avaliar como esta é distribuída temporalmente. Basicamente, uma análise pertence à camada R se ela não agrega informações do objeto requisitado e só utiliza unidades de análises de tal camada, que nos casos são os atributos dos *logs* de acesso.

Conforme dito anteriormente, na Seção 3.1, esta camada apresenta a peculiaridade de ser um componente comum à *hierarquia de acesso*. As análises são as mesmas, não existe distinção lógica para tal camada em nenhuma das duas hierarquias. Mas é importante ressaltar que tal camada é o elo de ligação entre os servidores e conteúdo, informação essencial para que se continue as análises nos resto da metodologia.

3.2.2 Análises na Camada R

A seguir mostraremos uma lista, não exaustiva, de análises que propomos para este nível de abstração. A forma como tais análises estão representadas são verbosas intencionalmente. O objetivo desta seção é servir como um guia ao tentar seguir esta metodologia para futuros estudos.

R_1: Taxa de Resposta (*Throughput*) do Servidor

Objetivo : Realizar um estudo da carga no servidor avaliando o número de requisições simultâneas e sua capacidade de resposta em diferentes escalas de tempo.

Exemplos de Representações :

- Um gráfico da distribuição cumulativa (cumulative distribuction function, *CDF*) e/ou sua complementar (complementary cumulative distribution function,*CCDF*) das requisições simultâneas do período total estudado.
- Um histograma do número de requisições simultâneas.
- Um gráfico da média do número de requisições simultâneas por hora de todo o período.
- Análise do comportamento períodico do número de requisições simultâneas. Sendo t o intervalo de análise avaliar $t \ mod \ p$ onde p é o período que deseja se analisar, como mensal, semanal, diário, etc ...
- Um gráfico da distribuição média das requisições simultâneas durante as horas do dia, para cada dia da semana.

R_2 : Tempo entre Chegadas de Requisições no Servidor

Objetivo : Através do tempo entre chegadas de transferências concorrentes IAT (Inter Arrival Time) no servidor iremos avaliar a variação da carga do servidor em diferentes intervalos em tempo. Essa análise possibilita o reconhecimento de períodos de alta e baixa atividade no servidor.

Exemplos de Representações :

- A distribuição dos IATs através da *CDF* e/ou *CCDF*.
- Um histograma dos IATs durante todo o período.
- A média dos IATs por diferentes períodos de tempo, mensal, semanal, diário, por dias de semana ou por finais de semana.

R_3 : Tempo de Resposta do Servidor

Objetivo : Avaliar o desempenho do servidor através da distribuição do tempos de respostas apara avaliar o impacto de tal variável nas características do tráfego gerado na rede e nos servidores.

Exemplos de Representações :

- A distribuição dos tempos de respostas através da *CDF* e/ou *CCDF*.
- Um histograma dos tempos de respostas durante todo o período.
- A média dos tempos de respostas por diferentes períodos de tempo, mensal, semanal, diário, por período da semana (dias de semana e *finais de semana*).

R_4 : Largura de Banda das Requisições

Objetivo : Análise da largura de banda utilizada pelas requisições. Nos permite avaliar a velocidade média em que os usuários se conectam aos servidores e as condições de rede durante as transferências.

Exemplos de Representações :

- A distribuição da largura de banda através da *CDF* e/ou *CCDF*.
- Um histograma da largura de banda durante todo o período.
- A média da largura de banda por diferentes períodos de tempo, mensal, semanal, diário, por período da semana (dias de semana e *finais de semana*).

R_5 : Tamanho das Transferências

Objetivo : Analisar a distribuição do tamanho da transferência de cada requisição. Compreendendo melhor qual o impacto desta variável no desempenho do servidor.

Exemplos de Representações :

- A distribuição do tamanho da transferência através da *CDF* e/ou *CCDF*.
- Um histograma do tamanho da transferência durante todo o período.
- A média do tamanho da transferência por diferentes períodos de tempo, mensal, semanal, diário, por período da semana (dias de semana e *finais de semana*).

R_6 : Análise dos Endereços de Origem das Requisições

Objetivo : Caracterizar as origens das requisições e suas respectivas popularidades. Tais informações podem ser obtidas através de ferramentas que mapeiam os *IPs* das requisições para regiões geográficas ou utilizando outra fonte como cadastro de usuários por exemplo.

Exemplos de Representações :

- Tabelas que apresentem análises comparativas com os acessos segmentados por países, regiões e cidades.
- Um histograma do número de acessos únicos por dia durante todo o período.
- Uma *CCDF* das visualizações por cidade, para avaliar a popularidade

R_7 : Análise dos Agentes de Origem das Requisições

Objetivo : Compreender de que forma os acessos são realizados avaliando os diferentes agentes de acesso (navegadores, robôs, scripts) e quais plataformas são utilizadas para gerar as requisições.

Exemplos de Representações :

- Um gráfico comparativo entre o acesso via plataformas móveis versus *desktops*
- Um gráfico comparativo entre o acesso através dos diferentes sistemas operacionais.
- Um gráfico comparativo entre o acesso através dos diferentes navegadores.

Ressaltamos que tais análises são apenas uma ilustração da potencialidade das análises para essa camada. A metodologia pode ser estendida de acordo com o estudo de caso. Por exemplo, análises conjugadas que relacionam duas unidades de análise são viáveis. Por exemplo, uma análise adicional poderia incluir a relação entre a largura de banda e o tempo de resposta.

Através das análises mostradas nesta seção é possível avaliar a forma de como os acessos estão distribuídos nos servidores multimídia e identificar análises que necessitam investigações mais detalhadas no resto da metodologia. A apresentamos a camada de objetos do primeiro nível de análise da nossa metodologia.

3.3 (O) Objetos

Os objetos são as entidades alvo das requisições. Em um serviço de distribuição de conteúdo multimídia tal entidade pode representar uma página *Web*, um arquivo texto, uma imagem, um *streaming* de vídeo, um vídeo armazenado, um aplicativo flash, dentre outras tipos. Neste nível de abstração, o foco das análises são as propriedades desses objetos, as características dos acessos segmentados por tipo de objeto e as diferenças nas análises que essas segmentações proporcionam.

Para que as análises agregadas da camada **R** sejam viáveis, é necessário que se possa identificar o objeto e seu tipo através das informações do acesso. Tal identificação irá variar de acordo com cada aplicação, por exemplo a URL sendo requisitada ou um identificador do objeto no registro de acesso.

As unidades de análise intrínsecas a esta camada são os objetos, os respectivos tipos e suas propriedades. Entretanto, como existem vários tipos de objetos que variam de acordo com a aplicação em avaliação é impossível prever todas propriedades envolvidas. Dessa forma, tentamos sumarizar de forma generalizada tais informações na Tabela 3.2.

Introduzimos nessa camada a ideia de uma **projeção**, que consiste basicamente em uma segmentação de análises de camadas anteriores utilizando informações/dados pertencentes à camada atual. Por exemplo, as análises de tempo de resposta realizadas em R_1 podem ser refinadas na camada O através da segmentação dos diferentes tipos de mídias (vídeo, áudio, imagem) traçando por exemplo curvas de *throughput* distintas para cada tipo de mídia. A relevância de algumas dessas análises pode ser questionada, porém acreditamos que essa prerrogativa será sempre um critério que deve ser avaliado pelo autor do estudo com base no conjunto de dados sendo analisado e nos objetivos que se almeja alcançar. Cabe a metodologia incluí-las e deixar ao estudo de caso adotá-

Identificador do Objeto	Campo que possui a informação que identifique o objeto. Em caso de acesso HTTP, a URL pode ser suficiente para identificar o objeto. O único requisito deste campo é a identificação única do objeto.
Identificador do Tipo do Objeto	Campo que identifica o tipo do objeto. Por exemplo, texto, imagem, áudio, vídeo, *stream*, etc.. Pode-se utilizar o campo *content-type* da requisição HTTP, fazer um *parsing* da url e inferir o tipo do objeto ou mesmo usar um identificador explícito dependendo da aplicação.
Descritor das Propriedades dos Objetos	Os objetos possuem propriedades físicas como tamanho, formato, codificação e podem possuir propriedades específicas de cada tipo de objeto. Por exemplo: *character encoding* para textos, resolução e tamanho para imagens, duração e *kbps* para áudios e vídeos e *codecs* e compressão para alguns vídeos. É necessário uma forma de acesso a tais propriedades ou extrair estas diretamente do vídeo ou de alguma base de informação a tais dados dos objetos.

Tabela 3.2. Informações complementares para as análises na camada **O**

las quando pertinente. Observe também que a primeira proposta de análise é sempre um sumário das informações a serem abordadas nessa camada.

3.3.1 Análises na Camada O

As análises que sugerimos neste nível de abstração são:

O_1 a O_7 : Projeções da camada R:

O_1: Taxa de Resposta (Throughput) por tipo de Objeto

O_2: Tempo entre Chegadas de Requisição por tipo de Objeto

O_3: Tempo de Resposta do Servidor por tipo de Objeto

O_4: Largura de Banda por tipo de Objeto

O_5: Tamanho das Transferências por tipo de Objeto

O_6: Análise dos Endereços de Origem por tipo de Objeto

O_7: Análise dos Agentes de Origem por tipo de Objeto

Objetivo : Segmentar as análises da camada anterior **R** pelos diferentes tipos de objetos caso as análises sejam pertinentes. Por exemplo uma análise em **R** pode ter sido inconclusiva.

Exemplos de Representações :

- As mesmas representações da camada **R**, só que segmentado por tipo de objeto

O_8 : Distribuição das Propriedades de objetos

Objetivo : Avaliar a popularidade e distribuição das propriedades dos Objetos pela coleção. Valer ressaltar que nesta visão, diferentes tipos de objetos podem demandar análises específicas aos diferentes objetos. Por exemplo, somente para vídeo podemos analisar a distribuição de *encoders* e *containers*.

Exemplos de Representações :

- Uma *CDF* de cada propriedade dos objetos.
- Um histograma das propriedades destes.

O_9 : Popularidade de objetos

Objetivo : Avaliar a variação da popularidade dos objetos ao longo do tempo, usando diferentes escalas de tempos (p. ex, hora, dia, semana, mês) de acordo com os registros de acesso disponíveis.

Exemplos de Representações :

- Uma *CCDF* dos objetos ordenados por popularidade versus a fração agregada das suas requisições.
- Uma análise de *Zipf*. Um gráfico em escala *log-log* do ranking dos vídeos (ordenados) versus sua popularidade.

O_{10} : Distribuição da idade de objetos

Objetivo : Avaliar o tempo de vida dos objetos multimídia utilizando a sua data de inserção e o tempo em que este permanece ativo (alvo de requisições), ou caso exista, análise do tempo de sua remoção na base. Também é interessante a análise da taxa de inserção dos objetos na base.

Exemplos de Representações :

- Um histograma por todo período mostrando a taxa de inserção de objetos por dia, e uma projeção do número médio de inserção de objetos por dia da semana.
- Distribuição do tempo de vida do objeto (número de dias de acesso)

O_{11} : Relação idade versus popularidade de objetos

Objetivo : Investigar a relação da popularidade dos objetos de acordo com seu tempo de vida.

Exemplos de Representações :

- Um gráfico que mostra o percentual das requisições dos objetos agrupadas por um certo período de tempo, dias por exemplo, para todos os objetos da base. Grupos de popularidade podem ser agrupados por cores por exemplo.
- Um gráfico me mostra a taxa de requisições dos objetos a partir do dia de publicação deste.
- Uma análise do número de visualizações por dia.

Mais uma vez chamamos a atenção ao fato das análises definidas aqui servirem apenas como uma ilustração das análises possíveis para esta camada. A metodologia foi desenvolvida para ser estendida de acordo com o estudo de caso.

As análises desta camada possibilitam a avaliação dos vários tipos de objetos existentes, suas propriedades e a forma como as requisições se segmentam entre estes objetos. Tais análises possibilitam uma melhor compreensão dos servidores multimídia. Os aspectos avaliados com as análises desta camada podem demandar outras investigações mais detalhadas no resto da metodologia. A seguir detalhamos a camada de conteúdo do primeiro nível da nossa metodologia e elencamos algumas análises possíveis tal camada.

3.4 (C) Conteúdo

Cada objeto, seja um vídeo, uma imagem, uma página *Web*, etc., representa algum tipo de informação que o *Produtor de Conteúdo* deseja passar para seu público alvo. Segundo o dicionário de língua portuguesa Michaelis, a palavra *conteúdo* pode ser definido como "*Assunto, tema, matéria de carta, livro etc.; teor, texto*". Com base nesta definição e no caráter informativo que todos objetos multimídia possuem, podemos dizer que todo objeto possui um conteúdo relacionado.

O que irá definir o conteúdo dependerá unicamente do objeto em questão. De fato, o conteúdo de um objeto multimídia é, em última instância, o próprio objeto. O que comumente se faz é generalizar o conteúdo do objeto de forma que este se enquadre em um conjunto similar de objetos que nos remetem a um conceito único. Por exemplo, uma reportagem a respeito dos problemas locais de uma cidade pode ser generalizado com um conteúdo jornalístico. Entretanto, o nível de detalhamento em que se pode definir tal generalização pode ser cada vez maior. De conteúdo jornalístico pode-se derivar uma hierarquia que se especializa cada vez mais até definir detalhadamente o teor do objeto.

A extração do conteúdo de um objeto multimídia diretamente de tal objeto pode não ser simples de ser obtida de forma automática. Desta forma, pode-se utilizar metadados para auxiliar nesta tarefa. Tais metadados podem ser textuais, como título, descrição, categoria, *tags* ou outros metadados não textuais. Por exemplo, para o caso de áudios musicais teríamos o timbre, ritmo, intensidade, pulsação (tempo) e brilho, já no caso de imagens a luminosidade, espectro de cores, brilho, contraste e formas.

Apesar dessa metodologia poder ser generalizada para qualquer tipo de metadados, ela foi primariamente *concebida com o foco em metadados textuais*. Assim, um vídeo de uma reportagem, por exemplo, idealmente irá possuir um título e descrição que sumarizam o seu conteúdo de natureza jornalística, podendo ainda sugerir que ele pertença a uma subcategoria específica, como economia ou política.

O objetivo da camada de conteúdo é permitir análises semânticas dos objetos, definindo como informação relevante o conteúdo dos objetos. Com isso, as informações relativas a forma como tais objetos são representados ou entregues tornam-se fonte de dados secundárias. Em nossa metodologia, definimos que esse conjunto de dados descritivos e complementares às propriedades de cada objeto será chamado de metadados.

Vale observar que a medida que mudamos de camada na hierarquia, o conjunto de análises possíveis cresce exponencialmente, pois o refinamento contínuo da granularidade dos objetos permite que uma requisição possa ser vista, por exemplo, como um vídeo, uma imagem ou um *streaming* de vídeo. Um vídeo, por sua vez, pode ser visto como uma reportagem, um clipe musical ou um evento esportivo. Por outro lado, essa maior especificidade pode tornar a metodologia restritiva, tornando a definição de atributos e propriedades relevantes em camadas mais baixas uma tarefa mais difícil, pois tais informações são frequentemente dependentes dos dados sendo processados e da aplicação que os gerencia.

Ao contrário de como foi feito na camada **R** e **O**, não tentaremos generalizar os insumos possíveis a serem analisados. Dessa forma, sugerimos que ao se instanciar esta camada é importante que se defina detalhadamente quais os metadados que servem de insumo para o estudo. Como dissemos anteriormente, esta metodologia foi concebida com o foco em metadados textuais, assim o conjunto de metadados básicos a serem considerados foram definidos com base na forma como, comumente, os objetos multimídia são apresentados na *Web* atual. Tais metadados podem ser vistos na Tabela 3.3.

Ressaltamos que **a metodologia não restringe sua aplicação aos campos citados**. Campos adicionais devem ser adotados, assim como campos definidos podem ser ignorados. Por exemplo, podemos ter propriedades dos vídeos relacionadas ao conteúdo que não sejam textuais. Espectro de cor, padrão de imagens e cenas, objetos identificados nos vídeos, características e propriedades do áudio dentre outras que

Metadado	Descrição
Título	Descrição em poucas palavras sobre o que se trata o conteúdo do objeto.
Descrição	Um sumário a respeito do objeto. Pode conter uma descrição detalhada do conteúdo podendo ter tamanho relativamente grande.
Tags	Conjunto de termos que representam temas associados com o conteúdo do objeto.
Categoria	Grupo ao qual o objeto pertence por apresentar características comuns aos outros integrantes.

Tabela 3.3. Informações - camada **C**

podem ser utilizadas nessa camada para gerar outras análises.

Entretanto, é importante que exista algum conceito que segmente os conteúdos como, por exemplo, categoria, pois a ideia de segmentação por conteúdo utilizada nos itens dessa camada é intimamente assocada com a ideia de segmentação por categoria.

Devido a flexibilidade de uso dessa camada, é interessante fornecer uma visão inicial dos metadados disponíveis a fim de guiar as análises subsequentes. O conceito que define o conceito de conteúdo e que segmenta os dados como, por exemplo, categorias deve ser devidamente definido e explicado para evitar dúvidas, confusões, ambiguidades e estabelecer as diretrizes que definem a instanciação do estudo de caso. Cada metadado relevante deve ser enumerado e seu domínio explicitado. Para campos textuais, como título e descrição, o procedimento de extração dos termos desses campos deve ser descrito, mencionando se qualquer tipo de processamento adicional foi utilizado, como remoção de *stopwords* (palavras repetitivas com pouco valor informacional, ex.: "a, as, de, dos, que, um", etc) conversão dos termos para seus radicais (*stemming*, por exemplo, interessante, interesses → interess), etc.

3.4.1 Análises na Camada C

A seguir sugerimos algumas análises passíveis de realização nesse camada. Mas ressaltamos que esta não é uma lista exaustiva. Caso seja conveniente esta lista deve ser estendida ou ignorada, o importante é adaptá-la aos interesses e objetivos do estudo que será conduzido.

C_1 a C_{13} : Projeções da Camada O

C_1: Taxa de Resposta (Throughput) por Conteúdo

C_2: Tempo entre Chegadas de Requisição por Conteúdo

C_3: Tempo de Resposta do Servidor por Conteúdo
C_4: Largura de Banda por Conteúdo
C_5: Tamanho das Transferências por Conteúdo
C_6: Análise dos Endereços de Origem por Conteúdo
C_7: Análise dos Agentes de Origem por Conteúdo
C_8: Distribuição das Propriedades dos Objetos por Conteúdo
C_9: Popularidade dos Conteúdos
C_{10}: Distribuição da idade dos objetos por Conteúdo
C_{11}: Relação Idade do Objeto versus Popularidade do Conteúdo

Objetivo : segmentação das análises anteriores utilizando os valores dos metadados dessa camada. Por exemplo, C_9 pode ser representado como múltiplas curvas de popularidade ao longo do tempo de vida de uma mídia onde cada curva representa uma categoria de conteúdo, como Esportes, Notícias, etc.

Exemplos de Representações :

- Se aplicam as mesmas representações das camadas anteriores.(**R** e **O**).

C_{12} : Popularidade de metadados

Objetivo : Avaliar a popularidade e distribuição dos metadados dos objetos.

Exemplos de Representações :

- *CDF* e/ou *CCDF* da popularidade dos termos por tipo de metadado
- Nuvem de termos para os metadados. (Wordle)

Nesta seção introduzimos os conceito relativos a camada de conteúdo. Explicamos a importância da definição do conceito conteúdo no estudo de caso. Explicamos que esta camada foi originalmente proposta com base em metadados textuais e possui uma alta afinidade com o conceito de segmentação de objetos em categorias. Com estas análises terminamos o primeiro nível da nossa metodologia fechando a hierarquia da informações. Na seção seguinte explicaremos o segundo nível de análises que é caracterizado pelo processo de **KDD** aplicado às unidades de análises da hierarquia da informação.

3.5 (K) Conhecimento

Até aqui, a metodologia apresentou análises diretas que podem ser realizadas nos vários níveis de acordo com o foco de avaliação de cada um deles. Os registros obtidos de

aplicações, muitas vezes, representam apenas dados e não conhecimento. Este é o caso dos *logs* de servidores que apenas registram as informações das requisições a medida que estas são realizadas. Visando transformar estes dados em conhecimento, surge o processo chamado de Descoberta de Conhecimento em Bancos de Dados (Knowledge Discovery in Databases - **KDD**), que [Fayyad et al., 1996] definem como sendo *"o processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados"*. Tal processo tem como objetivo aproveitar o potencial de análises mais complexas que podem ser desenvolvidas com base nas informações disponíveis a partir do uso de técnicas de inteligência computacional [Han, 2005, Witten & Frank, 2005], tais como técnicas de classificação e predição, agrupamento (*clustering*), regras de associação, inteligência artificial e aprendizado de máquina, dentre outras.

O processo de **KDD** é o conjunto de atividades contínuas que compartilham o conhecimento descoberto a partir de bases de dados. O **KDD** é composto pelas etapas de seleção de dados, pré-processamento e limpeza, transformação, Mineração de Dados e interpretação dos resultados. Esse é um tópico recente em ciência da computação mas utiliza várias técnicas da estatística, recuperação de informação,inteligência artificial e reconhecimento de padrões.

O objetivo deste do nível **K** é usar processo de descoberta de conhecimento (**KDD**) para obter padrões específicos das unidades de análise com o objetivo de obter conhecimento, que podem ser empregados em diferentes aplicações dependendo do cenário em estudo (p. ex.: personalização, recomendação, previsão etc.).

Apesar de não existir *unidades de análise* que sirvam de insumos para tais avaliações, consideramos ser fundamental descrever este nível da metodologia como uma extensão da hierarquia que complementa as análises anteriormente realizadas. Assim, o objetivo deste nível na metodologia é possibilitar a identificação de padrões que podem ser obtidos a partir das informações das camadas anteriores e outras que podem ser extraídas da interação com o usuário e seu comportamento. Esses padrões possibilitam identificar **conhecimentos** que podem contribuir para enriquecer a caracterização e servir de base para muitos serviços de valor agregado, como personalização de serviços, recomendação de conteúdo multimídia, novas estratégias de publicidade digital, etc.

Portanto, pode-se observar que esta camada é a mais ampla da metodologia em potencial, uma vez que existe uma gama de técnicas que podem ser aplicadas de diferentes formas às informações. Apesar de não ser possível listar precisamente o potencial de análises a serem desenvolvidas, indicamos a seguir algumas análises e procedimentos que consideramos bem aplicáveis à diferentes contextos de aplicações *Web*, em especial a serviços online de distribuição de conteúdo multimídia:

K_1 : Agrupamento de objetos

Objetivo :Identificar objetos com propriedades afins para propor uma melhor organização da infraestrutura de distribuição; Identificar propriedade dos objetos que são pouco requisitados para tentar melhorar sua popularidade.

K_2 : Agrupamento de conteúdos

Objetivo : Identificar conteúdos afins para recomendação a usuários, propor uma categorização dos mesmos ou sugerir mudanças na forma como tais conteúdos são exibidos ao usuário nos aplicativos dos *Produtores de Conteúdo*.

K_3 : Classificação de metadados

Objetivo : utilizar técnicas de classificação para inferir classes de conteúdos, que pode ser usado para sugerir adição de metadados aos provedores de conteúdo, como sugestão de *tags*, categorizar os conteúdos em classes mais afins e recomendação a usuários.

K_4 : Regras de associação:

Objetivo : identificar associações no uso ou consumo de conteúdos para inferir perfil de uso dos usuários, o que pode ser utilizado em recomendação de conteúdo, personalização do serviço ou mesmo para uso em campanhas publicitárias online.

K_5 : Avaliação da credibilidade do conteúdo:

Objetivo : modelar a credibilidade do conteúdo para classificação (*ranking*) do conteúdo a partir de informações como votos (*ratings*) associados ao conteúdo, comentários e sugestões dos usuários.

De uma maneira geral é possível aplicar várias técnicas de descoberta de conhecimento em banco de dados nos vários insumos da hierarquia de informação. Entretanto, predizer todas as aplicações e abordagens não é tarefa simples, é preciso muita experimentação. Contudo, acreditamos que a aplicação de tais técnicas podem melhorar a compreensão de tais meios de distribuição e sugerir melhorias em todo processo.

A metodologia aqui proposta foi aplicada as informações disponíveis da plataforma de vídeos onlines e o resultado está apresentado no capítulo 5.

Capítulo 4

Arcabouço Experimental

Estamos vivendo a era da informação e a produção desta cresce diariamente. A quantidade de dados gerados pelas transações eletrônicas, e-mail, blogs rede sociais é difícil de se mensurar mas estimativa-se que neste ano (2011) atingiremos algo em torno de 1.8 bilhão de terabytes [Gantz, 2008]. Um estudo feito por uma empresa que monitora serviços online [Pingdom, 2010] fornece os seguintes números de 2010 :

- 107 trilhões de e-mails enviado na Internet em 2010. (89% são spam);
- 2.9 bilhões de contas de e-mails no mundo todo;
- 255 milhões de Websites em Dezembro de 2010;
- 152 milhões de blogs;
- 175 milhões de pessoas no Twitter em setembro 2010;
- 600 milhões de pessoas no Facebook;
- 2 bilhões vídeos assistidos por dia no YouTube;

Quando o volume de dados a ser analisado é tão grande que os requisitos de tempo, armazenamento e capacidade computacional não podem ser satisfeitos por um único servidor, o processamento dessa massa de dados torna-se muito complexa que mesmo computações simples podem se tornar extremamente difíceis de serem realizadas num ambiente distribuído entre vários servidores. Conforme veremos no Capítulo 5 o volume de informações do presente trabalho se enquadram nessas condições e para a viabilização das análises realizadas aqui configuramos um arcabouço experimental distribuído capaz de processar terabytes de dados com eficiência.

Neste capítulo apresentamos arcabouço experimental que criamos para abordar o processamento do grande volume de dados do presente trabalho e sobre as tecnologias por trás desse arcabouço. Inicialmente, na Seção 4.1 falaremos do *MapReduce* um modelo de programação desenvolvido no Google que aborda o problema de processamento de grandes massas de dados, em seguida, na Seção 4.2, apresentamos o Apache Hadoop (e projetos relacionados), que se tornou a implementação de código aberto mais famosa do modelo desenvolvido no Google. E finalmente na Seção 4.3 mostramos o arcabouço experimental utilizado para auxiliar no desenvolvimento do presente trabalho.

4.1 MapReduce

MapReduce é um modelo de programação acompanhado de uma respectiva implementação que surgiu do trabalho de Jeffrey Dean e Sanjay Ghemawat [Dean et al., 2004] no Google. O objetivo é facilitar o processamento de grandes massas de dados distribuindo transparentemente a computação através de vários servidores e cuidar para que tudo ocorra da melhor forma possível abstraindo detalhes complexos como o de tolerância a falhas, distribuição de dados através da rede e balanceamento de carga

Antes do framework existir, por mais simples que fosse a computação a ser realizada, esta era paralelizada de forma *ad-hoc* para que pudesse ser aplicadas a grandes volumes de dados e pudesse ser executada em um cluster com várias máquinas. Isso aumentava a complexidade dos programas que deveriam tratar problemas como a distribuição dos dados e da computação, tratamento de falhas, comunicação entre os processos dentre outros.

O conceito por trás do método é o mesmo das funções de *map/reduce* das linguagens funcionais. Assim a computação é expresso através de duas funções: *Map* e *Reduce*. Ambas recebem como entrada um conjunto de pares *chave/valor* CV_e e produzem como resultado outro conjunto também de *chave/valor* CV_s . A função *Map* recebe o pares de entrada e produz um conjunto intermediário de *chave/valor* CV_i. O framework *MapReduce* agrupa todas os valores pertencentes a mesma chave e os passa a etapa de *Reduce*. Esta, por sua vez, agrega todos os valores que foram fornecidos como entrada e realiza a computação final.

A tipagem de dados entre as funções é a seguinte:

- **map** : $(k1, v1) \rightarrow list(k2, v2)$
- **reduce**: $(k2, list(v2)) \rightarrow list(v2)$

Ou seja, a etapa de *Map* trabalha com chaves e valores arbitrários definidos pelo usuário que pode ser de um domínio diferente do conjunto chave/valor de saída. Entretanto o domínio das chaves e valores intermediários são o mesmo do conjunto de saída.

4.1.1 Resumo da Execução

A execução de uma computação, que chamaremos daqui para frente de *Job*, é distribuída automaticamente através de um particionamento automático dos dados de entrada em um conjunto de M pedaços (*splits*). Estes pedaços podem ser processados em paralelos em diferentes máquinas e a cada um é atribuído um processo *Mapper*. O número de processos de *Reducer* e a instanciação destes são distribuídos pelo cluster de máquinas através de uma função de particionamento que pode ser fornecida pelo usuário, por exemplo $hash(chave)\ mod\ R$.

A Figura 4.1, retirada de [Dean et al., 2004], descreve uma visão geral do fluxo de dados e computação em um *Job* MapReduce. Os números da lista abaixo corresponde a aos números da Figura 4.1.

1. A biblioteca divide a entrada em **M** *splits* e inicia o processamento instanciando o mesmo número de Mappers.

2. Uma instância mestre é criada e é responsável por assinalar os processos **M** *mappers* e **R** *reducers* aos trabalhadores ociosos.

3. Cada *mapper* lê seu conjunto de dados, faz o parser das chaves/valores, passa à função *mapper* que produz os pares chave/valor intermediário e passa a biblioteca que as armazena temporariamente em memória.

4. Periodicamente os dados em memória são escritos em disco e particionados em R regiões pela função de partição. Estes arquivos intermediários particionados serão manipulados pelo processo mestre que entregarão ao *Reducers*.

5. Quando um *Reducer* é notificado pelo processo mestre ele começa a ler remotamente os dados de sua partição e então começa a ordená-los para que todas as ocorrências da mesma chave sejam agrupadas juntas.

6. Finalmente o *Reducer* começa a iterar sobre os valores ordenados passando-os para a função de *reduce*. O resultado é adicionado a um arquivo de saída.

Para que toda a orquestração de tarefas como instanciação de trabalhadores (*Mappers* e *Reducers*), gestão de arquivos intermediários, progresso das computações,

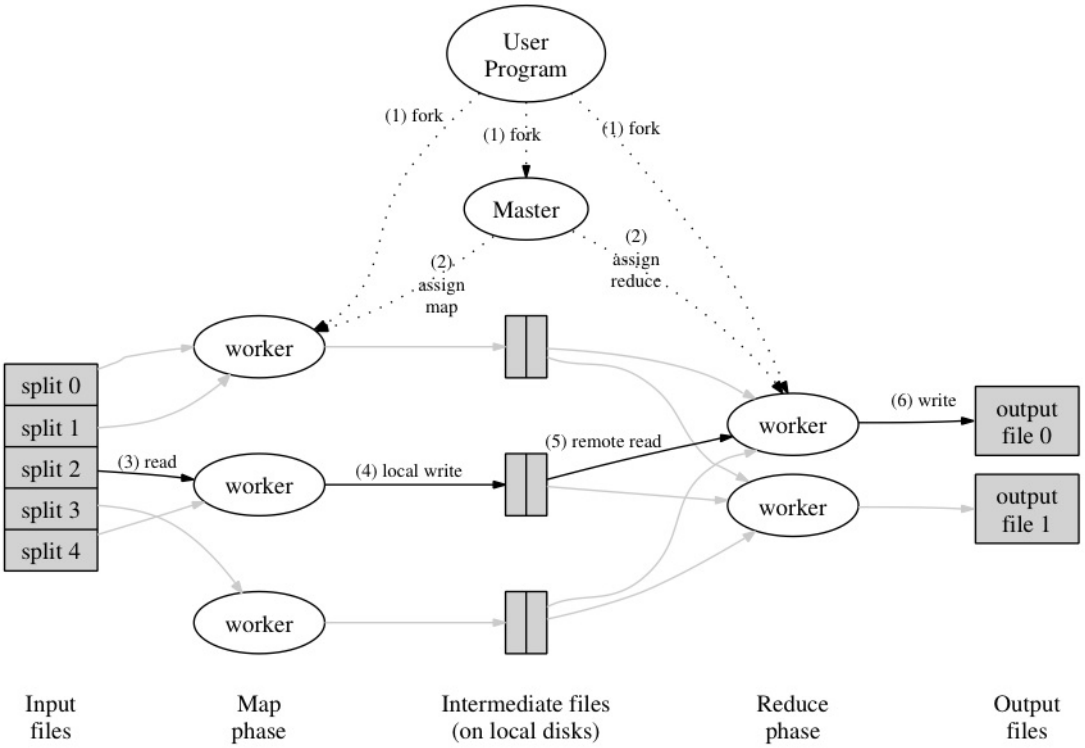


Figura 4.1. Fluxo de execução do modelo de *MapReduce*

etc, o processo Mestre armazena várias estruturas de dados. Por exemplo, para cada tarefa ele armazena o estado da tarefa (idle, in-progress ou completa) e a identificação da máquina trabalhadora. Além disso outro papel importante da instância Mestre é que ela atua como o conduíte através do qual os dados intermediários são propagados através do cluster.

4.1.2 Propriedades

Uma vez que o framework de MapReduce foi desenvolvido para ajudar no processamento de petabytes de dados, utilizando centenas a milhares de servidores [Dean & Ghemawat, 2008] a biblioteca que a implementar precisa tratar falhas de forma transparente.

MapReduce trata tolerância a falhas por reexecução. O processo mestre da um ping a cada trabalhador periodicamente, caso algum não responda, ele marca o seu estado como falho. Qualquer tarefa da fase *map* que tenha sido completada pelo trabalhador falho são ignoradas se tornando aptas ao reescalonamento em outros trabalhadores. Tarefas *map* são reexecutadas mesmo que seu trabalho tenha sido terminado pois os dados ficam armazenados localmente o que pode impedir o acesso aos dados

pelos *Reducers*. O mesmo ocorre para um *reducer*, como uma diferença que para o *reducer*, se este já estiver concluído seu trabalho, os dados estão armazenados num sistema de arquivos global que falaremos adiante na Seção 4.2.

Um ponto interessante sobre a biblioteca é que existe uma execução de tarefas backups. Os autores notaram que um dos principais causadores de lentidão em uma operação de MapReduce pode ser os trabalhadores *Retardatários*, que seriam servidores que gastam um tempo muito grande para terminar uma das tarefas finais do *Job*. Esses retardatário podem ser causados por diversos fatores, por exemplo, problemas nos discos tornando a operação de *IO* mais lenta, outros processos rodando nos servidores, o que aumenta a competição por recursos da máquina, ou até mesmo algum bug inesperado (os autores relataram um bug na inicialização de seus servidores que desabilitavam o cache do processador gerando uma lentidão no processamento de um fator de 100 vezes o normal).

Quando uma operação de MapReduce está próxima de completar, o framework coloca os processos pendentes para serem executados por máquinas ociosas visando evitar problemas de gargalos de máquinas. Sendo assim, se alguma máquina apresentar algum tipo de lentidão causada por qualquer que seja a origem, ela não irá impactar tão profundamente no tempo de execução total como poderia. Isso também aproveita a capacidade computacional do cluster uma vez que o tempo de ociosidade do cluster diminui como um todo.

Outro aspecto importante que precisa ser citado é a disponibilidade de dados e como o MapReduce a trata. Como o ambiente de execução é todo distribuído, os dados precisam estar disponíveis para todos os o trabalhadores caso estes precisem, além disso uma vez que um *reducer* finalizou seu trabalho, o resultado precisa ser gravado de forma definitiva e confiável. Estas propriedades são conseguidas através de um sistema de arquivos distribuído, no caso da Google, o *Google File System* (GFS).

O GFS é um sistema de arquivo distribuído que foi desenvolvido inicialmente para as necessidades da máquina de busca da empresa. Arquivos são armazenados em *chunks* (pedaços) de 64MB e raramente são sobrescritos ou diminuem de tamanho, a casos de usos mais comuns são leitura ou a concatenação de novas informações. A confiabilidade do GFS é obtida através de replicação. Cada *chunk* são replicados em pelo menos 3 servidores. Os metadados de cada arquivo são armazenados em um servidor mestre que também faz a coordenação de acesso.

No trabalho de [Dean & Ghemawat, 2008] é apresentado dois exemplos grep / sort, mostrando os ganhos de performance e é falado da experiência do emprego de MapReduce nos programas cotidianos da Google e como este modelo de programação transformou a empresa.

4.2 Apache Hadoop

Com o sucesso do trabalho de [Dean et al., 2004] Doug Cutting deu início em 2004 [White, 2009] ao desenvolvimento do Projeto Hadoop quando este estava trabalhando no desenvolvimento de uma máquina de busca de código aberto. O trabalho foi espelhado no trabalho de Dean e acabou virando uma implementação de código aberto do *MapReduce*. Em 2008 este projeto transformou-se em um projeto de primeira linha da Apache passando a ser conhecido como Apache Hadoop [Apache Foundation, 2011].

O Hadoop é subdividido em duas grandes abstrações: (1) um sistema de arquivos distribuído conhecido como *Hadoop File System* (HDFS); e (2) arcabouço de programação em MapReduce para o processamento de grandes bases de dados.

O *Hadoop File System* [Shvachko et al., 2010] foi desenvolvido para aplicações de processamento em lote que necessitam acesso contínuo a arquivos muito grandes que podem estar distribuídos em várias máquinas. O HDFS possui quatro diretivas básicas que influenciaram bastante em seu *design* :

1. Mover a computação é mais barato que mover os dados. Assume-se que não existe separação entre nós de armazenamento e nós de processamento

2. Flexibilidade na semântica do acesso de arquivos. O HDFS sacrifica a algumas consistências definidas pela API POSIX em favor de melhor desempenho. Por exemplo, não existe mecanismo de trava (lock) para controle de concorrência.

3. Grandes arquivos com permissão leitura somente. A configuração do HDFS é otimizada para grandes leituras sequenciais de arquivos.

4. Tratar falhas de hardware. Tratamento de erros no HDFS é realizado através de checksums e replicação de dados.

No HDFS existem dois tipos básicos de servidor, o *namenode* que controla os metadados dos sistema de arquivo e os *datadones* que armazenam os blocos de dados. Um cluster de de HDFS é composto por um *namenode* e vários *datanodes*. As operações de gerência do sistema (alocação de bloco de armazenamento, manipulação de arquivos, etc.) são controladas pelo *namenode* enquanto o *datanode* faz a gestão dos seus discos através do *namenode* mestre, e serve as requisições de leitura e escrita de arquivos dos clientes. Dessa forma o *namenode* é um ponto único de falha, entretanto não é um gargalo.

Sobre o HDFS trabalha o segundo grande componente do Hadoop é o arcabouço de programação *MapReduce* que segue os mesmos princípios de arquitetura e execução

do trabalho inicial de [Dean et al., 2004]. A implementação é em Java e provê uma API para que se crie programas nesta mesma linguagem. Este módulo consiste basicamente de um *JobTracker*, um gestor de trabalhos e tarefas ao qual aplicações clientes submetem os *Jobs MapReduce*, e um *TaskTracker* que é um *deamon* responsável por executar o *Job* sobre um volume de dados especificado pelo *Job*. O *JobTracker* envia tarefas para os nós *TaskTrackers* distribuídos no cluster, utilizando heurísticas para evitar que dados sejam migrados entre nós no cluster evitando sobrecarregar a rede. Caso ocorra alguma falha em alguma tarefa local este trabalho é reescalonado para outro *TaskTracker* ativo. O escalonamento de *Jobs* através do cluster é feito da forma mais simples possível utilizando um algoritmo *FIFO* para decidir a precedência de cada trabalho.

A Figura 4.2, também retirada de [Dean et al., 2004], mostra como é distribuído arquiteturalmente os componentes do framework Hadoop em um cluster. Existirá sempre um servidor mestre, que hospedará o *namenode* e o *jobtracker*, e todas as demais máquinas do cluster (o que também pode incluir o servidor mestre) terão apenas o *datanode* e o *tasktracker*.

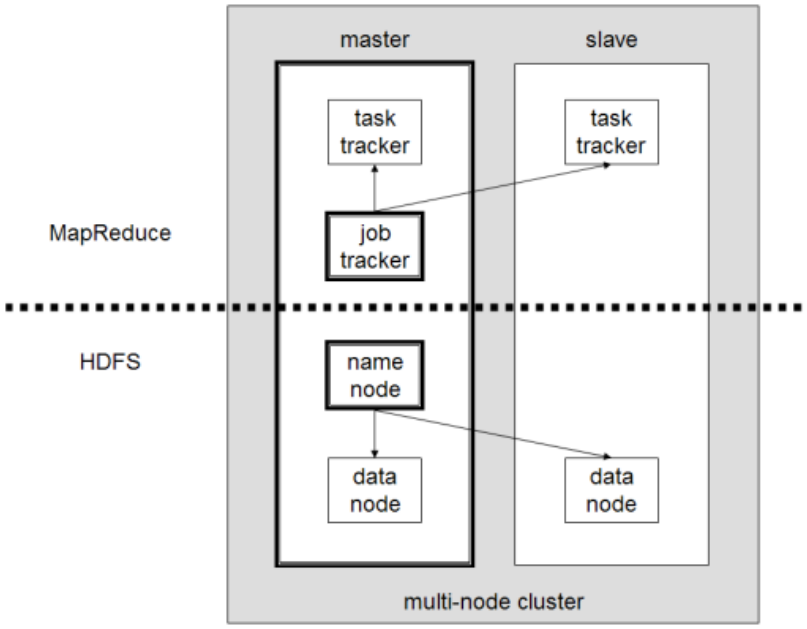


Figura 4.2. Esquema da arquitetura de um cluster *Hadoop*

4.2.1 Projetos Relacionados

Embora o Hadoop seja comumente conhecido pela sua implementação do MapReduce e pelo HDFS, o termo também é muito utilizado por uma família de projetos rela-

cionados que utilizam os fundamentos da infraestrutura para computação distribuída e processamento de dados em larga escala. A maior parte destes projetos fazem parte da *Apache Software Foundation* que provê suporte para a comunidade de projetos de código aberto. A medida que o ecossistema Hadoop cresce mais projetos tendem a aparecer para prover projetos complementares ou para adicionar abstrações de alto nível. Alguns dos projetos serão brevemente introduzidos aqui para que se conheça parte desse ecossistema que vem cada dia ganhando mais força e adoção pela comunidade que precisa desenvolver plataformas de forma distribuída.

Avro (`http://avro.apache.org/`)

Apache Avro é um sistema de serialização que provê estrutura de dados ricas, um formato binário de dados compacto e rápido, chamada de procedimentos remotas (RPC), um contêiner para armazenamento persistente e uma integração com várias linguagens, inclusive dinâmicas.

Pig (`http://pig.apache.org/`)

Pig é uma plataforma para análise de grandes bases de dados com uma linguagem de fluxo de dados e um ambiente de execução que permite a exploração de bases de dados muito grandes. Pig roda em cima de clusters Hadoop e provê uma linguagem própria (Pig Latin) simples, intuitiva e extensível. Falaremos um pouco mais do Pig na Subseção 4.2.2.

Hive (`http://hive.apache.org/`)

Um armazém de dados distribuído que gerencia dados armazenados no HDFS e provê uma linguagem de query baseada em SQL chamada HiveQL que é traduzida em tempo de execução em tarefas *MapReduce*. Adicionalmente é possível estender tal linguagem utilizando *mappers* e *reducers* personalizados.

HBase (`http://hbase.apache.org/`)

A distributed, column-oriented database. HBase uses HDFS for its underlying storage, and supports both batch-style computations using MapReduce and point queries (random reads).

Apache Mahout (http://mahout.apache.org/)

Uma biblioteca para algoritmos de Aprendizado de Máquina e Mineração de dados que escala para base de dados grandes.[1]. Possui algoritmos de agrupamento, classificação, regras de associação dentre outros, todos implementados utilizando Hadoop.

4.2.2 Pig

Apesar de ter se mostrado eficiente, o Hadoop necessita que o desenvolvedor aprenda a programar em um novo paradigma, quebrando suas tarefas em vários fluxos de execuções de tarefas *map* e *reduce*. Além disso, é necessário dominar uma linguagem de programação, tipicamente Java, o que proíbe que pessoas sem conhecimentos de desenvolvimento possam se beneficiar de um ambiente tão escalável quanto o proporcionado pelo Hadoop. Pensando nessas limitações o trabalho [Olston et al., 2008] propõe o Pig, um arcabouço de execuções em cima do Hadoop que possibilita pessoas que não entendam de programação possam realizar algumas análises em cima de grandes volumes de dados.

Pig provê um novo nível de abstração para o processamento de grandes bases de dados. Apesar do MapReduce prover uma interface programática, é preciso que você modele suas análises no paradigma o que pode ser desafiador e pode requerer vários estágios de map/reduce. Com o Pig as estruturas de dados são muito mais ricas (tipicamente sendo multivaloradas e aninhadas) e o conjunto de transformações que podem ser aplicadas aos dados são muito mais ricas e poderosas incluindo uniões (joins) o que não é uma tarefa simples em mapreduce.

Pig é composto de duas partes :

- A linguagem para expressar as computações, chamada de Pig Latin.
- O ambiente de execução para rodar os programas em Pig Latin. Este ambiente pode ser local, ideal para pequenos testes e validações, ou distribuido num cluster Hadoop.

Um programa em Pig Latin é composto de uma série de operação, ou transformações, que são aplicadas a um conjunto de dados de entrada para produzir um conjunto de dados de saída. Vista como um todo, as operações descrevem um fluxo de dados no qual o ambiente de execução do Pig traduz em uma representação executável e só então esta é de fato executada. Em baixo nível, o que o Pig faz é converter as transformações numa série de *Jobs* MapReduce, entretanto, o programador não chega

[1]Segundo a documentação *reasonably large datasets*

a tomar conhecimento desse processo, o que faz com que este possa focar nos dados ao invés da natureza da execução.

Pig é uma linguagem de script para a exploração de grande conjuntos de dados. Um criticismo do MapReduce é que o ciclo de desenvolvimento é muito longo. A escrita de *mappers* e *reducers*, o empacotamento do código, a submissão dos *Jobs* e a recuperação dos resultados é uma tarefa demorada. Com o Pig é possível processar terabytes de dados simplesmente inserido algumas poucas linhas através do terminal *Grunt*. De fato, Pig foi desenvolvida no Yahoo! para facilitar a vida de pesquisadores e engenheiros a minerar os enormes conjunto de dados lá.

A ferramenta foi desenvolvida para ser extensível. Virtualmente todas as partes do fluxo de dados são customizáveis: o carregamento e armazenamento dos dados, as filtragens, agrupamentos e junções podem todas serem alteradas por funções definidas pelos usuários. UDFs (User Defined Functions). Outro benefício é que tais UDFs tendem a ser mais reusáveis do que bibliotecas desenvolvidas para escrever programas MapReduce.

