CARACTERIZANDO VARIÁVEIS DE INTERATIVIDADE
DOS ALUNOS DO CURSO DE COMPUTAÇÃO DO CEDERJ,
BASEADO NO SERVIDOR MULTIMÍDIA RIO

Bruno César Barbosa Alves

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Aprovada por:

Prof. Edmundo Albuquerque de Souza e Silva, Ph.D.

Profa. Rosa Maria Meri Leão, Dr.

Profa. Jussara Marques de Almeida, Ph.D.

Prof. Henrique Luiz Cukierman, D.Sc.

RIO DE JANEIRO, RJ - BRASIL
JULHO DE 2007

ALVES, BRUNO CÉSAR BARBOSA

Caracterizando variáveis de interatividade dos alunos do curso de computação do CEDERJ, baseado no servidor multimídia RIO. [Rio de Janeiro] 2007

XV, 78 p. 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia de Sistemas e Computação, 2007)

Dissertação - Universidade Federal do Rio de Janeiro, COPPE

1. Caracterização
2. Vídeo sob Demanda
3. Transmissão Multimídia
4. Interatividade
5. Redes de computadores

I. COPPE/UFRJ II. Título (Série)

Dedico este trabalho aos meus pais
Luiz César Alves e Denise Maria de Paula Barbosa,
às minhas irmãs, minha noiva, familiares e
todos que contribuíram para a conclusão do mesmo.

Agradecimentos

Agradeço aos meus pais que se dedicaram ao meu lado para que este trabalho fosse concretizado. À minha noiva, minhas irmãs, meus avós, familiares e amigos que contribuíram para que eu me concentrasse em meus objetivos e os pudesse concretizar.

Não posso me esquecer da valiosa atenção de meus colegas do LAND, aos membros da equipe do projeto DIVERGE/GIGA, à Carol e aos professores Edmundo e Rosa pelo grande apoio e dedicação. Agradeço em especial ao colega Bernardo Netto pela disponibilidade incondicional e à colega Carolina Vielmond pela valiosa ajuda.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

CARACTERIZANDO VARIÁVEIS DE INTERATIVIDADE
DOS ALUNOS DO CURSO DE COMPUTAÇÃO DO CEDERJ,
BASEADO NO SERVIDOR MULTIMÍDIA RIO

Bruno César Barbosa Alves

Julho/2007

Orientadores: Edmundo Albuquerque de Souza e Silva
Rosa Maria Meri Leão
Programa: Engenharia de Sistemas e Computação

O serviço de vídeo sob demanda com transmissão das mídias em tempo real implica em uma série de desafios devido à grande quantidade de recursos de rede e computacionais necessários. Para projetar um sistema que atenda um grande número de clientes com uma qualidade de serviço satisfatória se faz necessário a utilização de técnicas de compartilhamento de recursos, onde conhecer bem as características da carga de trabalho gerada pelos usuários ao servidor é um fator fundamental para o sucesso destes métodos.

Com o objetivo de conhecer melhor o comportamento interativo dos usuários do sistema multimídia utilizado no curso de Tecnologia de Sistemas de Computação do consórcio CEDERJ, neste trabalho é feita uma caracterização de variáveis de interatividade dos alunos do curso através de registros reais de ações gerados durante suas sessões. São analisados diversos aspectos sobre a interação dos alunos com o sistema. Características como o tempo em que o aluno permanece ininterruptamente assistindo a aula, padrões de movimentação dentro do conteúdo, duração das sessões, tempos de inatividade, entre outras, são analisadas. Para algumas métricas, é feita uma busca por modelos de distribuições conhecidas que possam representá-las com a maior fidelidade possível.

Baseado neste conhecimento um modelo matemático pôde ser construído [24] para gerar carga sintética a fim de estudar o servidor. Este estudo permite avaliar o desempenho e qualidade do serviço prestado, bem como definir parâmetros para utilização ou implementação de técnicas que permitam melhorar a escalabilidade do sistema.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

CHARACTERIZING INTERACTIVITY VARIABLES OF CEDERJ'S COMPUTER SCIENCE STUDENTS, USING RIO MULTIMEDIA SERVER.

Bruno César Barbosa Alves

July/2007

Advisors: Edmundo Albuquerque de Souza e Silva
Rosa Maria Meri Leão
Department: Computer and System Engineering

Real time video streaming applications have stringent computational requirements. In order to design a system capable of serving a large number of clients with satisfactory quality it is necessary to utilize resource sharing techniques. These in turn heavily depend on the characteristics of the workload generated by the users.

In this work we study the workload generated by the users of a streaming video server currently employed in the CEDERJ consortium of public universities in the state of Rio (the RIO server). The users of the system have a lot of flexibility to perform interactive operations that interrupt the flow of the video being transmitted. These are, for example, pause, fast forward, jump to a new slide of the lecture being presented or to a different topic. We characterize the workload focusing on the user interactivity, that certainly affects the performance of any resource sharing algorithm. We analyze characteristics such as the interval of time the students watch a video-lecture before performing an operation such as stop or jump; moving patterns in the video-lectures; session duration, among others. We also try to determine the bests distribution functions that can be used to represent the collected metrics with acceptable accuracy. We compared the results obtained in our study with those that analyze workloads generated by similar applications. Our workload however have unique features that contrast with those in the literature.

Our work was the foundation for the research presented in [24], that addressed the construction of a mathematical model of the students behavior while watching the video-lectures. The mathematical model is currently being used to generated synthetic load in order to study the performance of the RIO video server used in the CEDERJ consortium and its distributed version as part of the DIVERGE project sponsored by RNP/FINEP.