Entretanto, Pig não é adequada para qualquer tipo de processamento de dados. Como o MapReduce, ele foi desenvolvido para processamento de dados em lote. Se a tarefa a ser executada somente utiliza uma pequena porção de dados em um grande conjunto de dados, então Pig não será tão eficiente uma vez que este é configurado para ler todo o conjunto de dados, ou pelo menos a maior porção deste. Em alguns casos os programas em Pig não possuem um desempenho tão satisfatório quando programas *ad hoc* em MapReduce. Entretanto, esta diferença de performance está se estreitando a cada nova *release* da biblioteca. Mas em geral, escrever scripts em Pig Latin irá poupar muito mais tempo para a maior parte das análises do que se aventurar a escrever fluxos map/reduce puramente.

Para ilustrar a facilidade de usar o Pig, a Figura 4.3 mostra um dos scripts usados nessa dissertação. Tal script foi usado para gerar o número de requisições concorrentes, sua média e máxima agregadas por dias do mês (1-31). Veja que para isso só foi preciso um script de 9 linhas.

```
1 REGISTER /home/speed/cdh-hadoop/MscPigScripts/jar/MscPigUtils.jar
2
3 DEFINE dom msc.pig.ExtractTime('dd','America/Sao_Paulo');
4
5 raw = LOAD '$input' AS (ts:long ,num:long , b:long);
6 dom_in = GROUP raw BY dom(ts);
7 dom_req = FOREACH dom_in GENERATE FLATTEN(group), SUM(raw.num), AVG(raw.num), MAX(raw.num);
8 sorted = ORDER dom_req BY group;
9 store sorted into '$output';
```

Figura 4.3. Exemplo de um script em *Pig Latin*

Como podemos ver, Pig é uma framework de processamento de grandes volumes de dados feito para otimizar o trabalho poupando o máximo de tempo. Desta forma escolhemos tal ferramenta para nos ajudar com as análises do presente trabalho. O que mais nos motivou a adotar o Pig foi a baixa curva de aprendizado e por prover um nível maior de abstração evitando que se necessite conhecer a fundo MapReduce para realizar tarefas simples de análise de dados.

4.3 Ambiente de Experimentação

Em [White, 2009] existem vários casos de usos em que o Hadoop foi aplicado com sucesso em grandes empresas. Na maior parte destas empresas o caso de uso padrão é processamento de logs de acesso. Baseado nessas evidência decidimos adotar o *Hadoop* e o *Pig* para facilitar o trabalho desta dissertação e possibilitar as análises do grande volume de dados envolvido.

Para o presente trabalho contamos com um cluster de 16 máquinas, contendo cada uma processadores Intel Core 2 Duo de 2.13 GHz de frequência, 2 GB de memória RAM, 200GB de disco, conectadas por uma rede 10/100 *Megabit Ethernet*. Note que para os padrões atuais de servidores, estas não são servidores modernos[2] mas são suficientes para poder rodar mais de um processo map/reduce por máquina.

A instalação e configuração de um ambiente Hadoop para desenvolvimento e depuração com apenas um nodo é bastante simples. A Cloudera [Cloudera, 2010] oferece imagens pré-configuradas com sua própria distribuição Linux prontas para rodar Hadoop. Ela também disponibiliza repositórios para instalações em sistemas Linux baseados em Debian e Redhat. Esses cenários são interessantes para execuções pseudo-distribuídas, onde explora-se todos os *cores* ou núcleos de processamento de um servidor com um esforço mínimo de configuração.

Entretanto, uma instalação distribuída entre múltiplas máquinas em uma mesma rede, como o nosso ambiente, não foi simples, exigindo um esforço considerável e horas de trabalho e pesquisa. Apesar de existir uma bastante documentação na Web nenhum documento contemplava todos os aspectos necessários para a configuração e podemos concluir que a ferramenta ainda não é de simples instalação e configuração.

A configuração do *HDFS* talvez seja a mais propensa a falhas. O HDFS é a base do desempenho, confiabilidade e alta disponibilidade do arcabouço, portanto deve ser configurado corretamente para garantir o funcionamento ótimo do Hadoop. Um dos

[2]Na realidade tais servidores são meio que *desprezados* no laboratório e nao foi difícil utilizá-los

problemas mais persistentes deparados durante a instalação foi uma falha intermitente do Hadoop ao replicar os dados para todos os nodos que a configuração previa.

Uma observação importante durante o processo é que o *tunning* de um cluster Hadoop não é simples, existem aproximadamente 200 parâmetros que guiam a execução. Inicialmente é interessante ignorar configurações minuciosas para minimizar os possíveis problemas, porém eventualmente, quando alto desempenho for necessário, os parâmetros devem ser estudados e testados experimentalmente para que se obtenha uma configuração satisfatória. Os exemplos que acompanham a instalação rodam sem grandes dificuldades, porém uma pequena investigação mostra que eles utilizam de 20 a 30 parâmetros de execução. Um programa simples como o *WordCount*, que vem como exemplo na biblioteca, não completa sua execução com uma base grande de entrada caso os parâmetros *io.sort.mb* e *io.sort.record.percent* não forem definidos corretamente.

Uma última observação é referente à adição de novas máquinas. Uma vez configurado o cluster, a adição de novas máquinas é uma tarefa relativamente simples. Deve-se configurá-la como um *single node*, em seguida adicioná-la na lista de nodos da configuração do *Master* se necessário, configurar o SSH sem senha e iniciar os *daemons* locais. É interessante executar um comando de rebalanceamento dos nodos do HDFS, garantindo uma distribuição mais homogênea dos dados entre os novos nodos.

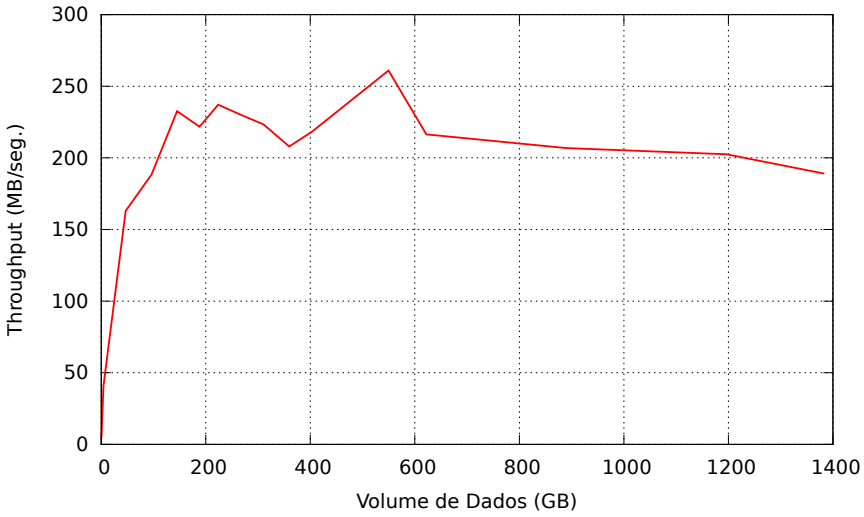


Figura 4.4. *Throughput* do ambiente de experimentação

0

Alguns pequenos experimentos foram realizados para avaliar a escalabilidade do ambiente que estava sendo configurado. Notamos que o *speedup* do arcabouço, utilizando um ambiente configurado com Hadoop e Pig foi sublinear na proporção

$\{(1;1),(2;2),(4;3,5),(8;6,1),(16;11,8)\}$. Ainda assim, o ambiente se mostrou escalável a baixo custo de desenvolvimento.

Como o nosso principal problema era o volume de dados, a maior parte das análises a serem realizadas aqui teriam o tempo de execução dominadas pelo tempo de *IO*. Assim realizamos um teste para avaliar a vazão de dados (*throughput*). A Figura 4.4 mostra os valores de *throughput* para o nosso ambiente a medida que o tamanho da entrada cresce rodando com 16 instâncias. Observa-se um crescimento acentuado com volumes de dados inferiores a 149 GB, e a partir desse ponto o *throughput* tende a se estabilizar.

O ambiente manteve valores estáveis de *throughput* para entradas até 1.4 TB, sugerindo ser robusto e escalável. Em especial, a execução com 1.4 TB levou 124 minutos e apresentou um *throughput* médio de 190 MB/s.

Neste capítulo apresentamos o arcabouço utilizado para a execução do presente trabalho. Tal arcabouço foi essencial para que pudéssemos analisar a quantidade de dados disponível de forma eficiente e proporcional a dinamicidade adequada para um trabalho experimental como este uma vez que o ciclo de hipóteses, análises e resultados é muito dinâmico e frequente. Além disso, erros e dúvidas são comuns e a reexecução e revalidação dos resultados seriam desgastantes e trabalhos se tal arcabouço não tivesse sido construído.

Capítulo 5

Estudo de Caso: Plataforma de Conteúdos Multimídia

Neste capítulo aplicaremos a metodologia proposta no Capítulo 3 para o estudo e análise de um cenário real que representa parte significativa do tráfego de vídeos corporativos no Brasil.

Os dados utilizados para tal estudo de caso são originados da empresa líder no setor de plataformas comerciais de vídeos online no Brasil, a Samba Tech [SambaTech, 2011b], uma empresa que atua no mercado de vídeos corporativos no Brasil e que trabalha nos moldes de computação em nuvem provendo infraestrutura, gerenciamento e distribuição de vídeos em canal digital. Dentre seus clientes, podemos citar o Portal R7, GloboSat, SBT, Grupo Abril, IG, Bandeirantes, RBS, e Lancenet.

O objetivo deste estudo de caso é aplicar a metodologia em dados reais e demonstrar a capacidade analítica do processo. Não pretendemos de maneira nenhuma ser uma instanciação exaustiva da metodologia. Avaliaremos a plataforma de distribuição de conteúdos multimídias com o foco na forma em que se consome tais conteúdos, como se distribui a popularidade dos diferentes objetos e conteúdos e de que maneira pode-se explorar a potencialidade de uma plataforma com tais características de distribuição.

É importante ressaltar que esse estudo de caso foi desenvolvido aplicando-se a metodologia, e não o contrário. Ou seja, a metodologia tem um escopo mais amplo e o nosso estudo de caso se utiliza desta e não contempla todos os aspectos metodológicos possíveis. A escolha de quais análises devem ser realizadas deve se basear no foco do estudo de caso. Um estudo que tem como objetivo melhorar o modelo de cobrança de um sistema multimídia deve encarar uma segmentação do uso de banda por conteúdo como uma análise importante. Já uma análise dos recursos de infraestrutura é agnóstico aos diferentes conteúdos trafegados. Desta forma, nem todas as análises elencadas no

Capítulo 3 serão contempladas neste capítulo.

5.1 Arquitetura da Plataforma

Antes de passar para a análise dos dados, apresentaremos a arquitetura da Plataforma Gerenciadora de Conteúdo que será alvo do nosso estudo.

O nosso ambiente de avaliação, experimentação e estudo se trata de uma empresa de *logística digital* que atua no mercado latino-americano de gestão de mídias digitais através do modelo *Platform as a Service* (*PaaS*), provendo infraestrutura, serviços de gerenciamento e distribuição de conteúdo multimídia para grandes corporações, em sua maioria brasileiras.

A empresa no caso é a *Samba Tech*[SambaTech, 2011b], que trabalha com grandes grupos da mídia brasileira e atualmente é responsável pela gestão de um grande conteúdo de mídias ricas, conforme podemos ver na Tabela 5.1 que mostra as informações as quais tivemos acesso sobre a plataforma de vídeos da empresa.

Estatísticas Gerais Disponíveis	
Produtores de Conteúdo	57
Armazenamento	10 TB
Vídeos	152 Mil
Mídias	482 Mil
Média Mensal 2010	
Tráfego	246 TB
Videos Adicionados	10800
Requisições	243 Milhões
Visualizações de Vídeos	19 Milhões
Usuários Únicos	6.1 Milhões

Tabela 5.1. Informações disponibilizadas em Janeiro de 2011

O serviço de logística digital é prestado através da plataforma de soluções para gestão de mídias. Tal plataforma é composta de vários componentes, dentre eles:

- um *Software as a Service* (*SaaS*), que dispensa a necessidade de infraestrutura como um computador para se instalar o programa. Toda interação é online via Navegador web, dispensando um cliente de instalação;

- um conjunto de Interfaces de Programação de Aplicativos (APIs) que possibilitam a gestão programática dos conteúdos cadastrados;

- uma série de serviços que são disponibilizados online e que não demandam infraestrutura por parte do cliente.

A plataforma em questão é conhecida como *Liquid*^{TM} [SambaTech, 2011a].

Assim, uma empresa que trabalha com mídias ricas, por exemplo uma agência de notícias que possui um portal com páginas dinâmicas e grande quantidade de vídeos, pode utilizar a plataforma da empresa para gerenciar suas mídias ricas (músicas, imagens, vídeos).

O trabalho de distribuição sob demanda, com infra-estrutura redundante, acordo de nível de serviço (SLA) 24/7 [1] e capacidade elástica de entrega de conteúdo é terceirizado através de redes de distribuição de conteúdo CDN (Content Delivery Network).

Os produtores de conteúdo utilizam a solução de logística digital *Liquid*^{TM} para gerenciar seus conteúdos. A *Liquid*^{TM} é uma plataforma *Web*, sendo seu principal objetivo oferecer serviços voltados para soluções de logística digital com destaque para vídeos. Alguns dos serviços oferecidos pela plataforma são o armazenamento de mídias em CDN's ou Storages de baixo custo, transcodificação/*encoding* das mídias (ex.: de Flash flv para o formato MP4), *geoblocking* (publicação restrita a regiões geográficas), gerenciamento de publicação de conteúdo em diversas redes sociais, gestão de publicidade nos conteúdos e um conjunto de APIs para que as empresas clientes possam programaticamente interagir com seus conteúdos permitindo flexibilidade ao cliente.

As funções principais da API são: a contagem de mídias armazenadas, a listagem de mídias, a obtenção dos meta dados de uma mídia pelo seu identificador, a listagem e contagem de categorias. Porém, também existem outras funcionalidades como busca de vídeos relacionados, inserção de informações relativas a mídia como votações de *feedback* de usuários (*ratings*), comentários dentre outros[2]

A carga de trabalho da *Liquid*^{TM} mostra-se expressiva e interessante para estudo pelo fato de que os clientes que utilizam a plataforma estarem entre os maiores grupos de mídias do Brasil, tais como: SBT, Rede Record com o Portal R7, IG, Bandeirantes, Grupo Abril, Lancenet, Oi TV, o Grupo Associados (Estado de Minas, TV Alterosa e Portal UAI), Rede Minas, RBS além de outros clientes que atuam nos mais diversos seguimentos como O Boticário, Água de Cheiro, Cruzeiro Esporte Clube, Clube Atlético Mineiro, Internacional Futebol Clube e vários outros.

[1] 24 horas por dia 7 dias por semana

[2] Veja a documentação completa em: `http://docs.liquidplatform.com/`

A *Liquid*^{TM} é utilizada pelos produtores de conteúdo para o registro de novas mídias na plataforma para que estas sejam controladas pela plataforma. O cliente insere o conteúdo (vídeo, imagem ou áudio) e os metadados como título, subtítulo, categoria, descrições e tags , estes então são armazenados na plataforma compondo os conteúdos que serão distribuídos para o usuário final através dos sites corporativos dos clientes.

A dinâmica de como um conteúdo está interligado entre a *Liquid*^{TM}, produtor de conteúdo, serviços e usuário final pode ser vista na Figura 5.1

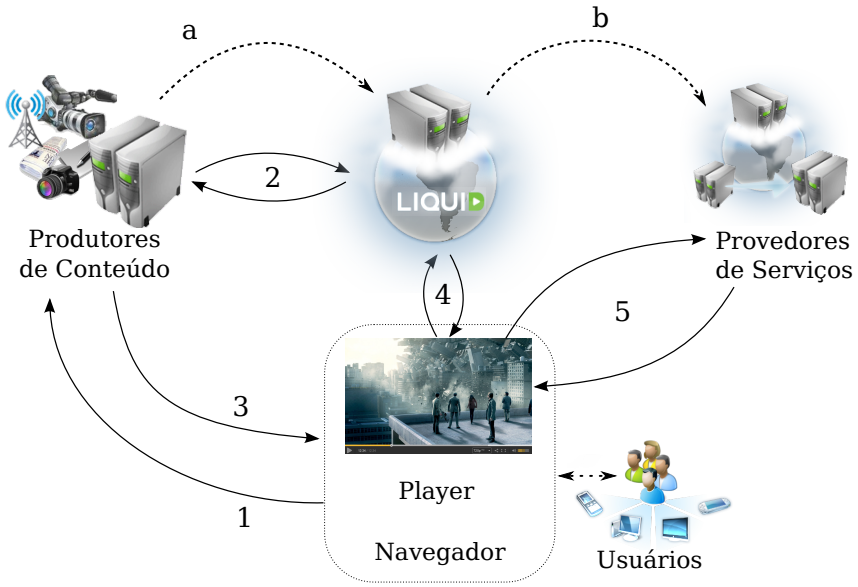


Figura 5.1. Esquema de distribuição de conteúdos

A *Liquid*^{TM} pode ser utilizada de várias maneiras dependendo do caso de uso do *Produtor de Conteúdo* em questão. Pode-se utilizar vários *Provedores de Serviços* como *encoding* online (Encoding.com, HeySpread, Zencoder), CDN (Amazon, RackSpace, Akamai), ou até mesmo outros serviços de publicação de vídeo (Youtube, Vimeo, etc). A seguir explicaremos os dois casos de uso mais utilizados e algumas das interações essenciais entre a plataforma e os agentes que compõem o sistema.

As setas pontilhadas da Figura 5.1 e elencadas com letras compõem o caso de uso de publicação.

Seta a : O processo se inicia com o *Produtor de Conteúdo* cadastrando suas mídias na *Liquid*^{TM}. Nesse passo o encarregado da publicação, normalmente um editor ou jornalista, cadastra suas mídias no sistema preenchendo os respectivos metadados e construindo relações semânticas entre as diversas mídias do seu editorial.

Seta b : Cada *Produtor de Conteúdo* pode possuir diversos projetos dentro da *Liquid*™. Cada projeto possuirá uma série de configurações que irá determinar quais Serviços utilizar, quais formatos de saída, se existe algum compartilhamento automático para outros meios de divulgação, dentre outras configurações. Assim, para cada mídia inserida no projeto, a plataforma irá realizar uma série de operações utilizando diversos *Provedores de Serviços* para tal fim.

Uma vez realizado o fluxo de publicação, o *Produtor de Conteúdo* passa a utilizar a *Liquid*™ como ferramenta centralizadora para a publicação e distribuição de seus conteúdos. Para a publicação, o caso de uso mais comum é mostrado na Figura 5.1, através das setas contínuas e numeradas.

Seta 1 : Através de um navegador, o usuário final acessa o aplicativo (site/-porta/aplicativo móvel, etc..) do *Produtor de Conteúdo* utilizando qualquer plataforma de acesso.

Seta 2 : O aplicativo em questão consulta a API da plataforma para montar a página requisitada. Tal operação visa buscar metadados de mídias, conteúdos relacionados, imagens, opiniões e comentários de usuários.

Seta 3 : O aplicativo é retornado para o usuário com os conteúdos organizados além do player da plataforma.

Seta 4 : O player consulta novamente a plataforma através da API para buscar o endereço da mídia, registrar visualizações, obter informações de publicidade, verificar se o usuário pode assistir tal mídia.

Seta 5 : Ao exibir a mídia, o player se encarrega de salvar o mídia em questão do CDN e exibi-lá para o usuário. Outras operações possíveis no player é o compartilhamento com redes sociais, *feedback* do conteúdo e carregar vídeos relacionados.

Como podemos ver na explicação anterior a arquitetura do serviço disponibilizado pela plataforma é muito complexa e cheia de interações que podem ser estudadas separadamente. No presente trabalho iremos estudar e caracterizar o **fluxo de distribuição** ilustrado na Figura 5.1, utilizando para isso a metodologia *ROCK*, já apresentada no Capítulo 3.

5.2 Descrição dos Dados

Nesta seção descrevemos os dados que foram utilizados para o presente estudo, como estes foram coletados e de que forma estes dados foram preparados para a realização das análises realizadas neste trabalho.

Diferente de trabalhos como [Gill et al., 2007, Veloso et al., 2006a, Almeida et al., 2001], que coletaram os dados dos usuários ou através de um proxy, as informações foram disponibilizadas, diretamente pelo distribuidor de conteúdo, através de parte dos *logs* de acesso e também dos metadados das mídias relacionadas a tais *logs*. Não tivemos acesso a conteúdo ou informação confidencial dos clientes da Samba Tech. Com exceção dos *logs* de acesso, todas as informações disponibilizadas pela empresa estão publicadas nos portais de seus clientes o que poderia ser facilmente coletado na Web.

5.2.1 Logs de Acesso

As informações relativas aos acessos das mídias dos *Produtores de Conteúdo* foram obtidas dos registros de acesso dos servidores HTTP dos CDNs utilizados pela empresa. Tais CDNs são os responsáveis pelo armazenamento e entrega elástica dos conteúdos. A plataforma em análise aceita a utilização de diversos CDNs para a distribuição do conteúdo, fazendo com que os registros de acesso sejam de diferentes origens. Tais *logs* são gerados nos vários servidores de cada CDN que estão distribuídos em diferentes regiões geográficas. Um sistema central agrega os vários registros individuais que são entregues a plataforma para que essa possa extrair informações relativas aos acessos e tráfego.

Atualmente a escolha de qual CDN utilizar para trafegar o conteúdo ainda não é dinâmica, e a maior parte do conteúdo trafegado pela plataforma é feita somente por um CDN, somente em casos de demandas específicas outros CDNs são utilizados. As informações relativas aos logs que tivemos acesso podem ser vistas na Tabela 5.2:

Logs dos CDNs			
Período	**Tráfego (TB)**	**Requisições (Milhões)**	**Tamanho (GB)** [3]
17 meses	3.493,72	3.378,6	1.376,36

Tabela 5.2. Algumas informações relativas aos logs dos CDNs estudados. Os 17 meses são de agosto de 2009 à dezembro de 2010, o tamanho é relativo a quantidade de dados a serem processados contida nos logs.

Os registros dos CDNs analisados se apresentam no formato estendido da W3C[4], porém cada CDN provê um conjunto de campos distintos. A seguir descrevemos os principais campos que se encontram nesses logs e que serão utilizados para as análises que virão:

1. **timestamp:** Tempo em segundos da requisição. Representado em segundos desde o *epoch time* (00:00:00 UTC de 1º de Janeiro de 1970)

2. **time-taken:** Tempo em milissegundos gasto escrevendo os dados requisitados para o cliente. (Não necessariamente o tempo gasto entregando o conteúdo.)

3. **c-ip:** Endereço IP do cliente.

4. **s-ip:** Endereço IP do servidor.

5. **s-port:** Número da Porta utilizada no servidor (80 para http, 443 para https)

6. **sc-status:** Status do servidor-cliente, incluindo status do cache (TCP_HIT, TCP_MISS, etc.) e o código status de http (200, 304, 404, etc.)

7. **sc-bytes:** Número de bytes enviados do servidor ao cliente.

8. **cs-method:** Método http da requisição cliente-servidor (GET, HEAD, POST, etc.)

9. **cs-uri-stem:** Url requisitada pelo cliente ao servidor.

10. **c-referrer:** O cabeçalho http *Referer* da requisição enviado pelo cliente. O *referer* é qualquer coisa que direciona o visitante ao site : uma página, uma publicidade, um link, etc.

11. **c-user-agent:** O cabeçalho http da requisição que identifica o agente de acesso *User-Agent*. É importante ressaltar que este campo pode ser facilmente burlado.

Com os campos acima é possível fazer diversas análises relativas ao acesso, aproximação da largura de banda, quantidade de dados trafegada, localidade temporal e geográfica [5] do acesso, dentre várias outras. É importante ressaltar que o campo URL provê uma chave que identifica o conteúdo trafegado junto ao bando de dados que será descrito na próxima sessão.

[4] http://www.w3.org/TR/WD-logfile.html
[5] Utilizando o IP como uma aproximação

5.2.2 Metadados Textuais

Os *logs* dos CDNs nos permite ter informações relativas ao tráfego e acesso. Além disso, a URL possibilita que se determine qual objeto e conteúdo está sendo acessados através de um identificador único. Com isso é possível fazer análises que irão correlacionar os significados semânticos aos acessos registrados nos *logs*.

Foi-nos concedido alguns registros dos metadados relativos as URLs de acesso contidas nos registros dos CDNs. Tais metadados possuíam os seguintes campos:

Título : Nome atribuído à mídia relacionada pelo editor da mídia.

Descrição : Descrição detalhada do Conteúdo representado pela mídia em questão.

Categoria : Uma categoria (que pode ser hierárquica ou não) específica do *Produtor de Conteúdo* que agrupa as mídias comum a um fim específico. Por exemplo, o *Produtor de Conteúdo* SBT, tem algumas categorias relativas a seus programas, *bom dia e companhia*, *programa Sílvio Santos*, *Eliana*, *jornal do sbt*, etc.

Gênero : Um gênero global (não específico por *Produtor de Conteúdo*) com os possíveis valores:

1. Comédia
2. Entretenimento
3. Filmes
4. Música
5. Política
6. Pessoas
7. Animais
8. Ciência
9. Esportes

Tags : Lista de palavras que são usadas em associação ao conteúdo.

Ratings : Informações relativas a votação dos usuários em uma dada mídia. A votação é categórica em valores de 1 à 5.

Diversos : Alguns outros campos também existiam na base. Valores binários para as mídias que: (1) possuíam publicidade, (2) foram compartilhadas com o *YouTube*, e que (3) tiveram destaque no site.

5.2.3 Tratamento de Logs e Metadados

Como já dito anteriormente, dispomos de acesso a dados reais de tráfego e conteúdo de vários *Produtores de Conteúdo* Brasileiros. Uma das implicações desse fato é que que nem toda informação é passível de utilização diretamente. São necessárias algumas etapas de pré-processamento, tratamento dos dados, transformações e estudo do domínio dos dados para que seja possível extrair de forma eficaz as análises que se deseja.

Cada etapa será discutida aqui e retomada se apropriado em seções futuras.

Processamento dos Logs

Como foram disponibilizadas *logs* de diferentes CDNs foi preciso inicialmente gerar um parser para cada formato para extrair os campos necessários para as análises da metodologia proposta. Além disso, era preciso segmentar os *logs* de forma que erros, requisições inválidas e outras transações pouco relevantes pudessem ser separadas e contabilizadas.

Outro desafio enfrentado durante o processamento dos *logs* foi o grande volume de dados. Para tratar tal desafio utilizamos o Arcabouço explicado no Capítulo 4 que viabilizou as análises em tempo hábil. Todas as funções e scripts de processamento foram implementadas utilizando a plataforma de análise de dados Pig [Olston et al., 2008]. Todos os parsers e funções de agregação foram desenvolvidas utilizando a extensibilidade de Pig através das *UDFs* (User Defined Functions) em java.

Tratamento e Pré-Processamento dos Metadados

Por se tratar de uma base real, o banco de dados a que tivemos acesso possui uma certa quantidade de detalhes a serem tratados. Por exemplo, a existência de mais de um *encoding* de texto para diferentes *Produtores de Conteúdo*; diferença na utilização dos metadados para um determinado *Produtor de Conteúdo* (categoria era o título e o título era a descrição); diferentes separadores de *Tags*; além de vários outros pequenos detalhes. Para outras análises foi necessário realizar um pré-processamento dos Metadados como remoção de *stop-words*, *stemming*, categorização de valores numéricos, etc.

Limitação dos Dados

As informações de geolocalização são obtidas através de um mapa entre IP e localidade. Utilizamos a biblioteca livre de *GeoIp* da MaxMind[6] para realizar a conversão de IP

[6] http://www.maxmind.com/app/geolitecity

para localidade geográfica. Tal biblioteca permite a identificação do país, região (algo semelhante aos nossos estados) e cidade do IP de origem. Segundo o portal da empresa, a biblioteca possui uma acurácia média de 95% para a identificação de países e de 79% para a identificação de cidades.

Outro ponto é a informação do *User Agent*, o agente que dá origem as requisições. Apesar de ser facilmente alterado por um robô ou programas específicos, acreditamos que devido ao grande volume de requisições que possuímos não será significativo qualquer impacto de possíveis alterações.

É importante frisar que nos dados **não há** uma identificação específica do usuário. Não há os registros um identificador único de usuário. O campo que mais poderia nos aproximar de um usuário é o campo de endereço IP, contudo sabemos das limitações dessa abordagem uma vez que a maior parte das redes atualmente se escondem atrás de uma tecnologia que compartilha um único endereço IP para vários computadores em uma mesma rede.

Os registros das requisições fornecidas possuem resolução em segundos. Apesar de ser uma resolução mais que suficiente para diversos estudos ela é bastante limitante para análises mais detalhadas como tempo entre requisições, tempo de respostas e quaisquer métricas que utilizem o tempo de resposta. Por exemplo, em conteúdos muito populares temos que um segundo podem haver milhares de requisições em 1 segundo. Outro caso que limitou nossas análises é o tempo de resposta. Como os servidores só registram as respostas em segundos, temos que requisições atendidas na grandeza de micro ou milissegundos, são registradas como 0 segundos de resposta. Alguns servidores modernos já registram as transações em milissegundos, o que pode trazer novas evidências em futuros estudos.

5.3 Análises da Camada de Requisições (R)

Analisaremos aqui a plataforma em estudo sob o ponto de vista mais baixo na hierarquia de informação. Tal análise envolve as informações de maior resolução e maior granulação que são as informações relativas às requisições individuais realizadas nos servidores. Descreveremos cada análise em um sub-tópico específico de acordo com a metodologia apresentada na sessão 3.2. Como a maior parte do tráfego gerado é proveniente do Brasil adotaremos o fuso horário oficial de Brasília (*GMT-3*) para todas as análises temporais que envolvem resolução de tempo em horas.

5.3.1 Descrição das Requisições

Para conhecer melhor o escopo do estudo, iremos apresentar um sumário das informações de requisições contidas nos *logs* de acesso nesta etapa. A Figura 5.2 mostra o agregado das requisições e tráfego dos *logs* analisados. São mais de 3.7 bilhões de requisições, o que corresponde a mais de 3 petabytes de dados trafegados num período de 17 meses de análise. O gráfico da Figura 5.2 mostra o perfil de utilização da plataforma e o crescimento de acesso no período analisado. O que se destaca nesse gráfico é o pico de acesso existente no final do ano de 2010 e a súbita queda no período relativo as festas de fim de ano.

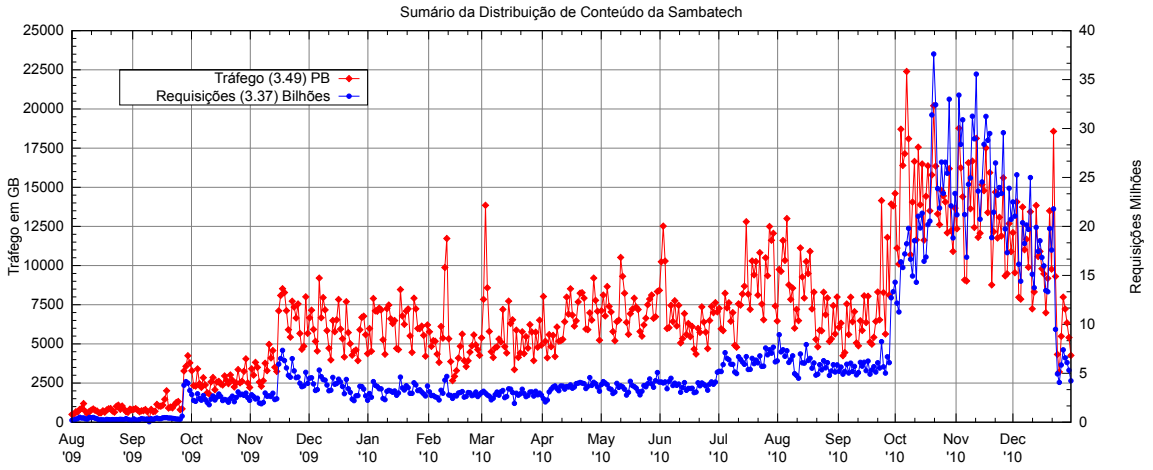


Figura 5.2. Informações relativas a distribuição de conteúdo multimídia pela Samba Tech contida nos logs disponibilizados.

A Tabela 5.3 mostra o resumo dos métodos de requisições existentes nos *logs* analisados. Conforme esperado, podemos ver a predominância do método GET, uma vez que os *logs* são referentes ao *download* dos conteúdos direto da rede de distribuição da *Liquid*^{TM}. Outros métodos são insignificantes por exemplo: POST e OPTIONS com, respectivamente, 11 e OPTIONS 30 requisições.

Metodo	% Req.	% Tráfego
GET	99.99%	100.00%
HEAD	0.01%	0.00%
OTHERS	0.00%	0.00%

Tabela 5.3. Sumário dos métodos das requisições nos *logs*

Por sua vez, na Tabela 5.4 mostramos o sumário dos códigos de retorno (*status*) HTTP das requisições. O código 200 indica que a requisição foi respondida com sucesso pelo servidor, já o código 206 indica retorno com sucesso de uma requisição parcial de conteúdo, por exemplo, a requisição de uma parte específica de um vídeo. Essas duas requisições são responsáveis pela grande maioria do tráfego e também das requisições na rede.

Uma resposta com código 304 é enviada para requisições de validação de objeto. Um GET é requisitado ao servidor condicionado a data da versão do objeto em questão, caso esse não tenha sido alterado, uma resposta desse tipo é retornada o que evita tráfego desnecessário. Podemos ver que o número de requisições condicionais é expressiva, isto porque a maioria dos navegadores modernos possuem um *cache* de imagens pequenas como *thumbnails* fazendo a requisição condicional para diminuir o uso da rede no cliente.

Código de Status	**% Req.**	**# Req.**	**% Tráfego**	**Tráfego GB**
200 (OK)	77.12%	2.541.463.776	93.37%	3.204.893,79
206 (Partial Content)	1.93%	63.745.627	6.62%	227.541,94
304 (Not Modified)	18.79%	619.236.365	0.01%	178,34
4xx (Client Errors)	2.12%	69.987.493	0.00%	34,96
5xx (Server Errors)	0.03%	978.036	0.00%	0,50

Tabela 5.4. Sumário dos códigos de retorno das requisições HTTP

Se compararmos os dados das Tabelas 5.4 e 5.3 com os dados presente em [Gill et al., 2007], vemos que as proporções são bastante semelhantes: predominância de GET, 75% para código 200, 17% para código 304, 1.29% para código 206.

A Tabela 5.5 detalha os códigos de erros das séries 4xx e 5xx, que são relativos aos erros de clientes e servidores respectivamente. É importante notar que somente 0.03% dos erros são de origem no servidor, um taxa bem aceitável que representaria um SLA de 99.97% por parte dos servidores da *Liquid*^{TM} se esses *logs* representassem continuamente o total período de uso da plataforma. A maior parte (95.78%) dos erros do servidor foram os 504 (*Gateway Timeout*), isso porque os *logs* são dos servidores do CDN que no caso de um *cache miss* requisitam diretamente no servidor primário, e nestes casos, o servidor primário não respondeu a requisição.

Com relação aos erros por parte dos clientes, a maior parte são erros 403, que acontecem quando o cliente requisitante não tem permissão de acessar o conteúdo. No

Erros de Servidores	% das Requisições
500 (Internal Server Error)	0.32%
501 (Not Implemented)	1.91%
502 (Bad Gateway)	0.23%
503 (Service Unavailable)	1.75%
504 (Gateway Timeout)	95.78%

Erros de Clientes	% das Requisições
400 (Bad Request)	0,01%
403 (Forbidden)	83,19%
404 (Not Found)	13,49%
416 (Req. Range Not Satisf.)	3,30%

Tabela 5.5. Descrição dos erros segmentada por cliente e servidor

caso da *Liquid*^{TM} isso ocorre devido ao fato de alguns conteúdos distribuídos possuem *copyright* restrito ao Brasil. Outro erro expressivo (13,45%) é o 404 que representa uma requisição a um objeto que não existe.

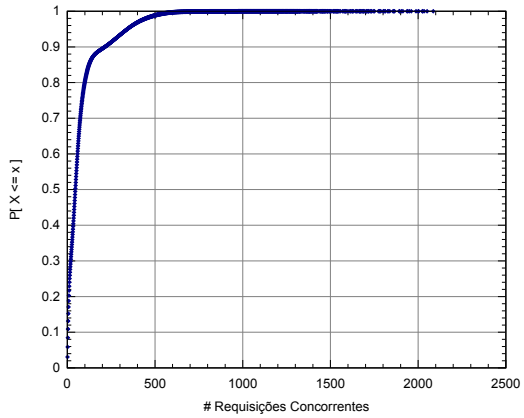
De uma forma geral, as requisições com erros são poucas, representam cerca de 2.5%. Ainda assim, a maior parte destas não são semanticamente um erro e sim uma proibição do acesso ao conteúdo. Temos assim uma massa de dados estatisticamente significativa, de aproximadamente 17 meses de *logs* e que em sua maioria representam transações válidas e que serão analisadas nas próximas etapas. Todas as análises seguintes foram realizadas considerando **somente as requisições válidas**, ou seja, aquelas com código de retorno 200, 206 ou 304.

5.3.2 R_1: Taxa de Resposta (*Throughput*) do Servidor

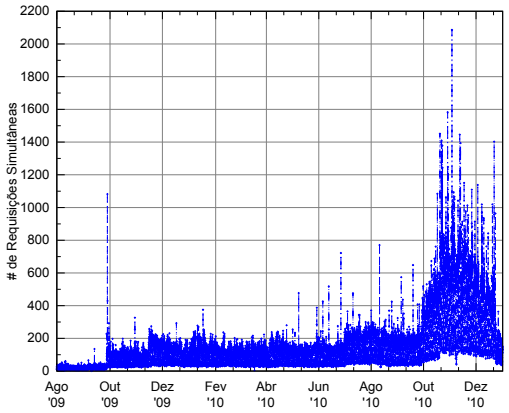
A cada instante de tempo t existe um número de requisições ativas no servidor $req(t)$. Nesta seção, iremos avaliar as distribuições de $req(t)$ e avaliaremos também vários comportamentos periódicos para $req(t \ mod \ p)$ onde p determina o período em que agregaremos as requisições, podendo ser mensal, semanal, diário ou qualquer outro período desejado.

Na Figura 5.3 mostra-se um sumário da distribuição das requisições do período analisado. No gráfico da Figura 5.3(a) temos a distribuição cumulativa (cumulative distribuction function, *CDF*) que mostra que a frequência de mais de 500 requisições por segundo é muito baixa, apenas 1,2% das do tempo é que se verifica presença de mais de 500 requisições simultâneas. Para avaliar a carga nos servidores, informação muito útil para fazer o planejamento de capacidade do serviço, a Figura 5.3(b) mostra o número máximo de requisições durante todo o período de estudo agregados em períodos de uma hora. Podemos ver o crescimento do uso da plataforma de vídeos no intervalo. O pico em outubro de 2009 representa o lançamento do portal de um grande *Produtor de Conteúdo*, o que gerou uma característica atípica nas requisições. Com exceção desse pico, vemos que o número de requisições cresce com o tempo chegando a picos

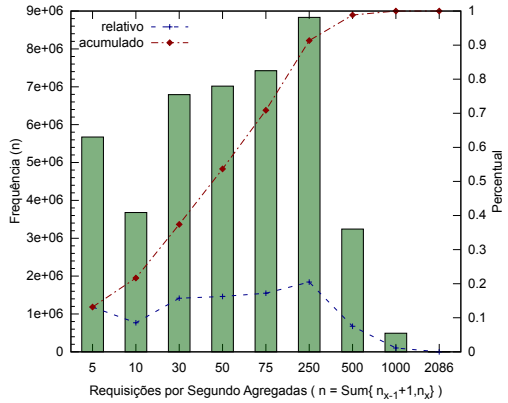
de mais de 2200 requisições por segundo.



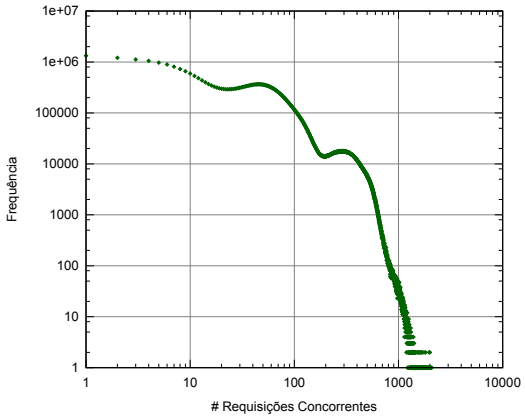
(a) Distribuição cumulativa das *req./s*



(b) Número máximo de requisições simultâneas por hora.



(c) Histograma de requisições concorrente sobre todo o período



(d) Distribuição das Requisições pelo tempo

Figura 5.3. Distribuição das requisições sobre todo tempo de estudo.

A distribuição completa das requisições pode ser vista na Figura 5.3(d). Para simplificar esta informação mostramos o histograma das requisições por segundo agrupadas em intervalos fixos na Figura 5.3(c). Para cada valor de x temos a frequência com que as requisições por segundo n ocorrem no intervalo entre $x-1$ e x. O valor da frequência é definida pelo somatório $\sum_{n_{x-1}+1}^{n_x}$ onde n_x é a frequência do número de requisições por segundo com x requisições simultâneas. Nessa figura mostramos também o percentual das requisições do intervalo e a sua acumulada. Podemos ver que 90% das requisições ocorrem com no máximo 250 requisições simultâneas, ainda assim, o intervalo de requisições por segundo de grandeza $250 < x \leq 2086$ correspondem a quase 4 milhões de requisições.

Na tentativa de descobrir um padrão comportamental periódico mensal plotamos a distribuição das requisições concorrente $req(t \ mod \ p)$ onde p são os dias do mês. O gráfico da Figura 5.4 mostra a distribuição de $req(t)$ máximo e médio por minuto. A análise dos picos máximos servem para avaliarmos os casos atípicos de acesso na plataforma para fins de planejamento de capacidade. Para determinar padrões periódicos utilizamos as médias das requisições.

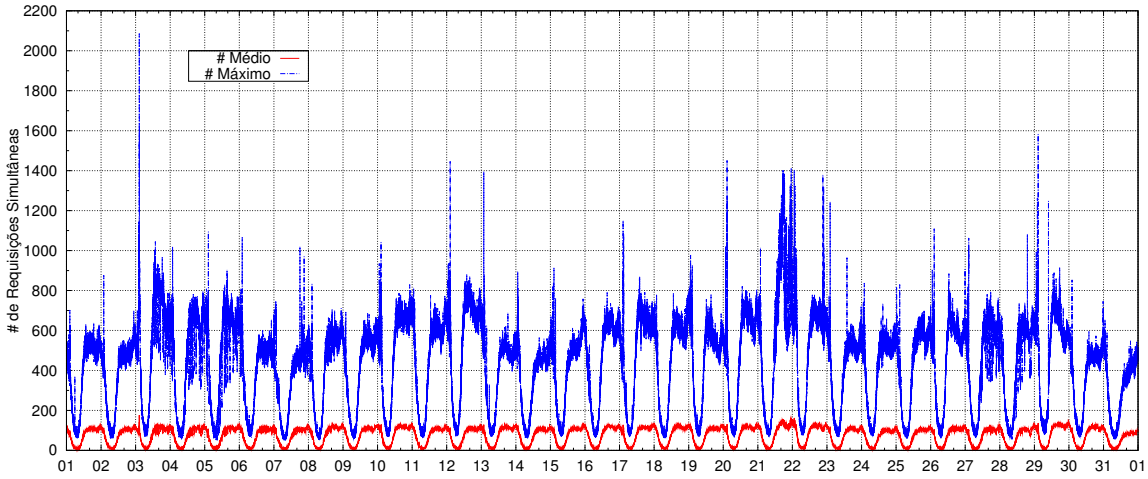


Figura 5.4. Requisições máximas e médias sob a perspectiva mensal do período e resolução por minuto.

Não há evidências de comportamento periódicos com amplitude mensal, as médias das requisições simultâneas de uma maneira geral comportam-se de forma similar. Casos que fogem um pouco do comportamento são justamente onde se encontram padrões atípicos máximos (dias 3,4,5 e 21) o que nos faz acreditar que houve uma distorção da média devido ao comportamento atípico.

Para a avaliação de um padrão periódico semanal temos a Figura 5.5 com $req(t \ mod \ p)$ onde p são os respectivos 7 dias da semana. Notamos que os casos atípicos geralmente ocorrem entre Terça-Feira e Sexta-Feira não ocorrendo picos nos finais de semana. De uma maneira geral existe um comportamento similar médio para os dias da semana que difere do comportamento dos sábados e domingos.

Para avaliar os comportamentos distintos entre dias de semana e finais de semanas fizemos uma nova análise de $req(t \ mod \ p)$ onde p são somente dias de semana (Segunda à Sexta) ou Sábado e Domingo para os fins de semana. Tal análise foi feita computando a média de todas as requisições no mesmo período de tempo através de todo o período (17 meses) analisado e pode ser vista na Figura 5.19

A primeira análise pode ser vista na Figura 5.6(a). Nesta figura, temos o comportamento esperado para um dia de semana sem eventos atípicos. Vemos um pequena

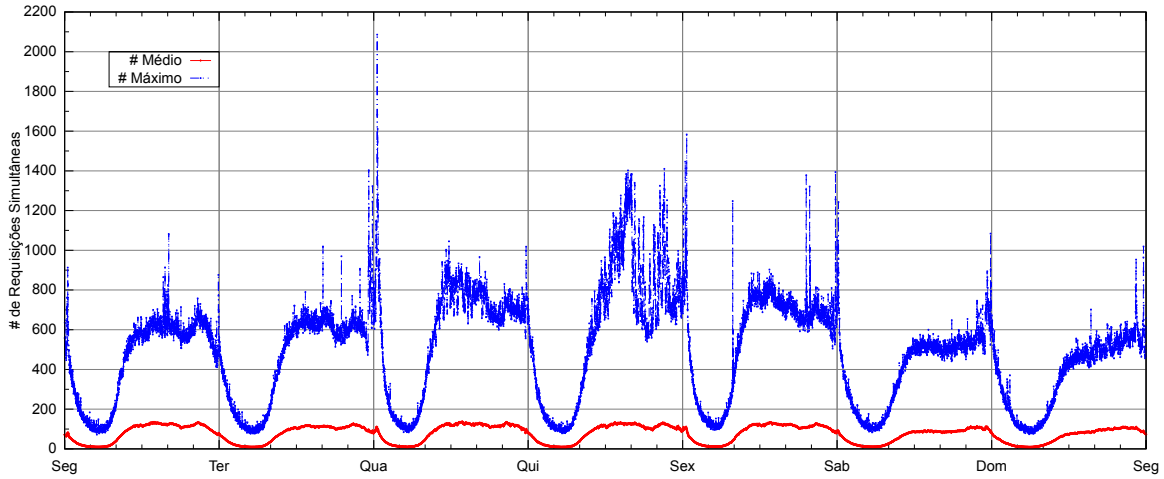
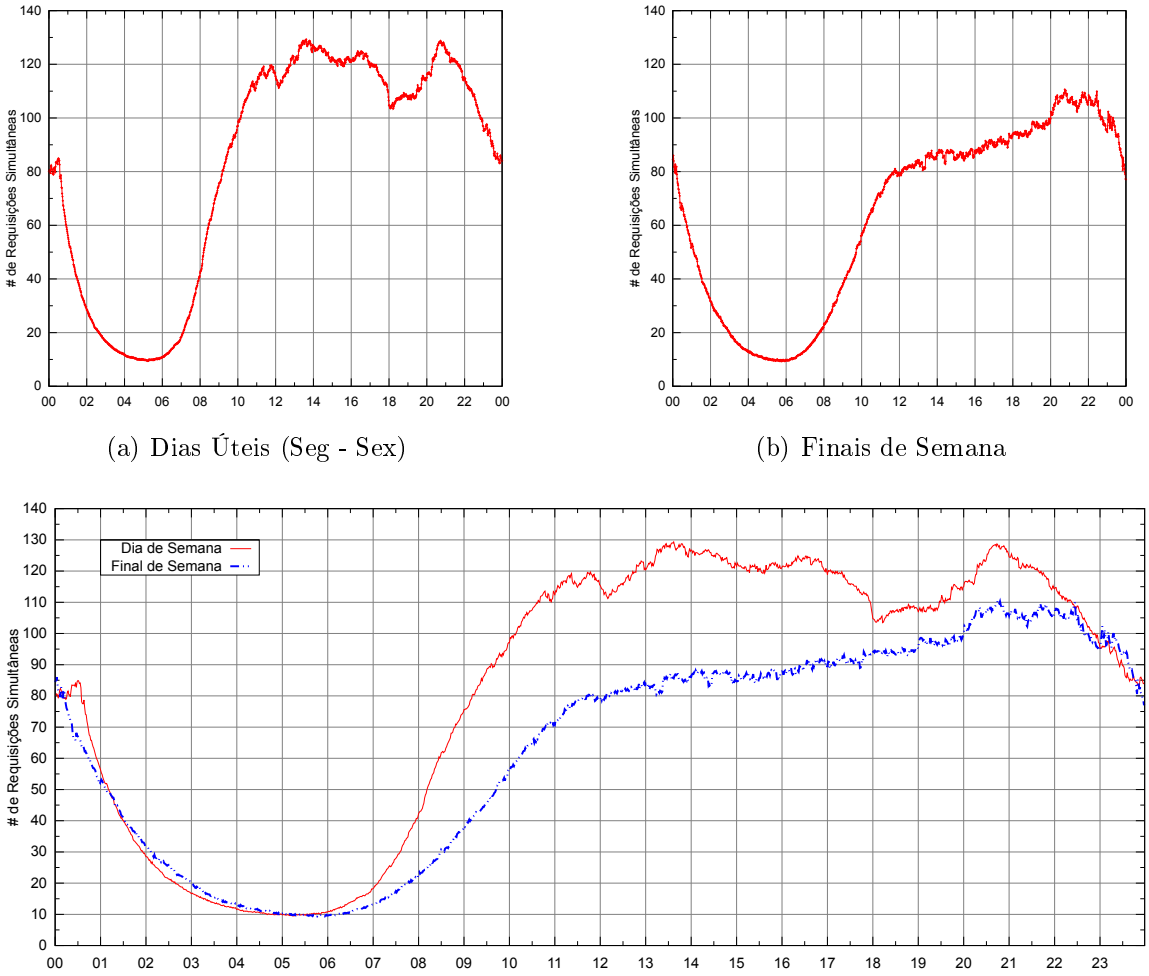


Figura 5.5. Requisições médias e máximas em periodicidade semanal e resolução de minutos

variação nos primeiros minutos do dia seguida de uma constante queda nas requisições atingindo seu menor valor (9 *req./s*) por volta das 5 horas da manhã. O intervalo de 6h às 11h é onde ocorre o maior crescimento levando as requisições de uma taxa de 10 à 110 *req./s*. No período de 11h as 18h a taxa de requisições flutua entre 110 a 130 *req./s* com pico máximo por volta das 13h 30m e uma queda acentuada exatamente às 18h, provavelmente devido ao fato de ser esta hora o fim do expediente na maioria das empresas brasileiras. Após a queda das 18h o número de requisições cresce a um ritmo moderado até atingir um novo pico por volta das 21h. Após tal pico ocorre uma queda forte levando a taxa de 130 *req./s* para 80 *req./s*.

O comportamento de finais de semana encontra-se no gráfico da Figura 5.6(b). De maneira geral a taxa de acesso dos finais de semana é menor se comparada aos dias de semana. O período da madrugada possui uma queda que vai de 80 à 10 *req./s* no período entre as 0h e 6h, seguido de um crescimento de requisições até voltar a taxa de 80 *req./s* por volta das 12h. Após esse horário ocorre um lento aumento na taxa de requisições até 95 *req./s* por volta de 20h. O período mais popular dos finais de semana é o entre as 20h e 22h30 onde a taxa varia entre 95 e 110 *req./s*. Em seguida observa-se a queda acentuada até as 0h atingir novamente a taxa de 80 *req./s*.

Ao contrastarmos as duas distribuições como feito na Figura 5.6(c) vemos algumas diferenças interessantes e podemos até fazer suposições a respeito do comportamento dos usuários em geral. Vemos que o as requisições nos finais de semanas são um pouco maiores na madrugada, consequentemente a taxa de crescimento das requisições no período da manhã é menos acentuada do que nos dias de semana. No período da tarde não existe uma flutuação tão grande quanto nos dias de semanas o que pode estar



(a) Dias Úteis (Seg - Sex)

(b) Finais de Semana

(c) Contraste entre os dois períodos

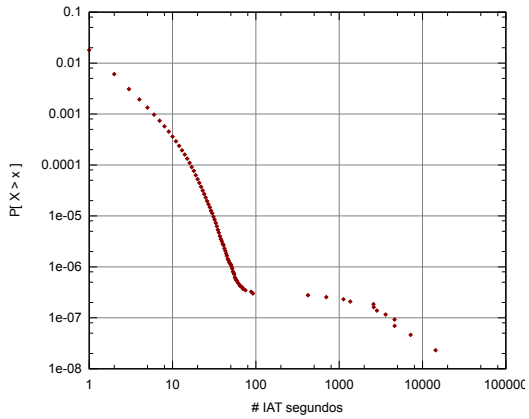
Figura 5.6. Contraste entre o número de requisições média (amostra por minuto) dos dias úteis com finais de semana durante todo o período estudado.

associado ao fato dos usuário não estarem com acesso à Internet, longe do trabalho ou de suas casas uma vez que a maioria dos acessos não são realizados por dispositivos móveis como veremos na Seção 5.3.8.

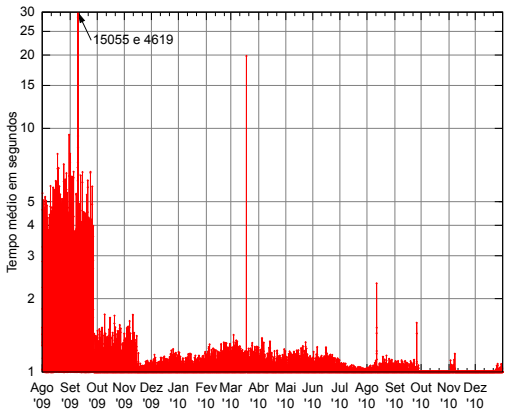
Se compararmos a distribuição das requisições por hora com a literatura vemos que o padrão comum reflete uma atividade diurna maior que a noturna, o que é esperado e intuitivo. Em [Gill et al., 2007] temos uma distribuição parecida com a nossa de dias úteis. Já em [Almeida et al., 2001] o padrão diverge. Nesse trabalho o padrão de uso está concentrado no período de 13h às 17h. Assim, podemos perceber que apesar do padrão predominante diurno o perfil de consumo varia de acordo com a aplicação e conteúdo fim.

As implicações desses padrões temporais ajudam os *Produtor de Conteúdo* a se planejarem com relação a horários de picos e outros horários que necessitem de estratégias para atrair mais usuários. Também ajuda na precificação e venda de publicidade. Por outro lado, os serviços de infraestrutura que provêm a entrega dos conteúdos podem se programar para melhores horários de manutenção assim como gerenciar a capacidade dos servidores.

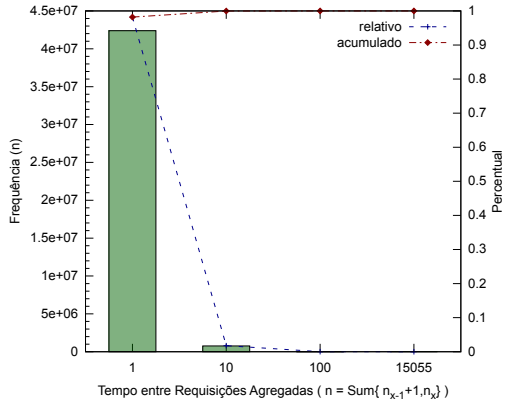
5.3.3 R_2: Tempo entre Chegadas de Requisições no Servidor



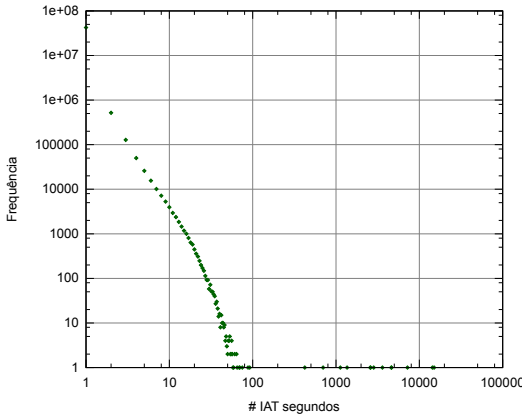
(a) Distribuição cumulativa complementar.



(b) Distribuição da média em uma dada hora do *IAT* por todo período.



(c) Histograma do *IAT*.



(d) Distribuição do *IAT*.

Figura 5.7. Informações sobre o tempo entre requisições (IAT) do período de estudo.

Em qualquer instante t de tempo uma requisição pode chegar no servidor. Seja $t(j)$ o instante de chegada da j-ésima requisição no histórico de acessos das requisições do servidor. Definimos o tempo entre chegadas de requisições IAT (*Inter Arrival Time*)

em um dado servidor como tempo decorrido entre duas requisições consecutivas, ou seja, $iat(j) = t(j+1) - t(j)$.

Através do tempo entre chegadas das requisições no servidor podemos avaliar o processo de chegada das requisições no servidor em diferentes intervalos em tempo. Essa análise possibilita o reconhecimento de períodos de alta e baixa atividade nos servidor.

Podemos ver um sumário das informações dos IAT na Figura 5.7. A função cumulativa complementar dos IAT pode ser vista no gráfico da Figura 5.7(a). Podemos ver que mais de 90% dos intervalos são de 1 segundo. Isso se deve ao fato da alta popularidade dos conteúdos da coleção avaliada e, principalmente, do fato da resolução temporal das requisições ser apenas de segundos o que limita nossa capacidade de análise. É interessante notar a popularidade de conteúdos multimídia tem aumentado. O tempo entre chegadas em servidores de mídia educacionais [Almeida et al., 2001] possuía tamanhos muito maiores, 90% se encontravam com IAT menores que 100s. Em [Cha et al., 2007] observou-se que aproximadamente 95% dos vídeos possuem um IAT maior que 10 minutos. Ou seja, somente 5% dos vídeos são frequentemente acessados. As implicações, segundo os autores, é que a distribuição de tais conteúdos pouco se beneficiaria de estratégias como p2p uma vez que a maioria dos arquivos não são frequentemente requisitados.

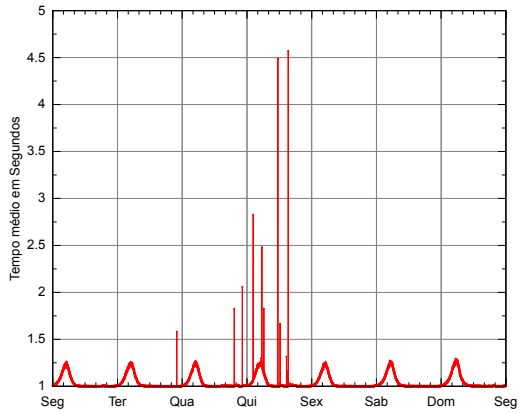
Os casos de tempo entre requisições maiores que 1000 segundos, acreditamos ser devidos a perda de registros de acessos uma vez que não há como ter certeza da completude destes *logs* devido a grande quantidades de dados. Outra possibilidade são manutenções preventivas por parte da equipe da empresa.

Vemos que o intervalo entre requisições é inversamente proporcional à popularidade do conteúdo sendo requisitado. Um comportamento intuitivamente óbvio e que pode ser confirmado na distribuição dos intervalos sobre todo o período de estudo da Figura 5.7(b). Nesta Figura existem dois casos atípicos de IAT por volta de setembro de 2009, 15055 segundos ($\pm$ 4 horas) e 4619 segundos ($\pm$ 1.5 hora). Como, segundo a empresa, não houve uma indisponibilidade desse tipo no período esses dois casos devem estar relacionados a uma perda das informações dos registros de acesso.

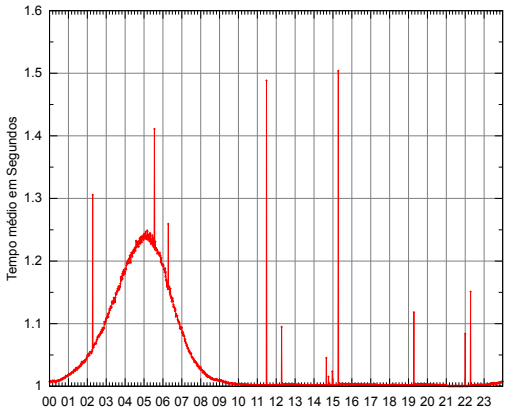
Os gráficos das Figuras 5.7(c) e 5.7(d) mostram, respectivamente, a distribuição dos intervalos agregados por tamanho destes e um histograma que agrega tais intervalos para apresentar de forma mais legível as proporções de tais intervalos.

Com o intuito de avaliar padrões comportamentais periódicos aplicamos a análise do IAT para diferentes períodos p de tempo. Após avaliação não foram detectados nenhum padrão comportamental periódico mensal, sendo assim, omitimos tal análise.

Na Figura 5.8(a) temos a análise do $iat(t \ mod \ p)$ onde p compreende uma semana.



(a) Tempo médio entre requisições visualizado por dias da semana



(b) Tempo médio entre requisições visualizado por dias da semana

Figura 5.8. Avaliação do IAT para diferentes períodos de tempo.

Os casos de exceções provocam distorções no gráfico, entretanto com menor intensidade. Como não existe nenhuma evidência forte de um padrão comportamental periódico mensal ou semanal podemos afirmar que, em um dia típico, o tempo entre as requisições para o sistema em estudo é descrito pelo gráfico da Figura 5.8(b).

Diferente do trabalho de [Veloso et al., 2006a], que encontrou uma distribuição Lognormal para o tempo entre chegadas nos servidores, processo de chegadas das requisições, avaliadas do ponto de vista somente das requisições, possui uma característica bimodal. É possível observar na Figura 5.7(a) que existe uma curva que descreve as chegadas até um IAT de 90s e outra curva que descreve as IATs com intervalos maiores. Acreditamos que a primeira descreve o comportamento normal de operação e a segunda pode se enquadrar como situações excepcionais, como falhas nos servidores ou sites de *Produtor de Conteúdo* ou perda de registros.

Por fim acreditamos que a resolução de segundo para estudo do processo de chegadas das requisições pode ser um fator limitante e encobrir características interessantes. Atualmente já existem alguns servidores (Nginx[7] por exemplo) que já registram as requisições de seus processos em resolução de milissegundos.

5.3.4 R_3: Tempo de Resposta do Servidor

Para cada requisição atendida existirá um período T_r, denominado tempo de resposta, que representa o tempo que servidor levou para atender tal requisição. O tempo de

[7] http://nginx.net/

resposta nos permite avaliar o desempenho do servidor e o impacto de tal variável nas características do tráfego gerado na rede e nos servidores.

A Figura 5.9 apresenta uma visão geral do tempo de resposta do servidor sobre todo o período de estudo. Os gráficos das Figuras 5.9(a) e 5.9(b) apresentam a *CDF* e a *CCDF* respectivamente. É importante observar que por questões de legibilidade os valores do eixo y foram restringido aos valores do domínio. Nestes gráficos vemos que 92.3% das requisições se encontram com tempo inferior a 1 segundo. Entretanto, existem também valores bem alto de tempo de resposta como 86116 segundos que é o tempo de resposta mais alto na coleção.

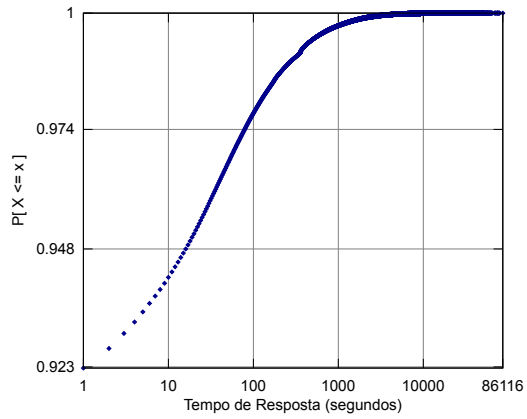
Para melhor comparar os intervalos do T_r, o gráfico da Figura 5.9(c) mostra a proporção de requisições segmentadas por T_r. Notem que $\sim$ 90% das requisições são da ordem de milissegundos, mas os registros só computam tempo na grandeza de segundos, o que limita a análise. Assim, tais registros ficam como se o tempo de resposta fosse imediato, ou seja, 0 segundos. Se a analisarmos somente requisições maiores que 1 segundo, existirá uma concentração maior no intervalo de 10 a 100 segundos. Tal ocorrência sugere um indício de requisições de objetos maiores ou então um performance comprometida dos servidores.

A distribuição do T_r pode ser vista na Figura 5.9(d). O que chama atenção neste gráfico é o pico no intervalo aproximado de 360 segundos. Acreditamos se tratar de um objeto (ou um grupo de objetos) particularmente popular que podem estar causando tal comportamento. Para analisar melhor este comportamento será necessário distinguir os diferentes tipos de objetos, o que veremos na seção 5.4.3.

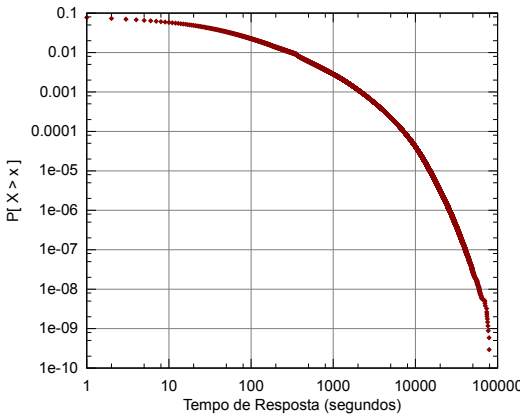
Para uma melhor visão do comportamento do tempo de resposta durante todo o período, a Figura 5.10 mostra a distribuição média e máxima agregada por hora durante todo o período de estudo. Notem que a escala é diferente para cada curva. O T_r médio está em segundos, enquanto que o T_r máximo está na escala de minutos. Tal abordagem foi adotada para facilitar a visualização do gráfico. Nota-se que em média o T_r é baixo, entretanto, durante o dia ocorrem valores muito altos para os T_r o que demanda análises mais adequadas (seção 5.4.3) para esclarecer tais valores.

Novamente, a análise de periodicidade mensal, $T_r(t \ mod \ p)$ onde p é um período mensal, não trouxe *insights* relevantes para a análise de eventos periódicos. Entretanto, apesar de mais de 90% das requisições possuírem tempo de resposta inferior a 1 segundo o tempo médio da distribuição mensal ficou em torno 15 segundos o que sugere um grande desvio devido aos T_r maiores.

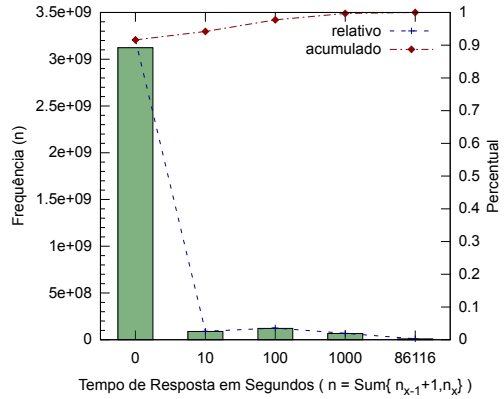
Com relação ao comportamento periódico semanal, a Figura 5.11 apresenta o gráfico da função $T_r(t \ mod \ p)$, onde p corresponde aos dias da semana. Nota-se um padrão de tempo médio de resposta ligeiramente similar durante todos os dias da



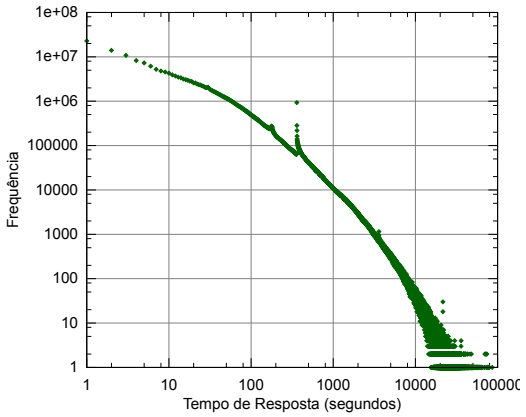
(a) Distribuição cumulativa



(b) Complemento da distribuição cumulativa



(c) Histograma do T_r



(d) Distribuição do T_r por duração.

Figura 5.9. Distribuiçao cumulativa do tempo de resposta das requisições e sua complementar.

semana. Entretanto, a variabilidade destes valores não permite inferir um padrão comportamental forte.

O comportamento dos dias da semana (segunda à sexta) segue o padrão do gráfico da Figura 5.12(a). O T_r médio varia de 10 a 21 segundos. O padrão é composto de três picos de T_r médio, um as 3h ($\sim$ 18s), um as 15h($\sim$ 16s) e outro as 21h ($\sim$ 16.5s). O período com menor T_r médio ocorre no intervalo de 6 às 8h da manhã. Apesar de similar, o comportamento do tempo de resposta médio para os período do final de semana (sábado e domingo) é mais estável. Através da Figura 5.12(b), observamos que as variações no T_r médio são menos acentuadas e nota-se que o padrão noturno do tempo médio de resposta é menos expressivo.

De uma forma geral, os valores para o tempo de resposta são muito amplos (Figura 5.9(d)). Mais de 90% das requisições são menores que 1 segundo mas ainda

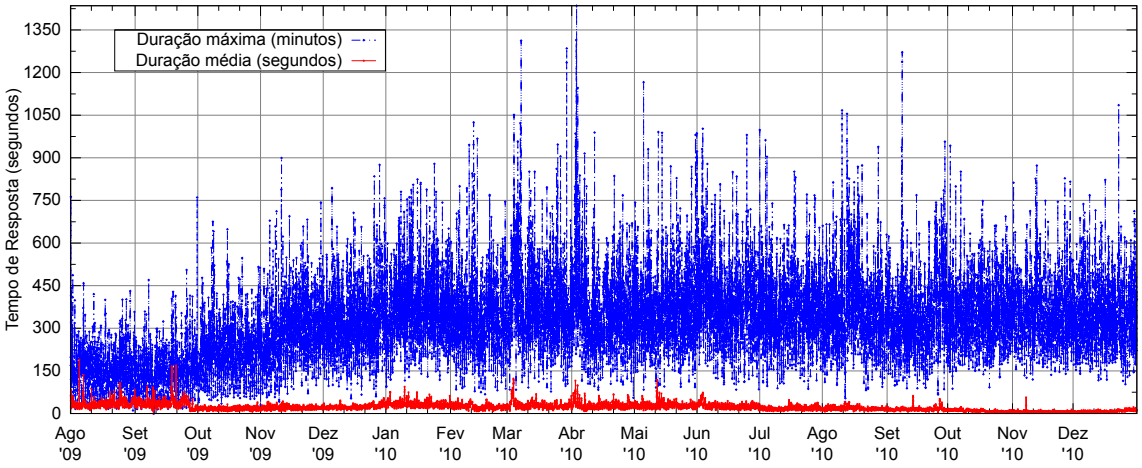


Figura 5.10. Tempo médio (em segundos) e máximo (em **minutos**) do T_r das requisições por todo período analisado.

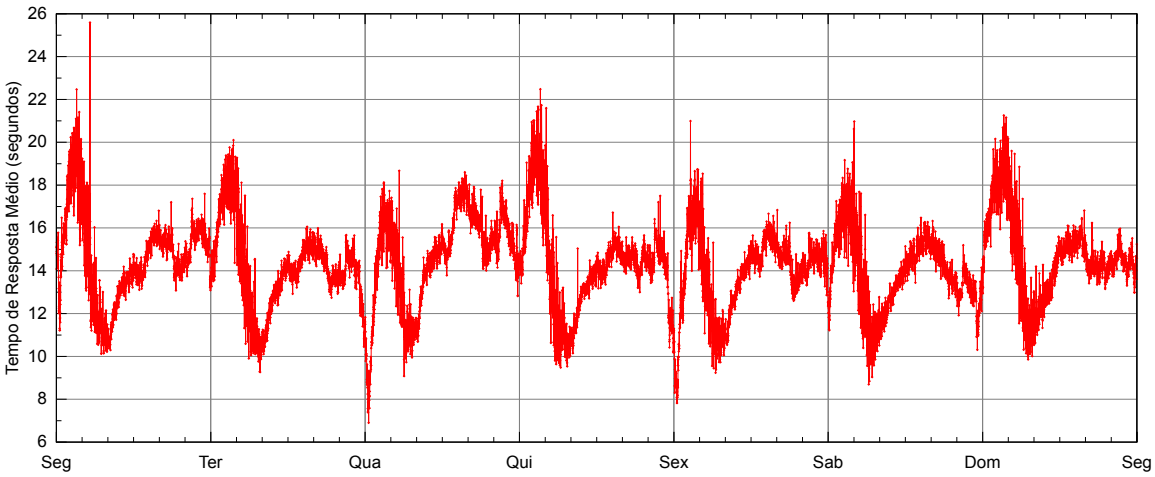
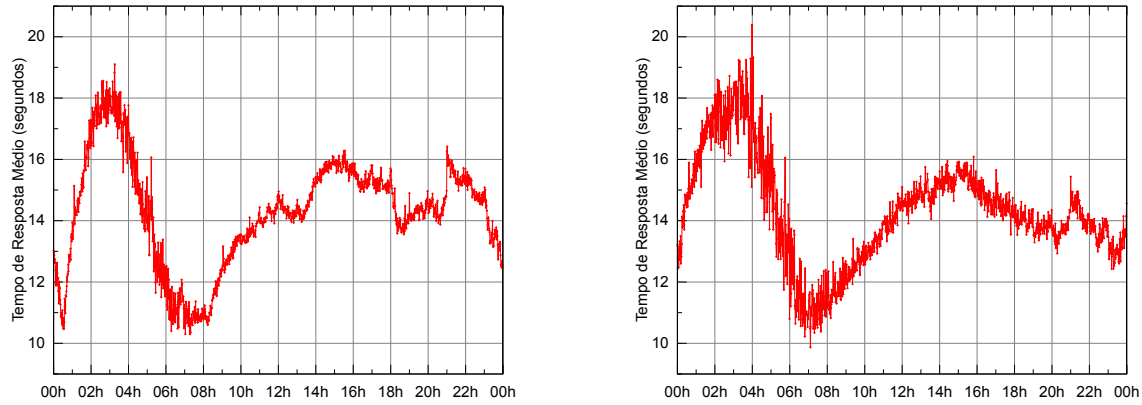


Figura 5.11. T_r médio em segundos por dia da semana

assim existem requisições que demoraram quase 24 horas para se completar. Os padrões temporais para diversos intervalos não mostraram-se forte. Tais evidências sugerem que o universo de objetos requisitados sejam analisados em grupos segmentados para prover melhor compreensão do dados.

5.3.5 R_4: Largura de Banda das Requisições

A largura de banda é uma medida da taxa de bits que são consumidos/fornecidos em um canal digital de comunicação. Usualmente expressa em kilobits por segundo (kbps), a largura de banda é também uma medida de capacidade de transmissão das interfaces de redes do sistema.



(a) T_r para dias de semana

(b) T_r para finais de semana

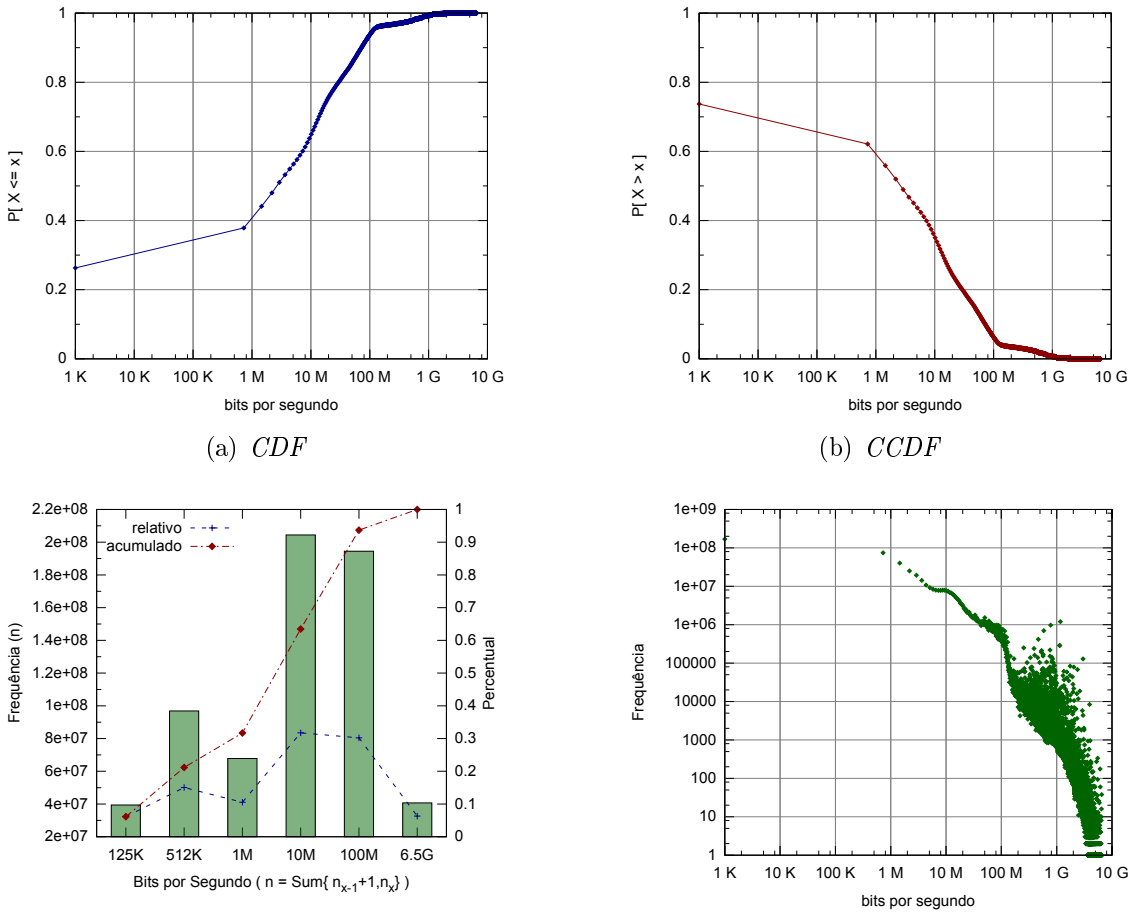
Figura 5.12. Tempo de resposta por período da semana.

Os *logs* de acesso fornecem informações relativa à quantidade de dados trafegado em uma requisição e o tempo gasto pelo servidor escrevendo as informações para o cliente. Tais informações nos permite calcular um valor aproximado para largura de banda média consumida nas transferências das requisições. Os gráficos da Figura 5.13 mostram as informações utilizando essa aproximação.

As funções cumulativas da largura de banda e sua complementar podem ser vistas respectivamente nas Figuras 5.13(a) e 5.13(b). Em tais gráficos nota-se que mais de 60% das requisições possuem largura de banda compatível com velocidades de conexão de banda larga (1Mbps à 100 Mbps).

O histograma da Figura 5.13(c) mostra a frequência, o percentual acumulado e relativo da largura de banda, segmentados por intervalos pré-definidos. Ao analisar tais gráficos o que mais nos chama a atenção é a porcentagem significativa de valores muito altos. Cerca de 7% das requisições se encontram numa faixa de largura de banda superior 100 Mbps. Tal faixa está num limite de velocidade maior que as suportadas pela maior parte das interfaces de redes comumente utilizadas por usuários domésticos e corporativos (Fast Ethernet, *IEEE 802.3u*). Mesmo para valores menores que 100Mbps o número de transferências com largura de banda maior que 10Mbps é muito alta para os padrões de velocidade de conexão praticados no ano de 2010 no Brasil.

O problema dessas análises é que estamos avaliando a largura de banda de requisições separadamente e como temos limitações com o tempo de resposta não temos precisão na análise de requisições de baixo tempo de resposta. Dessa forma, tais análises merecem um investigação mais detalhada, separando requisições de tamanhos diferentes, que será realizada em outro nível da hierarquia de caracterização (Seção 5.4.4).



(a) *CDF*

(b) *CCDF*

(c) Histograma

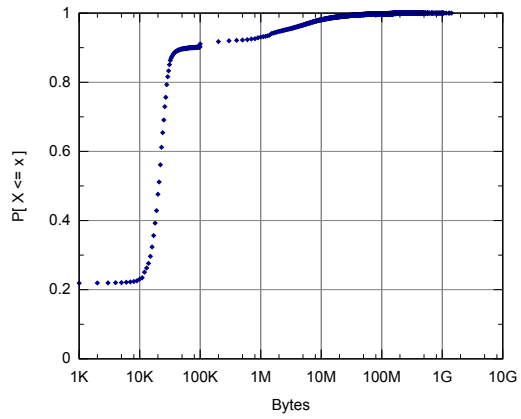
(d) Distribuição

Figura 5.13. Informações relativas a largura de banda das transferências

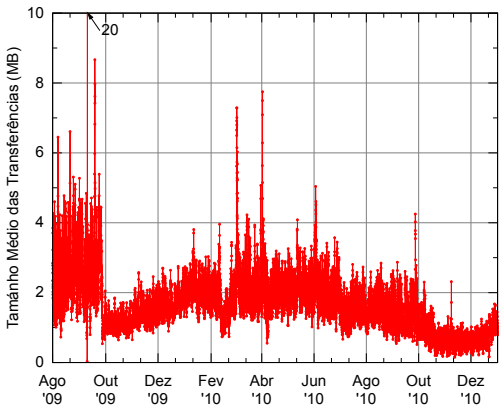
5.3.6 R_5: Tamanho das Transferências

A análise da distribuição do tamanho das transferências proporciona, de maneira limitada, o entendimento da distribuição dos tamanhos dos objetos sendo requisitados.

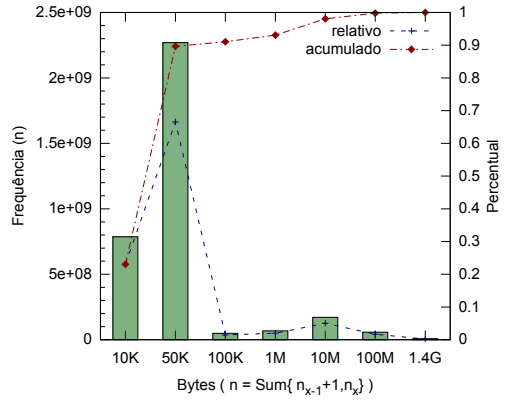
A função de distribuição cumulativa pode ser vista no gráfico da Figura 5.14(a). Em tal distribuição vemos que a variedade da quantidade de dados trafegadas por requisição é ampla, variando de alguns kilobytes até 1.5 gigabytes de informação por requisição. Cerca de 20% das requisições são menores que 10KB e se o tamanho da requisição for aumentado para 100KB a proporção passa a ser de aproximadamente 90% do total de requisições no sistema. Tais informações podem ser vistas, de forma mais, clara no histograma da Figura 5.14(c). Nesta figura também é possível ver que, se descartarmos as pequenas transmissões menores 100KB, existe uma dominância de objetos de tamanho médio, entre 1 e 10 MB.



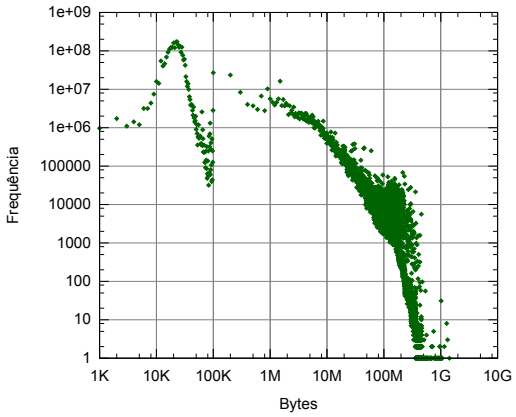
(a) *CDF*



(b) Tamanho médio por todo período.



(c) Histograma



(d) Distribuição

Figura 5.14. Sumário das informações relativas ao tamanho das transferências.

Uma visão geral da média da distribuição do tamanho das requisições pode ser visto na Figura 5.14(b). Nesta Figura foi realizado a média da quantidade de informação trafegada em uma hora, durante todo o período de estudo. Nota-se uma mudança abrupta no padrão da curva em Outubro de 2009, a média do tamanho das requisições passa a ser consideravelmente menor. Tal evento está associado a entrada de um *Produtor de Conteúdo* muito popular na plataforma em questão.

Analisando a distribuição do tamanho das requisições no gráfico da Figura 5.14(d) notamos que a variabilidade no tamanho das requisições de tamanhos maiores que 10MB é muito maior que a de tamanhos menores. Tal fato pode estar relacionado a maior variabilidade de objetos maiores ou a transferências incompletas destes. Tais conclusões só serão esclarecidas nas análises em camadas superiores.

Para a análise de padrões temporais, analisamos a periodicidade mensal, semanal e diária. Como todas as outras análises de periodicidade mensal, não foram encontradas

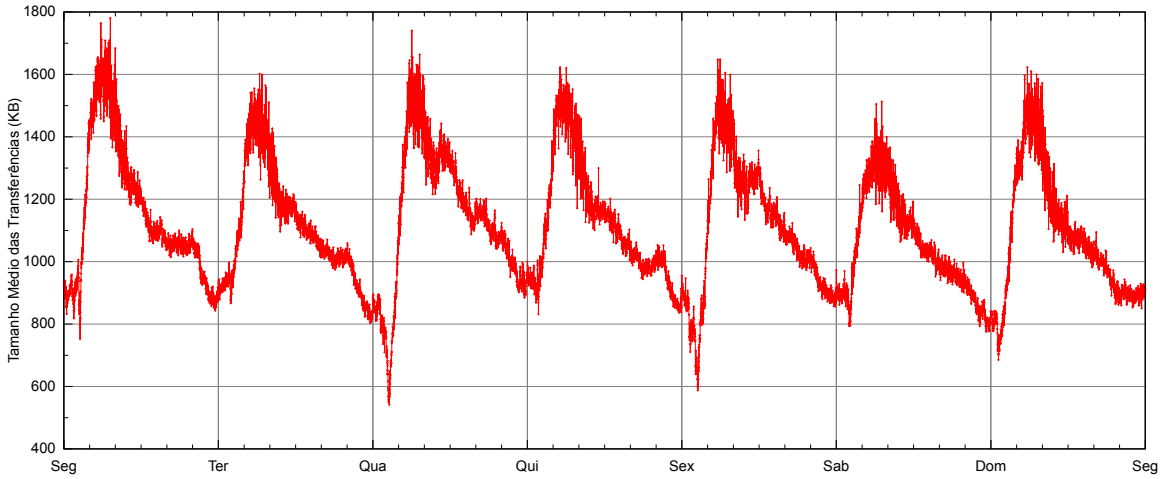


Figura 5.15. Tamanho médio (em kilobytes) das transferências por dia da semana

evidências de um padrão temporal mensal, somente um padrão diário.

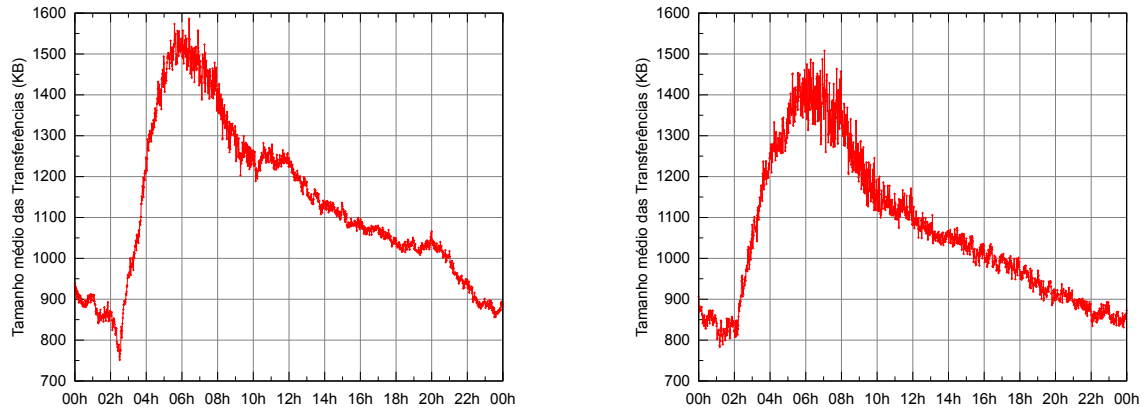
A Figura 5.15 mostra a análise para a periodicidade semanal. Para tal análise todos os dias possuem comportamento similar com pequenas variações, não existe algum padrão que diferencie algum dia especificamente.

De uma forma geral o comportamento temporal do tamanho médio das requisições sobre a plataforma da Samba Tech segue um padrão diário como podemos ver nos gráficos da Figura 5.16. Mesmo para os períodos de dia de semana e final de semana, que anteriormente apresentaram comportamento diferenciado, a diferença no padrão é muito sutil mostrando apenas que no período da manhã, que neste caso é o pico do tamanho médio, os valores são ligeiramente menores para o período do final de semana.

A análise do tamanho médio das transferências pode ser mais informativa. Como a variabilidade do tamanho das transferências é muito alta, a média de um conjunto deste tipo não agrega informações relevantes. Acreditamos que ao segmentar os diferentes tipos de objetos na camada de objetos (Seção 5.4.5), tal análise se mostrará mais relevante.

5.3.7 R_6: Análise dos Endereços de Origem das Requisições

Para cada requisição realizada na plataforma, existe um endereço IP associado a cada requisição. Utilizando o IP é possível identificar a região geográfica do agente que faz a requisição através de bibliotecas de terceiros. Nesta seção, iremos realizar um estudo das regiões geográficas de origem dos agentes que realizaram requisições no período de estudo para analisar a localidade geográfica das requisições em nossos servidores.



(a) Dias de semana

(b) Finais de semana

Figura 5.16. Tamanho das transferências por período da semana.

Para ilustrar a diversidade de IPs que acessam os conteúdos trafegados pelos vários *Produtores de Conteúdo*, a Figura 5.17 mostra a distribuição dos IPs únicos por dia que acessaram tais conteúdos. Notem que em períodos de muita atividade, registrou-se quase 500 mil IPs distintos acessando os conteúdos.

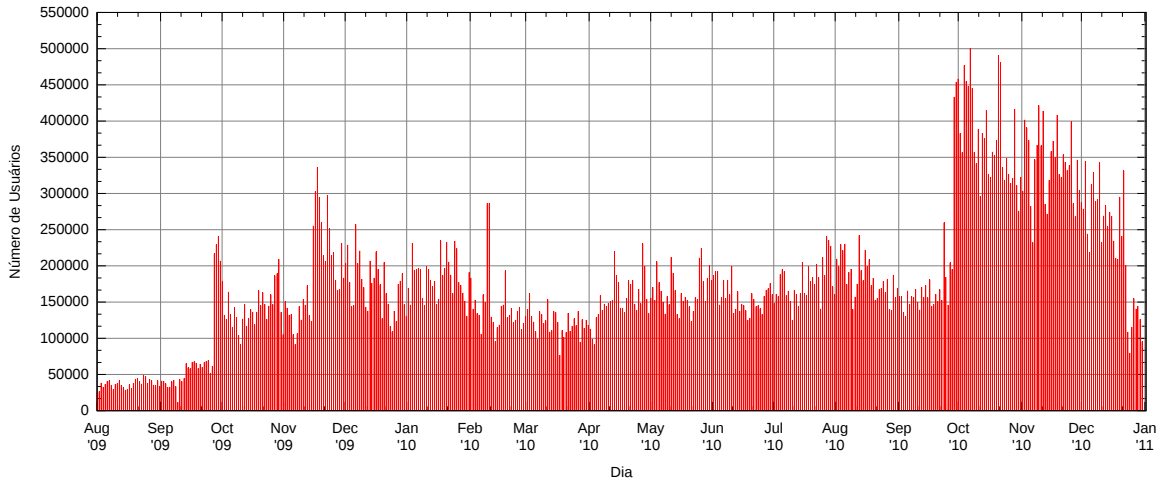


Figura 5.17. Números IPs únicos por dia no registros de acesso.

Utilizamos a biblioteca livre de *GeoIp* da MaxMind[8] para realizar a conversão de IP para localidade geográfica. Tal biblioteca permite a identificação do país, região (algo semelhante aos nossos estados) e cidade do IP de origem. Segundo o portal da empresa, a biblioteca possui uma acurácia média de 95% para a identificação de países e de 79% para a identificação de cidades.

[8]http://www.maxmind.com/app/geolitecity

A Tabela 5.6 apresenta uma síntese dos endereços de origem que acessaram a plataforma no período de estudo. Temos um total de 219 países, 1998 regiões distintas e mais de 40 mil cidades. Nesta Tabela também mostramos a cobertura do alcance da plataforma no Brasil. Existe aos cobertura de todos os estados brasileiros (26 estados mais Distrito Federal) e 1145 cidades.