Glossário de Redes

Largura de Banda	:	Capacidade de transmissão de dados.
QoS	:	Qualidade de Serviço (*Quality of Service*).
Multicast	:	Envio de uma mesma informação a um grupo específico.
TCP	:	Protocolo de controle de transmissão do nível da camada de transporte (*Transmission Control Protocol*).
UDP	:	Protocolo da camada de transporte sem conexão (*User Datagram Protocol*).
Prefetching	:	Armazenamento prévio de dados.
Cache	:	Armazenamento de dados que já foram transmitidos para reaproveitamento posterior.
Proxy	:	Servidor intermediário usado para reduzir a carga no servidor e na rede entre o servidor proxy e o servidor principal, além de reduzir o tempo de resposta ao usuário.

Glossário de Serviços Multimídia

VoD	:	Vídeo sob Demanda.
VCR	:	Interações do tipo Gravador de Vídeo Cassete.
Transmissão em tempo real	:	Transmissão de fluxos de dados onde o usuário visualiza o objeto a medida que este vai sendo transmitido.
Stream	:	Fluxo contínuo de dados.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

APLICAÇÕES de vídeo e voz têm se tornado cada vez mais comuns e utilizadas na internet. Tais aplicações demandam uma quantidade considerável de recursos de rede. Alguns vídeos, mesmo comprimidos em um determinado padrão, como o MPEG, um dos mais usados atualmente, consomem até vários megabits por segundo da taxa de transmissão da rede. Como exemplo, um vídeo codificado no padrão MPEG-2, com resolução de 320x240 pixels, consomem em média 1,5 megabits por segundo. Sendo assim, para transmitir cerca de 1000 destes objetos, seria necessária uma banda de aproximadamente 1500 megabits por segundo (Mbps). Para realizar uma transmissão com estas características, sem que uma grande quantidade de largura de banda seja necessária, representando um aumento no custo de implantação dos recursos, é necessário o emprego de métodos que permitam economia no uso de tais recursos, por exemplo pelo compartilhamento de banda [11, 8].

Outra forma de diminuir a carga e a demanda por recursos da rede e aumentar a escala de um sistema consiste na utilização de arquiteturas Peer-to-peer (P2P). Sistemas P2P vem sendo amplamente utilizados para o compartilhamento e transmissão de dados na internet. Alguns trabalhos [5, 3] procuram explorar serviços de vídeo sob demanda sobre arquiteturas P2P. O uso de arquiteturas P2P para prestar serviço de entrega de fluxos multimídia em tempo real possibilita um aumento na escalabilidade do sistema, uma vez que esta aumenta proporcionalmente com o número de usuários. Diversos algoritmos de agendamento para a entrega e armazenamento dos blocos das mídias, bem com técnicas para permitir o agrupamento de usuários em um mesmo fluxo são propostas e analisadas nestes trabalhos.

Além dos dois métodos citados, outra técnica consiste em fazer *cache* de parte dos dados presentes no servidor em proxies localizados próximos aos clientes [16, 2, 18, 25]. Isto pode evitar que os objetos mais populares tenham que ser transmitidos

desde o servidor até o cliente toda vez que são solicitados.

Quando o sistema permite que o usuário realize interações (como em um VCR) com o conteúdo que está assistindo, a aplicação de técnicas para prover maior escalabilidade se torna mais complexa isto porque, nestes casos, não se sabe ao certo como e quando os acessos são feitos às partes dos objetos disponibilizados pelo servidor.

Em ambientes onde é permitida a interatividade, conhecer e caracterizar o comportamento dos usuários que habitualmente acessam o sistema, permite identificar quais são as técnicas mais indicadas para prover maior escalabilidade e avaliar o comportamento do sistema quando submetido a uma carga de usuários interativos.

Além disso, uma vez que o comportamento dos usuários é caracterizado, ou seja, sabendo quais são as características da carga que é gerada pelos clientes, é possível realizar a geração de carga sintética (simulação de usuários) para observar como o sistema se comporta e estudar quantitativamente e qualitativamente a melhoria no desempenho obtido através do uso de técnicas para prover economia de recursos.

Um exemplo de ambiente com alta interatividade é um sistema multimídia para educação a distância, onde os alunos poe exemplo assistem o início do vídeo, para identificar a parte do vídeo que mais os interessa, e em seguida saltam para a parte de seu interesse. Com este tipo de comportamento acontecendo numa boa parte dos casos, armazenar previamente o início dos vídeos disponíveis no servidor em um servidor proxy, que esteja mais próximo do cliente que o servidor, pode agilizar o acesso do cliente ao objeto e diminuir o gargalo no servidor, além de reduzir o tráfego na rede do servidor até o proxy, uma vez que a parte do vídeo pré-armazenada não precisará ser transmitida do servidor até o proxy sempre que solicitada.

1.1 Objetivo

O principal objetivo deste trabalho é a caracterização das variáveis de interatividade de usuários que acessam o servidor multimídia RIO (Random I/O Storage System), usado no projeto CEDERJ (Centro de Educação Superior a Distância do Estado do Rio de Janeiro), para a distribuição de conteúdo educacional no curso de "Tecnologia em Sistemas de Computação", baseado nos logs de ações destes usuários.