Países Distintos	219	**Regiões do Brasil**	
Regiões Distintas	1.998	**Estados**	27
Cidades Distintas	40.617	**Cidades Distintas**	1.145

Tabela 5.6. Alcance geográfico da plataforma

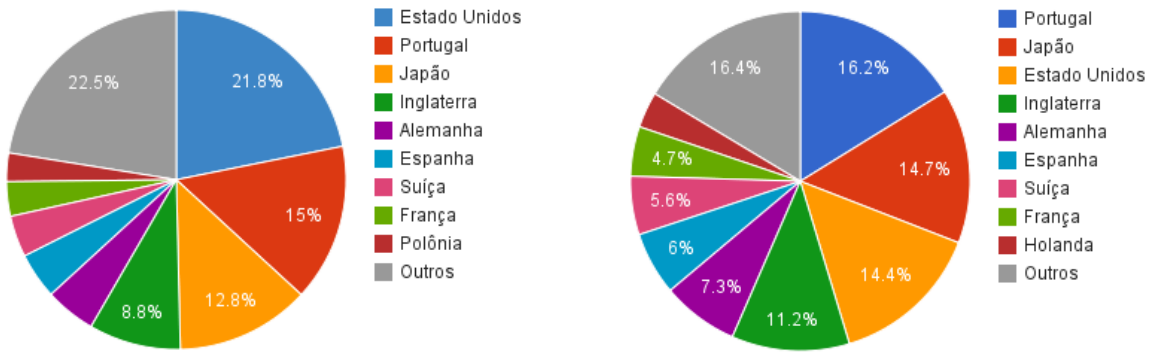
Realizamos um estudo contrastando os acessos realizados no Brasil versus os acesso de outros países. O objetivo era avaliar a penetração em demais países. A expectativa era de que a penetração fosse baixa uma vez que todos os conteúdos são exclusivamente em português, o que impede a ampla disseminação em países de outras línguas. A Tabela 5.7 contrasta o tráfego, e requisições do Brasil com demais países. Temos que 5% das requisições são originadas de países estrangeiros, tais requisições correspondem por aproximadamente 9% de todo o tráfego. Uma fração representativa dada a barreira da língua dos conteúdos. Cabe ressaltar também, que a porcentagem poderia ser ainda maior dado o grande número de requisições bloqueadas por restrição de *copyright* (Tabela 5.5)

Item	**Req.** ($\times 10^6$)	**% das Req.**	**Tráfego (TB)**	**% do Tráfego**
Brasil	281,71	94,92%	3.090,70	91,08%
Outros Países	15,08	5,08%	302,71	8,92%

Tabela 5.7. Sumário comparativo entre o uso do Brasil

Para a melhor análise da origem das requisições estrangeiras, segmentamos tais requisições por países e montamos o gráfico da Figura 5.18(a). Note que os principais países que mais requisitam os conteúdos são Estados Unidos, Portugal, Japão e Inglaterra. Entretanto, na Figura 5.18(b), os Estados Unidos não mantém a mesma proporção de dados trafegados. Acreditamos que esse fato isolado ocorre devido ao grande número de *bots* de origem dos IPs americanos.

Continuando a análise entre acessos domésticos e estrangeiros, analisamos a distribuição dos acessos entre as regiões brasileiras e estrangeiras. O conceito de *região* da biblioteca é um conceito pouco restritivo pois necessita lidar com diferentes subdivisões



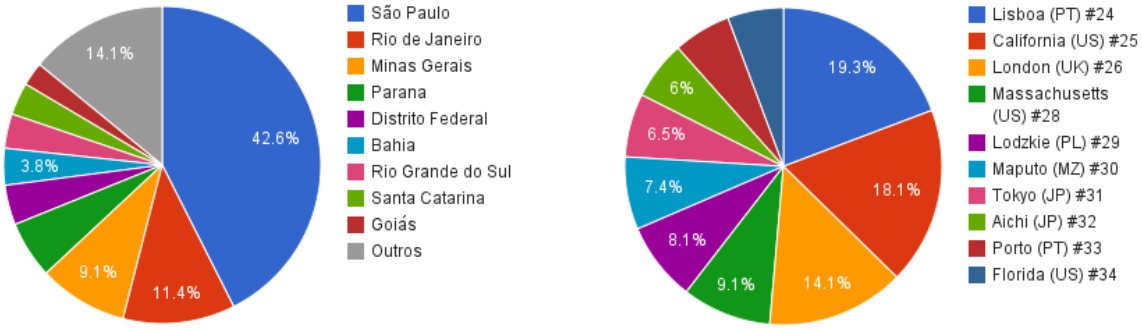
(a) Requisições de outros países

(b) Tráfego de outros países

Figura 5.18. Sumário do uso da plataforma distribuído por países

territoriais de vários países. Para não perder a generalidade, iremos considerar que tais regiões, quando se tratar de Brasil, estão relacionadas ao Estados Brasileiros.

A Figura 5.19(a) mostra a subdivisões das requisições entre os estados Brasileiros (não representamos a distribuição do tráfego pois a proporção é a mesma). O Estado de São Paulo é o detentor da maior parte do acesso, com cerca de 43% das requisições, seguido por Rio de Janeiro (11.4%) e Minas Gerais (9.1%). Se levarmos em consideração a ordem das regiões mais populares, as 23 regiões mais populares são estados brasileiros, seguidas das regiões da Figura 5.19(b).



(a) Requisições dos estados brasileiros

(b) Requisições de outras regiões do mundo.

Figura 5.19. Divisão do acesso por estados do Brasil e regiões do mundo

Interessante notar que das 10 regiões estrangeiras mais populares somente 3 destas são regiões onde se falam português. Acreditamos que tais regiões, mesmo as de língua portuguesa, abriguem uma comunidade expressiva de imigrantes brasileiros.

Uma análise estatística dos usuários do YouTube pode ser vista

em [Duarte et al., 2007a]. Diferente do presente trabalho, o foco é em usuários da América Latina. Com a exceção do Brasil nenhum dos países analisados em [Duarte et al., 2007a] se encontram em nosso estudo.

A Tabela 5.8 mostra uma análise detalhada do acesso das regiões brasileiras. A primeira coluna referencia o estado, a segunda (*Rank.*) mostra a ordem por popularidade de acesso do estado brasileiro considerando uma ordenação global, a terceira coluna é o número de cidades distintas do estado, que requisitaram algum conteúdo no período analisado. As duas colunas que seguem são respectivamente a porcentagem das requisições e tráfego global que é devido a tal estado e a última coluna, é a ordem de popularidade da capital de tal estado na *ranking* global por cidades.

Estado	Rank.	# Cidades	% das Req	% do Traf.	Rank. Capital
São Paulo	1	302	42.56%	45.80%	1
Rio de Jan.	2	67	11.42%	10.51%	2
Minas Gerais	3	159	9.07%	7.98%	3
Paraná	4	88	5.83%	6.21%	6
Dist. Federal	5	8	4.06%	4.24%	4
Bahia	6	52	3.75%	3.75%	5
Rio Gr. do Sul	7	102	3.51%	3.27%	11
Santa Catarina	8	80	3.30%	2.96%	18
Goias	9	28	2.42%	2.41%	9
Ceara	10	27	2.22%	2.07%	7
Pernambuco	11	31	2.02%	2.02%	10
Esp. Santo	12	26	1.92%	1.96%	17
Para	13	17	1.11%	0.95%	13
Paraíba	14	18	0.93%	0.83%	24
Mato Grosso	15	19	0.88%	0.81%	23
Mato G. do Sul	16	19	0.75%	0.70%	30
Rio G. do Norte	17	16	0.69%	0.58%	26
Alagoas	18	13	0.65%	0.57%	28
Maranhão	19	15	0.63%	0.53%	29
Tocantins	20	11	0.54%	0.50%	58
Sergipe	21	9	0.51%	0.43%	32
Piauí	22	11	0.51%	0.39%	33
Amazonas	23	11	0.41%	0.27%	39
Rondônia	**27**	9	0.23%	0.20%	68
Acre	**45**	1	0.06%	0.05%	194
Amapá	**88**	2	0.02%	0.01%	423
Roraima	**91**	4	0.02%	0.01%	534

Tabela 5.8. Distribuição da utilização por estados brasileiros

É importante notar a dominância do estado de São Paulo. Sozinho, ele é responsável por 42,6% de todo o tráfego da plataforma. Contribui com 302 cidades e a sua capital é a cidade que também mais requisita conteúdos. Nos 23 primeiros estados, por ordem de popularidade, não há muita surpresa, apenas que nem sempre a capital do

estado acompanha o ranking do seu respectivo estado. Note que o Distrito Federal é contado nesta tabela como estado. Apesar de não ser nem estado nem município possui, nesta Tabela, 8 cidades relacionadas. Na realidade estas cidades são as aglomerações urbanas denominadas regiões administrativas (ou cidades-satélites).

Um fato interessante que pode-se inferir da Tabela 5.8 é a discrepância de alguns estados com relação aos demais. Por exemplo, os estados da região Norte do Brasil possuem uma penetração muito inferior do que a de diversas cidades fora do País. O Estado de Rondônia possui uma audiência menor do que da região de Lisboa e da Califórnia. Em situações mais extremas se encontram o Acre, Amapá e Roraima. Tais regiões possui audiência muito menor que outras regiões de outros países, sendo que o ranking de suas capitais é praticamente desprezível. Acreditamos que tais evidências se devem a baixa inclusão digital em tais regiões por falta de infra-estrutura e e investimentos nas áreas de telecomunicações.

Item	Req. $\times 10^6$	Tráf.(TB)	% Req.	% do Tráf.	# Cidades
Capitais	142.33	1620.30	57.40%	58.18%	27
Interior	105.65	1164.69	42.60%	41.82%	1118

Tabela 5.9. Comparação entre o uso das capitais e cidades do interior

A Tabela 5.9 realiza uma comparação entre o tráfego doméstico brasileiro contrastando os acessos oriundos das capitais versus os originados nas cidades do interior. Podemos dizer que, atualmente, as capitais possuem o mesmo peso de acesso do que as cidades do interior do país. Não temos acesso a estudos similares, mas acreditamos que num passado recente tal diferença deveria ser maior, com as capitais dominando o acesso devido a disponibilidade de infra-estrutura de acesso. Acreditamos que em um breve período essa relação se inverta pois os recursos estarão cada vez mais disponíveis com os planos de inclusão digital do governo federal.

Top 10 Cidades do Brasil				Top 10 Cidades de Outros Países			
Cidade	**Req** $x10^6$	**Traf. TB**	**# R.**	**Cidade**	**Req** $x10^6$	**Traf. TB**	**# R.**
São Paulo	50.10	618.90	1	**London (UK)**	0.65	13.94	49
Rio De Jan.	21.06	219.71	2	**San Jose (US)**	0.62	2.38	52
Belo Horiz.	11.00	116.79	3	**Lisbon (PT)**	0.46	10.26	73
Brasília	9.42	110.84	4	**Lodz (PO)**	0.37	0.22	96
Salvador	7.41	85.33	5	**Maputo (MZ)**	0.34	2.87	106
Curitiba	7.13	85.87	6	**Tokyo (JP)**	0.27	3.81	121
Fortaleza	5.02	53.32	7	**Luanda (AO)**	0.26	1.75	125
Campinas	4.83	46.95	8	**B.Aires(AR)**	0.17	1.39	186
Goiânia	4.71	54.27	9	**Porto (PT)**	0.11	2.42	235
Recife	3.81	43.75	10	**Madrid (SP)**	0.11	2.85	250

Tabela 5.10. Relação entre cidades brasileiras e estrangeiras com maior acesso à plataforma

A Tabela 5.10 mostras as top 10 cidades nacionais e estrangeiras, as respectivas porcentagens de tráfego e requisições além do *ranking* entre as cidades. Notamos que nas top 10 cidades brasileiras encontramos somente uma cidade que não é uma capital, a cidade de Campinas. Com relação as cidades estrangeiras temos a presença de cidades de países que não possuem tanta expressão separadamente como Angola, Argentina e Espanha.

Para analisar a distribuição do popularidade das cidades com relação ao quanto uma cidade acessa os conteúdos da rede de distribuição da Samba Tech fizemos uma análise de pareto que pode ser vista na Figura 5.20. Nesta figura constatamos que existe uma relação de pareto 90-10, onde 90% das requisições são originadas de apenas 10% das cidades.

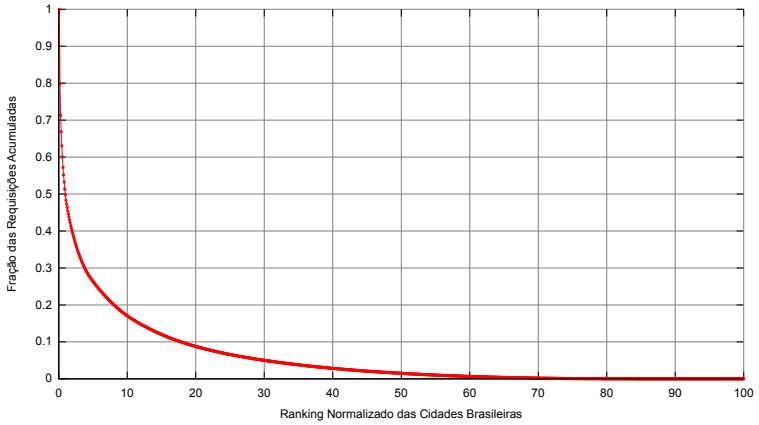


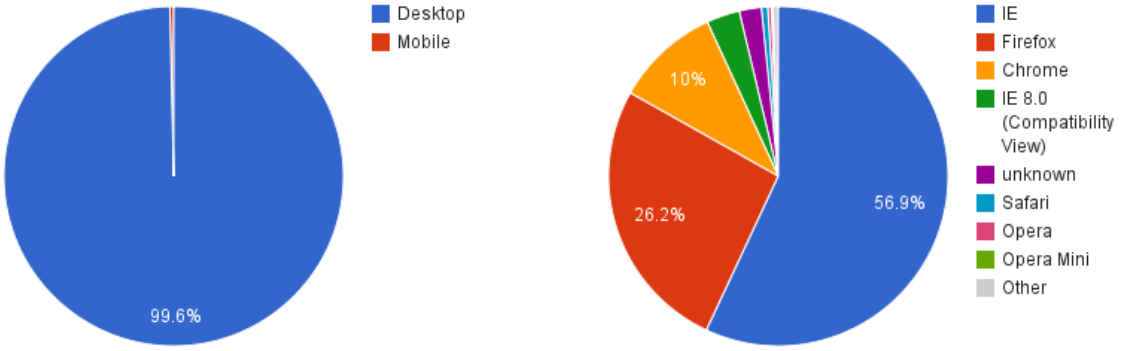
Figura 5.20. Assimetria do alcance do conteúdo nas cidades brasileiras

5.3.8 R_7: Análise dos Agentes de Origem das Requisições

Nesta seção iremos abordar a forma com que as requisições são realizadas na plataforma de estudo. Tal acesso por ser realizado por um usuário acessando um navegador *Web*, ou robô (*bot*) através de vários programas com diversos fins específicos ou até mesmo das formas de mais *ad-hoc* como agentes que monitoram o sistemas, scripts maliciosos, etc.

O objetivo é que possamos identificar quais sistemas operacionais estão utilizando e quais Navegadores (se algum) mais utilizados. Tal análise é possível pois para cada requisição temos os registros do cabeçalho HTTP *User Agent* que informa diversas informações como qual sistema operacional, se é um navegador ou um *bot* além de outras informações como versão, família do SO e navegadores. Sabemos que tal infor-

mação pode ser manipulada e modificada, mas acreditamos que a fração destes dados alterados, se existente é desprezível.



(a) Acesso móvel x Desktop (b) *User Agent*

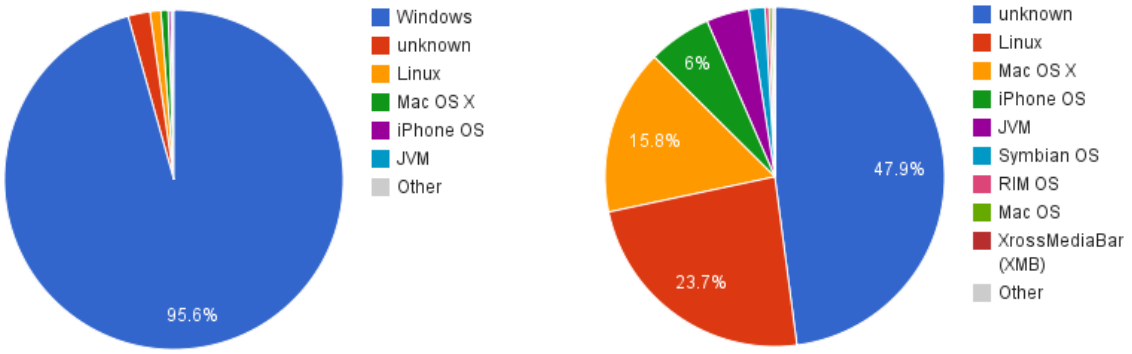
Figura 5.21. Distribuição das formas de acesso.

A Figura 5.21 mostra a distribuição das formas de acesso. A comparação entre dispositivos móveis e desktops pode ser vista na Figura 5.21(a). Vemos que utilização de dispositivos móveis para acessar os conteúdos é extremamente incipiente, menos de 1% das requisições na plataforma utilizam tais dispositivos para acessar o conteúdo. Isso se deve ao despreparo dos sites dos *Produtores de Conteúdo* de produzirem seus sites adequadamente para as necessidades de dispositivos móveis e também à inadequação das tecnologias da web tradicional como tamanho de imagem, *codecs* de vídeos etc.

Com relação aos navegadores/agentes de acesso a Figura 5.21(b) mostra a distribuição dos navegadores. Podemos ver que o *Internet Explorer* domina a forma de acesso com quase 57% das totalidades da forma de acesso, seguido pelo *Firefox* com 26% e *Chrome* com 10% do total.

A Figura 5.22 mostra informações relativas ao Sistema Operacional utilizado pelo agente de acesso.

Conforme esperado, o sistema operacional predominante com 95.6% é o *Windows*, sendo que o sistema operacional em segundo lugar é o Linux com apenas 1%, seguido do Max OS X com 0.7%. Note que a plataforma do segundo colocado com 2.1% são agentes em que não é possível associar um SO por se tratarem de agentes autônomos de acesso, comumente chamado de *bots* e/ou da incapacidade da biblioteca utilizada de identificar o SO de origem.



(a) Sistemas operacionais (b) SO menos windows

Figura 5.22. Distribuição dos sistemas operacionais hospedeiro dos agentes de acesso.

5.3.9 Conclusões da Camada R

Nesta subseção avaliamos vários aspectos da plataforma de distribuição da Samba Tech sobre o ponto de vista das requisições. A análise desta camada mostrou alguns *insights* interessantes para o planejamento de capacidade, uso de rede, e períodos mais propícios para a realização de tarefas que podem causar algum impacto ao usuário final.

Do ponto de vista de origem dos acessos vimos que cerca de 9% do tráfego da rede tem origem internacional. Avaliamos a segmentação da origem dos acessos através dos diferentes estados e municípios e mostramos que 90% das requisições são originadas a partir de somente 17,81% das cidades.

Nas análises desse nível da hierarquia não houve indícios ou evidências de algum padrão de comportamento com periodicidade mensal. Todas as análises para tal período somente indicam uma tendencia diária de acesso. Acreditamos que isso ocorre por se tratar de um período de análise muito extenso (quase 17 meses). Dessa forma, situações atípicas de acesso podem distorcer as análises sobre o ponto de vista de picos máximos em análises periódicas como mensais ou semanais.

Em vários casos, as análises sob o ponto de vista somente das requisições limitam a capacidade de interpretar corretamente os dados e esclarecer o porque de alguns comportamentos. As poucas evidências limitam a possibilidade de concluir o motivo de vários comportamentos. Isso porque as análises isoladas nesta camada não são informativas o suficiente para se compreender completamente todos os fenômenos observados assim, é preciso especializar melhor as análises. Tal fato fortalece a adequação da nossa abordagem metodológica e a medida que trabalharemos nas camadas seguintes da hierarquia, revisitaremos alguns pontos que ficaram abertos nesta seção.

5.4 Análises da Camada dos Objetos (O)

Abordaremos nesta camada os dados da plataforma de distribuição de conteúdo do ponto de vista dos objetos sendo requisitados. Nesta camada da hierarquia pode-se repetir as mesmas análises da camada R porém, segmentando pelos diferentes objetos em estudo.

Inicialmente iremos avaliar o domínio de estudo realizando um sumário das informações relativas a esta camada. Em seguida iremos rever **algumas** das análises da Camada R segmentando tais análises pelos diferentes tipos de objetos. Alguns estudos não serão revisitados uma vez que já sabemos que uma segmentação por tipo de objeto não acrescentará nenhuma informação nova ou melhorará a capacidade de análise sobre os dados. Finalizaremos esta seção com as novas análises pertinente a esta camada apresentadas na metodologia da Seção 3.3.

5.4.1 Sumário das informações sobre a camada de Objetos

A Samba Tech provê uma plataforma de gerenciamento de conteúdos multimídias, a proporção do tipos de objetos na plataforma pode ser vista na Tabela 5.11. Podemos ver que a empresa administra uma amostra expressiva de objetos multimídia sendo seu enfoque principal a gestão de vídeos.

Apesar do Thumbnail ser apenas uma imagem representativa de um dado conteúdo, como por exemplo, um vídeo, imagem ou até mesmo um áudio, iremos considerar, nesse nível da hierarquia, como um objeto autônomo. Lembrando que na camada **O** sabemos apenas que existem objetos distintos, mas não há um significado semântico para eles como é o caso do Thumbnail, uma espécie de metadado de outra mídia.

Tipo de Objeto	(#)	(%)
THUMBNAIL	239.805	59.70%
VÍDEO	188.574	37.93%
IMAGE	6.733	1.68%
ÁUDIO	2.778	0.69%
TOTAL	401.664	100.00%

Tabela 5.11. Distribuição de tipos de objetos na **base de dados**.

Ressaltamos que nesta etapa os thumbnails se diferenciam das imagens, apesar de serem um imagem. Esse fato se deve de propriedades diferentes como uso, relacionamentos com o vídeos e dimensões peculiares . A predominância de thumbnails

na Tabela 5.11 é devido a uma particularidade da base de dados fornecida pela Samba Tech. Como o thumbnail é uma mídia associada a outro objeto, a relação em geral é de 1 para 1. Entretanto, existe um conceito de *media folder* na plataforma que permite associar vários formatos de objetos diferentes (*media file*) sobre o mesmo arquivo lógico. Por exemplo, um vídeo v (*media folder*) pode ter várias versões do mesmo vídeo com diversos formatos (*media file*) como alta resolução, baixa resolução, formato para dispositivos móveis, etc. Para a Tabela 5.11 consideramos cada vídeo, imagem ou áudio lógico como apenas um objeto mas não fizemos a mesma segmentação para os thumbnails. Dessa forma um único vídeo pode ter mais de 1 thumbnail na contagem desta Tabela.

Com relação a distribuição dos tipos de objetos nos registros de acessos a Tabela 5.12 apresenta a divisão destes. Note que a diferença entre a Tabela 5.11 e a Tabela 5.12 são a origem da análise, na primeira se avalia os dados existentes nos sistemas dos *Produtor de Conteúdo*, já na ultima são os registros que de fato tiveram acessos. A primeira coisa que necessita de explicação é a presença de um número maior de imagens nos logs do que existente na base de dados. Isso ocorre devido a mudança da modelagem da base de dados no início de 2010 o que separou logicamente os thumbnails de imagens. Antes desse período não existia a distinção lógica, assim, a maior parte das imagens contabilizadas nesta tabela são thumbnails. Notem que o mesmo ocorre para os vídeos, mas a justificativa para esse caso é devido ao modelo de negócios da empresa. Como dito anteriormente, existem uma distinção entre *media folder* e *media file* e no caso desta tabela estamos considerando todos os vídeos como *media file*. Assim, é natural que exista esse número maior nesta Tabela.

Tipo de Objeto	(#)	(%)
ÁUDIO	2.018	0.54 %
IMAGE	14.377	3.88 %
VIDEO	160.361	43.23 %
THUMBNAIL	194.219	52.53 %
TOTAL	370.975	100.00%

Tabela 5.12. Distribuição de tipos de objetos nos **registros de acesso**

A Tabela 5.13 mostra a proporção de tráfego e requisições que são devidos a cada tipo de objeto. Os dois tipos de objetos predominantes são os thumbnails com relação as requisições e os vídeos com relação ao tráfego. Dado a natureza desses dois objetos era de se esperar este comportamento. Como os vídeos são objetos naturalmente maiores

que os outros, não é surpresa que estes dominam o tráfego. E sendo os thumbnails a representação estática dos demais objetos, amplamente usados em páginas iniciais de portais e em sistemas de recomendação, também é natural que estes possuam mais requisições.

Item	Thumbnail	Imagem	Vídeo	Áudio	Total
Req. ($\times 10^6$)	2953.00	96.20	301.12	0.22	**3350.55**
Tráf. (GB)	56619.46	2972.90	3422523.38	4.63	**3482120.38**
% Req.	88.13%	2.87%	8.99%	0.01%	**100.00%**
%Tráf.	1.63%	0.09%	98.29%	0.00%	**100.00%**

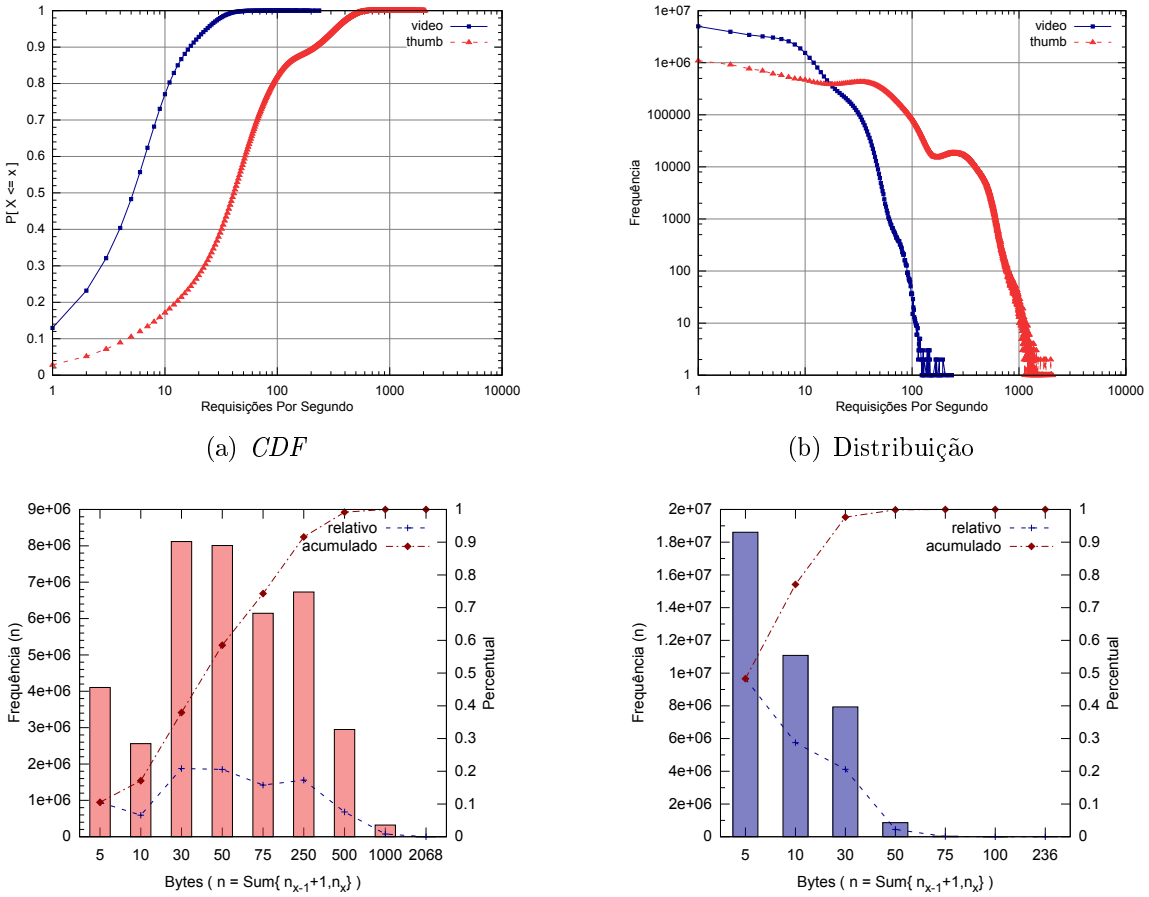
Tabela 5.13. Discriminação dos acessos por tipo de conteúdo

Como parte dos dados que estão contabilizados como imagem, são na realidade thumbnails não iremos estudar, nas demais camadas, o objeto Imagem pois não há um distinção clara com relação aos thumbnails, nem uma representatividade significativa na proporção do tráfego/requisições que generalize o comportamento do acesso fazendo válida as análises que serão realizadas. O mesmo argumento se aplica para áudio devido a baixa utilização deste na plataforma. Assim, nas próximas seções iremos realizar as análises dos **thumbnails** e **vídeos** para estudar os diferentes padrões que cada um desses tipos de objetos geram na plataforma.

5.4.2 O_1: Taxa de Resposta (*Throughput*) do Servidor por tipo de Objeto

Nesta seção revisitamos os estudos realizados na Seção 5.3.2 para realizar o mesmo estudo porém segmentando os dados de acordo com o objeto alvo das requisições. O objetivo de tal estudo é comparar a carga gerada no servidor para cada tipo de objeto e analisar o padrão de acesso para cada tipo de objeto. Tais análises podem ser úteis para o planejamento de capacidade e aprimoramento do sistema de entrega que pode ser otimizado somente nos pontos problemáticos caso existam.

Na Figura 5.23 encontramos algumas das análises segmentadas por tipo de objeto. As funções cumulativas podem ser vistas na Figura 5.23(a). Em tal gráfico vemos que a demanda por thumbnails é uma ordem de grandeza maior que a de vídeos, entretando a distribuição das requisições (Figura 5.23(b)) seguem o mesmo padrão. Podemos inferir que em média, para cada vídeo requisitado, 10 thumbnails são requisitados através das páginas de acesso ao conteúdo.



(a) *CDF*

(b) Distribuição

(c) Histograma thumb.

(d) Histograma vídeo

Figura 5.23. Informações gerais sobre *Throughput* por objeto.

Comparado ao trabalho de [Veloso et al., 2006a] temos nos dois trabalhos uma distribuição do throughput com variações de até um pouco mais de 2000 requisições por segundos. Para o nosso caso aproximadamente 65% da frequência das requisições são inferiores a 70 requisições simultâneas, já para o outro estudo são 400 requisições simultâneas. Se aumentarmos o limite para 90% temos no nosso um valor de 250 e para o estudo de [Veloso et al., 2006a] são aproximadamente 800.

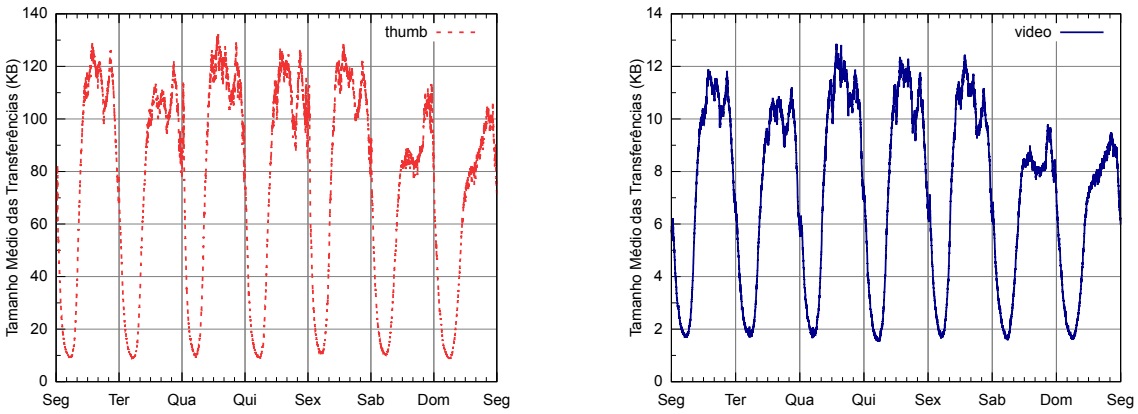
A Figura 5.23(c) apresenta o histograma da taxa de resposta para os Thumbnails. Se compararmos este gráfico e o da Figura 5.3(c) vemos que a maior parte das requisições no servidor é gerada por causa dos thumbnails e não outro objeto. Vemos que 80% das requisições estão na taxa entre $30 req./s$ e $250 req./s$.

O histograma mostrado na Figura 5.23(d) apresenta a distribuição de faixas de requisições por segundo exclusivamente para vídeos. Veja que a taxa de resposta é bem menor que a dos thumbnails como era de se esperar. Em mais de 90% do tempo a taxa

é inferior a 30 *req./s* sendo que taxas maiores que essas podem chegar a 236 *req./s*.

Era de se esperar uma taxa maior de requisições de thumbnails por ser este um recurso muito utilizado na composição das páginas dos *Produtores de Conteúdo* em geral. Outro ponto, é que um usuário visitando aleatoriamente as páginas web de um *Produtor de Conteúdo*, irá gerar várias requisições de thumbnails o que não acontece para o caso dos vídeos. É importante notar também que o vídeo possui uma duração não nula, o que demanda um tempo de interação com o usuário diminuindo assim a taxa de resposta.

Nos gráficos da Figura 5.24 apresentamos o padrão temporal semanal para *throughput* médio agregado em períodos de 1 minuto. O padrão relativo aos thumbnails e aos vídeos podem ser vistos nas Figura 5.24(a) e 5.24(b), respectivamente. Podemos ver em tais gráficos que a distribuição das requisições tanto para vídeos quanto para thumbnails seguem a mesma distribuição, porém em proporções diferente. Podemos dizer que existe uma relação aproximada de 10:1 requisições entre thumbnails para vídeos.



(a) Thumbnail

(b) Vídeo

Figura 5.24. Padrão temporal semanal da taxa de resposta por objeto.

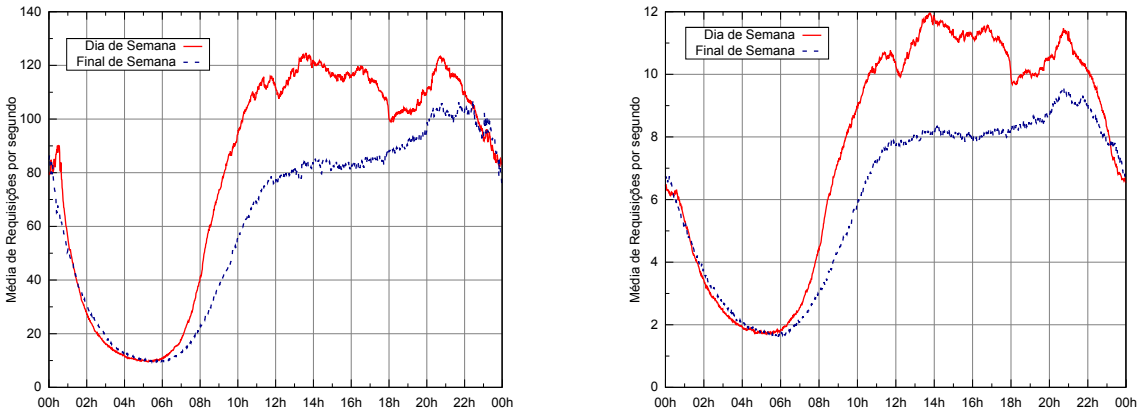
É interessante notar que em trabalhos anteriores [Gill et al., 2007, Veloso et al., 2006a] existem padrões similares de requisições de vídeos por dia da semana. Uma queda de requisições nos finais de semana e uma intensificação entre quarta, quinta e sexta-feira.

Com relação ao padrão semanal vemos que o padrão diário durante os dias úteis de semana são bem similares. Existe uma queda das requisições nas primeiras horas do dia seguida atingindo seu valor mínimo por volta das 4/5 horas da manhã seguida do crescimento matinal das requisições que irá atingir o primeiro pico máximo por volta

das 14 horas, seguido de uma queda brusca por volta das 18h e o último pico máximo as 21h.

Nota-se que os dias da semana são os que mais geram requisições sendo segunda e terça-feira os dias com menores tráfegos, sendo o último o menor de todos. O throughput do final de semana cai consideravelmente sendo que no sábado existem dois picos de acesso por volta das 14h e 21h e no domingo o há um pico máximo ocorrendo novamente as 21h.

Na Figura 5.25 apresentamos o padrão temporal semanal para *throughput* médio agregado em períodos de 1 minuto. Nos gráficos da Figura 5.25(a) e 5.25(b) temos o padrão diário médio para os dias de semana e também o padrão médio para sábado e domingo dos thumbnails e vídeos nessa ordem. Pela forma como os thumbnail se relaciona com o vídeo, ou seja, como um metadado visual, e de se esperar que exista alguma correlação entre eles. Note que a distribuição visualmente similar dos dois tipos de objetos na Figura 5.25 é outro indício dessa correlação.



(a) Thumbnail (b) Vídeo

Figura 5.25. Padrão temporal diário da taxa de resposta por objeto.

No trabalho [Gill et al., 2007] o padrão de acesso a vídeos por hora do dia é um pouco diferente. O pico de acesso é realmente por volta das 15h, entretanto, a partir desse horário segue-se uma progressiva queda até atingir valores mínimos as 5h da manhã. Comparando os valores da Figura 5.25(b) com o trabalho de [Gill et al., 2007] vemos que temos um padrão noturno que não foi caracterizado anteriormente.

O foco básico deste estudo de caso são páginas de conteúdo de vídeo. Atualmente tais portais são montados de forma que vários thumbnails sejam a uma das chamadas para o conteúdo em si. Sendo assim é razoável afirmar que existe uma relação entre o número de vídeos requisitados e o número de thumbnails requisitados. O

comportamento das curvas de requisições de tais objetos devem ser similares, ou seja, seguir o mesmo padrão. Se em algum momento o comportamento das duas divergirem consequentemente a proporção será maior ou menor dependendo do caso.

Para o caso desse estudo de caso, é natural que se espere, sempre, número maior de requisições de thumbnails do que vídeo, pois os sites são montados de forma a conter vários thumbnails de diferentes vídeos por página. Por exemplo, em um site de um vídeo qualquer, terá vários vídeos relacionados, sendo um thumbnail por vídeo.

Como não sabemos qual é exatamente a relação de thumbnails por vídeo, e nem se esta relação é constante no tempo, realizamos uma relativização da proporção de média de requisições por minuto para poder comparar o throughput do vídeo e thumbnail. O objetivo aqui era comparar, relativamente, sem usar valores absolutos, a taxa de requisições dos thumbnails com as dos vídeos. A pergunta era saber se em alguns momentos as requisições de thumbnails ou vídeos eram proporcionalmente maior ou menor.

Para isso normalizamos os valores de *req./s* para um intervalo entre 0 e 1.[9].Utilizamos os valores máximo e mínimo da distribuição para determinar o intervalo médio possível de *req./s* para cada curva. Assim, este intervalo (máximo - mínimo) era o valor máximo percentual das requisições para o objeto, e para cada minuto do dia (intervalo de avaliação) calcular o valor percentual relativo daquelas *req./s* no dado instante. Que em termos algébricos seria: $vr = \frac{x-min}{max}$ com vr sendo o valor relativo, o x o valor de *req./s* no dado momento e max e min como os intervalos máximos e mínimos.

Com os valores de throughput normalizados plotamos a razão entre o throughput do vídeo sobre o do thumbnail para os períodos de finais de semana e dias de semana na Figura 5.26. Essa análise foi realizada utilizando todos os valores do período de estudo e computando a média para cada minuto no dia. Se esta curva fosse $f(x) = 1$ existiria uma proporção constante e a distribuição entre as curvas seria idêntica.

Se um dado instante t do dia a proporção entre o throughput médio entre vídeos e thumbnails for :

$f(t) = 1$: A proporção é constante, o número de vídeos sendo requisitados é proporcional ao número de thumbnails sendo requisitado.

$f(t) > 1$: A fração de **vídeos** sendo requisitados é **proporcionalmente** maior

[9]Na realidade normalizamos no intervalo [1:2], pois se utilizarmos valores percentuais entre [0:1], quando os valores ficam muito pequenos a razão entre esses dois números próximo de zero pode gerar valores bem maiores que 1 ou até mesmo infinito (divisão por zero), o que não ocorre para o intervalo [1:2]. Por exemplo, para o intervalo [0:1] os valores 0.00185 e 0.00007 dariam a razão 26.42 enquanto que para o intervalo [1:2] teríamos 1.00185 e 1.00007 com razão de 1.0017.

$f(t) < 1$: A fração de **thumbnails** sendo requisitadas é **proporcionalmente** maior

Mas é importante ressaltar que o throughput do thumbnails é **sempre** maior que a de vídeo como podemos ver na Figura 5.25. Por isso é necessário a normalização. Se fossemos somente avaliar a razão de um pelo outro teríamos somente a grandeza dessa diferença, e não uma avaliação relativa de proporcionalidade sobre o período do dia. Como o thumbnail é uma espécie de cartão de visita para os vídeos, se muitos thumbnails são requisitados e a proporção de vídeos pode ser interpretada como uma espécie de eficácia desse cartão de visitas, ou mesmo uma metáfora da disponibilidade/interesse do usuário que está navegando no site de se engajar no consumo de vídeos.

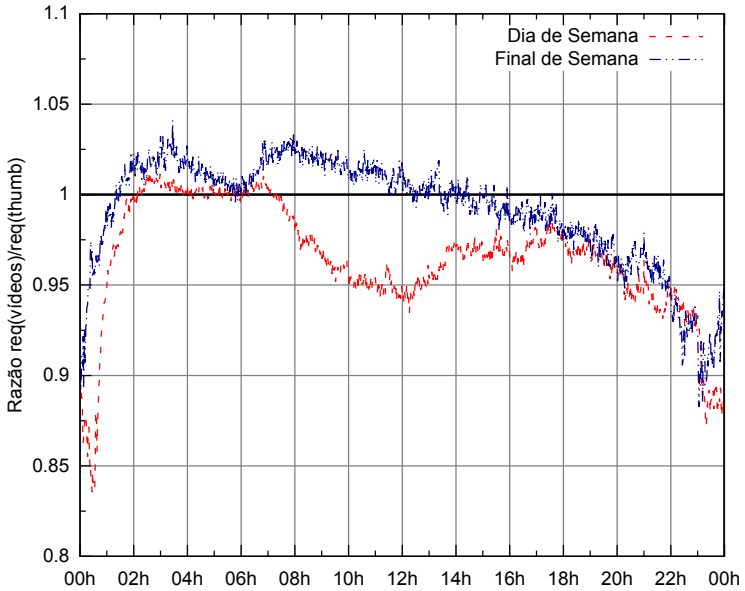


Figura 5.26. Relação entre as requisições de vídeos sobre as de *Thumbnail*

Dessa forma podemos ver na Figura 5.26 que em geral, nos finais de semana engajamento dos usuários em assistir os vídeos é maior que nos dias de semana, especialmente no período da madrugada, manhã e início da tarde.

Os dias de semana possuem dois períodos que se destacam pela falta de engajamento dos usuários em consumir vídeos. Tais períodos são o período comercial, e o período comumente conhecido como *fim de noite* entre 23h e 01h. Tal comportamento pode ser atribuído ao horário de trabalho, no período diurno as pessoas estão engajadas em seus trabalhos e no período noturno precisam dormir para acordar cedo. Mas se realmente for esse o caso, também é verdade que o tráfego de tal rede de conteúdo é dominado por pessoas acessando a Internet do trabalho, o que também caracteriza o perfil da audiência de tais conteúdos como altamente apropriada para publicidade.

Tais evidências tem várias implicações para os *Produtor de Conteúdo* que podem utilizar tal informação para potencializar a comercialização de publicidade. De forma complementar para o distribuidor, Samba Tech, essa informação é interessante para monitorar áreas de picos e como melhor provisionar o serviço.

5.4.3 O_3: Tempo de Resposta do Servidor por tipo de Objeto

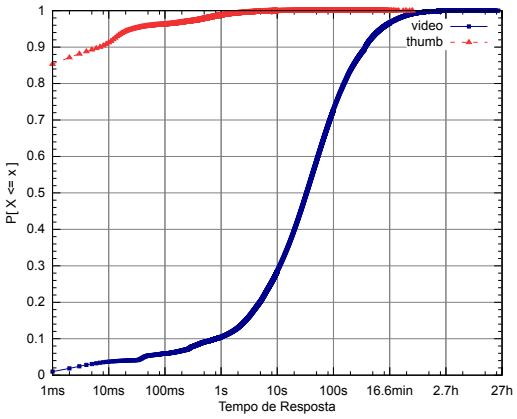
Nesta seção revisitamos a análise do Tempo de Resposta (T_r) do servidor feita na seção 5.3.4 porém, segmentamos entre os dois principais tipos de objetos.

Podemos ver pelos gráfico da *CDF* na Figura 5.27(a) que a diferença entre os tipos de objetos impacta muito nessa análise uma vez que os tipos possuem, em geral, tamanhos muito diferentes o que impacta diretamente no tempo de resposta do servidor.

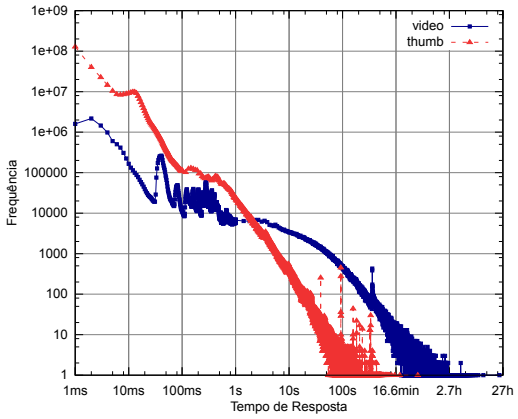
Vemos que praticamente todas as requisições de thumbnails possuem T_r inferior a 1 segundo, comportamento compatível com este tipo de objeto. O que chama mais a atenção é existência de valores muito altos para o T_r de thumbnails, como podemos ver nos gráficos da Figura 5.27. No histograma da Figura 5.27(c) uma pequena porção das requisições podem possuir de 1 segundo à 42 minutos de duração, o que nos parece ser um erro dado que tais objetos deveriam ser pequenos, e mesmo para entrega em períodos de congestionamento deveriam possuir no máximo alguns poucos segundos. Entretanto, como o período de análise é muito longo e não existe distinção entre requisições de produção e/ou de testes, casos excepcionais podem acontecer.

No trabalho de [Veloso et al., 2006a] foi feito um estudo similar ao nosso para tempo de resposta avaliando um caso de transmissões ao vivo. Nessa situação temos uma distribuição do tempo de resposta muito maior, o que é esperado por se tratar de um *streaming* de vídeo com tamanho indeterminado a priori. No nosso caso temos vídeos sobre demanda (VOD) e o objeto tem duração predeterminada. Comportamento similar ao nosso acontece no *YouTube* como mostrado em [Duarte et al., 2007a, Gill et al., 2007], porém com uma predominância de tempos de resposta menores que 600s, o que está relacionado com a limitação no tamanho dos vídeos do *YouTube*.

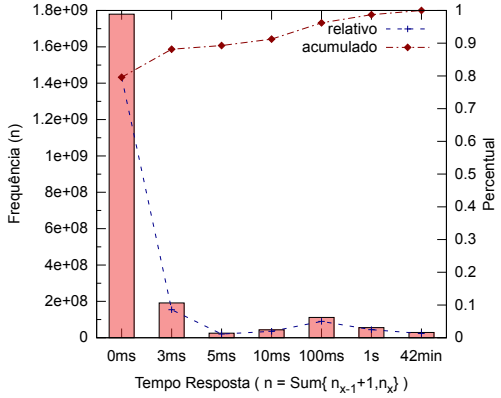
Com relação a segmentação do T_r por objetos do tipo vídeos temos que pequenos tempos de resposta são pouco frequentes compondo menos de 20% das requisições. Por outro lado, a maior parte dos vídeos ($\sim$ 66%) requisitados tiveram um tempo de resposta inferior a 1 minuto. A requisições com tempo de resposta de até 10 min é o segundo intervalo mais frequente para vídeos compondo mais de 30% das requisições destes. Apesar de 10min ser um tempo de resposta muito alto, ainda existem requisições muito longas que podem chegar a quase 24h de duração como podemos ver na



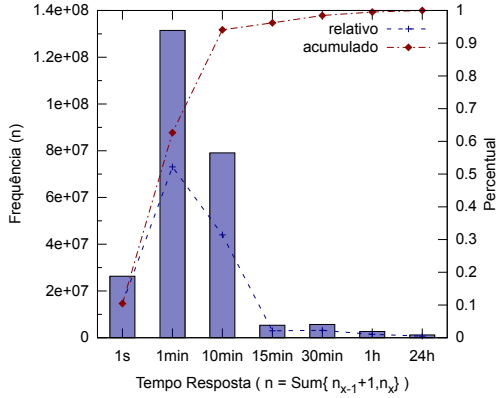
(a) *CDF*



(b) Distribuição



(c) Histograma thumb.



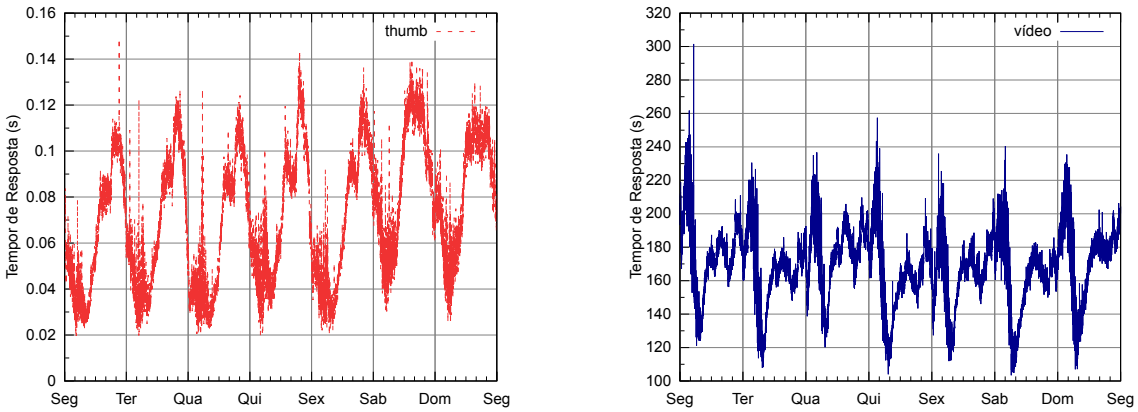
(d) Histograma vídeo

Figura 5.27. Informações gerais sobre tempo de resposta por objeto.

Figura 5.27(d).

Comparativamente, pelo gráfico da Figura 5.27(b) podemos ver que as distribuições dos T_r possuem duas regiões com comportamentos distintos. A primeira, com os T_r pequenos ($< 1s$) onde a frequência é maior para os thumbnails e uma variabilidade alta é observada para os vídeos. A segunda região, com T_r mais alto ($> 10s$), de comportamento semelhantes para os dois tipos de objetos, porém, os T_r para os vídeos é em geral uma ordem de grandeza maior.

Com relação aos padrões temporais comportamentais dos tempos de resposta, avaliamos somente as funções $T_r(t \ mod \ p)$, onde p corresponde aos períodos semanais e diários. O padrão temporal médio semanal dos T_r para os thumbnails pode ser visto na Figura 5.28(a). Notamos que a média é baixa, varia de 20ms à 150ms. Existem 2 padrões bem distintos, um presente nos dias de semana (seg. a sex.) e outro para os finais de semana.



(a) T_r por thumbnail

(b) T_r por vídeo

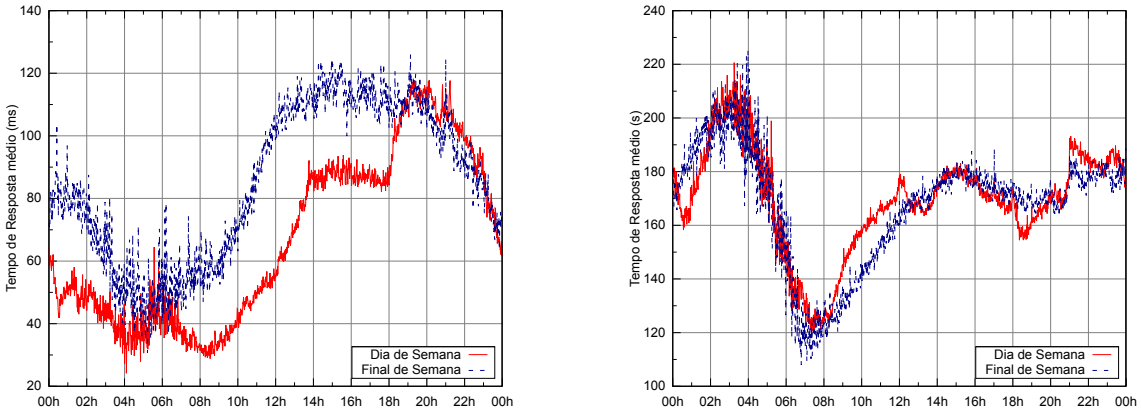
Figura 5.28. Padrão temporal semanal do T_r por objeto.

Já o padrão semanal para os vídeos pode ser visto no gráfico da Figura 5.28(b). Para este caso a variabilidade de tempo é alta, o tempo médio de resposta varia de 100s à 300s. De forma similar ao thumbnail, existem dois padrões para períodos distintos, um para os dias de semana e outro para os finais de semana. Os tempos de resposta tendem a ser maiores no meio da madrugada sendo que o vale característico de baixa atividade só ocorre por volta das 7h da manhã.

Para melhor análise dos dois períodos diários distintos existentes na semana, os gráficos da Figura 5.29 apresentam os padrões diários para dias úteis e finais de semana segmentados por thumbnails e Vídeos.

O padrão diário do tempo de resposta do thumbnail para os dias úteis é composto por um vale de baixo T_r no período da madrugada (4h) e outro no início da manhã (8h). Estranhamente, por volta das 6h da manhã existe uma variação dos T_r que culmina numa pequena elevação deste. O final da manhã e inicio da tarde é composto por um aumento gradativo que vai das 8h às 14h onde o T_r se estabiliza num valor médio por volta dos 90ms. Este período de estabilização do valor do T_r dura até as 18h e é o período mais estável de todo o dia. A partir das 18h os valores para os tempo de resposta voltam a crescer de forma acelerada até as 19h, mantêm um valor alto até por volta das 21h e vai decaindo a partir deste horário até chegar no vale das 4h da madrugada.

O padrão para o final de semana de thumbnails, também visto na Figura 5.29(a), segue o mesmo comportamento do padrão semanal, mas com o T_r médio, ligeiramente elevado no período da madrugada até o final da tarde (18h). A diferença marcante entre os dois períodos semanais, é que para o final de semana, não existe a estabilização dos



(a) T_r por thumbnail (milisegundos)

(b) T_r por vídeo (segundos)

Figura 5.29. Padrão temporal diário do T_r por objeto.

tempos de resposta no período de 14h as 18h. A curva para o final de semana pode ser descrita como uma senoide de vale as 6h e um pico as 16h.

A Figura 5.28(b) mostra o comportamento periódico diário para os T_r de vídeo para os dias úteis e finais de semana. A distribuição diária do T_r para os vídeos possui comportamento completamente distinto dos thumbnails. O período inicial da madrugada é composto por um crescimento do T_r até as 3h, seguido de uma queda com valor mínimo por volta das 7h da manhã. Isso para os dias úteis e finais de semana, entretanto, para os dias úteis, existe uma pequena queda no T_r na primeira hora do dia compondo um pequeno vale nesse período. O período de 8h às 0h possui comportamento similar para ambos os períodos entretanto a variabilidade do tempo de resposta para os dias úteis é muito maior. Os valores do final de semana possuem um crescimento gradativo de 8h às 15h, seguida de uma leve queda do T_r até as 21h quando o valor volta a crescer variando o seu valor entre 170s e 175s até o final do dia.

Para o padrão dos dias úteis, o crescimento do período da manhã (8h às 12h) é mais forte e possui um pico acentuado por volta do meio dia. O período de 12:15 às 14h existe uma queda no T_r, que logo após as 14h, volta a crescer atingindo um pico por volta das 15h quando começa a decair gradativamente. Às 18h ocorre uma queda brusca no T_r, seu valor retorna a crescer lentamente a partir das 18:30 e às 21h ocorre um crescimento abrupto. A partir de então o valor volta a cair sendo que entre as 23h e as 0h ocorre o último, e pequeno, pico de T_r.

Vimos nessa seção a analise do tempo de resposta segmentada por tipos de objetos. A distribuição cumulativas por objetos mostrou que é necessário segmentar tal análise por tipo de objeto pois particularidades de cada objeto afetam diretamente o

T_r das requisições. A presença de T_r muito altos sugere que existam alguns erros na identificação de alguns objetos possa ter ocorrido, ou casos excepcionais possa, por motivos desconhecidos, ter sido forçados.

No geral a análise dos padrões periódicos temporais do tempo de resposta não é muito esclarecedora pois vários aspectos podem afetar tal parâmetro. O tamanho do objeto sendo requisitado, o congestionamento na rede, a proporção do objeto requisitado para o caso dos vídeos. Talvez a análise da relação entre o T_r e o tamanho do objeto sendo requisitado possa prover melhores *insights* a respeito do T_r do servidor.

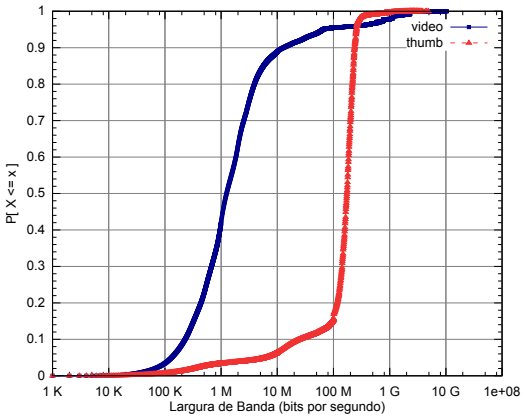
5.4.4 O_4: Largura de Banda por tipo de Objeto

Na Seção 5.3.5 fizemos uma análise geral da largura de banda (*badwidth*) para as requisições em geral. Em tal seção notamos que os valores para tal propriedade estavam muito altos e que não correspondiam a realidade das velocidades praticadas para a Internet nos dias de hoje. Pretendemos na presente seção avaliar, de forma segmentada por tipos objetos, tal propriedade e investigar o porque dos valores tão altos na Seção 5.3.5.

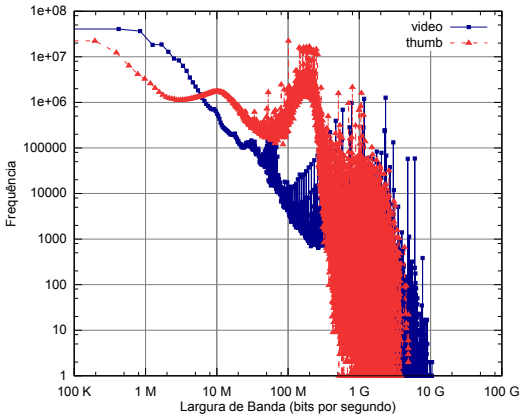
Os gráficos da Figura 5.30 mostram as informações gerais sobre o largura de banda segmentadas por objetos. Na distribuição cumulativa da Figura 5.30(a) vemos que a maior parte da largura de banda para as requisições dos thumbnails possuem valores superiores a 100mbps. Podemos confirmar tais valores através do histograma da Figura 5.30(c) onde vemos que mais de 90% das requisições possuíram valores relativos de largura de banda entre 100mbps e 4.8Gbps. Tais valores não podem ocorrer numa plataforma de distribuição de conteúdo pela Internet uma vez que tais velocidades de transmissão de dados não é a realidade atualmente praticada pelos chamadas *Provedores de Internet*.

Tais valores absurdamente altos ocorrem devido a natureza dos dados sendo utilizados para o cálculo de tal propriedade e a forma que estes são gerados. Na realidade, o cálculo da largura de banda deveria ser realizado contabilizando a quantidade total de bits trafegados sobre o tempo total de transmissão excluindo-se a latência. Entretanto isso não é o que fazemos aqui, pois somente calculamos a aproximação da largura de banda através da razão entre a quantidade de dados trafegados pelo tempo em que o servidor gastou para escrever as informações para o cliente. Não há uma contabilização de latência, nem mesmo do tempo total de transmissão, pois se o objeto for pequeno o suficiente para caber no *buffer* de rede do servidor o tempo decorrido será somente o tempo gasto para escrever os dados no *buffer*. Uma vez que tais *buffers* podem ser de alguns poucos megabytes em servidores de alta performance, objetos pequenos

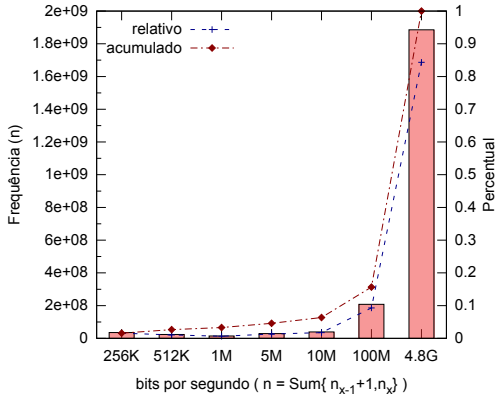
como thumbnails não podem ser utilizados para o cálculo de tal valor aproximado de largura de banda por distorcer demais tal propriedade pois os tempos de escrita são muito pequenos. Sendo assim, os valores relativos a thumbnails da Figura 5.30 não são confiáveis e não devem ser considerados para as análises da largura de banda.



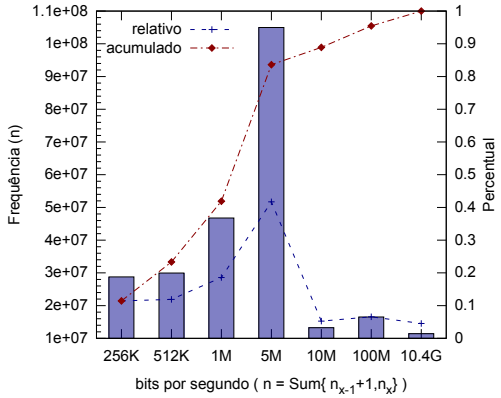
(a) *CDF*



(b) Distribuição



(c) Histograma thumb.



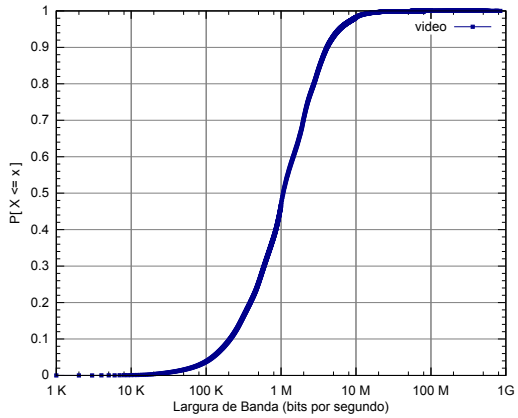
(d) Histograma vídeo

Figura 5.30. Informações gerais sobre a largura de banda por objeto.

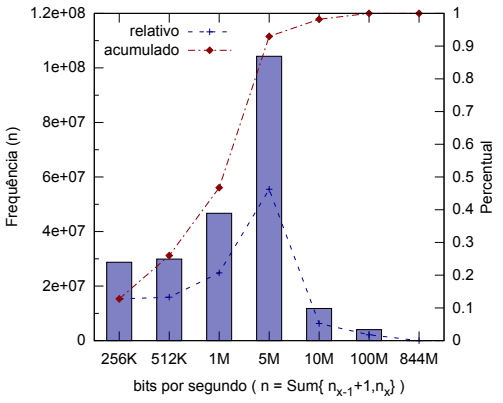
Com relação a largura de banda para os objetos do tipo vídeo podemos ver que os valores correspondem, proporcionalmente, a valores mais razoáveis de velocidade de transmissão praticados pelos *Provedores de Internet* no Brasil. Na distribuição cumulativa da Figura 5.30(a) podemos ver que cerca de 40% das requisições possuem largura de banda menor que 1Mbps, sendo que tal valor sobe para 88% se levarmos em conta os valores inferiores a 10Mbps. Ainda assim, existe um percentual elevado de valores acima de 10Mbps (12%) chegando a valores de 10.4Gbps. Apesar de existirem redes de alta velocidade comerciais (100mbps) e redes ligadas a POPs e ISP com

velocidades compatíveis a tais velocidades, sabemos que esses valores não são reais pois ainda estão entrando no cálculo tempo de respostas muito pequenos.

Para corrigir tal análise, foram descartados todas as requisições para vídeos que possuíam um T_r menor que 1s, tempo suficiente para incluir o tempo de ida e volta (RTT *Rond-Trip Time*) em uma rede com latência alta além de excluir objetos pequenos o suficiente para caber no *buffer* de rede dos servidores. Feito isso geramos os gráficos da Figura 5.31 que são uma versão da *CDF* e do histograma para a largura de banda média para os vídeos onde o T_r é maior que 1 segundo.



(a) *CDF* para $band(T_r > 1s)$



(b) Histograma vídeo para $band(T_r > 1s)$

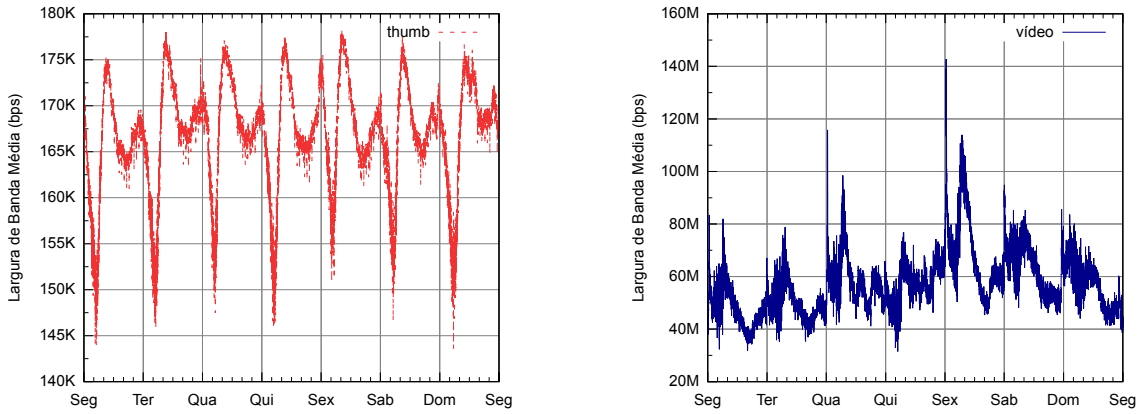
Figura 5.31. Comportamento semanal da largura de banda de entrega da plataforma

Com essa nova abordagem temos a distribuição cumulativa da Figura 5.31(a). A proporção para os valores comummente praticados por *Provedores de Internet* pode ser visto de maneira mais legível no histograma da Figura 5.31(b). Valores de largura de banda entre 1Mbps e 5Mbps são os que dominam as requisições de vídeos com 46% das requisições ao sistema. Taxa de transferências maiores que 5Mbps ainda são pouco comuns ($\sim$ 7%) mas existem. De uma forma geral, os gráficos da Figura 5.31 mostra que a disseminação das redes de banda larga estão ocorrendo e que o perfil do usuários que consomem vídeos são em sua maioria (52,82%) usuários de redes de alta velocidade ($1Mbps \leq x$).

Ressaltamos que aqui o termo usuário é empregado como sendo um usuário genérico, não conhecido, que origina as requisições e que pode ser até mesmo um *bot*. O importante nessa seção é a largura de banda utilizada nas requisições que representa um perfil do usuário brasileiro.

A Figura 5.32 mostra a largura de banda média semanal, agregada por minuto,

que a plataforma em análise utilizaria em média para entregar os conteúdos segmentados, caso existisse um canal de entrega para cada tipo de objeto.



(a) Thumbnail (b) Vídeo

Figura 5.32. Comportamento semanal da largura de banda de entrega da plataforma

O padrão semanal da largura de banda para os objetos do tipo thumbnail (Figura 5.32(a)) forte e semelhante durante toda a semana. Somente na Segunda-Feira existe uma diminuição da taxa. O padrão semanal de vídeos é mostrado na Figura 5.32(b). Vemos que nos dois primeiros dias úteis da semana existe um consumo moderado de tais objetos representado pelas menores taxas de largura de banda. Quarta e Quinta-Feira apresenta uma alta variação do consumo de banda sendo que a Sexta-Feira apresenta os maiores picos de consumo de banda. O final de semana é o período com a menor variabilidade e o maior consumo distribuído por todo o dia.

Apesar do cálculo da largura de banda utilizada nesta seção ser aproximada, por não levar em consideração os valores de latência, foi possível extrair alguns insights a respeito do uso da banda média em geral, além do consumo médio da plataforma como um todo. Vimos que objetos pequenos, como thumbnails, não podem ser considerados nas avaliações de largura de banda por requisição, uma vez que distorcem os valores desta propriedade. Notamos que as requisições de vídeos, possuem em sua grande maioria, velocidades de conexão entre 1 e 5 Mbps. Entretanto a proporção de requisições com baixa velocidade ainda é alta ($\sim$25%). Outra observação é que o padrão de consumo de banda no início da semana é em média menor que a dos outros dias da semana.

5.4.5 O_5: Tamanho das Transferências por tipo de Objeto

Daremos sequência agora ao estudo do tamanho das transferências (requisições) segmentando estas pelo tipo de objeto, que neste estudo de caso são os vídeos e os thumbnails. O objetivo aqui é aprofundar as análises e complementar as outras realizadas na Seção 5.3.6 segmentando os diferentes objetos para elucidar a grande variabilidade nos tamanhos de objetos, que como pudemos ver na Figura 5.14(a) possuíam tamanhos com 3 ordem de grandezas de diferença.

Notem que a análise das transferências dos objetos é diferente da análise do tamanho dos objetos. Esta está relacionada à propriedade do objeto em si, aquela, ao consumo do objetos, que não necessariamente serão requisitados/consumidos por completo.

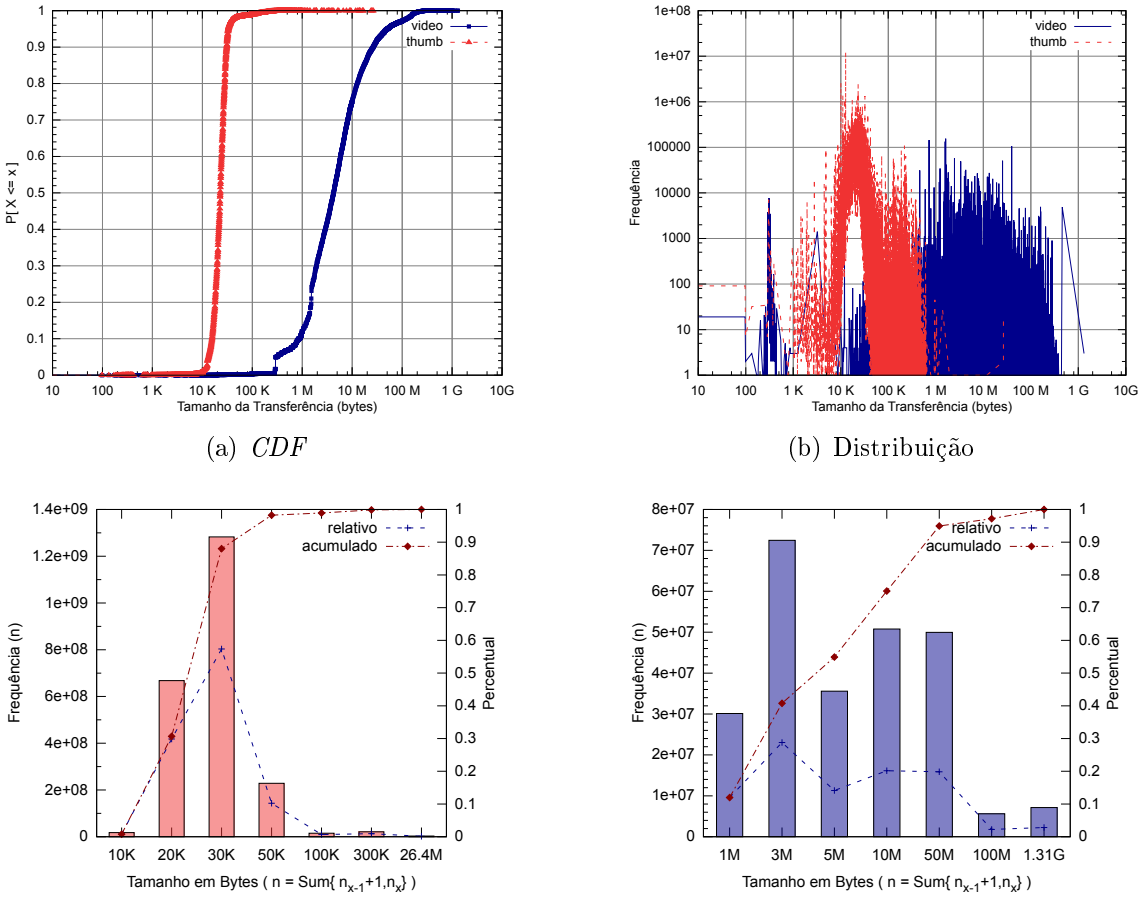
Faremos agora uma análise da distribuição dos tamanhos dos objetos através da funções cumulativas de distribuição e também por histogramas como podemos ver na Figura 5.33.

As funções de distribuição cumulativas podem ser vistas na Figura 5.33(a). Vemos que a distribuição dos tamanho das requisições para os thumbnails corresponde a faixa de poucas dezenas de kilobytes. O histograma da Figura 5.33(c) mostra que a grande maioria das transferências são menores a 50 KB, e quase 90% dessas transferências são menores que 30 KB. Entretanto se reparamos na distribuição da Figura 5.33(b) podemos ver que existem transferências de objetos do tipo thumbnail com tamanho maior que 100KB. Isto pode indicar algum erro na infraestrutura de distribuição de conteúdos ou o mal uso de tal recurso por parte dos *Produtor de Conteúdo* e deve ser investigado.

Com relação as transferências de objetos do tipo vídeo vemos que a variabilidade do tamanho das transferências se distribui por 4 ordens de grandeza. Tal comportamento demostra que existem vídeos grandes em tamanho e que tal objeto é o principal responsável pela utilização de banda de transferência dos servidores. Através do histograma na Figura 5.33(d) vemos que a faixa de transferência dominante é a entre 1 e 3 MB com 28,79% das transferências. Ainda assim, a diferença na distribuição entre as 5 faixas inicias $(0-1], (1-3], (3-5], (5-10], (10-50]$ não é tão grande.

Ao compararmos nosso trabalho com outros da literatura vemos que os padrões do tamanho das transferências se assemelham. No trabalho de [Gill et al., 2007] temos distribuições muito similares para vídeos e para imagens, se compararmos nossos thumbnails como imagens.

A distribuição temporal semanal para o tamanho das transferências ($size(t \ mod \ p)$ com p período semanal) pode ser vista na Figura 5.34. Com relação



(a) *CDF*

(b) Distribuição

(c) Histograma thumb.

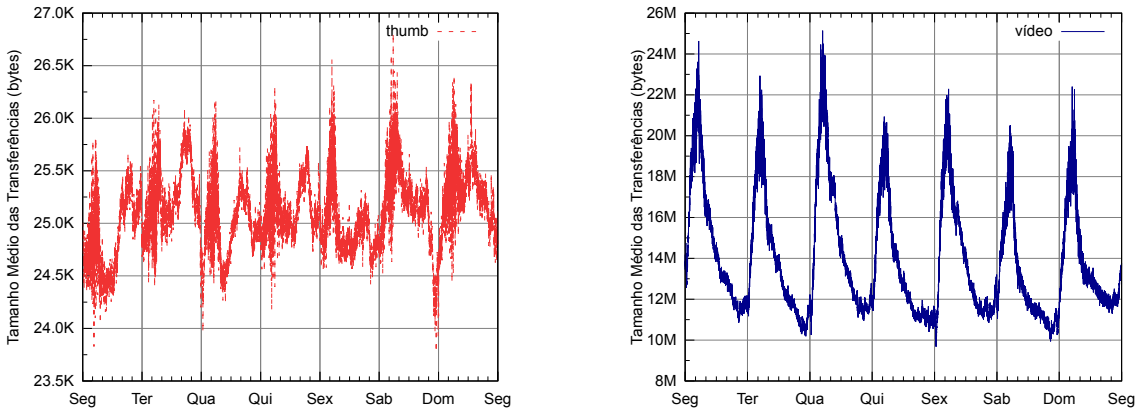
(d) Histograma vídeo

Figura 5.33. Informações gerais sobre o tamanho da transferência por objeto.

ao padrão temporal semanal do thumbnail (Figura 5.34(a)) ocorre uma alta variabilidade no comportamento sem um padrão forte. Destaca-se apenas o vale acentuado na madrugada de domingo para segunda. Já para o caso do vídeo notamos nitidamente um padrão comportamental forte onde o tamanho médio das transferências dos vídeo é maior no período da madrugada e esta média tende a cair gradativamente durante o decorrer do dia.

Novamente, segmentamos as análises por períodos diários distinguindo os períodos de finais de semana com o período útil semanal, tais informações com relação ao tamanho médio das transferências pode ser vistas na Figura 5.35.

Como dito na subseção 5.4.2 o thumbnail é uma mídia relacionada ao vídeo e geralmente compõe as páginas de conteúdo dos grandes portais de mídia com objetivo de atrair o espectador para outro conteúdo. Assim, quando mais páginas se visitam mais thumbnails são requisitados. Notamos em nossas análises que apesar do objeto



(a) T_r por *thumbnail* (b) T_r por vídeo

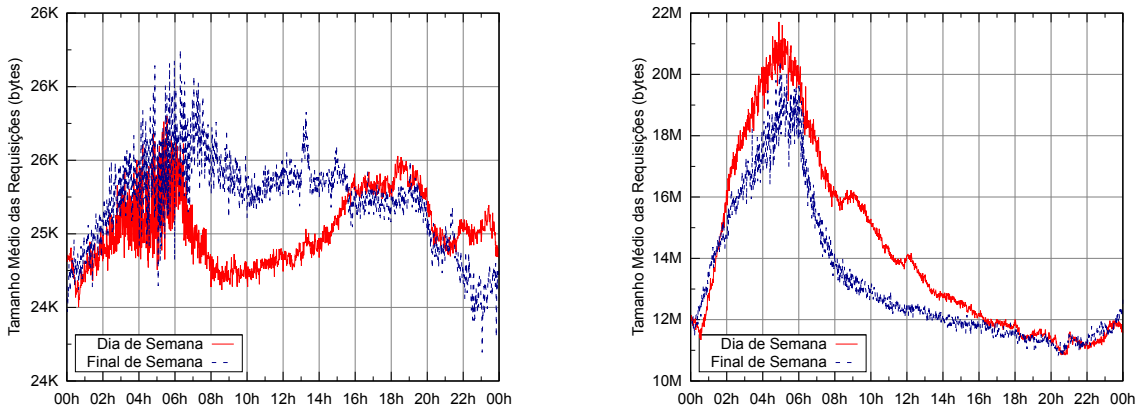
Figura 5.34. Padrão temporal semanal do T_r por objeto.

thumbnail ser pequeno, muitas vezes a sua requisição é incompleta, isto é, somente parte do conteúdo é salvo. Isto pode ser interpretado como a ação do usuário de mudar de página (clique em um link) antes mesmo que esta se carregue por completo. Tal comportamento pode ser interpretado como a ação do usuário de navegação entre páginas a procura de conteúdo (busca por conteúdo). Acreditamos que esta busca por conteúdo é a razão do vale característico da Figura 5.35(a) no horário comercial, e sua ausência um indício do maior engajamento do usuário nos conteúdos nos períodos da madrugada e entre 16 e 20h.

Outro indício interessante nesta mesma e notado também na subseção 5.4.2 é a mudança de comportamento que ocorre sempre as 18hs. Acreditamos que este indício é resultado da grande quantidade de pessoas acessando conteúdos multimídia direto das empresas onde trabalham, por isso a mudanção de comportamento às 18hs, horário que geralmente se encerra o expediente na maior parte das empresas brasileiras.

Com relação aos vídeos, a distribuição diária pode ser vista na Figura 5.35(b). O padrão de transferência média para o caso dos vídeos é muito forte. Porém o que podemos concluir diretamente é que as requisições do período da madrugada geralmente trafegam mais informações. Qual é a real causa desse comportamento demandam análises mais elaboradas tipicamente pertencente a camada **C** ou **K**. Avaliaremos tais detalhes nas camadas superiores.

Uma hipótese que surge ao se analisar o tamanho das transferências é se a distribuição desta é uma aproximação da distribuição do tamanho dos objetos na base de dados? Tal resposta será respondida na seção de análise das propriedades dos objetos a frente. Discutiremos a relação entre estas duas medidas e responderemos a tal hipótese.



(a) Tamanho da transferência por *thumbnail*

(b) Tamanho da Transferêcia por vídeo

Figura 5.35. Padrão temporal diário do tamanho da transferência por objeto.

5.4.6 O_8: Características dos objetos

Nesta Seção apresentamos as análises das propriedades dos objetos em estudo. Tais propriedades podem ser comuns aos objetos ou específicas. Como exemplo da primeira temos o tamanho do objeto e da segunda o *bitrate* dos vídeos. Analisaremos as seguintes propriedades:

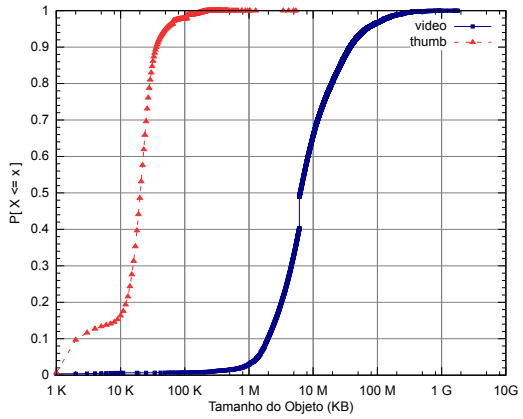
Vídeo : tamanho, duração, canais de áudio, codec de áudio, sample rate de áudio, bitrate, framerate, resolução (altura x largura) e codec de vídeo

Thumbnail : tamanho, resolução, formato.

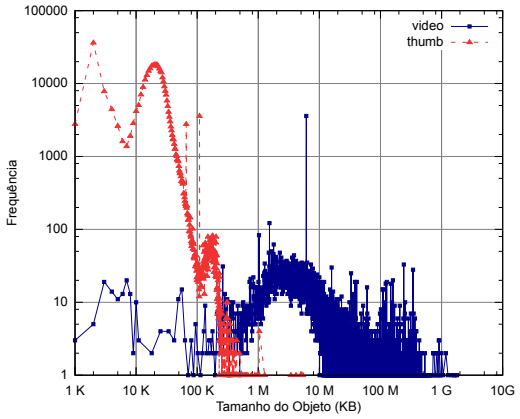
Tamanho dos Objetos

Inicialmente iremos analisar o tamanho dos objetos pois esta análise é comum aos dois tipos. A Figura 5.36 apresenta a *CDF* e o histograma dos tamanhos dos Thumbnails e dos vídeos. Os thumbnails entre 20 e 30KB são os mais frequentes na coleção, mas é interessante notar que existem thumbnails grandes de até 1MB (e maiores) que podem degradar a experiência do usuário e são desaconselháveis.

Já para os vídeos, menos de 3% destes possuem tamanhos inferiores a 1MB. Se aumentarmos o valor para 10MB a taxa passa a ser de 70%. É interessante notar o aumento na disponibilidade de vídeos de tamanho grande, 26% dos vídeos estão entre 10 e 100MB, e quase 4% maiores que 100MB, chegando a existir vídeos de 1.8GB de informação, o que mesmo para os padrões de banda larga é um valor muito alto.



(a) *CDF* do tamanho dos objetos



(b) Distribuição do tamanho dos objetos

Figura 5.36. Informações sobre a distribuição dos tamanhos dos objetos

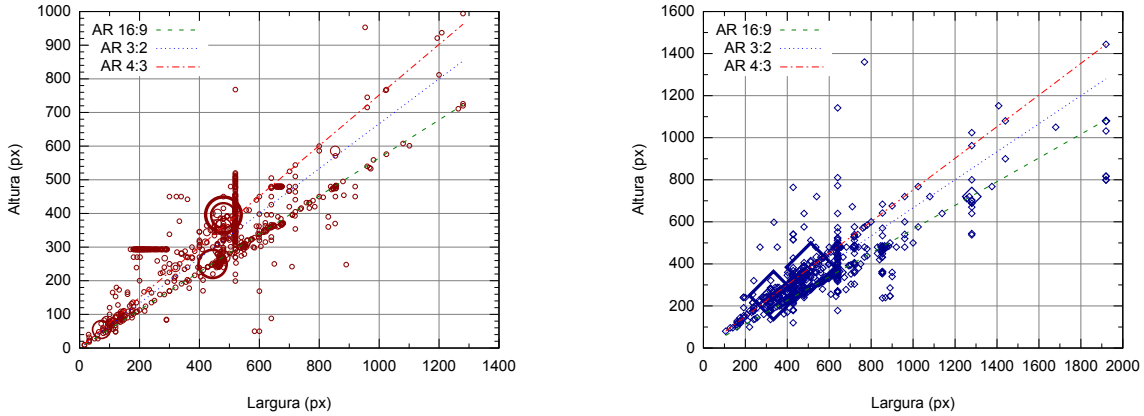
Com relação a discussão sobre a aproximação que a distribuição do tamanho das requisições pode ser da distribuição do tamanho dos objetos, comparando as Figuras 5.36(a) e 5.33(a) podemos ver que para o caso dos thumbnails a aproximação é mais razoável. Apesar dos 18% de objetos menores que 10KB não estarem representados na distribuição das transferências o comportamento em que quase todos os objetos são menores que 100KB é consistente. Entretanto tal aproximação não se aplica nas mesmas proporções para o caso dos vídeos. Os objetos tendem a ser maiores que os tamanhos das requisições fazendo com a distribuição aproximada apresente um erro. Para os objetos de 10MB de tamanho a *CDF* das transferência apresenta uma diferença de 10%. Assim é necessário ter em mente que a distribuição aproximada tende a descrever os objetos menores do que ele são.

Dimensão dos Objetos

Avaliaremos agora a dimensão (resolução espacial) *Largura x Altura*, dos vídeos e thumbnail. Para o caso dos vídeos, tal propriedade é uma alusão à capacidade dos dispositivos de exibição de vídeo (monitores, televisores, projetores, etc) de exibir pontos em suas telas. Várias siglas e códigos utilizados por vários fabricantes de hardware [10] caíram em uso popular no jargão da Internet entretanto não existe nenhum padrão definido para a nomenclatura e a dimensão exata depende também do *aspect ratio* que é a relação entre a largura e a altura da imagem e normalmente impacta muito na apresentação do objeto em questão.

[10] `http://en.wikipedia.org/wiki/Display_resolution`

Os gráficos da Figura 5.37 precisa de uma explicação. Foi contabilizado as resoluções dos objetos e registrado no formato {*largura x altura*, frequência}. Para cada ocorrência registrou-se o ponto no plano cartesiano da figura. O tamanho do ponto representa (proporcionalmente) a frequência de cada resolução. Adicionalmente registramos as constantes dos 3 *aspect ratio* (AR) mais frequentes na base que são o *4:3* que é conhecido como padrão TV, o *3:2* padrão de filmes fotográficos e muito comum também em dispositivos ditais como câmeras e filmadoras amadoras e o *16:9 widescreen* padrão $HDTV$.



(a) Distribuição da resolução dos *thumbnails*

(b) Distribuição da resolução dos vídeos

Figura 5.37. Informações sobre a distribuição dos tamanhos dos objetos. Gráfico de dispersão das dimensões dos objetos. O tamanho do ponto é proporcional a frequência da respectiva resolução.

Na Figura 5.37(a) vemos que as dimensões dos thumbnails não seguem um padrão rígido (a ferramenta possibilita essa flexibilidade). Entretanto percebe-se que existe uma tendência a seguir o AR *16:9* através da linha de frequência que se forma junto a essa reta. As dimensões mais frequentes e seus AR aproximado são :

AR 3:2 :$\{586x853; 12137\}, \{290x460; 10487\}$

AR 4:3 : $\{240x320; 10493\}, \{360x480; 29389\}, \{55x73; 33088\}, \{394x480; 50255\}, \{396x480; 77635\}$

AR 16:9 :$\{270x480; 21161\}, \{250x444; 58821\}$

Com relação aos vídeos podemos ver as dimensões na Figura 5.37(b). No caso dos vídeos também existe uma variação grande das dimensões, porém, em menor proporção do que os thumbnails. A plataforma utiliza por padrão 5 formatos que fixam a altura

do vídeo no número de linhas para uniformizar os formatos 204p,360p,480p,720p,1080p. O *Produtor de Conteúdo* pode escolher forçar o formato a seguir o AR 16:9 ou manter o AR. Outro fato importante é que os dados analisados cobre duas versões distintas da plataforma, sendo que a versão anterior não definia as dimensões dos vídeos fazendo com que os *Produtores de Conteúdo* pudessem customizar o formato.

As dimensões dominantes estão enumeradas na lista abaixo de acordo com cada AR:

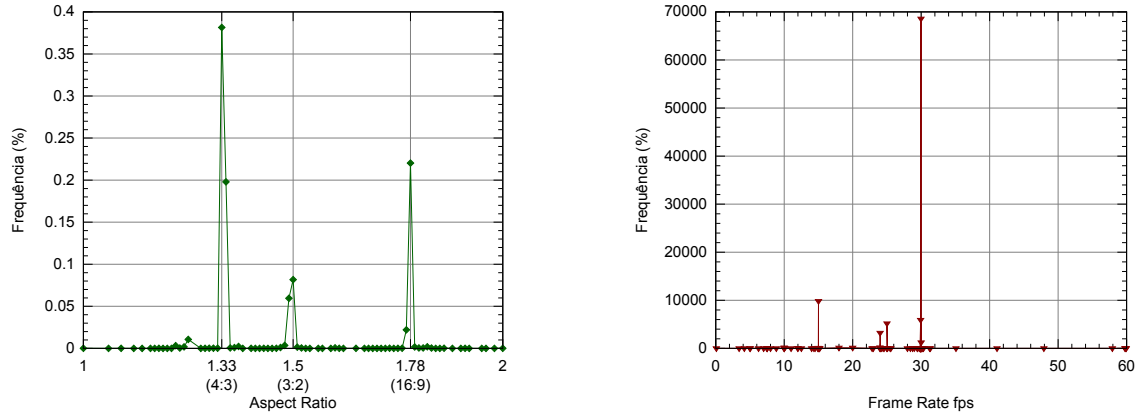
AR 3:2 : $\{316x472; 16977\}, \{320x480; 14806\}$

AR 4:3 : $\{384x512; 60603\}, \{250x334; 59936\}, \{240x320; 24081\}$, $\{360x480; 17840\}, \{480x640; 7089\}$,

AR 16:9 : $\{720x1280; 17365\}, \{360x640; 15438\}, \{270x480; 12981\}$

Podemos ver que existe nitidamente uma predominância dos formatos de `AR` *4:3*, entretanto a tendência a utilização dos formatos *widescreen* tendem a dominar principalmente com o surgimento das *connected TVs* que são televisores que se conectarão diretamente a Internet.

Com relação ao AR dos vídeos plotamos no gráfico da Figura 5.38(a) a frequência da razão da largura pela altura de todos os vídeos. Como podemos ver o AR de *4:3*] é dominante presente em quase 40% dos vídeos, entretanto o formato *{16:9}* está se consolidando devido a popularização de monitores *widescreen* e dispositivos neste formato.



(a) *Aspect Ratio* (b) *Frame Rate*

Figura 5.38. Distribuição de *bitrate*. Comparação com os conteúdos publicados e não publicados.

Formato dos Objetos

A Tabela 5.14 mostra a distribuição das extensões dos arquivos na base. Para o caso dos thumbnails conseguimos traduzir diretamente para os formatos associados: bmp → Bitmat File Format, tiff → *Tagged Image File Format*, jpeg → formato JPEG[11], png → *Portable Network Graphics*. Entretanto para o caso dos vídeos as extensões se relacionam ao contêiner do vídeo mas somente a extensão não revela as informações completas a respeito do contêiner.

Temos para o caso dos thumbnails o domínio do formato JPEG que geralmente provê uma imagem de tamanho menor que os outros formatos e por isso é amplamente adotado.

Thumbnail		**Vídeo**			
Formato	***%***	***Formato***	***%***	***Formato***	***%***
bmp	0.00%	mp4	80.38%	mov	0.62%
tif	0.00%	flv	15.96%	avi	0.22%
jpeg	97.13%	3gp	1.49%	asf	0.17%
png	2.87%	wmv	1.13%	mxf,m4v,f4v	0.04%

Tabela 5.14. Formatos dos objetos e sua distribuição na base

No caso dos vídeos o contêiner mais utilizado é o MPEG-4 (mp4) seguido do flash vídeo *flv*. Isto se deve mais por aceitação e compatibilidade dos navegadores atuais. O formato 3gp foi muito utilizado para dispositivos móveis mas já esta sendo substituído pelo formato mp4 nos novos aparelhos móveis. De uma forma geral os contêineres suportados estão diretamente relacionados com o *player* da plataforma e a sua padronização de *transcoding*. A opção padrão da nova versão da plataforma é utilizar o mp4.