Os resultados obtidos estão sendo utilizados, para a definição de um modelo [24] e geração de carga sintética. A carga sintética vem sendo utilizada em outros trabalhos em desenvolvimento no laboratório LAND (projeto DIVERGE - Distri-

buição de Vídeo em Larga Escala sobre Redes Giga, com Aplicações a Educação). Os experimentos executados no projeto Diverge visam analisar e aperfeiçoar o serviço prestado pelo servidor RIO implantado em um ambiente de alta velocidade de transmissão de dados. O uso da carga sintética permitirá que o sistema possa ser avaliado antes mesmo que os usuários tenham acesso ao ambiente que está sendo projetado.

1.2 Contribuições

As principais contribuições deste trabalho são:

- caracterização da carga gerada no sistema (*workload characterization*);
- criação de um ambiente de coleta, análise e geração de estatísticas baseados nos logs de acesso;
- análise de informações a respeito do acesso dos alunos do curso de ensino a distância, como que aulas e disciplinas possuem mais acessos, a duração das sessões, a ocorrência de interações do aluno dentro da sessão, etc;

1.3 Organização do Texto

Este trabalho está organizado da seguinte forma. Para entender melhor o ambiente usado pelos alunos, cujo comportamento interativo é estudado, uma descrição detalhada da configuração, sistemas utilizados e dos projetos envolvidos é feita no capítulo 2. Alguns trabalhos presentes na literatura têm uma relação direta com este trabalho e são apresentados no capítulo 3. O capítulo 4 resume todo estudo feito e os resultados obtidos. As conclusões a respeito destes resultados são discutidas no capítulo 5.

Capítulo 2

Ambiente Estudado

PARA entender melhor o ambiente estudado neste trabalho, neste capítulo serão descritos o sistema utilizado, a configuração do ambiente e os projetos que estão envolvidos. Serão apresentadas inicialmente algumas definições.

Todo estudo aqui apresentado foi baseado na coleta e análise de arquivos de log gerados pelo sistema de distribuição multimídia RIO ([11, 12, 9]) utilizado no projeto CEDERJ. Os resultados obtidos estão sendo utilizados em trabalhos em desenvolvimento para o projeto DIVERGE/GIGA.

2.1 Definições

Algumas definições sobre termos adotados para descrever as interações dos clientes com o servidor serão apresentadas nesta seção.

Um **objeto multimídia**, representa um objeto de vídeo, voz ou slides, integrados e sincronizados. Os objetos multimídia são previamente armazenados no **servidor multimídia** e são entregues aos **clientes** em partes de mesmo tamanho chamadas de **blocos**.

Um **usuário** consiste de um aluno, tutor ou professor que acessa o sistema através de um software definido como **cliente**. O cliente trata as interações realizadas pelo usuário e envia os pedidos (**requisições**) relacionados a estas ações ao servidor. Ao receber a resposta de uma requisição, o cliente prepara a informação a ser exibida ao usuário. Um **buffer** é alocado do lado do cliente para receber e armazenar temporariamente os blocos que chegam do servidor. O cliente só inicia a exibição do vídeo quando o *buffer* está cheio. Desta forma, se algum bloco for perdido ou sofrer

atraso durante a transmissão, a exibição do vídeo não pára. O tamanho deste tipo de *buffer*, conhecido como ***playout buffer***, não deve ser muito grande de forma que o usuário não tenha que esperar muito para ter a exibição do vídeo iniciada, porém, não pode ser tão pequeno de forma que uma seqüência de atraso ou perda de alguns blocos não prejudique a exibição para o usuário.

Cada vez que um usuário se conecta (realiza a operação de *logon*) no sistema através do cliente, uma **sessão** é iniciada. Ao finalizar a visualização e encerrar a sessão (*logoff*), as **ações do usuário** que foram registradas pelo cliente vão gerar um arquivo de **log** que será enviado ao servidor para armazenamento.

2.2 Servidor RIO

O servidor RIO é um servidor de distribuição de conteúdo multimídia desenvolvido e mantido pelo Laboratório de ANálise, modelagem e Desenvolvimento de sistemas de computação e comunicação (LAND), localizado na Universidade Federal do Rio de Janeiro (UFRJ).

Trata-se de um servidor de objetos multimídia para transmissão em tempo real que possui, em sua implementação, diversas técnicas de otimização como compartilhamento de banda, balanceamento de carga, alocação aleatória de dados nos discos e controle de admissão de usuários. A implementação de cada uma das características do servidor são originadas de trabalhos de pesquisa realizados neste sistema pelo laboratório LAND.

O servidor RIO é composto por um servidor central que recebe e controla as requisições geradas pelos clientes para acesso ao conteúdo disponível. Além do servidor central existem servidores de armazenamento que podem ser instalados de forma distribuída pela rede, com a função de controlar o armazenamento dos objetos e o envio dos mesmos aos clientes quando solicitados pelo servidor central. Cada servidor de armazenamento pode controlar diversos discos rígidos usados para armazenar os objetos. O software cliente realiza pedidos ao servidor principal que os controla e repassa ao(s) servidor(es) de armazenamento segundo uma política de balanceamento de carga (caso haja replicação) para que nenhum deles seja sobrecarregado. Os processos de armazenamento e leitura dos objetos são descritos a seguir.

- **Armazenamento** - Quando um objeto é passado pelo cliente ao servidor para armazenamento, o servidor principal distribui de forma aleatória os blocos para serem armazenados nos discos dos servidores de armazenamento. Para

aumentar a confiabilidade do sistema e permitir o balanceamento da carga os blocos podem ser replicados pelos servidores de armazenamento existentes.