Propriedades Exclusiva dos Vídeos

Começaremos as análises das propriedades intrínsecas dos vídeos pela taxa de imagens por segundo conhecida como *framerate*. A Figura 5.38(b) mostra a distribuição da taxa de imagens por segundo. As taxas mais frequentes são a de 29.97 (e similares), 15, 25 e 24 imagens por segundo. Com exceção da taxa de 15 imagens por segundo que é empregada para diminuir o tamanho final do vídeo (com perda na qualidade) os outros formatos são padrões na indústria cinematográfica.

[11] o acrônimo vem do nome do grupo que criou o formato, *Joint Photographic Experts Group*

Com relação aos codecs de áudio utilizado nos vídos da base, o padrão dominante em 83.37% dos vídeos é o AAC (Advanced Audio Coding) que foi desenvolvido para ser o sucessor do mp3, que por sua vez, está presente em 11.39% dos vídeos. 3.15%. O restante dos codecs de áudio se dividem em uma miscelânea de codecs pouco difundidos que não vale a pena citar. Para os canais de áudio dos vídeos, temos a proporção é de 59.32% de vídeos com apenas 1 canal de áudio contra 40.67% de vídeos estereofônicos. Apesar de insignificante já existem vídeos com mais de 2 canais de áudio dentre os vídeos da coleção.

O último aspecto relativo ao áudio nos vídeos é o *sampling rate* (*SR*), que é a amostragem que se realiza num sinal contínuo (no caso o áudio) para a conversão em sinal digital. Usualmente a taxa é por segundos, assim um *SR* de 40000 significa que são coletadas 40 mil amostras da onda contínua num intervalo de 1 segundo. A Tabela 5.14(b). Os valores mais frequentes são respectivamente o de 44100 com 46.82% e 22050 com 21.13%. O primeiro valor corresponde a qualidade de amostragem dos CDs de áudio e o último é uma recomendação de amostragem para MPEG vídeos.

(a) Distribuição dos *codecs* de vídeos

Vídeo Codec			
Codec	**%**	**Codec**	**%**
H264	83.36%	MPEG4	0.27%
FLV1	8.74%	WMV2	0.03%
VP6F	4.48%	MPEG1	0.03%
H263	2.35%	WMV1	0.02%
WMV3	0.67%	OUTROS	0.07%

(b) Distribuição do *Sample Rate* dos áudios dos vídeos

Vídeo Áudio *Sample Rate*			
SR	**%**	**SR**	**%**
48000	13.04%	22050	21.13%
44100	46.82%	16000	2.01%
32000	3.78%	11025	1.91%
24000	10.06%	8000	1.25%

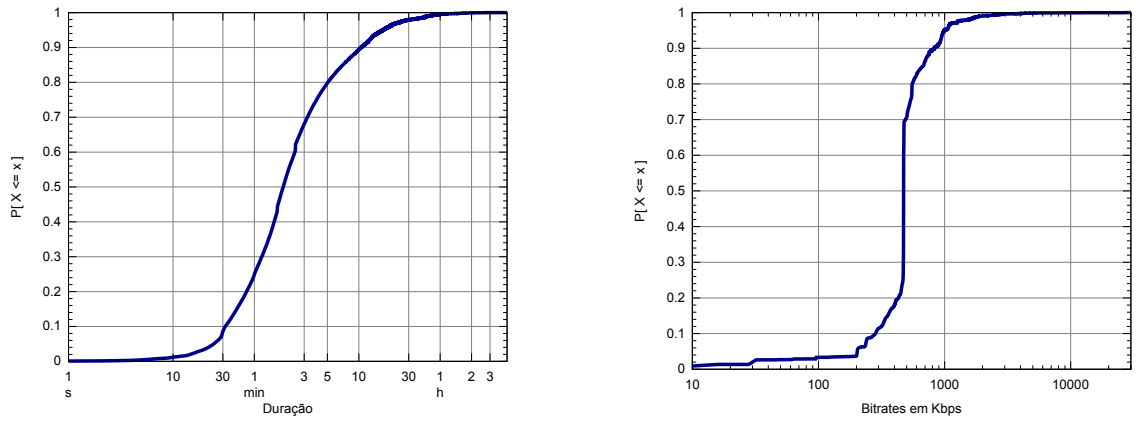
Tabela 5.15. Informações da distribuição dos *codecs* de vídeos e também do *Sample Rate* de áudio utilizado nos mesmos

Um ponto importante para se investigar quando se avaliando o consumo de vídeos pela Internet é que tipo de vídeo está sendo consumido pelo usuários em geral. Para avaliar o aspecto de duração do vídeo a Figura 5.39(a) apresenta a *CDF* do tempo de duração dos vídeos da base. Neste gráfico o eixo das abscissas está numa escala de tempo dividida entre segundos, minutos e horas.Podemos ver que aproximadamente 24.9% da base é composta por vídeos muito curtos, menores que 1 minuto de duração. A faixa de vídeos curtos, entre 1 e 3 minutos compõe a maioria dos vídeos 43.1% da base. O restante, estão divididos na seguinte forma duração entre 3 e 5 minutos

correspondem a 11.7%, 5 e 10 minutos 9.5%, 10 e 30 min 8.5% e maiores que 10 minutos 2.3%. Assim existe uma predominância de vídeos curtos, entre 1 e 3 minutos mas vídeos maiores estão cada vez mais presentes. Temos que 0.5% dos vídeos da base são vídeos com mais de 1h hora de duração, o que neste universo são mais de 2000 vídeos.

É intuitivo acreditar que a duração dos vídeos esteja relacionada com o propósito desse e características dos seus produtores. Mas também com alguma característica de seu distribuidor. Por exemplo, no caso do YouTube temos a restrição de 10 minutos para tamanho dos vídeos. Em [Duarte et al., 2007a], em que se analisa vídeos do YouTube para diferentes regiões como América Latina, EUA e outros países, os vídeos analisados possuem em sua maioria ($\approx$ 80%) são inferiores a 5 minutos, sendo que essa proporção chega a 99% se levarmos em conta uma duração de 10 minutos. O mesmo se verifica para trabalhos relacionados ao YouTube como em [Gill et al., 2007] e [Maia & Virgilio Almeida, 2009].

Já no trabalho de [Cherkasova & Gupta, 2002] os vídeos avaliados, de servidores corporativos, já possuem duração mais similares aos padrões dos vídeos analisados neste estudo. Vídeos de até 2h10. A distribuição para uma das corporações avaliadas é balanceada, vídeos de diversos tamanhos sem nenhuma dominância de duração, mas para outra a 60% dos vídeos são maiores que 1h. No trabalho de [Almeida et al., 2001], mídias de universidades, as durações que se destacam são vídeos curtos, de até 5 minutos e vídeos grandes, $\approx$ 60 min.



(a) *CDF* das durações vídeos

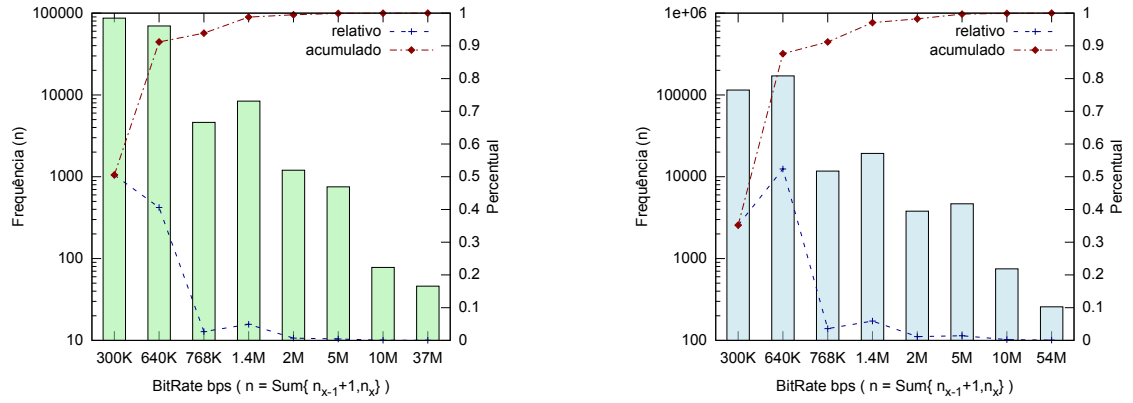
(b) *CDF* do bitrate

Figura 5.39. *CDF* do tempo de duração dos vídeos e do *bitrate* dos vídeos.

A transmissão de vídeos pela Internet envolve um compromisso entre aumentar a qualidade do vídeos e manter seu tamanho aceitável para os padrões de velocidade de

Internet utilizado pelos usuários em geral. Dois aspectos que influenciam na qualidade do vídeo são o *codec* de vídeo utilizado para a compressão dos dados e a taxa de bits por segundo no vídeo conhecido como *bitrate*. É comum se associar a qualidade do vídeo diretamente com o *bitrate* mas atualmente a qualidade final do vídeo não é uma função linear tão simples e depende de uma série de fatores. Abstraindo tais detalhes faremos um levantamento agora do *bitrate* dos vídeos e mostraremos os principais *codecs* utilizados nos vídeos em análise. Os *codecs* de vídeos podem ser vistos na Tabela 5.14(a). O padrão H264 é o mais utilizado em 83% dos vídeos e é o adotado como padrão na plataforma. Os outros codecs são casos pontuais e a tendência é que fiquem cada vez mais raros.

A distribuição cumulativa do *bitrate* dos vídeos pode ser vista no gráfico da Figura 5.39(b). Podemos ver que os valores se concentram entre 300 e 600 Kbps que é a faixa de *bitrate* mais utilizada para vídeos na Internet.



(a) Vídeos publicados

(b) Toda a base

Figura 5.40. Distribuição do bitrate. Comparação com os conteúdos publicados e não publicados.

A Figura 5.40 apresentamos um contraste entre todas as mídias da base e as mídias publicadas. As mídias publicadas são aquelas que efetivamente estão sendo expostas para os usuários através meios de publicação digitais. Os valores em tais figuras são relativos aos padrões utilizados comumente na plataforma, 300kbps para o padrão 240p, 640kpbs para 360p, 768kbps para 480p, 1408kbps para 720p e 2048 para 1080p.

Na Figura 5.40(a) podemos ver que a maior parte das mídias ($\sim$ 50%) possuem *bitrate* de 300Kbps. Já na Figura 5.40(b) vemos que essa não é a proporção dos vídeos existentes na base. Outro ponto é que proporcionalmente, os vídeos publicados possuem

bitrate menores. Este é um indício de que os *Produtores de Conteúdo* ainda não utilizam as mídias de maior, bitrate para distribuir seus conteúdos com uma qualidade mais alta. Tal fato pode estar a dificuldade de distribuir seus conteúdos em alta qualidade (baixa aceitação dos usuários) e também ao custo dessa distribuição.

Na literatura a distribuição do bitrate também foi estudada em [Maia & Virgilio Almeida, 2009, Gill et al., 2007]. No trabalho de [Maia & Virgilio Almeida, 2009] vemos uma distribuição pouco concentrada para os vídeos de um sistema de compartilhamento de vídeos (Veoh) chegando a taxas de até 1600 kbps e para outro (YouTube) mais de 80% se concentram entre 300kbps e 320kbps. As conclusões sobre o Youtube se repetem em outros estudos como em [Gill et al., 2007].

5.4.7 O_9: Popularidade de objetos

A popularidade dos objetos possui implicações importantes no planejamento e desenvolvimento dos sistemas para quem distribui os conteúdos e possibilita a melhor compreensão do consumo de seus conteúdos para quem produz tais conteúdos. A análise de popularidade de vídeo é uma análise muito comum na literatura [Gill et al., 2007, Cha et al., 2007, Cherkasova & Gupta, 2002, Almeida et al., 2001, Yu et al., 2006, Duarte et al., 2007a, Silva et al., 2009, Maia & Virgilio Almeida, 2009]. A abordagem varia entre análise da distribuição, se pode ser modelada como uma Zipf de um ou vários modos, ou pelo estudo da concentração das visualizações, onde se concentram a maior parte das visualizações. Nesta seção iremos avaliar a popularidade dos objetos de vídeos e thumbnail utilizando duas abordagens diferentes a análise de Zipf e o estudo de concentração.

A lei de Zipf diz que forem ordenados de acordo com sua frequência de ocorrência, com o mais popular sendo o primeiro, o segundo mais popular o segundo e assim sucessivamente, então a frequência de ocorrência F está relacionada com a posição do objeto R de acordo com a seguinte relação,

$$F \sim R^{-\beta} \tag{5.1}$$

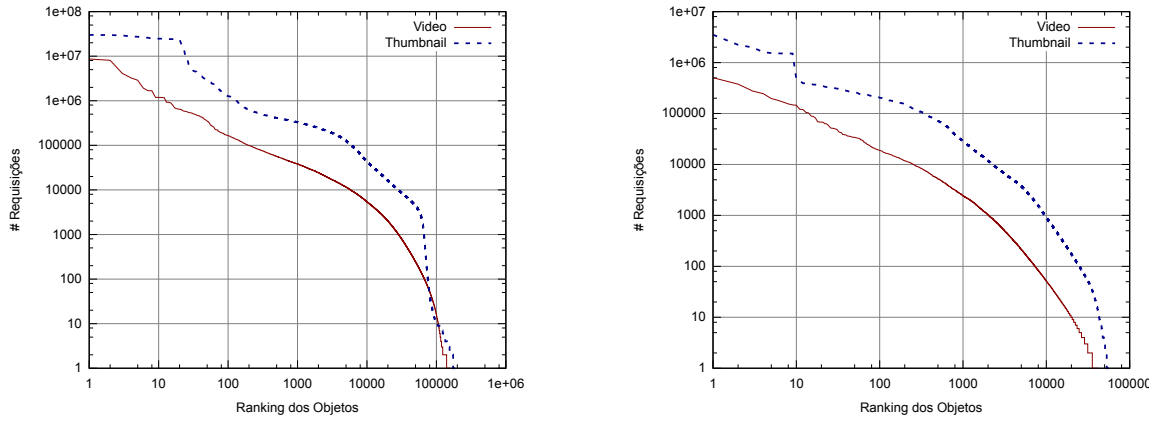
sendo que a constante β é perto de 1. [Zipf, 1949].

É comum se observar na literatura o fornecimento dos parâmetros de *fitting* da curva observada e (β e R^2) para que estes valores possam ser utilizados em simuladores e geradores de carga sintética. Por exemplo, valores de β são fornecidos em [Maia & Virgilio Almeida, 2009](0.66), [Gill et al., 2007](0.56),

[Cherkasova & Gupta, 2002](1.6) e [Silva et al., 2009](1,7 e 2,6).

Contudo, uma forma simplificada da verificação da aplicabilidade da lei de Zipf é plotar os objetos ordenados versus sua respectiva frequência numa escala log-log. Uma observação de uma linha reta é um indicador que a lei de Zipf se aplica. Tal análise foi realizada para todos os vídeos e thumbnails na base, por todo o período de estudo, e pode ser visto na Figura 5.41(a). Nessa figura vemos que a curva de popularidade dos thumbnails é muito irregular e não segue um modelo linear. Poderíamos, contudo, aproximar a curva de popularidade dos vídeos em duas semirretas, uma para os vídeos mais populares ($\approx$ 6000), e outra para os demais. Seria uma aproximação por duas distribuições de Zipf distintas. O mesmo comportamento foi observado nos trabalhos de [Silva et al., 2009] e [Almeida et al., 2001].

Na tentativa de averiguar se a amostragem muito grande de tempo e objetos pudesse prejudicar a análise, realizamos a mesma avaliação com a amostra de apenas 1 mês. Sem perda de generalidade escolhemos o mês de agosto de 2010 por não haver nenhum evento em especial que pudesse influenciar na análise. Tal análise (Figura 5.41(b)) apresentou os mesmos resultados. A distribuição da curva de popularidade dos vídeos poderia ser descrito por uma Zipf de duas componentes. Uma para os vídeos mais populares ($\approx$ 2000) e outra para as demais.



(a) Todo o período de estudo

(b) Agosto de 2010

Figura 5.41. Avaliação da aplicabilidade da Lei de Zipf aos objetos da Base

É interessante notar, em ambos os gráficos da Figura 5.41, que para o caso dos thumbnails, os objetos mais populares, algo em torno dos 10% mais populares, se destacam sendo muito semelhantes suas popularidades. Tal comportamento é visto na distorção nos thumbnails mais populares em tais gráficos. Acreditamos que tal

comportamento se deve ao recurso de destaque nos sites dos portais dos *Produtores de Conteúdo*.

Através da literatura vemos que o comportamento da distribuição das visualizações, ou seja a popularidade do conjunto dos vídeos em análise, independe da magnitude das visualizações e da coleção. Mesmo em trabalhos específicos e com poucas visualizações, como vídeos ao vivo e somente centenas de visualizações [Silva et al., 2009], ou servidores corporativos com algumas dezenas de milhares de visualizações [Cherkasova & Gupta, 2002], o comportamento pode ser modelado como uma distribuição de Zipf com uma ou várias componentes.

Outra abordagem para compreender como está a distribuição das requisições para os objetos da base é através da análise de concentração. O objetivo desta análise é determinar a fração das referências totais contabilizadas para os objetos mais populares. O gráfico da Figura 5.42 apresenta a ordenação dos objetos, normalizado entre 0 e 100, pela sua popularidade no eixo das abscissas versus a fração acumulada dos n-objetos menos populares. Através deste gráfico podemos ver apenas 10% dos objetos são responsáveis por acumular quase 90% das requisições. Constatamos assim que para os objetos dessa coleção aplica-se o princípio de Pareto na proporção 90-10.

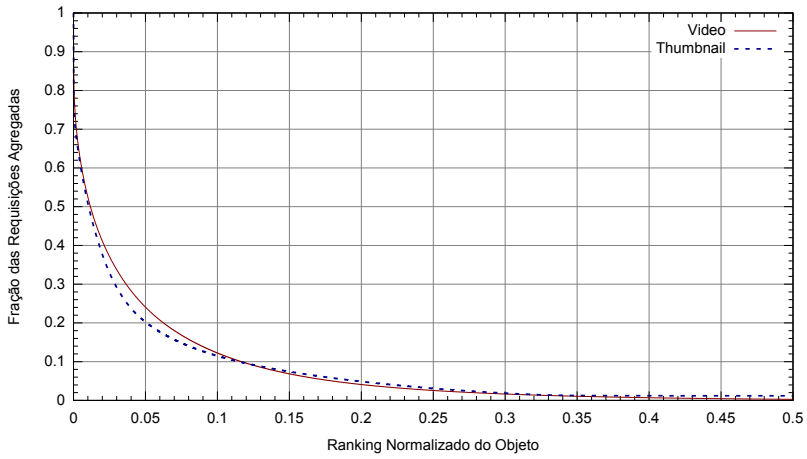


Figura 5.42. Análise da concentração das requisições dos objetos

A análise de concentração varia de acordo com o ambiente de estudo. Em [Gill et al., 2007] o estudo realizado foi num contexto muito específico e a coleta refletia somente o universo do campus universitário, por isso o princípio de pareto não foi verificado, 10% dos vídeos tiveram 39.7% das visualizações, aumentando a quantidade de vídeos para 20%, ainda assim, as visualizações acumuladas sobem para 52.4%.

Em [Yu et al., 2006] a fração de visualizações também não segue pareto mas não chega a ser tão baixa. Com 10% dos vídeos recebendo 60% das visualizações (subindo

para 23% de vídeos temos 80% de visualizações). A princípio os autores creditam esse comportamento devido a grande diversidade de vídeos na biblioteca, contudo, no presente estudo temos uma diversidade muito grande de vídeos e não temos o mesmo problema.

Valores mais semelhantes ao nosso são visto nos trabalhos de [Maia & Virgilio Almeida, 2009] (10% → 85% e 20% → 93.4%), [Cha et al., 2007] (10% → 80% e 20% → 90%), [Cherkasova & Gupta, 2002](14% → 90%), e [Duarte et al., 2007a] (10% → 76%)

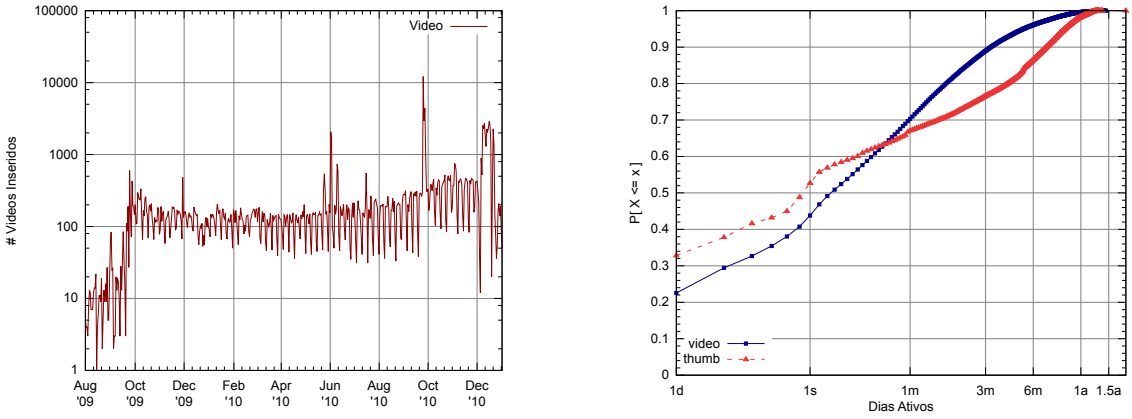
5.4.8 O_{10} : Distribuição da idade de objetos

Nesta seção iremos avaliar aspectos relativos a idade do objeto na base. A idade do objeto é o tempo decorrido entre a inserção do objeto na base e a sua remoção. Entretanto, não é usual na plataforma em análise a remoção dos objetos. Assim, iremos considerar o tempo de vida do objeto como o tempo em que este permaneceu ativo, ou seja, recebendo requisições. Para isso, realizamos uma análise do número de dias de requisições por objeto e plotamos a sua função de distribuição cumulativa (Figura 5.43(b)).

Notamos que a taxa de objetos que possui tempo de vida de apenas um dia é muito alta ($\sim$ 23% para vídeos e $\sim$ 33% para thumbnails) indicando que é necessário otimizar a produção de conteúdo. Metade dos objetos possuem tempo de vida curto, para os vídeos o período é de 10 dias, para os thumbnails 1 semana. Mas é interessante notar que existem muitos objetos que possuem tempo de vida longo, maior que 1 mês, a taxa para vídeos é de 30% e para os thumbnails 35%.

É possível perceber também que existe uma relação inversa entre os thumbnails e vídeos com relação ao tempo de atividade dos objetos. Enquanto os vídeos possuem, proporcionalmente, menos objetos pouco ativos (com menor tempo de vida), os thumbnails com maior tempo de atividade, permanecem em geral mais ativos que os vídeos. Acredita-se que este comportamento é devido à diagramação dos portais de mídias online que possuem uma seção de vídeos em destaque. Como o algoritmo de recomendação de vídeos em destaque segue uma heurística muito simples, mostrando sempre os objetos mais populares, os thumbnails desses conteúdos permanecem sempre ativos.

Na literatura só encontramos o trabalho de [Gill et al., 2007] que faz uma avaliação da idade dos vídeos no YouTube. Nesse estudo, a abordagem para a idade dos vídeos é o tempo desde a inserção no sistema, o que difere da nossa avaliação, contudo os resultados mostram que os vídeos mais populares, se avaliando um período de ape-



(a) Distribuição da inserção de vídeos

(b) *CDF* do Tempo de vida dos objetos

Figura 5.43. Informações a respeito da idade dos objetos.

nas um dia possuem, geralmente, idade menor que 3 dias. O mesmo se observa para períodos de semanas e meses, ou seja, o vídeo mais popular da semana/mês tende a ter menos de uma semana/mês de vida.

A taxa de inserção de objetos novos na base também é uma métrica interessante de se avaliar. As informações a respeito da taxa de inserção dos objetos do tipo thumbnail só estavam disponíveis para um pequeno intervalo dentro do período avaliado. Desta forma optamos por avaliar somente a taxa de inserção dos vídeos. Avaliando a proporção entre a taxa de inserção dos dois objetos (para o período disponível) observamos que o número de thumbnails inseridos na base, é em geral maior que a de vídeos.

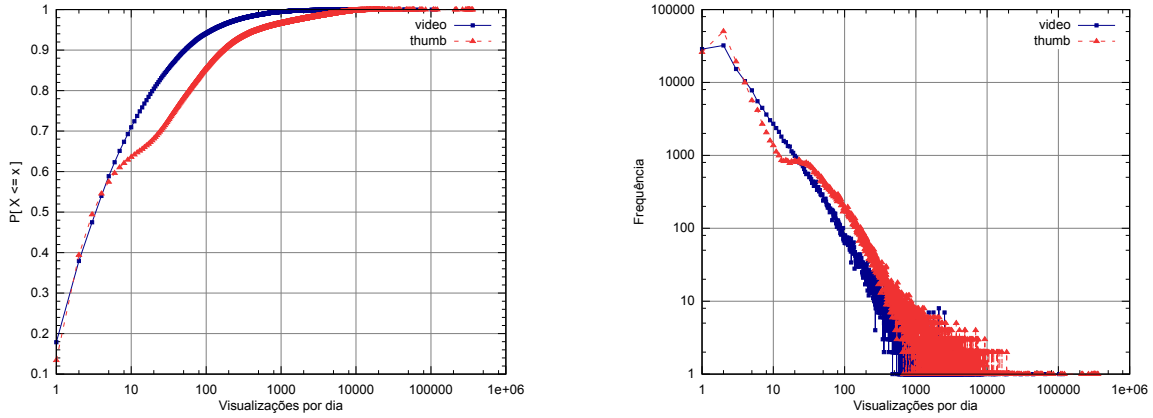
O gráfico da Figura 5.43(a) mostra a distribuição da inserção de vídeos na plataforma. Todos os picos de inserção, por exemplo, final de setembro de 2009, junho de 2010, outubro de 2010, são marcados por uma alta taxa de inserção de vídeos na plataforma devido a adoção da ferramenta por novos *Produtores de Conteúdo*, período conhecido como migração de dados. Vemos que os períodos marcados por um aumento na audiência, por exemplo, Novembro também é marcado pela maior inserção de vídeos. O valor médio de inserção diária de vídeos na plataforma é de aproximadamente de 200.

5.4.9 O_{11}: Relação idade versus popularidade de objetos

Na seção anterior analisamos o tempo de vida dos objetos que é o tempo em que eles ficam ativos. Agora iremos avaliar a relação entre a idade do objeto (período

ativo) versus a popularidade, procurando entender, por exemplo, qual é a popularidade relativa ao número de dias de vida e como é feita a distribuição da fração agregada das visualizações dos objetos durante o intervalo de atividade.

Inicialmente iremos avaliar a popularidade relativa dos objetos. Os gráficos da Figura 5.44 foram criados computando a fração da popularidade total dos objetos pelo número de dias ativos. A distribuição cumulativa do número de visualizações por dia pode ser vista na Figura 5.44(a). Através deste gráfico podemos ver que mais de 70% dos vídeos não recebem mais que 10 acessos diários, se analisarmos conjuntamente a distribuição das visualizações diárias (Figura 5.44(b)) vemos que tais objetos compõe algumas dezenas de milhares de vídeos.



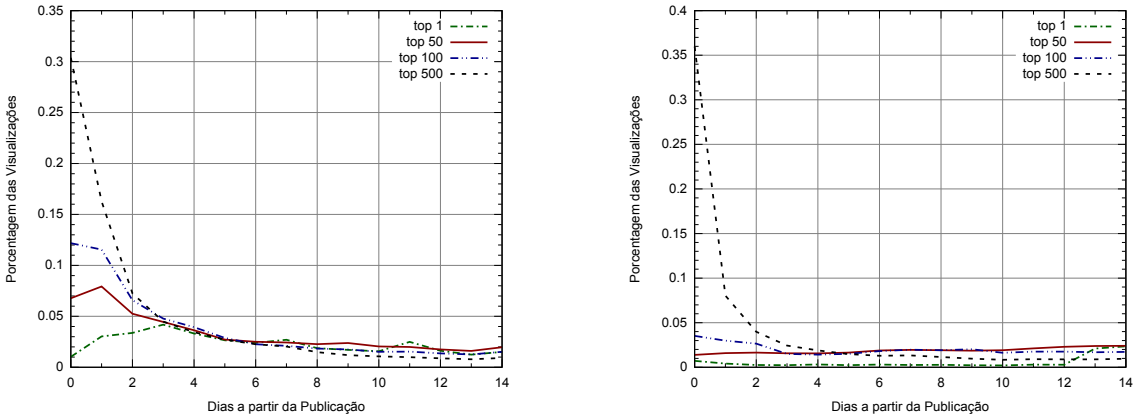
(a) *CDF* (b) Distribuição

Figura 5.44. Informações relativas do número médio de visualizações por dia dos objetos.

E novamente podemos observar que uma pequena parcela dos objetos são responsável pela maior parte das visualizações, vemos que somente 5% dos objetos da base recebem visualizações com ordem de grandeza que variam de algumas centenas a dezenas de milhares de visualizações.

Para analisar a popularidade dos objetos a medida que este envelhece, resolvemos analisar o percentual das visualizações agregadas dos objetos a partir da data de publicação. Para isto fizemos os gráficos da Figura 5.45:

Para ambos os casos, vídeos (Figura 5.45(a)) e thumbnails (Figura 5.45(b)), vemos que as mídias menos populares concentram a maior parte das suas visualizações nos primeiros dias de vida, em contrapartida, os objetos mais populares tendem a possuir suas visualizações mais dispersas durante todo seu período de vida. A única diferença que notamos é que os vídeos mais populares tendem a receber uma fração



(a) Vídeos (b) Thumbnail

Figura 5.45. Média das frações das requisições por dia, a partir do dia da publicação.

maior das visualizações nos primeiros dias se comparado aos thumbnails. Tal comportamento decorre do fato que os thumbnails são utilizados de maneira mais distribuídas em várias páginas e como divulgação de conteúdo, assim este continua recebendo requisições mesmo quando inseridos em contextos diferentes do seu conteúdo em que pertence.

5.4.10 Conclusões sobre a Camada O

Nesta seção, avaliamos vários aspectos da plataforma em estudo sob o ponto de vista dos objetos (vídeo e thumbnail). Inicialmente, realizamos uma projeção de algumas análises da camada **R** para os diferentes objetos desta camada e finalizamos realizando novas avaliações com as unidades de análises exclusivas dessa camada.

Avaliando a segmentação do número da taxa de resposta, vimos que as requisições de thumbnail são, em geral, uma ordem de grandeza maiores que as requisições para vídeos. Realizamos uma análise comparativa entre a taxa de requisições normalizada para os dois tipos de objeto e propomos uma métrica para medir o engajamento do usuário.

Com tal métrica, chegamos a conclusão de que o engajamento (disposição em assistir os vídeos) do usuário em vídeos é maior nos finais de semana. Nos dias úteis o usuário tende a *navegar* mais entre as páginas (busca por conteúdo) do que se dedicar a assistir os vídeos. Tal evidência mostra que a adoção de publicidades estáticas nos thumbnails pode ser uma estratégia efetiva para tal período. De forma complementar, nos horários em que esse engajamento aos vídeos é maior a probabilidade de conversão

da publicidade, o que pode ser utilizado para diferenciar o preço das publicidades.

Uma hipótese que levantamos é a alta taxa de acesso a objetos multimídia a partir do ambiente de trabalho. Uma das evidências de tal fato é a abrupta queda no consumo na rede em avaliação às 18h, período em que se termina o expediente na maior parte das empresas do país.

Avaliando a segmentação da largura de banda por objetos, concluímos que objetos pequenos distorcem a aproximação de tal métrica e devem ser descartados. Vimos que a maior parte dos usuários que efetivamente consomem vídeos ($\sim$ 45%) possuem velocidade de conexão entre 1 e 5 Mbps, mostrando a difusão da tecnologia de banda larga e também que tal tecnologia é essencial para a consolidação dos vídeos como principal mídia da Internet. Contudo, ainda existem $\sim$40% de consumidores de vídeo que possuem conexões com velocidades menores que 1 Mbps. Acreditamos que com uma velocidade de conexão maior tais usuários poderiam aumentar ainda mais tal consumo.

Na análise do tamanho das transferências dos objetos percebemos a presença de thumbnails com tamanho inadequados para tal tipo de objeto o que pode degradar a qualidade do serviço para o usuário. Através da avaliação da porcentagem do tamanho dos objetos que são efetivamente entregues, reforçamos a nossa evidência anterior da busca por conteúdo uma vez que existe uma porcentagem de thumbnails, objetos pequenos, que não são totalmente entregues. Mais uma vez com esta análise percebemos um maior engajamento à noite.

Avaliando as propriedades dos objetos mostramos que vídeos de alta resolução e formato *widescreen* estão sendo mais adotados recentemente. Outro ponto é a disponibilização de conteúdos com tamanhos e duração altos para Internet, 1,8GB e até 2h. Sendo que vídeos de longa duração estão cada vez mais frequentes.

Com relação a popularidade dos objetos, mostramos que a concentração das requisições segue o princípio de Pareto [Newman, 2005] numa proporção 90-10. Entretanto, a aplicabilidade da lei de Zipf não se verifica para o caso de thumbnails. Já para os vídeos tal aplicabilidade existe, mas não é tão forte como para sites *UGC* como youtube [Gill et al., 2007].

Com a análise do tempo de vida dos objetos concluímos que apenas 30% dos vídeos permanecem ativos por mais de 30 dias. A relação para os thumbnails é similar, mas, devido aos mecanismos de recomendação, existe uma fração maior destes que permanecem mais ativos por maiores período de tempo.

Ao cruzar as informação de popularidade e o tempo de vida dos objetos, as análises indicaram que as mídias mais populares possuem uma distribuição das requisições mais espalhadas durante seu tempo de vida. Mas à medida que as mídias ficam

menos populares, estas concentram maior parte das suas visualizações nos primeiros dias após sua publicação. Concluímos então que, em geral, a maior parcela das requisições já ocorreram após o quarto dia de publicação da mídia.

Com as análises realizadas para esta camada da hierarquia da informação, levantamos evidências capazes de compreender melhor a forma como os objetos são entregues pela rede da plataforma em estudo. Em seguida, avaliaremos os dados *Liquid*^{TM} utilizando os metadados dos objetos para compreender o impacto que a diferenciação por tipo de conteúdo pode causar.

5.5 Análises da Camada do Conteúdo (C)

Nesta seção realizaremos uma análise dos conteúdos presentes na plataforma utilizando como base as diretrizes da metodologia da Seção 3.4. Conforme dito anteriormente, o que define o conteúdo de um objeto depende unicamente deste. O que iremos fazer nesta parte do estudo de caso é avaliar os metadados textuais associados aos objetos, que fornecem indícios deste conteúdo, para melhor compreender as informações dos vídeos distribuídos pela plataforma em análise.

A princípio vamos avaliar o domínio de estudo desta camada realizando um sumário das metadados relacionados aos objetos da plataforma, explicando a semântica de cada um e como o cadastramento deste metadados é realizado. Em seguida iremos apresentar as análises desta seção que incluem algumas das análises da Camada **O** segmentanda pelos diferentes tipos de conteúdos.

Lembramos que o objetivo deste estudo de caso é apenas ilustrar a aplicabilidade da metodologia em dados reais e não uma aplicação extensiva da metodologia.

5.5.1 Descrição dos metadados dos Vídeos

A *Liquid*^{TM} é uma ferramenta para gestão de conteúdos multimídia e uma das suas funcionalidades é o cadastro de metadados para uma determinada mídia. Por exemplo, ao inserir um vídeo na plataforma, o editor cadastra os metadados do tal vídeo utilizando a interface mostrada na Figura 5.46. Na parte superior a esquerada é permitido visualizar o conteúdo (no caso de vídeos), logo abaixo é mostrado o thumbnail do conteúdo. O formulário da direita é utilizado para o cadastro dos metadados.

Nesta figura podemos ver que existem 7 metadados textuais que podem ser utilizados pelos editores no momento do cadastro de uma mídia na plataforma. Tais metadados são explicados na Tabela 5.16. Note que o thumbnail atua também como uma metainformação da mídia principal.

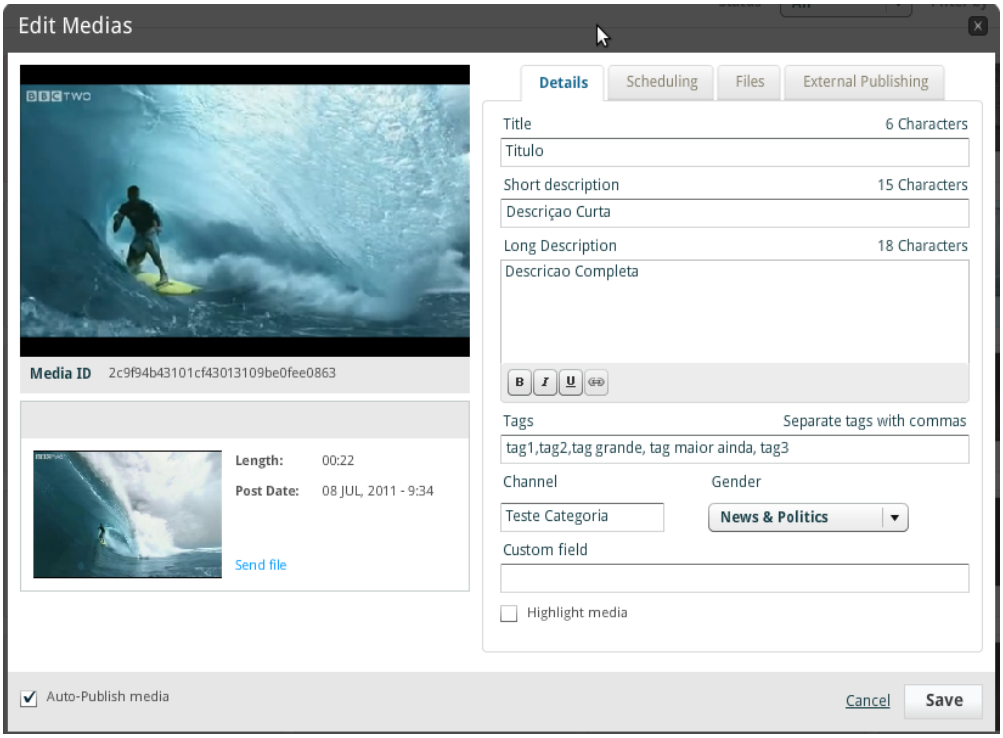


Figura 5.46. Interface de cadastro de metadados das mídias na *Liquid*™

Metadado	Descrição
Título	Descrição em poucas palavras sobre o que se trata o conteúdo do objeto.
Descrição	Um sumário a respeito da mídia. Pode conter uma descrição detalhada do conteúdo podendo possuir tamanho relativamente grande.
Descrição Curta	Um pequena descrição do objeto.
Tags	Conjunto de termos que representam temas associados com o conteúdo da mídia.
Categoria	Grupo ao qual o objeto pertence por apresentar características comuns aos outros integrantes. É um campo customizável pelo *Produtor de Conteúdo*. Geralmente agrupa os objetos em grupos semanticamente relacionados com os programas, canais e produtos do *Produtor de Conteúdo*.
Gênero	Possui a mesma semântica que a categoria, entretanto os valores são globalmente únicos e definidos pela plataforma.
Campo customizável	Campo genérico que pode ser utilizado para fins específicos dentro do modelo de negócio do *Produtor de Conteúdo*.

Tabela 5.16. Lista dos metadados textuais disponíveis na plataforma *Liquid*™.

A maneira com que os metadados são utilizados depende muito da aplicação do *Produtor de Conteúdo* e do projeto em questão. Dentre os metadados textuais alguns

são pouco utilizados como o campo customizável, que de fato, raramente é utilizado. Mas em geral informações como título, descrição, tags e categorias são frequentemente empregados.

É importante ressaltar que o thumbnail é um objeto associado a outro objeto, uma imagem que tenta atrair a atenção do usuário. Neste nível de abstração iremos considerar o thumbnail como uma meta informação do objeto princial e focaremos nos conteúdos em vídeos.

Na Tabela 5.17 mostramos a distribuição dos metadados entre os vídeos da base uma vez que nem todos os metadados estão presente em todos os vídeos. Veja que os campos que são pouco utilizados são apenas o gênero, que não se relaciona diretamente com o negócio do *Produtor de Conteúdo* e a descrição curta que muitos *Produtores de Conteúdo* não utilizam.

Metadado	**Percentual**
Título	99,93%
Descrição	87,59%
Descrição Curta	30,26%
Tags	71,26%
Categoria	93,44%
Gênero	5,42%

Tabela 5.17. Percentual dos vídeos que contém o respectivo metadado.

Apesar dos metadados como títulos, descrição, *tags* e categorias serem bastante empregados o conteúdo destes campos nem sempre são de qualidades e algumas vezes são má empregados. Por exemplo, um *Produtor de Conteúdo* específico faz uso de alguns dos metadados com outra semântica (categoria na descrição, e o título na descrição curta) e tratamos tais caso.

Na *Liquid*$^{\text{TM}}$, o metadado *categoria* pode ser criado pelo *Produtor de Conteúdo* para que este organize seus vídeos. Conforme dito na Tabela 5.16, a categoria é especificada pelo próprio *Produtor de Conteúdo* e geralmente está associada ao seu negócio. Em muitos casos os nomes adotados para as categorias se sobrepõem. Por exemplo, existem aproximadamente 4600 categorias em toda a plataforma, se removermos as que possuem nomes repetidos o número cai para 2800 categorias. Esse fato é devido ao uso de nomes de categorias amplamente adotadas como por exemplo Jornalismo, Entretenimento, Esportes, Notícias, Política, etc entre diferentes *Produtores de Conteúdo*. Uma restrição importante de se ressaltar é que uma mídia só pode estar contida

em uma única categoria, mas não é obrigatório associar uma categoria a uma mídia. As categorias podem se associar de forma hierárquica, como uma árvore, possibilitando subcategorias.

O metadado gênero foi adotado pela plataforma recentemente para que exista uma entidade que correlacione vídeos de diferentes *Produtor de Conteúdo* com uma mesma classe de equivalência de conteúdo. Os valores possíveis são: *Humor*; *Entretenimento*; *Filmes e Animação*; *Música*; *Notícias e Política*; *Pessoas e Blogs*, *Animais*, *Ciência e Tecnologia* e *Esportes*.

Entretanto, como pudemos ver na Tabela 5.16 a sua adoção é muito pequena para que possamos segmentar os conteúdos de nossa base de estudo utilizando este gênero. Para tratar esse problema decidimos derivar um gênero, ou seja uma categoria global entre diferentes *Produtores de Conteúdo*, a partir da categoria definida pelo próprio *Produtores de Conteúdo*. Adotaremos o termo ***gênero*** daqui para frente para diferenciar a *categoria derivada* deste processo da *categoria original* do vídeo.

A abordagem para a derivação do gênero foi manual. Listamos 1000 categorias, ordenadas de forma decrescente de acordo com número de vídeos e visualizações e derivamos os gêneros partir das seguintes informações:

- As categorias ancestrais, ou seja, em qual hierarquia esta categoria estava inserida. Por exemplo: *Jornalismo* $\rightarrow$ *Brasil* $\rightarrow$ *Rio de Janeiro* $\rightarrow$ *Boletim da Tarde*.

- As informações de qual *Produtor de Conteúdo* e qual projeto a categoria pertencia.

- Informações adicionais dos vídeos (metadados), caso as informações não fossem suficientes.

Categorias que possuíam poucos vídeos (< 50), fossem pouco populares (< 500 visualizações) e demandassem análises mais detalhadas eram simplesmente ignoradas. Um gênero não podia diferenciar da sua categoria se houvesse uma correlação igual na lista de gêneros, por exemplo, se a categoria raiz fosse *Esportes* esta deveria ser mantida. Em caso de análises dúbias ou nos casos em que a categoria não se enquadrasse dentro de um dos gêneros esta seria ignorada ficando sem gênero.

A cobertura final dos vídeos classificados foi expressiva. A lista de gêneros utilizados e o resultado dessa classificação manual podem ser vistos na Tabela 5.18

Os trabalhos de [Gill et al., 2007] e [Maia & Virgilio Almeida, 2009] realizam levantamentos relativos às categorias dos vídeos, o que equivale ao nosso *gênero*. Ambos trabalham com vídeos do YouTube e disponibilizaram a quantidade de vídeos

por categoria. Em [Gill et al., 2007] as categorias Entretenimento (23,97%), Música (22,35%), Comédia (13,60%) e Esportes (11,26%) são as que possuem mais vídeos. Já para [Maia & Virgilio Almeida, 2009], as categorias com mais vídeo são, Música (22,43%), Entretenimento (19,56%), Pessoas e Blogs (13,07%) e Comédia (11,91%). Se comparado com a distribuição existente no presente estudo vemos que a distribuição de vídeos tem um perfil diferente. Temos uma predominância de vídeos de Notícias e Política o que na literatura é baixo (3,34% e 3,8% respectivamente).

%	**Gênero**	%	**Gênero**
48,82%	Notícias e Política	0,92%	Filmes e Animação
16,90%	*Sem Gênero*	0,90%	Música
13,64%	Entretenimento	0,64%	Pessoas e Blogs
10,55%	Esportes	0,48%	Ciência e Tecnologia
4,05%	Variedades	0,36%	Negócios e Finanças
1,31%	Guias e Estilo	0,23%	Humor
1,13%	Saúde/Bem-estar	0,07%	Automóveis

Tabela 5.18. Distribuição dos gêneros e porcentagem dos vídeos nestes após classificação manual

A maior parte dos conteúdos estão relacionados ao gênero jornalístico que chamamos de *Notícias e Política*. Essa proporção é natural dado o perfil da maior parte dos projetos dos *Produtores de Conteúdo* que utilizam a plataforma. Adicionalmente, a geração de conteúdo jornalístico é mais fácil e ocorre naturalmente todos os dias. O segundo gênero mais popular é o *Entretenimento* seguido por *Esportes*. Novamente tal fato deriva do nicho de atuação dos *Produtores de Conteúdo* que em sua maioria são grupos de mídias que atraem a maior parte de seus acessos nesses conteúdos.

Aproximadamente 17% dos vídeos não receberam um gênero por simplificação ou simplesmente por não se enquadrarem em nenhum destes. Os vídeos publicitários são um bom exemplo para esse casos que deveriam ser enquadrados em um gênero separado.

A baixa frequência dos outros gêneros é natural devido à alta generalidade dos gêneros escolhidos, ao domínio dos *Produtores de Conteúdo* e pelo fato de termos simplificado a classificação dos vídeos limitando que um vídeo só possa estar associado a um gênero. Se uma notícia num jornal local fala de um assunto relacionado a saúde e bem estar das pessoas ele estará incluso dentro da categoria do jornal recebendo

automaticamente o gênero de *Notícias e Política* mas poderia também estar presente no gênero de *Saúde/Bem-estar*.

Como a maior parte dos vídeos ($\sim$ 70%) se enquadram nos três gêneros mais frequentes, para ilustrar a nossa metodologia, iremos segmentar o resto do nosso estudo nesses três conteúdos. Das análises descrita na Seção 3.4, escolhemos as mais interessantes para este estudo de caso e iremos mostrá-las nas próximas seções.

5.5.2 C_{10}: Distribuição das Propriedades dos Objetos por Conteúdo

Neste seção iremos avaliar a distribuição das propriedades dos vídeos segmentados por conteúdo. Dentre as várias propriedades dos vídeos vistas na subseção 5.4.6 a que pode sofrer maior impacto pelo tipo de conteúdo é a duração dos vídeos. As outras propriedades estão mais relacionadas as decisões dos *Produtores de Conteúdo* e a adoção de certos padrões para geração e distribuição dos conteúdos sendo menos impactadas pelo natureza do conteúdo.

O duração de um vídeo está intimamente ligado ao seu conteúdo, por exemplo, se um vídeo é um capítulo de uma novela na íntegra, é natural que este tenha mais que 30 minutos, entretanto esta duração é incompatível com um vídeo de notícia sobre a alta do dólar. Para avaliar tais aspectos as análises seguintes contrastam a duração de vídeos entre os diferentes conteúdos em estudo.

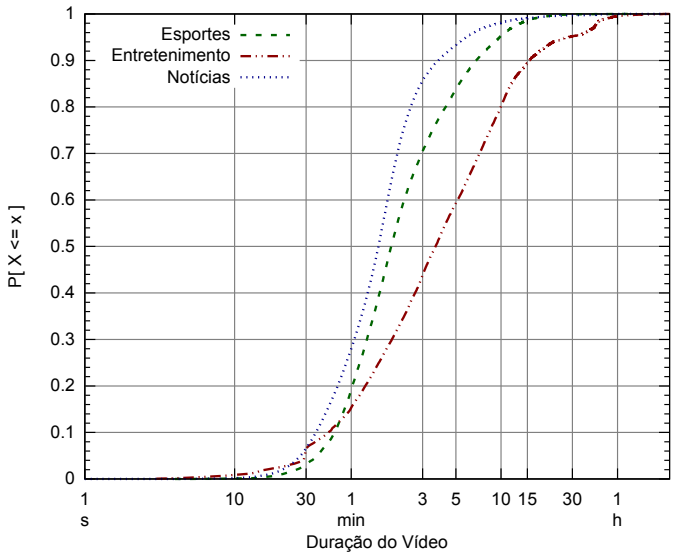
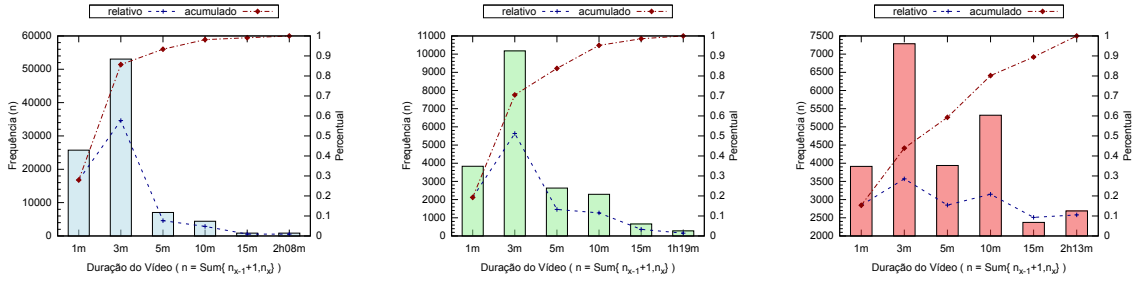


Figura 5.47. *CDF* da duração dos vídeos por *Gênero*

A Figura 5.47 contrasta a distribuição cumulativa da duração dos vídeos entre

os diferentes conteúdos. Nota-se neste gráfico que existe uma diferença clara entre a distribuição das durações dos vídeos e seu conteúdo. Em geral, os vídeos de notícias são mais curtos com quase 90% dos vídeos menores que 3 minutos. Em contrapartida, os vídeos de entretenimento são em geral mais longos. A fração menor que 3 minutos é aproximadamente 40% sendo que 80% possuem duração menor que 10 minutos. Já os vídeos de esportes são ligeiramente maiores que os de notícias mas seu comportamento em geral se assemelha com os vídeos jornalísticos.

Para possibilitar uma melhor comparação entre as durações dos três conteúdos diferentes a Figura 5.48 mostra os histograma para cada conteúdo para os mesmos intervalos de tempo. Os vídeos de notícias (Figura 5.48(a)) possuem durações curtas, características de vídeos informativos que geralmente são objetivos. A fração de vídeos maiores é cada vez menor a medida que aumenta-se a duração dos vídeos. Os vídeos maiores que dez minutos representam apenas 2%, sendo 1% para vídeos entre 10 e 15 minutos e mais 1% para vídeos maiores que 15 minutos.



(a) Notícias e Política (b) Esportes (c) Entretenimento

Figura 5.48. Comparativo entre a distribuição das durações através dos histogramas para cada gênero.

Já os vídeos de de esportes (Figura 5.48(b)) possuem durações maiores se comparado aos vídeos de notícias, sua distribuição para vídeos menores que 3 minutos ainda é alta, 70%, mas a fração de vídeos maiores é superior ao de notícias. Temos 13% de vídeos com duração entre 3 e 5 minutos, 11% de vídeos entre 5 e 10 minutos e mais de 4% dos vídeos com duração superior a 10 minutos. Acreditamos que essa semelhança de distribuição com notícias vem do fato que boa parte desses vídeos também possuem caráter jornalístico informacional com foco em esportes, o que é natural. Já os vídeos maiores podem estar associados a resumos de jogos, melhores momentos ou cobertura integral de algumas modalidades.

O conteúdo que possui vídeos com durações mais longas são os vídeos de entretenimento (Figura 5.48(a)). Tal fato se deve à característica de proporcionar diversão ao

usuário, que em geral, estará mais disposto a assistir um vídeo maior. Além disso tal conteúdo possui uma diversidade de subcategorias que aumenta a variedade do tamanho dos vídeos. Ao contrário dos outros conteúdos, a maior parte dos vídeos não está entre o intervalo de até 3 minutos. Os vídeos maiores que 5 minutos compõe 57% dos vídeos, sendo distribuídos da seguinte forma: 15% entre 3 e 5 minutos, 20% entre 5 e 10 minutos, 9% entre 10 e 15 minutos e 10% dos vídeos são maiores que 15 minutos podendo possuir durações maiores que 2h.

Com as análises feitas nesta seção, vimos que a duração dos vídeos está relacionado com o seu o conteúdo. Podemos encontrar vídeos de tamanhos variados de alguns poucos segundos até poucas horas de duração. Contudo a predominância por vídeos curtos ainda é muito forte na Internet Brasileira o que nos sugere que o engajamento por parte dos usuários ainda é baixa se comparada a televisão.

5.5.3 C_{11} : Popularidade dos Conteúdos

Avaliaremos nesta seção a popularidade dos diversos conteúdos que são distribuídos pela plataforma da Samba Tech. Na Tabela 5.19 mostramos um sumário da popularidade dos gêneros definidos na Seção 5.5.1. Nesta Tabela mostramos o número de mídias que cada gênero possui, o número de visualizações acumulada e uma média de visualizações por mídia. Os gêneros estão ordenados de acordo com número de mídia que cada um possui. Note que o gênero de *Notícias e Política* possui sozinho, mais mídias que os outros gêneros definidos. E mesmo com essa enorme quantidade de vídeos a média de visualizações por mídia ainda é relativamente alta.

Uma aproximação do grau de popularidade do gênero pode ser inferido através da média de visualizações por mídia. Esta taxa de visualizações por mídia possibilita a comparação do grau de interesse dos usuários entre os gêneros. Podemos ver que o gênero de *Entretenimento* é o que mais atrai visualizações dentre os conteúdos com mais de 9800 visualizações por mídia e mais de 253 milhões de visualizações. O gênero de *Automóveis* possui o menor número de vídeos e a segunda maior taxa de visualização por mídia com quase 3000 visualizações por mídia. Ja o gênero com menor taxa de visualização é o de *Saúde/Bem-estar* com 573 visualizações por mídia.

A coluna de visualizações médias permite ter uma noção da quantidade de visualizações o gênero tem em média por vídeo, entretanto tal informação não diz muito a respeito da popularidade dos vídeos individualmente. Para abordar este aspecto faremos a análise da distribuição acumulada de visualizações por mídia além de verificar a aplicabilidade da Lei de Zipf [Zipf, 1949] na distribuição de popularidade das mídias.

Para analisar a distribuição de visualizações através das mídias a Figura 5.49(a)

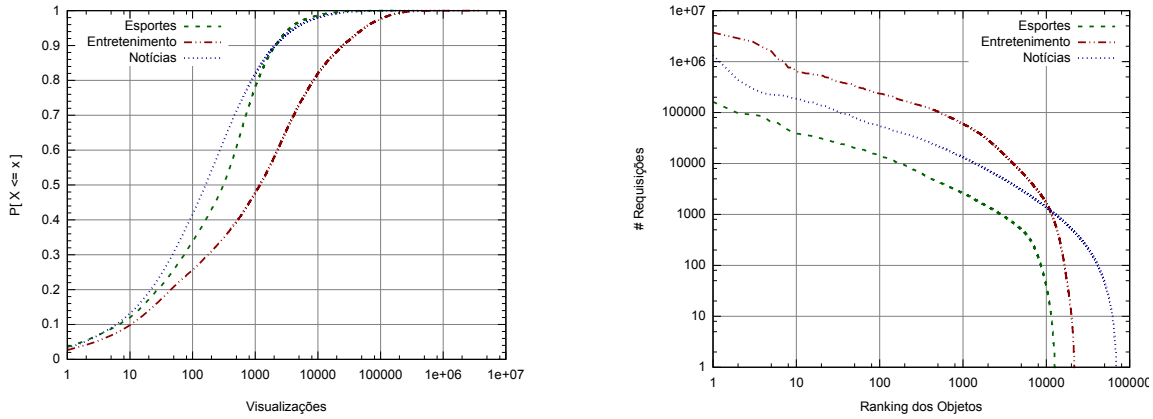
Gênero	# Mídias	Visualizações	Média Vis.
Notícias e Política	92.063	81.538.232	885,68
Sem Gênero	31.862	24.856.398	780,13
Entretenimento	25.715	253.441.068	9.855,77
Esportes	19.893	12.880.840	647,51
Variedades	7.636	19.878.036	2.603,20
Guias e Estilo	2.468	4.382.288	1.775,64
Saúde/Bem-estar	2.135	1.225.261	573,89
Filmes e Animação	1.743	4.515.876	2.590,86
Música	1.697	2.946.609	1.736,36
Pessoas e Blogs	1.202	4.588.544	3.817,42
Ciência e Tecnologia	914	1.372.906	1.502,09
Negócios e Finanças	674	427.246	633,90
Humor	442	795.277	1.799,27
Automóveis	130	382.572	2.942,86

Tabela 5.19. Distribuição das visualização das bases pelos diferentes gêneros

apresenta a distribuição cumulativa do número de visualizações totais por mídia. Contrastando os 3 conteúdos deste gráfico vemos que os vídeos de *Entretenimento* possuem uma maior distribuição de visualizações por mídias. Aproximadamente 50% das mídias possuem mais de 1000 visualização, sendo que desses 20% possuem mais de 10000 visualizações. Os conteúdos de *Esportes* e *Notícias* possuem distribuição semelhante, entretanto o conteúdo de *Esportes* possui uma taxa de visualizações por vídeo é superior a de *Notícias* entre os vídeos menos populares ($\sim$ 3000 visualizações) enquanto que na faixa de vídeos muito populares o gênero de *Notícias* e ligeiramente superior.

Para avaliar a aplicabilidade da lei de Zipf, conforme explicado na Seção 5.4.7 e em [Zipf, 1949], o gráfico da Figura 5.49(b) mostra o *ranking* dos vídeos ordenados por frequência numa escala logarítmica(log-log). É possível notar uma tendência linear para os vídeos mais populares ($\approx$ 1000) o que sugere uma possível aplicabilidade da lei de Zipf. Importante notar que as distribuições da Figura 5.49(b) possuem uma cauda pesada, tal comportamento é notado também em trabalhos como [Almeida et al., 2001] e [Yu et al., 2006] e sugerem a possibilidade da descrição por Zipf de duas componentes.

Dando seguimento na análise da popularidade dos conteúdos através das visu-



(a) *CDF* das visualizações

(b) Análise de Zipf

Figura 5.49. Avaliações de popularidade dos diferentes conteúdos, distribuição cumulativa dos víews e análise de Zipf

alizações de seus vídeos, a Figura 5.50 mostra a análise de concentração segmentada pelos conteúdos. Vemos que a relação de concentração das visualizações para os conteúdos de *Entretenimento* e *Notícias* possuem basicamente a mesma distribuição, onde verifica-se a aplicabilidade do princípio de Pareto com a proporção 90-20, para os dois casos, 90% das visualizações se acumulam em aproximadamente 20% dos vídeos mais populares. Já para o conteúdo de Esportes tal proporção é de 80-20.

A semelhança do comportamento entre *Entretenimento* e *Notícias* se deve ao fato de que em geral, as notícias e vídeos de entretenimento muito populares atraem a atenção do público como um todo, entretanto para o caso de esportes, existe uma divisão de preferência mais clara, como por exemplo o time do expectador ou esporte de preferência.

5.5.4 C_{12}: Distribuição da idade dos objetos por Conteúdo

Nesta seção iremos avaliar a distribuição da idade dos objetos segmentadas por conteúdo. O estudo tem como objetivos avaliar o tempo de vida média dos vídeos e sua distribuição além e avaliar a taxa de inserção de vídeos segmentado por cada tipo de conteúdo.

Durante os vários meses analisados nesses estudo contabilizou-se aproximadamente 150 mil inserções de vídeos na plataforma. Para avaliar a distribuição dos vídeos segmentada por diferentes conteúdos apresentamos os gráficos da Figura 5.51. Em tal figura pode-se observar a taxa de inserção de vídeos por conteúdo durante o período analisado. O gráfico da Figura 5.51(a) mostra a variação da inserção para os

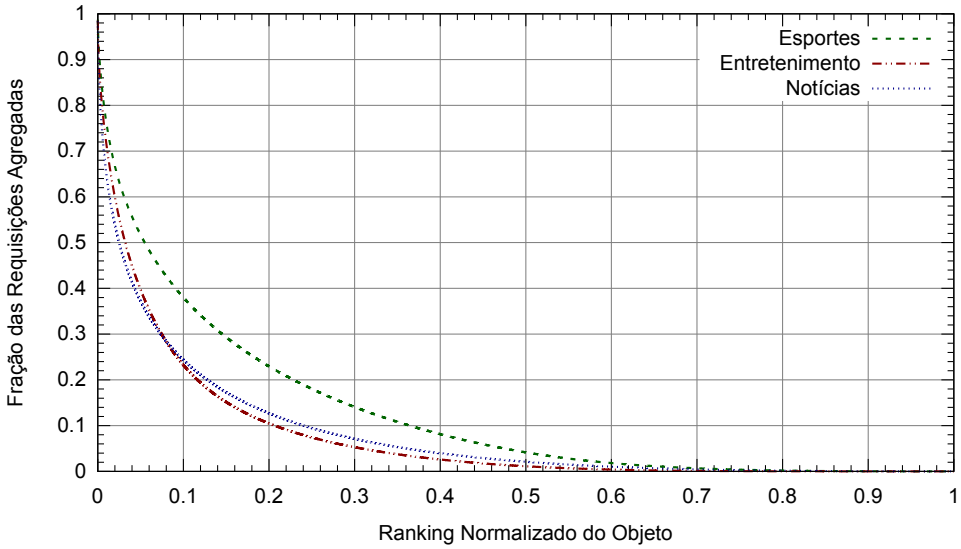
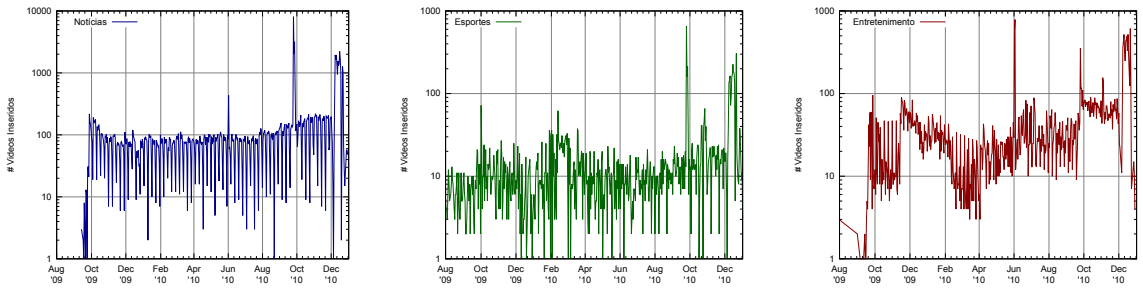


Figura 5.50. Análise de concentração para os difernetes conteúdos.

vídeos de notícias. A área vazia no período entre Agosto e Outubro de 2009 é devido ao fato dos vídeos que possuem o gênero de notícias (atribuídos manualmente) foram somente inseridos durante tal intervalo. Vemos que a taxa de inserção varia dependendo do período mas existe um comportamento cíclico semanal. Avaliando todo o período vemos que a partir de Agosto de 2010 houve um crescimento na taxa de inserção de notícias. Os picos de inserção de vídeos com valores muito altos (> 500) são atribuídos ao que chamamos de migração, que é a entrada de um novo *Produtor de Conteúdo* na plataforma com a respectiva inserção de conteúdo legado. A taxa de inserção média para os vídeos desse conteúdo foi de 150 vídeos diários



(a) Notícias e Política (b) Esportes (c) Entretenimento

Figura 5.51. Taxa de inserção de vídeos durante o período de estudo segmentadas por tipo de conteúdo. A escala do eixo y é logarítmica. Picos de inserção são devidos a entrada de novos *Produtores de Conteúdo*

As informações relativas ao conteúdo esportivo pode ser visto na Figura 5.51(b).

Vemos que neste gênero a variação é maior que a de jornalismo, adicionalmente o número de vídeos médios inseridos é bem inferior, cerca de 47 vídeos/dia.

E por último avaliamos a taxa de inserção de vídeos para os conteúdos de entretenimento (Figura 5.51(c)). Nota-se a falha inicial da curva, semelhante à de jornalismo e com a mesma justificativa. É dos três conteúdos avaliados a que possui maior variabilidade e isso se deve a sazonalidade de alguns conteúdos como séries, *reality shows* e outros conteúdos que geralmente são espelhados dos conteúdos televisivos e sofrem do mesmo caráter temporal. A taxa média de inserção para este conteúdo é de 47 vídeos diários.

É interessante notar que a taxa de vídeos que é efetivamente deletados pelos *Produtores de Conteúdo* é desprezível. Em geral o vídeo não é removido e permanece ativo por todo o tempo. Dessa forma para analisar o *tempo de vida médio* de um vídeo propomos uma métrica que chamamos de **dias ativos**. Os dias ativos é o número de dias distintos que a mídia efetivamente recebeu requisições. Dessa forma se uma mídia só recebeu requisições na Segunda-Feira e Sábado, naquela semana ela possui somente 2 dias ativos. Esta métrica não contabiliza os dias entre períodos de atividade.

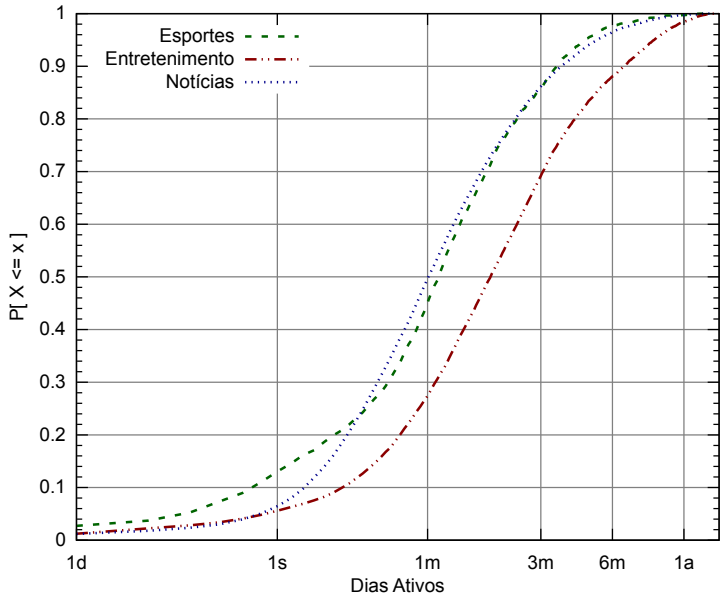


Figura 5.52. Distribuição cumulativa do dias de dias ativos para os conteúdos em análise

Com base nessa métrica o gráfico da Figura 5.52 mostra a distribuição dos dias ativos segmentada por diferentes conteúdos. Em tal gráfico podemos destacar o caráter mais duradouro das mídias de entretenimento, notamos que 30% dos vídeos desse conteúdo ficam ativos por períodos superiores a 3 meses enquanto que para jornalismo

e esportes esta proporção é pouco maior que 10%.

Com relação as mídias com poucos dias ativos vemos que o comportamento entre entretenimento e notícias é praticamente o mesmo para as mídias que só ficam ativas por uma semana, algo em torno de 5%. Já para os esportes a proporção aumenta consideravelmente, certa de 15% das mídias permanecem ativas por períodos inferiores a 1 semana.

Em geral o comportamento entre as mídias esportivas e jornalísticas são semelhantes diferenciando somente em períodos inferiores a 7 dias. Cerca de 50% das mídias permanecem ativas após 30 dias enquanto que para os vídeos de entretenimento essa proporção não chega a 30%.

Vimos nessa seção a distribuição das idades dos objetos segmentados por conteúdos. Mais uma vez constatamos que os conteúdos influenciam na forma como as mídias são consumidas e produzidas.

5.5.5 C_{13}: Relação Idade do Objeto versus Popularidade do Conteúdo

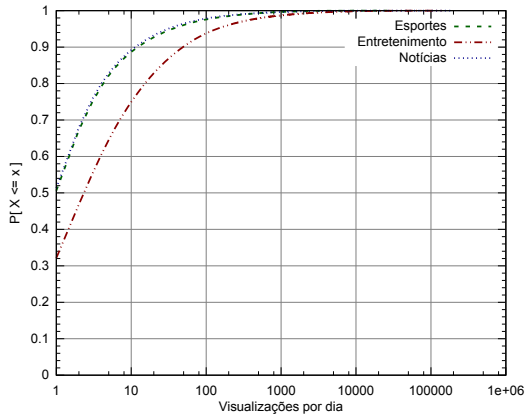
Nesta seção iremos correlacionar as análises das duas últimas seções. Cruzaremos as informações do tempo de vida das mídias (dias ativos) com as informações de popularidade com o objetivo de avaliar o comportamento das visualizações durante o tempo.

Inicialmente iremos avaliar a distribuição do número de visualizações por dia segmentada por conteúdo. No gráfico da Figura 5.53(a) temos a distribuição cumulativa das visualizações por dia. As informações desse gráfico podem ser interpretadas como a probabilidade de que uma mídia qualquer do gênero específico receba x visualizações em um dia. É interessante notar que a *CDF* das visualizações por dia para os gêneros de esporte e notícias possuem a mesma distribuição. A probabilidade que uma mídia receba somente uma visualização por dia é muito alta, 50%, se aumentarmos o número de visualizações para 10 a probabilidade é de 90%.

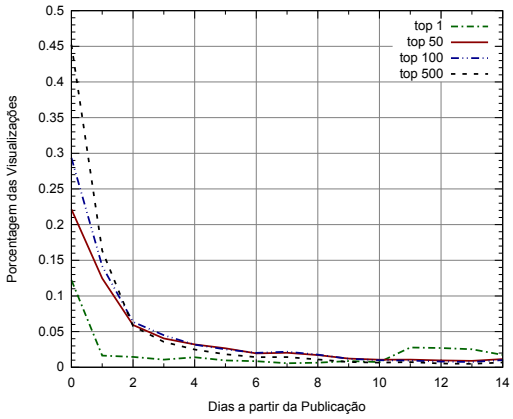
Para o gênero de entretenimento as mídias recebem em média mais visualizações por dia. A probabilidade de receber apenas uma visualização por dia cai para 30% e 10 visualizações por dia a probabilidade é de aproximadamente 75%.

Este resultado é uma outra maneira de avaliar a concentração de visualizações por mídia, apresentado na Figura 5.50 para todo o período. O que vemos aqui e a sua resolução alterada para dia. Apenas uma pequena parcela das mídias acumulam a maior parte das visualizações do todo.

A segunda análise que iremos realizar nesta seção é do comportamento das visualizações das mídias mais populares pelo tempo. Para fazer tal avaliação pegamos os



(a) *CDF* das Visualizações por Dia



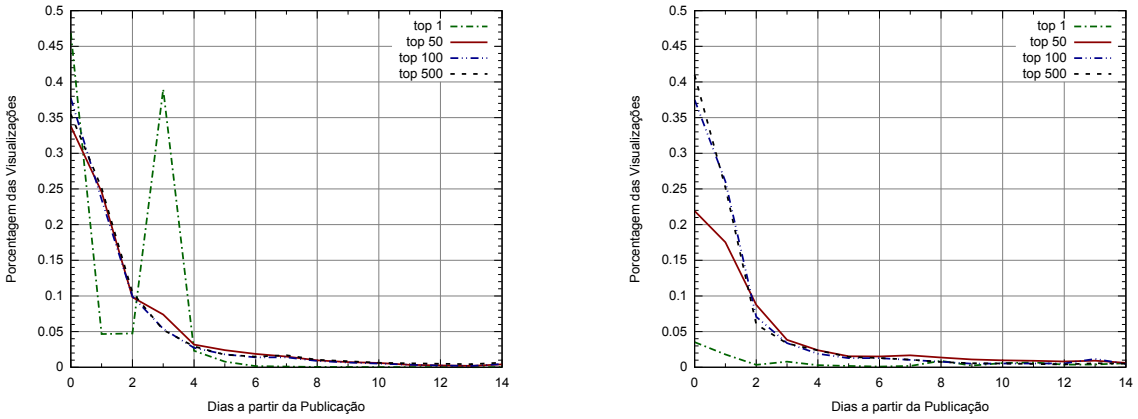
(b) Entretenimento

Figura 5.53. Distribuição do número de requisições diários e a média das frações das requisições diárias a partir do dia da publicação para o conteúdo de *Entretenimento*.

500 vídeos mais populares para cada gênero e fizemos uma análise da fração percentual de visualizações que estes recebem por dia a partir da sua data de publicação. Os gráficos que serão apresentados mostram as curvas para o vídeo mais popular, e depois agrega os vídeos mais populares em grupos de tamanho 50, 100 e 500 para estudar o comportamento médio.

A Figura 5.53(b) mostra o comportamento para o gênero de entretenimento. Vemos que o vídeo mais popular chega a receber 10% das suas visualizações no dia da sua publicação e depois a taxa de visualizações cai e só volta a subir depois do décimo dia, provavelmente quando este vídeo ganhou destaque por ser muito requisitada. Entretanto o comportamento do vídeo mais popular pode ser somente uma caso particular, estamos interessado no comportamento a medida que as mídias se tornam menos populares. A medida que o grupo de vídeos torna-se menos popular, ou seja, o grupo aumenta de tamanho, por exemplo, do mais popular para os 50 mais populares, as visualizações se acumulam perto da data de publicação. Com isso podemos concluir que para vídeos menos populares as visualizações se concentram perto da data de publicação e em contrapartida para os vídeos mais populares há uma melhor distribuição das suas visualizações durante o tempo. Para o caso do gênero de entretenimento vemos que há uma queda de visualizações nos dois primeiros dias após a publicação para os vídeo menos populares e um tendência das visualizações de se estabilizarem por volta do décimo dia.

Para o gênero esportes (Figura 5.54(a)) a mídia mais popular possui um comportamento muito atípico e será desconsiderada. Entretanto não há uma distinção



(a) Esportes (b) Notícias

Figura 5.54. Média das frações das requisições diárias a partir do dia da publicação para os conteúdos de *Esportes* e *Notícias*.

das mídias mais populares para as menas populares. Em geral todas as mídias do gênero de esporte agregam suas visualizações nos primeiros dias a partir da publicação sendo o maior decaimento nos primeiros 4 dias seguido de uma estabilização destas visualizações depois do décimo dia.

Para o caso dos vídeos de notícias a distribuição para os vídeos mais populares parece ser maior que as demais dado a progressão existente entre os grupos de vídeos. Entretanto o comportamento para os vídeos menos populares é similar aos anteriores. A medida que o grupo inclui mídias menos populares as visualizações se agregam nos primeiros dias de vida da mídia.

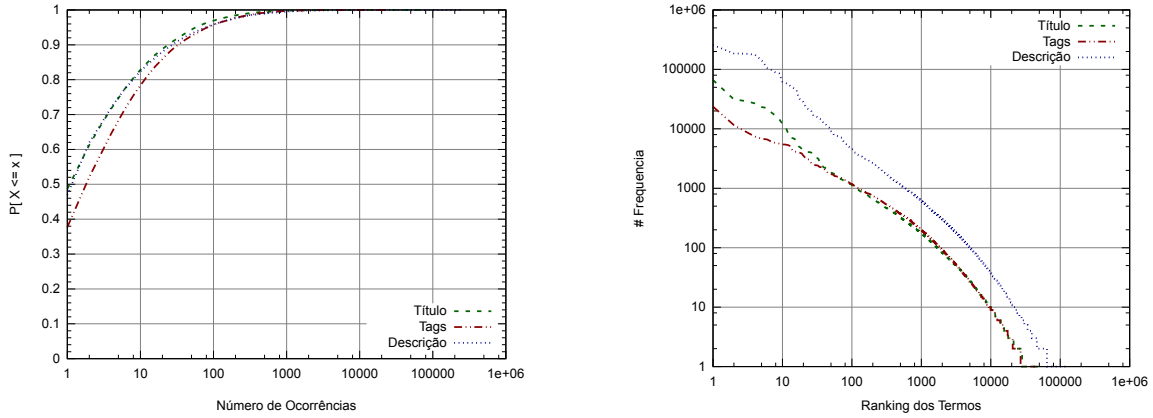
Vimos nessa seção a avaliação da popularidade dos vídeos pelo tempo. De forma semelhante visto na análise da Camada O as mídias menos populares agregam suas visualizações nos primeiros dias de publicação. Entretanto, vimos aqui que a distribuição das visualizações é impactada pelo gênero ao qual os vídeos estão inseridos.

5.5.6 C_{14}: Popularidade de metadados

Nesta seção iremos realizar uma avaliação da popularidade e distribuição dos principais metadados textuais associado aos vídeos da plataforma. Avaliaremos a popularidade dos termos dos títulos, descrições e *tags* através da distribuição cumulativa e um estudo da aplicabilidade da lei de Zipf. E em seguida mostraremos uma visualização dos termos mais frequentes para as *tags* e os nomes das categorias.

Para compreender a distribuição dos termos através dos metadados o gráfico da Figura 5.55(a) mostra a distribuição cumulativa do número de ocorrência dos termos

segmentados pelo título, *tags* e descrição dos vídeos. Vemos que a distribuição de termos possuem o mesmo comportamento, sendo que a distribuição para título e descrição são muito semelhantes. Para o caso de *tags* os termos são mais frequentes como era de se esperar. Geralmente os termos utilizados neste metadado são mais selecionados e descritivos, não há a presença de termos muito frequentes como preposições e artigos.



(a) *CDF* de ocorrência dos Termos (b) Análise de Zipf

Figura 5.55. Avaliação da frequência e popularidade dos termos nos metadados dos vídeos.

A Lei de Zipf foi originalmente verificada num corpus textual [Zipf, 1949] mostrando a relação inversa entre a ordem de popularidade dos termos com sua frequência. Para avaliar se o mesmo ocorre para um corpus textual controlado, como para metadados textuais dos vídeos dessa plataforma, o gráfico da Figura 5.55(b) traz a relação entre a frequência dos termos e a ordem desses termos na tabela de frequência. Podemos ver que para os três metadados textuais existe uma aplicabilidade da lei de Zipf mostrada através do comportamento das curvas. Para os três casos a distribuição da frequência dos termos segue, aproximadamente, uma reta numa escala logarítmica o que é uma forte evidência da aplicabilidade de Zipf.

Note que a distribuição entre *tags* e título são muito semelhantes. O que diferencia as duas distribuições é que para os termos de mais baixo ranking a frequência destes é muito mais elevada. Tal fato se deve a termos muito frequentes nos títulos como artigos, preposições que muito raramente aparecem nas *tags*.

Além de conhecer a distribuição dos termos por toda a base, é interessante saber precisamente quais termos são populares e em qual magnitude. Uma representação interessante para esse propósito é uma nuvem de *tags*, conforme mostrada na Figura 5.56, que é povoada pelas *tags* das mídias. Tais termos podem ocorrer até milhares de vezes

Figura 5.56. Representação da popularidade da tags através da *Nuvem de tags*

nas mídias conforme observado na Figura 5.49(a).

As *tag* são palavras-chave ou termos associado aos vídeos que tentam descrever estes e permitir uma classificação dos seus conteúdos baseada em palavras-chave. Note que ao observar a nuvem de *tags* da Figura 5.56 somos capazes de associar alguns tipos de conteúdos que podem estar sendo veiculados através da plataforma em análise.

De forma semelhante ao que foi feito para as *tags* a Figura 5.57 apresenta uma visualização dos nomes das categorias. É importante ressaltar que tais categorias não necessariamente estão expostas para os expectadores e podem representar classificações internas para cada *Produtor de Conteúdo*. Prova disso é a alta frequência do nome *teste* e *vídeos*. Ainda assim podemos identificar alguns nomes de categorias que são comuns entre diferentes *Produtores de Conteúdo* como por exemplo *notícias*, *futebol*, *saúde*, *jornalismo*, *esporte* dentre várias outras.

De uma forma geral, através da observação das Figuras 5.56 e 5.57 podemos perceber que é possível inferir conteúdos comuns aos vídeos que trafegam na plataforma através das *tags* e dos nomes categorias utilizadas pelos *Produtor de Conteúdo* para organizar seus vídeos.

5.5.7 Conclusões da Camada C

Nesta seção avaliamos os dados da plataforma segmentada por tipo de conteúdo. Utilizando como base os metadados dos vídeos fizemos uma partição destes agrupando-os em grupos coesos de acordo com o conteúdo associado.

Figura 5.57. Visualização da popularidade dos nomes das categorias

Mostramos que existe uma relação íntima entre a duração dos vídeos e o conteúdo ao qual estes vídeos pertencem. Por exemplo, vídeos esportivos e jornalísticos são menores que vídeos de entretenimento. Também pudemos perceber que, apesar de existirem vídeos com duração de horas, a grande maioria dos vídeos online ainda possuem durações curtas. Mas a disponibilização de vídeos de durações tão longas já é uma evidência que aos poucos a Internet está se consolidando como um meio de difusão de vídeos.

O perfil de popularidade de vídeos é distribuído de forma diferente entre os vários conteúdos. Constatamos que os vídeos de entretenimento são os mais populares, possuindo uma concentração enorme de visualizações ($\sim$ 9K visualizações/vídeo), mas os vídeos de notícias e esportes também possuem agregam visualizações consideráveis. Vimos que de forma semelhante a outros trabalhos [Gill et al., 2007], a popularidade dos vídeos possuem características típicas da distribuição de Zipf. Adicionalmente vimos que a concentração das visualizações dos vídeos também é afetada pelo tipo de conteúdo trafegado.

Analisando o tempo de vida dos objetos percebemos que a taxa de inserção para vídeos jornalísticos é muito maior que as demais. Notamos que em média 50% dos vídeos de jornalismo e esportes permanecem ativos por 30 dias, enquanto que a mesma proporção de vídeos de entretenimento permanece ativo por pelo menos 3 meses. Ainda existe uma pequena parcela dos vídeos que permanecem ativas por mais de 1 ano.

Ao cruzar as informações de popularidade com o tempo de vida dos vídeos analisamos a distribuição das visualizações por dia e a distribuição das visualizações a partir do dia de publicação. Os resultados mostraram que a probabilidade de uma mídia qualquer receber mais de 100 visualizações por dia é muito baixa mas que tal probabilidade varia de acordo com o conteúdo. Vimos que os vídeos menos populares tendem a concentrar suas visualizações nos primeiros dias a partir da publicação e que em geral a maior parte das visualizações para tais vídeos ocorrem nos primeiros 4 dias.

Avaliamos também a distribuição dos termos através dos diversos metadados textuais e verificamos que a distribuição entre os títulos e as *tags* são muito semelhantes. Notamos que as *tags* e as categorias nos remetem a conteúdos mais amplos, o que pode viabilizar a classificação em grandes grupos coesos visando segmentar os conteúdos distribuídos pela plataforma em grupos globalmente únicos.

5.6 Extração de Conhecimento (K)

Ilustraremos nesta seção a aplicação do nível **K** da metodologia nos dados da *Liquid*$^{\text{TM}}$, a plataforma de distribuição de vídeos online da Samba Tech. Para exemplificar a aplicabilidade das técnicas de **KDD** neste nível de análise, escolhemos avaliar alguns aspectos do universo de dados da plataforma utilizando *Regras de Associação* [Tan et al., 2005].

Inicialmente, neste capítulo mostraremos os objetivos das análises, em seguida explicaremos brevemente a técnica adotada e suas principais métricas. Por último, mostraremos a aplicação dos algoritmos de regras de associação, as definições e pré-processamentos utilizados e, finalmente, as conclusões a respeito de cada uma das análises.

5.6.1 Análises Propostas

O mercado de vídeos online no Brasil cresceu bastante nos últimos tempos. As empresas estão investindo cada vez mais nos conteúdos em vídeo pensando num mercado em potencial que ainda é incipiente. Atualmente, os investimentos em vídeos online ainda não são lucrativos se comparados à audiência que eles possuem. Pensando nisso, resolvemos ilustrar este nível da metodologia visando o aperfeiçoamento da distribuição de publicidade através deste meio. As análises que propomos aqui visam a melhor compreensão do conteúdo para prover mecanismos mais eficientes que possam aumentar a taxa de conversão de publicidade nesse meio e, consequentemente, aumentar o lucro com a distribuição dos vídeos online.

Com esse fim, decidimos utilizar regras de associação para tentar extrair relacionamentos entre os metadados dos vídeos, suas propriedades e suas requisições. Para isso utilizaremos como insumos alguns elementos da três camadas do primeiro nível da metodologia (R, O e C) para a realização das seguintes análises:

Popularidade do Vídeo : Iremos utilizar a duração do vídeo, a quantidade de requisições recebidas (discretizada em popularidade), o gênero ao qual o vídeo pertence para avaliar os relacionamentos existentes entre estas unidades de análise e extrair padrões de popularidade.

Taxa de Retenção dos Vídeos : Com base no percentual assistido do vídeo definiremos um critério de avaliação da taxa de retenção um vídeo. Aplicaremos tal critério para avaliar estas informações juntamente com outras unidades de análise. O objetivo é investigar se existe um padrão comum entre os vídeos com diferentes taxa de aceitação.

Predição de Gênero : A ideia dessa análise é avaliar as correlações existentes entre os metadados dos vídeos e os gêneros inferidos nas análises da camada **C**. O objetivo é avaliar a possibilidade de inferir o gênero automaticamente para prover uma segmentação do conteúdo da plataforma.

Em cada análise, as idéias, conceitos e objetivos serão detalhadamente descritos. A seguir introduziremos brevemente a técnica de mineração de dados conhecida como *Regras de Associação*.

5.6.2 Regras de Associações

Em mineração de dados a *Análise de Associações* é um campo de pesquisa muito estudado que visa descobrir relações interessantes e implícitas entre variáveis em grandes bases de dados [Tan et al., 2005]. A relação implícita pode ser representada na forma de *regras de associação* ou conjuntos de itens frequentes. A aplicação clássica dessa área é a análise de transações de compras em supermercados, sendo a regra de associação mais famosa a relação que mostra que quem compra fraldas geralmente também comprar cerveja. A representação de uma regra de associação neste texto será feita da seguinte forma :

$$\{\text{Fraldas}\} \longrightarrow \{\text{Cerveja}\}$$

O conjunto de itens a esquerda da seta é geralmente conhecido como **antecedente** e o conjunto de itens que co-ocorrem na relação (a direita da seta) é conhecido como **consequente**.

O trabalho que introduziu as regras de associação, e um dos mais citados da área de mineração de dados, é o trabalho de [Agrawal et al., 1993]. Neste artigo, os autores descrevem o problema de minerar regras de associação da seguinte forma:

> Seja $C = \{I_1, I_2, \ldots, I_m\}$ um conjunto de itens. Seja T um conjunto de transações, onde cada transação t é um conjunto de itens tal que $t \subseteq C$. Sejam A e B conjuntos de itens, uma transação t contém A se e somente se $A \subseteq t$. Uma **regra de associação** é uma implicação da forma $A \rightarrow B$, onde $A, B \subseteq C$, e $A \cap B = \emptyset$.

Em outras palavras, dado um conjunto de itens, por exemplo, características de acesso de vídeos online como hora do dia, categoria, popularidade, tamanho do vídeo, etc., e várias transações compostas exclusivamente com itens desse conjunto, no nosso caso várias requisições a diferentes vídeos. Uma regra de associação seria uma implicação do tipo *vídeos pequenos implicam em vídeos populares*.

Uma importante propriedade do um conjunto de itens, geralmente referido na área de mineração de dados como *Itemset*, é o **suporte**, ou contagem de *suporte*, que se refere ao número de transações que contém um conjunto de itens. Matematicamente, o suporte, $\sigma(X)$, para um *Itemset* é dado pela seguinte equação:

$$\sigma(X) = \big|\{t_i | X \subseteq t_i, t_i \in T\}\big|, \tag{5.2}$$

onde o operador $\big|X\big|$ representa o número de elementos em um conjunto.

Dada uma regra de associação do tipo $\{X\} \rightarrow \{Y\}$, a avaliação da força de tal regra é geralmente dada por duas métricas, o suporte e a confiança. Tais métricas são descritas pelas seguintes equações:

$$\text{Suporte} : s(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{N}; \tag{5.3}$$

$$\text{Confiança} : c(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}; \tag{5.4}$$

onde N é o número total de itens na base.

O suporte nos diz qual é a frequência da regra na base e é importante porque regras com suporte muito baixo podem ter ocorrido simplesmente por acaso ou representarem regras que não agregam valor muito alto para o negócio. Complementarmente, a confiança mede a intensidade desta inferência, ou seja o quão provável é Y ocorrer dada a ocorrência de X.

A geração das regras de associação a partir de um conjunto de transações consiste em um processo que pode ser dividido em duas partes.

Passo 1: Encontrar todos os conjuntos de itens (*Itemsets*) frequentes. Por definição, *Itemsets* frequentes são aqueles cuja frequência no conjunto de transações é igual ou superior a um suporte mínimo predeterminado.

Passo 2: Gerar, a partir dos *Itemset* frequentes todas as regras que satisfazem uma confiança mínima estipulada.

Existem diversos algoritmos para geração de regras de associação. Esses algoritmos diferem principalmente no primeiro passo, que visa determinar todos os *Itemset* frequentes. Isto ocorre principalmente porque, geralmente, o primeiro passo exige que a base de dados seja lida várias vezes, o que determina o desempenho global do algoritmo. O primeiro trabalho sobre mineração de regras de associação foi apresentado por [Agrawal et al., 1993], e, desde então, vários algoritmos para geração de regras de associação foram propostos, sendo o Apriori, descrito por [Agrawal & Srikant, 1994], o mais conhecido. Para o caso das análises realizadas neste trabalho, o software utilizado para a geração das regras de associação foi o Weka [Hall et al., 2009].

5.6.2.1 Outras Medidas de Interesse

Apesar de primariamente se avaliar as regras de associação pelo suporte e a confiança, as vezes essas métricas não são suficientes para afirmar o quão relevante é a regra em questão, já que a confiança pode produzir análises enviesadas. Existem outras métricas que tentam superar tais limitações. Iremos agora discutir as principais medidas de interesse utilizadas na avaliação das regras de associação no presente trabalho.

O ***lift***, também conhecido como *Interesse*, foi proposto por [Brin et al., 1997] e é uma medida que tenta levar em consideração o suporte do *Itemset* consequente. O *lift* de uma regra de associação é a confiança dessa regra dividida pelo suporte do *Itemset* consequente. Isto indica quão mais frequente é o *Itemset* consequente da regra quando o *Itemset* antecedente está presente.

$$\text{Lift} : lift(X \longrightarrow Y) = \frac{c(X \longrightarrow Y)}{s(Y)} \tag{5.5}$$

O *lift* maior que 1 é interessante pois este indica que o consequente é mais frequente quando o antecedente ocorre. Já para *lift* menor que 1, o consequente é mais frequente nas transações em que o antecedente não ocorre. Para o *lift* igual a 1, o consequente ocorre com a mesma frequência independente do antecedente ocorrer ou

não. A partir disso, pode-se notar que as regras que possuem *lift* maior que 1 são mais interessantes que as demais, sendo que, quanto maior o *lift* maior deverá ser a relação entre os dois lados da regra.

O ***leverage***, originalmente apresentado por [Piatetsky-Shapiro, 1991], quando utilizado numa regra de associação, representa o número de transações adicionais cobertas pela regra, além do esperado, caso o consequente e o antecedente fossem independentes um do outro. O *leverage* é definido pela diferença entre o suporte real e o esperado do antecedente e consequente em questão.

$$\text{Leverage} : lev(X \longrightarrow Y) = s(X \cup Y) - s(X) \times s(Y) \tag{5.6}$$

Pode-se verificar que um *leverage* maior que 0 indica que os dois lados da regra ocorreriam juntos em um número de transações maior que o esperado caso os itens encontrados na regras fossem completamente independentes. Ja para o *leverage* menor que 0, temos o inverso, os dois lados da regra ocorrem juntos menos que o esperado. Para o *leverage* igual a 0, os dois lados da regra ocorrem juntos exatamente o esperado indicando que os dois lados provavelmente são independentes. Deste modo, quanto maior o *leverage* mais interessante será a regra. Um problema do *leverage* é que ele não leva em consideração as proporções de uma regra para a outra, para evitar isso, uma forma interessante de se utilizar o *leverage*, quando este é positivo, é dividi-lo pelo suporte da regra. Desta forma pode-se obter o percentual do suporte que não ocorre por acaso.