- **Leitura** - Quando um cliente solicita um objeto ao servidor para visualização, o servidor central recebe a requisição do objeto desejado, coloca em uma fila com os outros pedidos e repassa o pedido de cada bloco que compõe o objeto ao servidor de armazenamento que o possui. O servidor de armazenamento envia, via um fluxo UDP (User Datagram Protocol), o bloco solicitado diretamente ao cliente que solicitou o objeto ao qual o bloco pertence.

Para um melhor entendimento do funcionamento do servidor RIO e sua interação com os clientes, um exemplo de sua configuração pode ser visto na Figura 2.1.

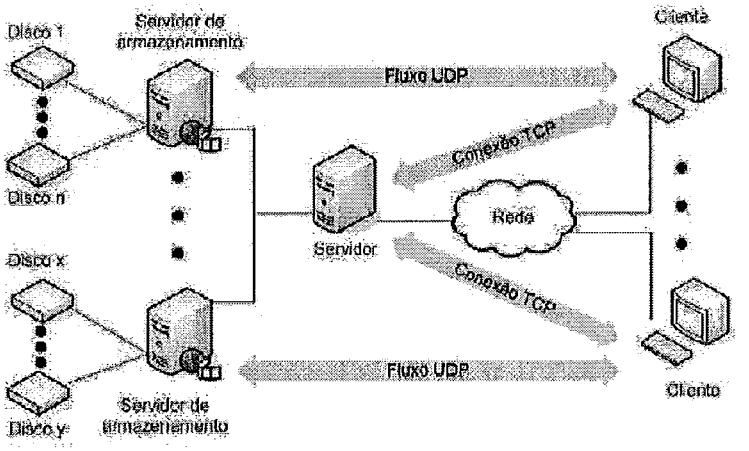


Figura 2.1: Servidor Multimídia Rio

2.2.1 O cliente de visualização

Para entender melhor o cliente utilizado pelos usuários para visualização e interação com os objetos contidos no servidor, serão descritas suas características e funcionalidades. A interface do cliente chamado **riommclient** pode ser vista na Figura 2.2.

O riommclient permite que os usuários solicitem ao servidor o objeto multimídia que desejam assistir e possam interagir com ele podendo realizar as seguintes ações: tocar (*play*), parar (*stop*), pausar (*pause*), avançar alguns blocos (*fast forward*), retroceder alguns blocos (*fast rewind*), acessar o índice de tópicos da aula ou mover a barra de progresso para a posição desejada. A sessão do usuário pode ser encerrada a qualquer instante através do botão *quit*.

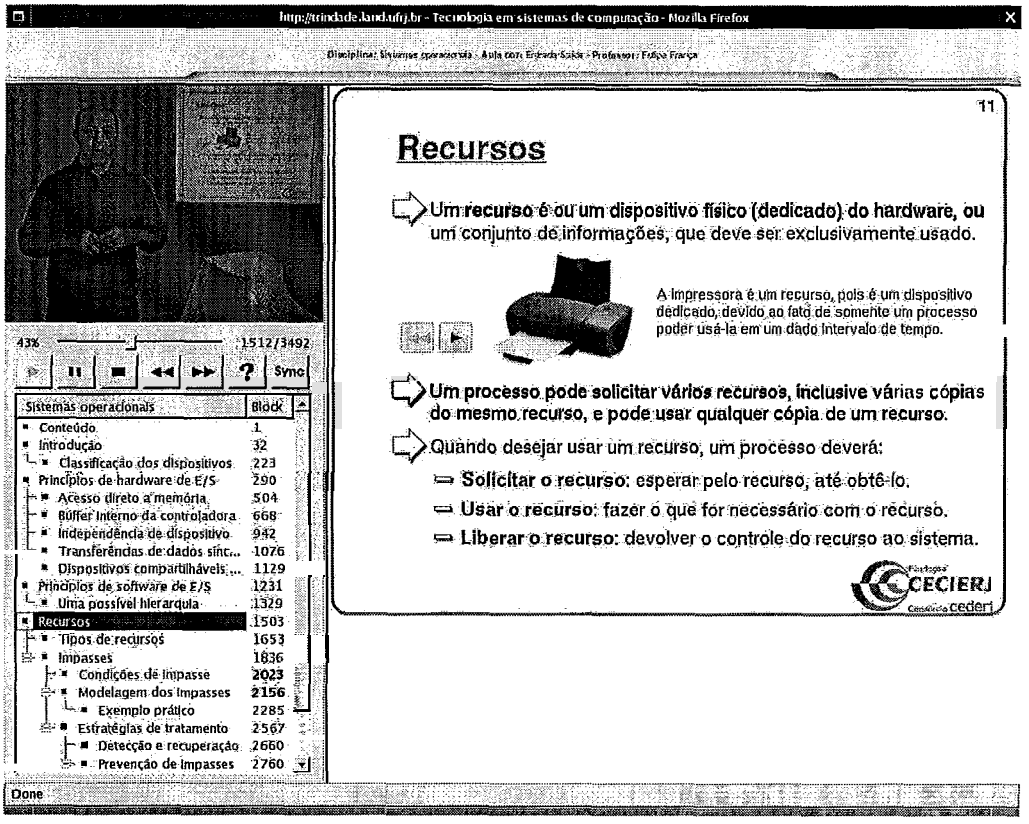


Figura 2.2: Cliente do servidor Rio

As aulas disponibilizadas no servidor são compostas por vídeo, áudio e slides sincronizados. O conteúdo das aulas é altamente interativo, vários exemplos, exercícios e animações interativas são freqüentes.