A ***convicção***, também apresentada por [Brin et al., 1997], parte da ideia de que, logicamente, $X \longrightarrow Y$ pode ser reescrito como $\neg(X \wedge \neg Y)$, então a convicção verifica o quanto $(X \wedge \neg Y)$ está distante da independência.

$$\text{Convicção} : \ conv(X \longrightarrow Y) = \frac{1 - s(Y)}{1 - c(X \longrightarrow Y)} \tag{5.7}$$

Ao contrário da confiança, a convicção tem valor 1 quando os *Itemsets* da regra não possuem nenhuma relação, sendo que quanto maior a convicção maior a relação entre X e Y, quando o valor é menor que 1 a relação entre os itens é negativa, ou seja, quando X ocorre, Y tende a não ocorrer.

Muitas regras podem ter correlações positivas fracas, enquanto a regra contendo a negação do consequente pode ser muito mais expressiva. Nesses casos, a convicção é o indício a ser considerado. A convicção mede o quanto a regra é mais forte em relação a sua regra complementar. Mais especificamente, ela quantifica o impacto da regra quando comparada com a sua regra complementar, ou seja, o conjunto de regras onde

o consequente é diferente.

Outro objetivo dessa métrica é cobrir uma falha da confiança, por exemplo, se 80% de clientes de um supermercado compram leite, e 2% compram salmão, existe uma grande chance de encontrarmos a regra $\{salmão\} \rightarrow \{leite\}$ com uma confiança bastante alta, como 80%, porém isto acontece porque o leite é um item muito frequente, e não porque os itens tenham alguma relação, isto pode ser verificado utilizando a convicção, já que $conv(\text{salmão} \longrightarrow \text{leite}) \cong 1$, o que significa que os itens não possuem nenhuma relação.

5.6.3 Pré-Processamento

Conforme dito anteriormente, uma das etapas do processo de **KDD** consiste no pré-processamento, limpeza e adequação dos dados para as análises. Neste trabalho, pré-processamentos foram necessários para que os metadados se adequassem às diferentes análises. Nesta seção, iremos abordar de forma sucinta as transformações e tratamentos aplicados aos dados para as análises deste nível.

5.6.3.1 Tratamento dos Metadados Textuais

Para a utilização dos metadados textuais é comum na área de mineração de dados e recuperação da informação a realização de algumas técnicas de pré-processamento como remoção de *stopwords* e acentos, além da aplicação da técnica de *stemming* (radicalização de palavras). Para o caso dos metadados textuais da nossa base, tais técnicas também foram aplicadas.

5.6.3.2 Duração dos Vídeos

Para possibilitar uma análise mais simples da duração dos vídeos segmentamos este atributo numérico contínuo em 5 categorias. Vídeos muito curtos, curtos, médios, longos e muito longos. Os intervalos que definem tais categorias foram baseados na análise da Figura 5.47 e na premissa de que são vídeos para Internet.

Seja a duração de um vídeo x dada pela função $dur(x)$ a categoria da duração é definida pela função $dur\_cat(x)$ definida como:

$$dur\_cat(x) = \begin{cases} \text{muito curto} & = \text{se} & & dur(x) & < 1 \text{ min.} \\ \text{curto} & = \text{se } 01 \text{ min.} & \leq & dur(x) & < 3 \text{ min.} \\ \text{médio} & = \text{se } 03 \text{ min.} & \leq & dur(x) & < 5 \text{ min.} \\ \text{longo} & = \text{se } 05 \text{ min.} & \leq & dur(x) & < 10 \text{ min.} \\ \text{muito longo} & = \text{se } 10 \text{ min.} & \leq & dur(x) & \end{cases} \tag{5.8}$$

Apesar de um vídeo de 10 minutos não ser longo se comparado a televisão, historicamente a Internet tem sido marcada por vídeos curtos e vídeos com durações maiores que 10 minutos são considerados vídeos longos.

No trabalho de [Cherkasova & Gupta, 2002] foi realizado segmentação semelhante a do nosso trabalho. Os autores dividem os vídeo em curtos, médios e longos. Contudo eles criam subconjuntos dessas classes. Três grupos de vídeos curtos 1) menores que 2 minutos, 2) entre 2 e 5 minutos e 3) entre 5 e 10 minutos. Um grupo de vídeos médios, entre 10 e 30 minutos. E dois grupos de vídeos longos, 1) entre 30 e 60 minutos e 2) maiores que 60 minutos. Tanto no nosso trabalho quanto na literatura a escolha dos limites é uma escolha arbitrária porém baseada no senso comum ou aplicabilidade do sistema.

5.6.3.3 Popularidade das Mídias

Não existe uma definição quantitativa de *popularidade* de um vídeo. Tal conceito é muito subjetivo e irá variar de acordo com o conjunto de mídias em análise. Apesar de existir consenso para os casos extremos, como vídeos que agregam milhares de visualizações, a diferença entre um vídeo popular ou não, pode ser sutil. O problema que enfrentamos ao analisar os nossos conteúdos pode ser descrito da seguinte forma:

> *Dado um vídeo escolhido aleatoriamente de um coleção de vídeos, como saber a popularidade desse vídeo a partir de seu número absoluto de visualizações?*

A resposta para tal hipótese pode ser bastante útil para o contexto de publicidade em vídeos onde se analisa a taxa de conversão de uma campanha para uma dada categoria ou vídeo. Saber da popularidade do vídeo pode levar a *insights* a respeito dos usuários que consomem tal vídeo.

Utilizando somente as informações exclusivas do vídeo, não é possível saber sua popularidade. Com base nisso utilizamos o estudo da análise de concentração (Figura 5.50) e determinamos que 25% das mídias com mais acessos entrariam no grupo das mídias *populares* e 25% das mídias com menos visualizações seriam consideradas *não populares*. O restante das mídias estariam em uma faixa de incerteza onde decidimos não fazer afirmações a respeito da popularidade, classificando esta como *popularidade desconhecida*. Essa abordagem permite trabalhar de forma discreta e simples numa área que é muito complexa. Claro que vídeos populares ou não populares com características interessantes ficam de fora, mas tal simplificação permite uma análise inicial sem muito margem para erros.

Entretanto, não podemos considerar todas as mídias, de diferentes *Produtores de Conteúdo* e com público alvo completamente diferentes, pertencentes a um mesmo grupo de classificação. Tal suposição faria com que *Produtores de Conteúdo* menores fossem automaticamente não populares. Para evitar essa limitação o conjunto de mídias a serem particionadas nos intervalos de populares, desconhecidas e não populares devem pertencer a um conjunto coeso e correlacionado, como por exemplo mídias exclusivas de um produtor de conteúdo ou mesmo mídias de uma mesma categoria dentro de um mesmo *Produtor de Conteúdo*.

Com base nesses conceitos e nas informações relativas aos *Produtores de Conteúdo* e seus respectivos projetos, fizemos uma tabela delimitando o intervalo de visualizações que classificaria as mídias nas 3 classes definidas. Tal tabela continha informação do intervalo definido de visualizações mínimas para classificar uma mídia dentre uma das classes de popularidade segmentada por projeto e *Produtor de Conteúdo*.

De posse dessa tabela, todas as mídias foram classificadas, resultando numa partição das mídias de acordo com a Tabela 5.20. Note que a distribuição diferente do fator 25%-50%-25% entre as classe de popularidade existe porque a divisão foi feita relativa aos projetos de cada *Produtor de Conteúdo* e não globalmente.

Classe	**# Mídias**	**%**
Não Popular	76.032	40.32%
Popularidade Desconhecida	56.482	29.95%
Popular	56.060	29.73%

Tabela 5.20. Distribuição dos vídeos de acordo com sua popularidade

5.6.3.4 Taxa de Retenção de uma Mídia

Nesta etapa de pré-processamento o problema é discretizar os vídeos em dois conjunto que chamamos de *grupo de sucesso* e *grupo de fracasso*. A premissa utilizada é que quando os *Produtores de Conteúdo* selecionam um conteúdo para colocar em seus portais, o objetivo é que o vídeo seja assistido por completo pela maioria de sua audiência. Partindo desse pré-suposto, definimos que se pelo menos 75% das pessoas assistir todo o vídeo, este pode ser considerado um **sucesso**. Em contra partida, se menos de 25% das pessoas não chegarem ao final do vídeo podemos dizer que este vídeo foi um **não sucesso**. Esta métrica complementar nos diz que existe uma alta taxa de desistência entre o início e o fim do vídeo. Note que o conceito de sucesso é somente uma metá-

fora, não está relacionado a popularidade, é simplesmente uma métrica de aceitação do vídeo.

De forma complementar, se um usuário desiste de assistir o vídeo logo no início, o vídeo não atingiu seus objetivos. Definimos então que se somente 25% da audiência de um vídeo chegou ao fim do primeiro quarto do vídeo este vídeo pode ser classificado como um **fracasso**. Adicionalmente, se mais de 75% da audiência do vídeo assistiu ao primeiro quartil do vídeo por completo, ele é um **não fracasso**. Novamente, este conceito não se relaciona a popularidade, é simplesmente a taxa de desistência durante o vídeo.

Note que para as duas métricas existe uma área de incerteza sobre a qual não é possível fazer afirmações sobre a aceitação. Em tais casos, a mídia será ignorada.

Com base nesses conceitos e nas informações de retenção dos vídeos, com a porcentagem de visualizações divididas por quartis, classificamos os vídeos nessas duas classes.

5.6.4 Popularidade do Vídeo

É muito importante compreender as características que fazem um vídeo ser popular ou impopular. Com posse de tais propriedades, é possível investir em melhorias na produção de conteúdos mais interessantes além de identificar os melhores vídeos para publicidade. Seria possível inclusive propor estratégias de diferenciação de preço de publicidade de acordo com a popularidade esperada do vídeo. No entanto, a geração de conteúdo é uma tarefa não-trivial, estratégias de criação de conteúdo podem ser complexas e até mesmo contraditórias.

Identificar exatamente se um vídeo específico será ou não popular é uma tarefa complexa. Uma tarefa menos desafiadora é descobrir quais tipos de conteúdos possuem vídeos mais populares e quais as características desses vídeos. Com esse objetivo, analisamos propriedades como **duração** e o **gênero** de um vídeo e tentamos relacioná-las com a sua **popularidade**.

Como uma primeira análise, avaliamos o impacto do tamanho do vídeo sob sua popularidade para diferentes gêneros. A pergunta básica para essa análise é: *Qual seria o tamanho de vídeo que tem mais chances de se tornar popular?* Em geral, acredita-se que vídeos curtos são mais efetivos nesse sentido. Entretanto, não foi esse o resultado que encontramos. Vejam as seguintes regras:

$$\{\text{Curto}\} \longrightarrow \{\text{Popular}\}\ (\ sup(0,12),\ conf(0.24),\ lift(0.9),\ lev(-0.01),\ conv(0.96)\ )$$
$$\{\text{Longo}\} \longrightarrow \{\text{Popular}\}\ (\ sup(0,03),\ conf(0.38),\ lift(1.4),\ lev(0.01),\ conv(1.18)\ )$$

Com esse resultado, temos que, apesar de existir uma relação de popularidade com vídeos curtos, em geral, os vídeos tendem a ser menos populares quando estes são curtos (*lift* e convicção menores que 1). Entretanto para a segunda regra, temos a relação inversa com melhores indicadores. O que mostra que a relação entre vídeos longos e popularidade é mais forte que a inversa, contrariando o senso de que vídeos curtos são mais populares.

Incluindo os conteúdos nas análises, encontramos dentre as regras geradas a relação de que notícias são menos populares que vídeos de entretenimento:

$$\{\text{Entretenimento}\} \longrightarrow \{\text{Popular}\}$$
$$(\ sup(0,07),\ conf(0.48),\ lift(1.75),\ lev(0.03),\ conv(1.39)\ )$$
$$\{\text{Notícias e Politica}\} \longrightarrow \{\text{Popular}\}$$
$$(\ sup(0,10),\ conf(0.18),\ lift(0.66),\ lev(-0.05),\ conv(0.89)\ )$$

Este comportamento é esperado, principalmente pelo carácter temporal e regionalizado das notícias. Além disso, não é estranho um usuário visitar mais de uma vez um vídeo de entretenimento, mas tal comportamento não é esperado para um vídeo jornalístico. Além disso, a própria análise feita na Figura 5.49 já mostra esses indícios.

Entretanto, *pode o tamanho do vídeo impactar na popularidade para esses dois tipos de conteúdo?* Analisando as regras para o caso de entretenimento temos que vídeos mais curtos são mais populares que os médios e longos. Porém, quando os vídeos são muito longos, tais vídeos recuperam a sua popularidade. Vejam as regras que evidenciam tal aspecto:

$$\{\text{Muito Curto e Entretenimento}\} \longrightarrow \{\text{Popular}\}$$
$$(\ sup(0,01),\ conf(0.46),\ lift(1.71),\ lev(0),\ conv(1.36)\ )$$
$$\{\text{Curto e Entretenimento}\} \longrightarrow \{\text{Popular}\}$$
$$(\ sup(0,02),\ conf(0.53),\ lift(1.95),\ lev(0.01),\ conv(1.55)\ )$$
$$\{\text{Médio e Entretenimento}\} \longrightarrow \{\text{Popular}\}$$
$$(\ sup(0,01),\ conf(0.44),\ lift(1.63),\ lev(0),\ conv(1.31)\ )$$
$$\{\text{Longo e Entretenimento}\} \longrightarrow \{\text{Popular}\}$$
$$(\ sup(0,01),\ conf(0.43),\ lift(1.57),\ lev(0.01),\ conv(1.27)\ )$$
$$\{\text{Muito Longo e Entretenimento}\} \longrightarrow \{\text{Popular}\}$$
$$(\ sup(0,01),\ conf(0.49),\ lift(1.79),\ lev(0.01),\ conv(1.42)\ )$$

Vemos que para o caso de entretenimento, a relação entre o tamanho do vídeo e sua popularida é uma relaçao forte. Todos os indicadores mostram que a relação enfraqueçe à medida que o vídeo aumenta. Excetuando no caso de vídeos muito longos, provavelmente devido a conteúdos mais elaborados como séries de Tv e capítulos de novelas.

Já para os conteúdos jornalísticos, esse comportamento não é observado. Ao contrário, a relação com a popularidade aumenta à medida que a duração do vídeo também aumenta.

$$\{\text{Curto e Notícias e Politica}\} \longrightarrow \{\text{Popular}\}$$
$$(\ sup(0,01),\ conf(0.17),\ lift(0.63),\ lev(-0.03),\ conv(0.88)\ )$$
$$\{\text{Médio e Notícias e Politica }\} \longrightarrow \{\text{Popular}\}$$
$$(\ sup(0,01),\ conf(0.23),\ lift(0.84),\ lev(0),\ conv(0.95)\ )$$
$$\{\text{Longo, Notícias e Politica}\} \longrightarrow \{\text{Popular}\}$$
$$(\ sup(0,008),\ conf(0.28),\ lift(1.04),\ lev(0),\ conv(1.01)\ )$$
$$\{\text{Muito Lonto, Notícias e Politica}\} \longrightarrow \{\text{Popular}\}$$
$$(\ sup(0,003),\ conf(0.3),\ lift(1.11),\ lev(0),\ conv(1.04)\ )$$

Com as informações vistas nessa seção, mostramos que a efetividade de uma estratégia para criação de vídeos pode depender do tipo de conteúdo a ser transmitido. Nesse caso, foi mostrado que vídeos mais curtos ou mais longos podem ser preferíveis dependendo do contexto. Uma estratégia de publicidade para vídeos de entretenimento nitidademente deve ser diferenciada de uma estratégia para a publicidade em vídeos de notícias.

5.6.5 Taxa de Retenção dos Vídeos

Atualmente os sites de vídeos online possuem vários segmentos. Os grandes portais de conteúdo como a TvIG, Terra e UOL possuem várias seções dedicadas aos mais diferentes tipos de conteúdo. Na disputa por audiência tais *Produtores de Conteúdo* têm procurado cada vez mais conteúdos que irão atrair mais público para seus portais. De uma forma geral, a escolha da publicação de um vídeo sempre tem o mesmo objetivo, atrair audiência e transmitir o conteúdo do vídeo.

Pensando nesses objetivos, pretendemos avaliar nesta análise as características de um vídeo que possui uma alta taxa de aceitação e também as características que podem levar um vídeo a ser muito rejeitado. Na tentativa de avaliar estas possíveis características definimos o conceito de *sucesso* e *fracasso* de uma mídia na subseção 5.6.3. Em síntese, um *sucesso* ocorre quando mais de 75% da audiência de um vídeo assistem tal vídeo por completo, se este valor é inferior a 25% temos o caso do *não-sucesso* que mostra que somente uma pequena parte da audiência termina de assistir o vídeo. Já o *fracasso* é quando menos de 25% da audiência de um vídeo desistem de assistir no primeiro quartil do vídeo. Adicionalmente, se esse percentual for maior que 75% temos o caso do *não-fracasso* que mostra que a audiência se mostra interessada no vídeo.

Primeiramente, analisaremos o perfil de um vídeo de *sucesso* ou *não-sucesso*. Saber que um vídeo tem uma taxa de aceitação alta também ajuda na escolha da forma de publicidade. Por exemplo quando usar ou não publicidade do tipo *pos-roll* (depois do vídeo) para esses casos.

A intuição é que os vídeos curtos terão uma taxa maior de sucesso, mas analisando estas duas propriedades isoladamente não encontramos uma correlação inversa forte. As regras a seguir evidenciam que mesmo para vídeos curtos, existe uma taxa alta de desistência.

$$\{\text{Muito Curto}\} \longrightarrow \{\text{Sucesso}\}\ (sup(0,08), conf(0.34), lift(2.04), lev(0.04), conv(1.26))$$
$$\{\text{Curto}\} \longrightarrow \{\text{Sucesso}\}\ (\ sup(0,07),\ conf(0.16),\ lift(0.95),\ lev(0),\ conv(0.99)\ )$$

De uma forma geral, o que notamos é que os casos de sucesso são pouco frequentes. Em sua grande maioria estão associados aos vídeos muito curtos e a sua frequência varia muito de acordo com o gênero.

$$\{\text{Muito Curto e Ciência e Tecnologia}\} \longrightarrow \{\text{Sucesso}\}$$
$$(\ sup(0,001),\ conf(0.49),\ lift(2.97),\ lev(0),\ conv(1.63)\ )$$
$$\{\text{Muito Curto e Notícias e Política}\} \longrightarrow \{\text{Sucesso}\}$$
$$(\ sup(0,04),\ conf(0.43),\ lift(2.61),\ lev(0.02),\ conv(1.47)\ )$$
$$\{\text{Muito Curto e Esportes}\} \longrightarrow \{\text{Sucesso}\}$$
$$(\ sup(0,008),\ conf(0.35),\ lift(2.1),\ lev(0),\ conv(1.28)\ )$$

Notamos que não existe uma correlação entre popularidade e o sucesso de um vídeo. Ao contrário, a relação que encontramos é que se um vídeo é um sucesso existe uma probabilidade maior que ele seja não popular.

$$\{\text{Sucesso}\} \longrightarrow \{\text{Não Popular}\}\ (sup(0,07), conf(0.43), lift(1.23), lev(0.01), conv(1.14))$$

Outra evidência que encontramos é de que apesar da popularidade de alguns vídeos, a taxa de retenção desses é baixa, o que pode indicar uma falha na geração desses conteúdos.

$$\{\text{Curto, Humor e Popular}\} \longrightarrow \{\text{Não Sucesso}\}$$
$$(\ sup(0,0001),\ conf(0.71),\ lift(2.73),\ lev(0),\ conv(2.1)\ )$$
$$\{\text{Música e Popular}\} \longrightarrow \{\text{Não Sucesso}\}$$
$$(\ sup(0,003),\ conf(0.5),\ lift(1.93),\ lev(0),\ conv(1.48)\ )$$
$$\{\text{Médio, Filmes e Animações e Popular}\} \longrightarrow \{\text{Não Sucesso}\}$$
$$(\ sup(0,0002),\ conf(0.46),\ lift(1.79),\ lev(0),\ conv(1.32)\ )$$

Acreditamos que a identificação desses casos pode ser benéfica para os produtores de conteúdo pois pode ser interpretada também como um *feedback* do conteúdo sendo transmitido. Por exemplo, talvez seja preciso investir mais em edição para aumentar a taxa de sucesso do vídeo.

Com relação ao conceito de *fracasso*, é natural que à medida que os vídeos cresça exista uma taxa maior de *fracasso*. Entretanto, a regras mais fortes que correlacionam *fracasso* e a duração dos vídeos nos mostram que é mais provável que o vídeo seja de curta duração caso este seja um *fracasso*.

$$\{\text{Fracasso}\} \longrightarrow \{\text{Curto}\}\ (\ sup(0,04),\ conf(0.35),\ lift(0.77),\ lev(-0.01),\ conv(0.84)\ )$$
$$\{\text{Fracasso}\} \longrightarrow \{\text{Muito Longo}\}\ (sup(0,02), conf(0.2), lift(2.28), lev(0.01), conv(1.14))$$

O tipo de conteúdo do vídeo influencia na taxa de retenção deste, podemos ver que o fato da segmentação dos vídeos muito longos aumenta a probabilidade de *fracasso* em algum gêneros como no caso de *Esporte* e *Pessoas e Blogs*.

$$\{\text{Muito Longo, Pessoas e Blogs e Popular}\} \longrightarrow \{\text{Fracasso}\}$$
$$(\ sup(0,0001),\ conf(0.64),\ lift(6.03),\ lev(0),\ conv(2.08)\ )$$
$$\{\text{Muito Longo e Esportes}\} \longrightarrow \{\text{Fracasso}\}$$
$$(\ sup(0,001),\ conf(0.38),\ lift(3.58),\ lev(0),\ conv(1.44)\ )$$
$$\{\text{Curto e Humor}\} \longrightarrow \{\text{Fracasso}\}$$
$$(\ sup(0,0001),\ conf(0.33),\ lift(3.13),\ lev(0),\ conv(1.29)\ )$$

Uma outra aplicação da métrica de *fracasso* é possibilitar a detecção de vídeos e conteúdos que estão sendo pouco efetivos. Qualquer regra que implique em *fracasso* e *não-popular* indica que a efetividade dos conteúdos é praticamente nula. Por exemplo, as regras abaixo mostram que vídeos muito longos do gênero música ou do gênero filmes e animação são muito pouco eficientes

$$\{\text{Muito Longo e } \textit{Filmes e Animação}\} \longrightarrow \{\text{Fracasso e Não-Popular}\}$$
$$(\ sup(0,0002),\ conf(0.27),\ lift(6.64),\ lev(0),\ conv(1.29)\ )$$
$$\{\text{Muito Longo e Música}\} \longrightarrow \{\text{Fracasso e Não-Popular}\}$$
$$(\ sup(0,0003),\ conf(0.19),\ lift(4.54),\ lev(0),\ conv(1.17)\ )$$

Em geral vimos que é possível explorar a taxa de retenção para, conjuntamente com outras unidades de análise, melhor compreender a forma como os vídeos são distribuídos na plataforma em estudo.

5.6.6 Predição Automática de Gênero

Um ponto importante na venda de publicidade em vídeos, seja na Internet ou não, é conhecer bem o perfil do expectador para poder mostrar a publicidade certa para o *consumidor* certo. De nada adianta mostrar uma propaganda de creme de barbear para o público feminino. Uma forma de minimizar esse erro é classificar os vídeos em gêneros que representem seu conteúdo, para com isso, exibir publicidades relacionadas ao conteúdo dos vídeos em questão.

Conforme vimos nas análises da Seção 5.5 as categorias dos vídeos estão intimamente ligadas ao nicho do *Produtor de Conteúdo* e não possibilitam na maioria das vezes uma segmentação categórica global. Na tentativa de abordar tal problema o metadado gênero foi proposto e adicionado à plataforma em questão. Entretanto, a sua adesão e utilização é baixa, somente 5,6%.

Uma forma de resolver esta limitação da adesão ao campo gênero, é predizer o gênero de um vídeo com base nos seus metadados. Essa abordagem melhoraria a categorização dos vídeos e possibilitaria à plataforma selecionar automaticamente uma publicidade e exibi-la ao usuário minimizando o risco de apresentar uma publicidade fora de contexto.

Para avaliar a viabilidade de atribuir um gênero automaticamente a um vídeo, utilizamos os dados já presentes nos vídeos para ver a relação entre os metadados e o gênero que foi atribuído manualmente. O objetivo dessa análise é utilizar a técnica de regras de associação para avaliar se a relação existente entre os metadados (título, descrição e *tags*) e o **gênero** é forte o suficiente para viabilizar tal automação. Neste caso, a categoria não foi usada pois derivamos os gêneros a partir desses metadados.

Para avaliar essa hipótese, utilizamos as informações dos títulos, descrições, *tags* e gêneros como entrada do algoritmo de regras de associação e procuramos por regras que implicassem no consequente gênero. Encontramos as seguintes regras associando a descrição com o conteúdo.

$$\{\text{policial}\} \longrightarrow \{\text{Notícias e Politica}\}$$
$$(\ sup(0,08),\ conf(0.95),\ lift(1.62),\ lev(0.03),\ conv(8.94)\ )$$
$$\{\text{lei, última, novidade}\} \longrightarrow \{\text{Notícias e Politica}\}$$
$$(\ sup(0,05),\ conf(0.86),\ lift(1.46),\ lev(0.02),\ conv(2.9)\ )$$

Tais regras mostram que a presença dos termos das regras implicam na pertinência dos vídeos no gênero de notícias bem acima do esperado. Veja que a confiança e os outros indicadores possuem valores acima do esperado indicando que existe uma correlação forte entre a descrição e seu gênero.

Como o campo de descrição é um campo textual livre e pode chegar a conter milhares de palavras, a coocorrência de vários termos, mesmo com o pré-processamento, domina as relações tornando difícil a avaliação das regras. Visando facilitar tal avaliação retiramos o campo de descrição e realizamos uma nova análise dos dados.

Avaliando somente os metadados de título e *tags*, vimos várias evidências que mostram, com alta confiança, que é possível derivar o gênero de forma automática a partir dos metadados. Uma regra que surge naturalmente forte é a associação entre o nome de um programa do *Produtor de Conteúdo* e seu gênero. Por exemplo:

$$\{\text{tags=(fazenda, fazenda 3)}\} \longrightarrow \{\text{Entretenimento}\}$$
$$(\ sup(0,008),\ conf(1),\ lift(6.13),\ lev(0.01),\ conv(849.54)\ )$$
$$\{\text{tag=(balanco geral)}\} \longrightarrow \{\text{Noticias e Politica}\}$$
$$(\ sup(0,006),\ conf(1),\ lift(1.7),\ lev(0),\ conv(313.34)\ )$$

Vemos que, nesse caso, a regra é óbvia, ocorrendo em 100% dos casos. Parte disso pode ser atribuído a forma como os gêneros foram atribuídos aos vídeos, porém a força da regra nos leva a ver que a utilização dos nomes dos programas como *tags* é um padrão adotado pelos editores nos vídeos de tais programas.

$$\{\text{titulo=(x),tag=(campeonato brasileiro)}\} \longrightarrow \{\text{Esporte}\}$$
$$(\ sup(0,003),\ conf(1),\ lift(9.48),\ lev(0),\ conv(333.64)\ )$$
$$\{\text{titulo=(veja), tag=(programa, episódio)}\} \longrightarrow \{\text{Entretenimento}\}$$
$$(\ sup(0,0008),\ conf(1),\ lift(6.13),\ lev(0),\ conv(83.7)\ )$$
$$\{\text{titulo=(polícia, prendeu), tag=(roubo)}\} \longrightarrow \{\text{Noticias e Politica}\}$$
$$(\ sup(0,0006),\ conf(1),\ lift(1.7),\ lev(0),\ conv(32.61)\ )$$
$$\{\text{tag=(musica, show, banda)}\} \longrightarrow \{\text{Música}\}$$
$$(\ sup(0,001),\ conf(0.98),\ lift(76.63),\ lev(0),\ conv(43.68)\ )$$
$$\{\text{tag=(cinema, filme, trailer)}\} \longrightarrow \{\text{Filmes \& Animação}\}$$
$$(\ sup(0,003),\ conf(0.87),\ lift(64.29),\ lev(0),\ conv(7.56)\ )$$

Como podemos ver, existem correlações extremamente fortes entre vários termos e os gêneros dos vídeos. A relação entre a *tag campeonato brasileiro* e a letra x no título é uma relação interessante que leva ao gênero esporte. A letra x nesse caso é comumente usada no título com o sentido de versus em vídeos futebolísticos.

Algumas regras chegam a ser óbvias para nós humanos, mas mostram o potencial de automação da inferência do gênero a partir dos seus metadados.

5.6.7 Conclusões

Nesta seção ilustramos a aplicação do nível de conhecimento aplicando regras de associação aos dados da plataforma de distribuição de conteúdo. Mostramos como é importante a etapa de pré-processamento que possibilita as análises mais diretas e simples e propomos alguns métodos para discriminar a popularidade dos vídeos e avaliar a taxa de retenção destes.

Aplicamos a técnica de regras de associação a três análises que produziram conclusões interessantes. Ao analisar a popularidade relativa dos vídeos e sua duração, vimos que ao contrário do senso comum, vídeos mais curtos tendem a ser menos populares que vídeos longos. Avaliamos também que a análise segmentada dos vídeos pode prover melhores *insights* a respeito da forma que esses vídeos são assistidos. Constatamos que o tamanho da mídia afeta a popularidade desta dependendo do conteúdo. Por exemplo, para o caso de vídeos de entretenimento, o fato da duração do vídeo estar entre 3 e 10 minutos influencia negativamente na popularidade se comparado aos vídeos com outras durações.

Analisando a taxa de retenção dos vídeos, pudemos perceber que, em geral, a taxa é baixa. Tal fato indica uma tendência de comportamento por parte dos usuários que não estão dispostos a se engajar nos vídeos muito provavelmente devido ao baixo interesse nos conteúdos exibidos. A baixa aceitação no conteúdo de humor mostra que tais conteúdos devem ser melhor elaborados.

Vimos também que não existe uma relação entre a popularidade de um vídeo e o seu *sucesso*. Na realidade existem evidências do contrário. Acreditamos que tais informações podem ser usadas pelos *Produtores de Conteúdo* para otimizar a entrega de seus conteúdos.

Outra constatação é que a segmentação dos vídeos em diferentes conteúdos influencia a taxa de retenção. Tal fato mostra que alguns conteúdos são mais adequados ao perfil da Internet que outros. De toda forma, a segmentação possibilita uma melhor análise e entendimento da dinâmica da distribuição dos vídeos.

Como nos dois casos a segmentação por conteúdos mostrou vantagens na análise e compreensão da forma como os vídeos são consumidos, seria ideal uma forma de segmentação automática dos vídeos. Pensando nisso, analisamos as correlações entre os metadados textuais e o gênero atribuído aos vídeos. Mostramos que exitem relações fortes entre tais campos textuais dos vídeos e seus gêneros. Tal fato mostra que é possível a automatização dessa classificação de vídeos, o que melhoraria a compreensão da distribuição dos vídeos online.

Capítulo 6

Conclusões e Trabalhos Futuros

Nesta dissertação foi proposta uma metodologia de avaliação e estudo para dados de servidores multimídia. Tal metodologia tem por objetivo principal estabelecer uma abordagem incremental que possibilite a análise dos conteúdos multimídias em dois níveis. O primeiro nível é hierárquico e utiliza as informações existentes nos objetos e seus conteúdos de forma a complementar a análise geral das requisições realizadas pelos usuários. Já o segundo nível, complementar ao primeiro, é marcado pela aplicação de um processo mais complexo de análise de dados visando a extração do conhecimento das informações avaliadas na camadas anteriores. Esta abordagem organiza as informações a serem avaliadas e guia o estudo de tais servidores através exemplos de análises. A metodologia aqui proposta foi aceita para ser apresentada no Simpósio Brasileiro de Sistemas Multimída e Web em outubro deste ano [Gonçalves et al., 2011a].

Também avaliou-se o uso do arcabouço experimental de computação distribuída, cuja arquitetura pode ser adotada em diferentes trabalhos de processamento, caracterização e análise de dados. Tal arcabouço mostrou-se adequado aos objetivos e escalável. O trabalho de avaliação desse arcabouço culminou na apresentação de um resumo extendido no SCECR [Gonçalves et al., 2011b].

Para validar a metodologia, desenvolvemos um estudo de caso aplicado a um cenário real através de um conjunto de informações obtido da maior empresa de distribuição de conteúdos multimídia corporativos da América Latina, que engloba parte significativa do tráfego dos vídeos online no Brasil.

Dentre os resultados obtidos através do estudo de caso, mostramos que as avaliações sobre o ponto das requisições permite extrair informações importantes para o planejamento de capacidade dos servidores. Não foram encontrado evidências de comportamentos periódicos temporais para intervalos maiores que uma semana. Entretanto, as análises levantaram questionamentos a serem abordados nas análises em out-

ras camadas.

Ao segmentar as análises através dos diferentes objetos, identificamos que era possível avaliar a o engajamento do usuário através da relação entre as requisições de thumbnails e vídeos. Concluímos que o período de final de semana o engajamento é maior que para os dias úteis. Vimos que o acesso aos conteúdos multimídia do trabalho impacta na rede de distribuição e que a maior parte dos usuários possuem velocidade de conexão entre 1 e 5 Mbps. Avaliando as propriedades dos objetos notamos que vídeos de duração maiores estão ficando mais frequentes. Com relação a popularidade dos objetos, vimos uma relação de concentração de requisições onde 90 das requisições são realizadas em somente 10% dos objetos. Além disso, a maior parte das requisições, ocorrem nos primeiros 4 dias a partir da publicação do conteúdo.

Ao abordar os dados com informações adicionais dos conteúdos concluímos que a segmentação dos objetos em diferentes conteúdos é essencial para a melhor compreensão do sistema. Em todas as análises, os resultados eram afetados pela diferenciação em conteúdos. As durações dos vídeos variam de acordo com o conteúdo, vídeos de entretenimento tendem a ser maiores que vídeo jornalísticos. O mesmo se aplica para popularidade, as visualizações de vídeos de entretenimento são muito maiores que a outro tipo de conteúdo.

No segundo nível de análises da metodologia, o nível de extração de conhecimento, fizemos algumas análises utilizando regras de associação. Focamos em três pequenas análises que avaliaram a popularidade dos vídeos, a taxa de retenção destes e a viabilidade da predição automática de gênero para os vídeos. Mostramos que a duração dos vídeos está relacionada com a sua popularidade. Por exemplo, os vídeos mais curtos tendem a ser menos populares, entretanto a avaliação da taxa de conversão mostrou que mesmo sendo populares, os vídeos maiores possuem uma alta taxa de desistência. Outro ponto avaliado foi que não existe uma relação entre a popularidade de um vídeo e o seu *sucesso* (retenção alta), na realidade as análises mostram o contrário.

Em geral vimos que a segmentação das análises ajuda a compreender melhor a forma como os objetos multimídia, principalmente vídeo para o nosso caso, são consumidos. A metodologia se mostrou efetiva pois a segmentação dos dados revelou informações que ajudaram a compreender a diferença entre os objetos e conteúdos. A aplicação das técnicas de **KDD** se mostraram promissoras aumentando o conhecimento do sistema em avaliação.

6.1 Trabalhos Futuros

A metodologia possui um escopo muito grande. Como foi dito anteriormente, a quantidade de análises que podem ser realizadas em cada camada cresce a medida que segmentamos os dados. Com base nisso, é possível especializar outras análises como as das camadas C e K e avaliar outros aspectos não contemplados neste trabalho.

Um dos indícios mais fortes que encontramos é que a segmentação dos objetos multimídia por diferentes tipos de conteúdo impacta na forma como os usuários consomem tais objetos. Com as análises do nível K vimos que é possível a inferência automática do gênero a partir dos metadados. Com isso, como trabalhos futuros pretendemos aplicar novos processos de **KDD** como classificação, agrupamento que segmente os conteúdos deforma automática e reaplicar algumas das análises para estudar aspectos de outros tipos de conteúdo.

Outro ponto que poderia ser explorado é a avaliação mais criteriosa de algumas das métricas propostas aqui como a mediada de engajamento dos usuários além da taxa de sucesso e fracasso. Tais métricas poderiam ser melhores detalhadas e abordadas com outras bases de dados.

Por fim acreditamos que uma abordagem comparativa com dados futuros possam demonstrar tendências e revelar comportamentos que podem ajudar na melhoria dos serviços de entregas de conteúdo multimídia.

Referências Bibliográficas

[Acharya & Smith, 2000] Acharya, S. & Smith, B. (2000). Characterizing user access to videos on the world wide web. In *IEEE Multimedia Computing and Networking (MMCN)*.

[Agrawal et al., 1993] Agrawal, R.; Imieliński, T. & Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, SIGMOD '93, pp. 207--216, New York, NY, USA. ACM.

[Agrawal & Srikant, 1994] Agrawal, R. & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pp. 487--499, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Almeida et al., 2001] Almeida, J. M.; Krueger, J.; Eager, D. L. & Vernon, M. K. (2001). Analysis of educational media server workloads. In *NOSSDAV '01: Proceedings of the 11th international workshop on Network and operating systems support for digital audio and video*, pp. 21--30, New York, NY, USA. ACM.

[Apache Foundation, 2011] Apache Foundation (2011). Hadoop: Apache software foundation project home page. `http://hadoop.apache.org/` acessado em Março de 2011.

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

[Benevenuto et al., 2010] 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):115–129.

[Benevenuto et al., 2009a] Benevenuto, F.; Rodrigues, T.; Almeida, V.; Almeida, J. & Gonçalves, M. (2009a). Detecting spammers and content promoters in online video social networks. In *Proc. of Int'l ACM SIGIR*, Boston, MA, USA.

[Benevenuto et al., 2009b] Benevenuto, F.; Rodrigues, T.; Almeida, V.; Almeida, J. & Ross, K. (2009b). Video interactions in online video social networks. *ACM Trans. Multimedia Comput. Commun. Appl.*, 5(4):1--25.

[Blog, 2010] Blog, T. O. Y. T. (2010). Oops pow surprise...24 hours of video all up in your eyes! `http://youtube-global.blogspot.com/2010/03/oops-pow-surprise24-hours-of-video-all.html` acessado em Março de 2011.

[Boll, 2007] Boll, S. (2007). Multitube–where web 2.0 and multimedia could meet. *IEEE MultiMedia*, 14(1):9--13.

[Brin et al., 1997] Brin, S.; Motwani, R.; Ullman, J. D. & Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, SIGMOD '97, pp. 255--264, New York, NY, USA. ACM.

[Cha et al., 2007] Cha, M.; Kwak, H.; Rodriguez, P.; Ahn, Y. & Moon, S. (2007). I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *Proc. Internet Measurement Conference (IMC)*.

[Cha et al., 2008] Cha, M.; Rodriguez, P.; Crowcroft, J.; Moon, S. & Amatriain, X. (2008). Watching television over an ip network. In *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, IMC '08, pp. 71--84, New York, NY, USA. ACM.

[Chatzopoulou et al., 2010] Chatzopoulou, G.; Sheng, C. & Faloutsos, M. (2010). A first step towards understanding popularity in YouTube. In *2010 INFOCOM IEEE Conference on Computer Communications Workshops*, pp. 1--6. IEEE.

[Cherkasova & Gupta, 2002] Cherkasova, L. & Gupta, M. (2002). Characterizing locality, evolution, and life span of accesses in enterprise media server workloads. In *NOSSDAV '02: Proceedings of the 12th international workshop on Network and operating systems support for digital audio and video*, pp. 33--42, New York, NY, USA. ACM.

[Chesire et al., 2001] Chesire, M.; Wolman, A.; Voelker, G. M. & Levy, H. M. (2001). Measurement and analysis of a streaming-media workload. In *Proceedings of the 3rd*

conference on USENIX Symposium on Internet Technologies and Systems - Volume 3, USITS'01, pp. 1--1, Berkeley, CA, USA. USENIX Association.

[Cloudera, 2010] Cloudera (2010). Cloudera. `http://www.cloudera.com/` acessado em Março de 2011.

[Costa et al., 2004] 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 *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pp. 534--543, New York, NY, USA. ACM.

[Crane & Sornette, 2008] Crane, R. & Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653.

[Dalmazo, 2011] Dalmazo, L. (2011). Sambatech começa a atender médias empresas. `http://exame.abril.com.br/blogs/zeros-e-uns/2011/02/16/sambatech-comeca-a-atender-medias-empresas/` acessado em Março de 2011.

[Dean & Ghemawat, 2008] Dean, J. & Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51:107--113.

[Dean et al., 2004] Dean, J.; Ghemawat, S. & Inc, G. (2004). Mapreduce: simplified data processing on large clusters. In *In OSDI'04: Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation*. USENIX Association.

[Duarte et al., 2007a] Duarte, F.; Benevenuto, F.; Almeida, V. & Almeida, J. (2007a). Geographical characterization of youtube: a latin american view. In *Proc. LA-Web Conference*.

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

[Fayyad et al., 1996] Fayyad, U.; Piatetsky-Shapiro, G. & Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. *Commun. ACM*, 39:27--34.

[Figueiredo et al., 2011] Figueiredo, F.; Benevenuto, F. & Almeida, J. (2011). The tube over time: Characterizing popularity growth of youtube videos. In *Proceedings of the 4th ACM International Conference of Web Search and Data Mining (WSDM'11)*.

[Gantz, 2008] Gantz, J. (2008). The diverse and exploding digital universe.

[Gill et al., 2007] Gill, P.; Arlitt, M.; Li, Z. & Mahanti, A. (2007). Youtube traffic characterization: a view from the edge. In *ACM SIGCOMM conference on Internet measurement (IMC)*.

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

[Gonçalves et al., 2011a] Gonçalves, C. F.; Totti, L. C.; Duarte, D.; Jr, W. M. & Pereira, A. C. M. (2011a). Rock: uma metodologia para caracterização de serviços web multimídia baseada na hierarquia da informação. In *Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)*, WebMedia 2011.

[Gonçalves et al., 2011b] Gonçalves, C. F.; Totti, L. C.; Duarte, D.; Pereira, A. C. M. & Jr, W. M. (2011b). An open large dataset processing framework for statistical analysis. In *Seventh Symposium on Statistical Challenges in Electronic Commerce Research*, SCECR 11.

[Hall et al., 2009] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P. & Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10--18.

[Han, 2005] Han, J. (2005). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Limelight, 2010] Limelight (2010). Limelight 2010 financial analyst meeting. `http://www.limelightnetworks.com/pdfs/2010/2010FinancialAnalystDay.pdf` acessado em Março de 2011.

[Maia & Virgilio Almeida, 2009] Maia, M. & Virgilio Almeida, J. A. (2009). Vídeo gerado por usuários: Caracterização de tráfego. In *XXVII Simpósio Brasileiro de Rede de Computadores*. SBC.

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

[Newman, 2005] Newman, M. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46.

[Olston et al., 2008] Olston, C.; Reed, B.; Srivastava, U.; Kumar, R. & Tomkins, A. (2008). Pig latin: a not-so-foreign language for data processing. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, SIGMOD '08, pp. 1099--1110, New York, NY, USA. ACM.

[Piatetsky-Shapiro, 1991] Piatetsky-Shapiro, G. (1991). *Discovery, Analysis, and Presentation of Strong Rules*. AAAI/MIT Press, Cambridge, MA.

[Pingdom, 2010] Pingdom (2010). Internet 2010 in numbers. `http://royal.pingdom.com/2011/01/12/internet-2010-in-numbers/` acessado em Março de 2011.

[SambaTech, 2011a] SambaTech (2011a). Liquid, gestão e distribuição profissional de vídeos. `http://www.sambatech.com/site/liquid-2/` e `http://webtv.liquidplatform.com/2.0/login` acessado em Março de 2011.

[SambaTech, 2011b] SambaTech (2011b). Samba tech, solução de vídeos online. `http://www.sambatech.com/` acessado em Março de 2011.

[Sar, 2011] Sar, E. V. D. (2011). Arrr! the music pirates are still here. `http://torrentfreak.com/arrr-the-music-pirates-are-still-here-110207/` acessado em Março de 2011.

[Saxena et al., 2008] 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*, NOSSDAV '08, pp. 39--44, New York, NY, USA. ACM.

[SBC, 2006] SBC (2006). Relatório preliminar dos grandes desafios da pesquisa em computação no brasil. `http://www.ic.unicamp.br/~cmbm/desafios_SBC/` acessado em Março de 2011.

[Shvachko et al., 2010] Shvachko, K.; Kuang, H.; Radia, S. & Chansler, R. (2010). The hadoop distributed file system. *Mass Storage Systems and Technologies, IEEE / NASA Goddard Conference on*, 0:1–10.

[Silva et al., 2009] Silva, T.; Mota, V.; Valadão, E.; Almeida, J. & Guedes, D. (2009). Caracterização do comportamento dos espectadores em transmissões de vídeo ao vivo geradas por usuários. In *27th Simpósio Brasileiro de Rede de Computadores*. SBC.

[Sripanidkulchai et al., 2004] Sripanidkulchai, K.; Maggs, B. & Zhang, H. (2004). An analysis of live streaming workloads on the internet. In *Proceedings of the 4th ACM*

SIGCOMM conference on Internet measurement, IMC '04, pp. 41--54, New York, NY, USA. ACM.

[StreamingMedia, 2011] StreamingMedia (2011). List of online videos platform. `http://www.streamingmedia.com/pdf/OnlineVideoPlatforms.pdf` acessado em Março de 2011.

[Tan et al., 2005] Tan, P.-N.; Steinbach, M. & Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[Veloso et al., 2006a] Veloso, E.; Almeida, V.; Meira, J. W.; Bestavros, A. & Jin, S. (2006a). A hierarchical characterization of a live streaming media workload. *IEEE/ACM Trans. Netw.*, 14(1):133--146.

[Veloso et al., 2006b] Veloso, E.; Almeida, V.; Meira, W.; Bestavros, A. & Ji, S. (2006b). A hierarchical characterization of a live streaming media workload. *IEEE/ACM Transactions on Networking*.

[White, 2009] White, T. (2009). *Hadoop: The Definitive Guide*. O'Reilly Media, 1 edição.

[Witten & Frank, 2005] Witten, I. H. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Yu et al., 2006] Yu, H.; Zheng, D.; Zhao, B. Y. & Zheng, W. (2006). Understanding user behavior in large-scale video-on-demand systems. *SIGOPS Oper. Syst. Rev.*, 40(4):333--344.

[Zink et al., 2008] Zink, M.; Suh, K.; Gu, Y. & Kurose, J. (2008). Watch global, cache local: Youtube network traces at a campus network - measurements and implications. In *IEEE Multimedia Computing and Networking (MMCN)*.

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