2.3 Projeto CEDERJ

O projeto CEDERJ é um projeto para ensino a distância junto ao governo do estado do Rio de Janeiro. Neste projeto os alunos matriculados em algum dos pólos de estudo distribuídos pelo estado, podem assistir às aulas disponibilizadas em um servidor localizado em cada pólo. A Figura 4.3 apresenta a distribuição dos pólos de ensino espalhados pelo estado do Rio de Janeiro. O consórcio CEDERJ de universidades públicas do estado do Rio de Janeiro conta com diversos cursos de graduação em diversas áreas de ensino. Alguns destes pólos possuem o curso de Tecnologia em Sistemas de Computação, que conta com vídeo-aulas gravadas e armazenadas previamente compostas por vídeo, áudio e slides sincronizados e disponibilizadas através do servidor multimídia RIO.

Figura 2.3: Pólos de ensino do consórcio CEDERJ

O curso para a formação de tecnólogos em sistemas de computação (que é estudado neste trabalho) é acompanhado pelos alunos de forma semi-presencial. Os alunos vão até o pólo onde estão matriculados e podem assistir às aulas através de clientes que se comunicam com o servidor implantado na mesma rede local (LAN). Além das aulas disponibilizadas nos servidores de todos os pólos, o aluno pode adquirir DVD´s contendo aulas do curso.

O ambiente de configuração do servidor RIO e seus clientes em cada um dos pólos pode ser observado na Figura 2.4.

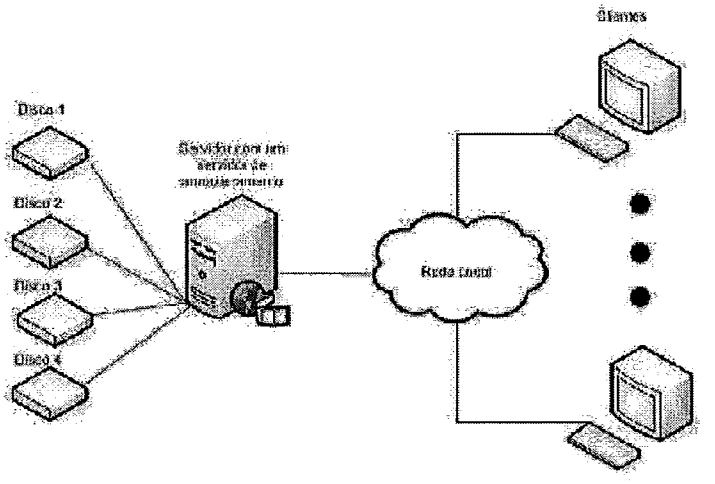


Figura 2.4: Pólo de ensino

Cada servidor armazena as ações de cada usuário em cada sessão por ele aberta em arquivos de log. Estes arquivos de log são coletados de todos os pólos e permitirão que o comportamento dos alunos em cada sessão possa ser analisado e caracterizado neste trabalho.

2.3.1 Arquivos de Log

Os arquivos que guardam informações de acesso dos usuários ao servidor multimídia RIO são os logs de sessão. Cada vez que um usuário realiza o *logon* através de um cliente do servidor, as ações deste usuário interagindo com o sistema, bem como outras informações sobre a sessão que foi aberta são gravadas em um arquivo deste tipo. Após o encerramento da sessão, o arquivo de log é enviado do cliente ao servidor para ser armazenado e posteriormente coletado. Um exemplo dos logs coletados e analisados é mostrado e explicado a seguir.

Na Figura 2.5 é mostrado o cabeçalho de um dos logs. Estão presentes nesta parte do arquivo as informações sobre a sessão que foi aberta. Informações como o instante em que a sessão foi aberta, *login* do usuário que iniciou a sessão, objeto (aula) que foi acessado e estado do índice são alguns dados existentes e que são importantes para a análise que será feita. Em uma parte seguinte do arquivo são mostradas informações sobre o arquivo de configuração que o cliente estava utilizando durante o acesso (ver Figura 2.6).

```
Log date: Fri Jul 15 20:57:37 2005
User login: a20051305065
Object path: /cederj/sistemas_comp/ead05005/Aula_013.mpg
Block size: 131072
Client IP Adrress: 10.5.1.50
Connected: Server
Index status: success
```

Figura 2.5: Cabeçalho do arquivo de log.

```
--------------------Config file--------------------
[riommclient]
RioMMClient/Buffers=2
RioMMClient/ConnectPI=NULL
RioMMClient/MainWindow=0
RioMMClient/PIHost=NULL
RioMMClient/PluginPort=0
RioMMClient/Sync=false
RioMMClient/TgifWindow=0
RioMMClient/UseCache=false
RioMMClient/UsePlugins=true
RioMMClient/User=guest
RioMMClient/Version=1.1
RioMMClient/VideoWindow=0
RioMMVideo/MPlayer/arguments=-quiet -vfm ffmpeg -cache 128 -
RioMMVideo/MPlayer/binary=mplayer
RioMMVideo/MPlayer/embeded=-wid %w
RioMMVideo/MPlayer/subtitle=-sub %s
RioMMVideo/Player=MPlayer
RioMMVideo/RIOxine/arguments=--vbuffer 100 --title %t --nokeyboard stdin://
RioMMVideo/RIOxine/binary=RIOxine
RioMMVideo/RIOxine/embeded=--rootWin %w
RioMMVideo/RIOxine/subtitle=
RioMMVideo/Version=1.0
RioMMVideo/makeLog=true
```

Figura 2.6: Arquivo de configuração do cliente.

As Figuras 2.7, 2.8, 2.9 e 2.10 trazem as ações realizadas pelo usuário durante o acesso ao servidor e ações que o software cliente realizou para atender as solicitações do usuário. O log completo é composto das partes mostradas nas Figuras 2.5, 2.6, 2.7, 2.8, 2.9, 2.10. A primeira coluna da tabela de ações contém o tempo, em micro segundos, desde zero horas de primeiro de janeiro de 1970. Esta data é chamada *Epoch*. A segunda coluna contém a ação do usuário ou do cliente para atender as solicitações deste, estas ações são descritas posteriormente. A terceira coluna mostra o número do bloco de vídeo que está sendo tocado, chegou ou foi pedido. Por fim, a última coluna mostra o tamanho do bloco de dados que foi tocado.

As possíveis ações do usuário e o identificador para cada tipo de ação são os seguintes.

1. BPLAY: Pressionar o botão *play*;

2. BPAUSE: Pressionar o botão *pause*;

3. BFFWRD: Pressionar o botão *fast forward*;

4. BFREW: Pressionar o botão *fast rewind*;

5. STOP: Pressionar o botão *stop*;

6. JMPBAR: Deslizar a barra de progresso do vídeo;

7. JMPTOP: Selecionar um dos tópicos da aula (index), avançando ou retrocedendo;

8. QUIT: Sair do Sistema.

As ações realizadas pelo cliente de visualização são identificadas da seguinte forma:

1. REQSERV: O cliente solicita um bloco do vídeo ao servidor;

2. ARRIVES: O bloco solicitado é recebido;

3. PLAYS: O bloco recebido é tocado.

```
------------------Actions log--------------------
1121465009605283 bplay    0
1121465009612235 reqserv  0
1121465009616389 reqserv  1
1121465009629172 arrives  0
1121465010099699 arrives  1
1121465010102332 plays    0   131072
1121465010543718 reqserv  2
1121465010546268 plays    1   131072
1121465010583398 arrives  2
1121465011188457 reqserv  3
1121465011191161 plays    2   131072
```

Figura 2.7: Início das ações do usuário.

Nas Figuras 2.7, 2.8, 2.9 e 2.10 as entradas que correspondem às ações do usuário estão marcadas em negrito, as demais são ações do cliente em resposta as solicitações do usuário. Na Figura 2.7 as ações do software cliente estão enfatizadas em itálico.

A primeira interação do usuário com o sistema pode ser notada na Figura 2.7 onde a entrada "bplay" indica o clique do usuário no botão *play* e o início da

```
1121465014009732 plays    5    131072
1121465014038597 arrives  6
1121465014064178 jmptop   202
1121465014067017 reqserv  202
1121465014077498 reqserv  203
1121465014493881 arrives  202
1121465014843651 reqserv  204
1121465014902943 arrives  203
1121465014905599 plays    203  131072
```

Figura 2.8: Ação de navegação pelo índice.

```
1121465180939424 plays    787  131072
1121465180971879 arrives  788
1121465181432490 bpause   786
1121465181976434 reqserv  789
1121465181979680 plays    788  131072
1121465182005350 arrives  789
1121465373429903 bplay    786
1121465373864403 reqserv  790
1121465373867024 plays    789  131072
```

Figura 2.9: Ação de pausa.

```
1121466507980012 plays    2809  131072
1121466508003307 arrives  2810
1121471857904063 quit
```

Figura 2.10: Finalização da sessão.

visualização da aula. A partir deste ponto, enquanto o usuário assiste a aula, o software cliente realiza ações de solicitação, controle de chegada e exibição dos blocos seqüenciais do vídeo. Pode ser observado nesta figura (o que é confirmado pelo arquivo de configuração do cliente) que o *buffer* do cliente comporta até dois blocos e este só inicia a exibição dos blocos quando o *buffer* estiver cheio.

Outros tipos de interação, incluindo a finalização da sessão com o comando *quit* podem ser vistas nas porções do arquivo de log mostradas nas Figuras 2.8, 2.9 e 2.10.

Capítulo 3

Trabalhos Relacionados

PARA uma melhor organização, os trabalhos serão classificados em duas categorias. Uma delas compreende trabalhos que realizaram uma caracterização de variáveis de sistemas de transmissão de video/voz. Uma outra categoria é composta por trabalhos cujos autores, além de caracterizar sistemas multimídia, desenvolveram modelos matemáticos para a representação do comportamento dos usuários desses sistemas.

3.1 Caracterização de variáveis aleatórias

3.1.1 Caracterização de usuários do Sistema MANIC

O primeiro trabalho apresentado nesta seção [15], estudou o sistema MANIC (*Multimedia Asynchronous Networked Individualized Courseware*) [19], que distribui aulas em áudio e slides HTML sincronizados para alunos da Universidade de Masssachusetts. As aulas são ministradas ao vivo em sala de aula e depois disponibilizadas no MANIC para serem posteriormente acessadas pelos alunos. A maior parte do áudio gravado nas aulas foi codificado a 14,4 kbps. No MANIC, toda vez que há uma interação do usuário, o áudio é parado e só é retomada a reprodução quando o usuário pressiona o botão *play*. Enquanto isso ele pode navegar livremente sobre os slides ou até mesmo permanecer inativo.

As possíveis ações do usuário no sistema englobam: tocar continuamente (*play*), tocar o slide atual, ir para o próximo, ir para o anterior, saltar pra frente (*fast foward*), saltar pra trás (*rewind*), usar o índice (*index*) e parar (*stop*).

O trabalho caracteriza a interatividade dos usuários do sistema e o processo de

abertura das sessões por estes usuários.

Um *fitting* de distribuições foi feito para verificar quais distribuições melhor se adequam aos dados empíricos. Para estimar os parâmetros para as distribuições analisadas foi usado o Método dos Momentos [21].

O principal objetivo do trabalho foi caracterizar o acesso ao conteúdo do sistema MANIC, de forma a identificar os melhores métodos de alocação de recursos para acesso ao conteúdo disponibilizado. Um exemplo é definir estratégias de admissão, técnicas de compressão, gerenciamento de buffers, etc.

A massa de logs analisada contou com 215 arquivos. Os resultados encontrados mostraram um alto grau de interatividade ao longo das sessões. Uma vez que o áudio começa a ser tocado, em 45% dos casos o usuário pausa (paraliza a reprodução) ou salta na aula para uma posição distante em até 3 minutos. O tempo médio de exibição do áudio antes de um salto foi de 8 minutos. A distribuição encontrada para representar as métricas analisadas tem cauda mais longa que a distribuição exponencial, geralmente usada em trabalhos anteriores. As distribuições gamma e lognormal são as que melhor se adequam às distribuições empíricas encontradas, onde a exponencial junto a estas duas são as três opções analisadas.

Os saltos para frente nas aulas foram 7 vezes mais comuns que saltos para trás, o que não era esperado (esperava-se que os alunos realizassem mais saltos para trás para a revisão de conteúdos). Um número significativo de saltos possui curto período de tempo, aproximadamente um terço dos saltos para frente foram para posições distante a menos de 3 minutos. Entretanto, o comprimento médio dos saltos é muito maior, mais de dois mil segundos.

Com relação ao comprimento das sessões, o comprimento médio é menor que meia hora (28,85 minutos), e a grande maioria (mais de 80%) das sessões são menores que uma hora.

O autor enfatiza que as medidas apresentadas são de interesse independente pois descrevem um sistema, uma população e um conteúdo específicos, porém são bons valores limites para outros estudos futuros.

Foi observado que a característica de acesso dos alunos que estavam assistindo a aula pela primeira vez era diferente dos que estavam revisando as aulas que tinham visto em sala de aula.

Para um maior nível de detalhe, algumas das medidas analisadas foram:

- Comprimento da sessão;

- Interatividade dos usuários nas sessões (tempos de atividade e inatividade);
- Tamanho e localidade dos saltos.

Em relação ao comprimento das sessões, como o sistema não possui esquema de *logoff*, um *Session Gap Threshold* (SGT) foi definido. Dado que um período sem interatividade (inatividade) do usuário (SGT) expira, a sessão é dada como encerrada. Assim é possível medir o comprimento das sessões abertas no sistema. O comprimento médio das sessões foi de 1731 segundos, com desvio padrão de 2662. O comprimento médio das sessões com áudio foi de 2827,22 segundos e o das sessões sem áudio de 304,52 segundos. As distribuições gamma e lognormal com cauda mais longa que a exponencial se aproximaram mais da distribuição empírica do tamanho das sessões.

Quanto à interatividade do usuário, observou-se que a utilização média de áudio dentro das sessões foi de 67,59% com desvio padrão de 27,50. Isto porque as aulas podem ser assistidas apenas por slides (sem áudio), a critério do usuário. O áudio só é tocado quando iniciado pelo usuário. Porém, os valores a respeito da utilização de áudio correspondem apenas às sessões onde algum áudio foi tocado. Uma vez que 50% das sessões não apresentaram visualização de áudio, a utilização média do áudio condiderando todas as sessões foi de 37,50% com desvio padrão de 39,34. No trabalho sendo descrito, o período *ON* se caracteriza pelo uso de áudio e o período *OFF* pela visualização da aula sem áudio, ou seja, período de silêncio. Mais de 45% dos períodos *ON* são menores que 180 segundos, o tempo médio em ON é de 509,50 segundos e o desvio padrão de 815,92. O tempo médio dos períodos de silêncio (*OFF*) foi de 171,91 segundos com desvio padrão de 392,24. As distribuições gamma e lognormal foram as mais adequadas para representar os tempos em ON e OFF.

O tamanho dos saltos foi analisado por duas razões principalmente. Uma delas, se os saltos para trás foram curtos, é possível aproveitar os dados que já estão no buffer sem ter que requisitá-los novamente ao servidor. Além disso, se os saltos para frente também forem curtos, realizar *prefetching* de parte dos dados a frente dos que estão sendo exibidos poderia diminuir a latência de resposta ao cliente. O tamanho médio dos saltos para frente foi de 2137,62 segundos e o tamanho médio dos saltos para trás foi de 3394,84 segundos. Aproximadamente 34,37% dos saltos para frente e 33,33% dos saltos para trás são menores que 3 minutos. Logo, há uma quantidade limitada de localidade nos saltos.

Algumas conclusões apresentadas a respeito das medidas obtidas foram que a alta interatividade pode fazer com que o uso de prefetching cause overhead e traga

poucos benefícios, uma vez que os dados carregados podem ser pouco acessados, em contrapartida pode reduzir o tempo de resposta ao usuário, quando estes dados forem aproveitados. Duas surpresas, segundo o autor, foram encontradas. Uma delas é a pequena existência de saltos para trás. Outra é o tamanho expressivo dos saltos. Alguns usuários mostraram certa localidade nos saltos, o que encoraja, nestes casos, o uso de prefetching. A característica *ON-OFF* na interatividade com o áudio, pode beneficiar o uso de multiplexação para aumentar a escalabilidade do sistema, ou seja, enquanto usuários estão no estado *OFF*, outros podem ser atendidos.

3.1.2 Estudo do acesso aos servidores BIBS e eTeach

No trabalho [10] é apresentado um estudo baseado nos sistemas BIBS (*Berkeley Internet Broadcast System*) e *eTeach* [22], que oferecem cursos pela internet a estudantes de duas grandes universidades dos Estados Unidos. O objetivo é analisar a carga no servidor, características de acesso a arquivos e a interatividade dos usuários com o sistema. Com estes resultados é possível avaliar o uso de estratégias de cache e compartilhamento de banda para melhorar o desempenho do sistema.

Os resultados obtidos são usados como base para a geração de carga sintética, visando o estudo do ambiente em questão, como o projeto de sua arquitetura, recursos e largura de banda necessários e estratégias de *cache* em proxies, que poderiam melhorar o desempenho.

Ambos servidores possuem vídeos de alta qualidade. O *eTeach* entrega aulas e demonstrações de laboratório para o curso de Ciência da Computação, não havendo aulas em classe. O BIBS entrega várias aulas de cursos como Biologia, Química, Computação, etc.

Alguns dos principais resultados, dentre os diversos obtidos no trabalho, são:

- Alto grau de similaridade das medidas obtidas dos dois servidores.
- O processo de chegada de clientes no BIBS, em períodos aproximadamente estacionários, pode ser aproximado por um processo de Poisson, enquanto o eTeach possui um processo mais bem modelado por uma distribuição Pareto.
- Em períodos de estabilidade razoável, a freqüência de acesso aos arquivos pode ser caracterizada pela concatenação de duas distribuições Zipf.
- Uma fração significativa dos arquivos que são acessados, não são acessados novamente por mais de oito horas. Segundo o autor, este dado incentiva a revisão do uso de cache na primeira falta.

- No eTeach, os segmentos (partes dos arquivos) dos filmes mais populares são acessados igualmente. Porém, em filmes de menor popularidade o acesso é muito mais freqüente aos primeiros segmentos.

- A distribuição da mídia entregue por sessão depende do tamanho da mídia. No BIBS, para vídeos curtos, a distribuição Lognormal é a que melhor se ajusta, para vídeos mais longos, a distribuição encontrada foi um híbrido de Gamma com Pareto. No eTeach, sessões com vídeos curtos tem distribuição Exponencial e com vídeos mais longos tem uma distribuição Weibull ou Pareto.

- O acesso parcial a arquivos (ou seja, não visualizando sempre continuamente a mídia do início ao fim e sim com interações onde qualquer parte pode ser acessada a qualquer instante) pode diminuir a eficiência de transmissões em multicast, entretanto as simulações mostraram uma redução de largura de banda necessária para o servidor de 40% a 60% para os vídeos mais acessados em períodos de alta carga.

As características de interatividade dentro das sessões foram analisadas para o servidor eTeach, cujos registros possuíam este tipo de informação. Foi observado que 90% dos tempos OFF (inatividade) foram menores que 4 minutos. Uma pequena fração (menos de 0,5% das interações) de operações de *fast forward* e *rewind* foi encontrada. As distribuições encontradas para modelar o tempo em OFF foram Pareto e Exponencial para vídeos de duração entre zero e 5 minutos e Pareto, Lognormal e Weibull para vídeos com duração entre 5 e 35 minutos. As distribuições Exponencial, Pareto e Weibull apresentaram o melhor ajuste para o tempo em ON dos usuários do eTeach, sendo a primeira encontrada para vídeos de até 5 minutos de duração e as outras duas para vídeos entre 5 e 35 minutos.

3.1.3 Caracterização de acessos a servidores de áudio e vídeo educacionais e de entretenimento

Em [7] um estudo é feito com registros de dois servidores de áudio (rádio/UOL e ISP/áudio) e um servidor de vídeos de curta duração (TV/UOL), com conteúdo de entretenimento, além de estudar o servidor de conteúdo educacional eTeach já analisado em [10].

Diversas características de acesso dos usuários foram analisadas. Segundo os autores, algumas observações importantes sobre os resultados da análise são:

- Uma pequena parte dos vídeos são significativamente mais acessados, e a duração do vídeo está relacionada com a quantidade de acessos. A fração de acessos decresce com o aumento do tamanho do arquivo;

- As características de interatividade dos usuários varia conforme os tipos de mídia que são transmitidos;

- Interações de pausa são as mais comuns. As interações realizadas são dependentes da interação anterior realizada mas não do número de interações desde o início da sessão. Além disso, as interações são mais freqüentemente seguidas de interações do mesmo tipo;

- Uma forte localidade espacial foi observada na movimentação dos clientes durante as sessões;

- A freqüência de acesso aos arquivos pode ser bem modelada com a concatenação de duas distribuições do tipo Zipf para objetos de áudio e vídeo educacional, enquanto que uma única distribuição do tipo Zipf representa um bom modelo no caso de vídeos de entretenimento de curta duração;

- As distribuições Weibull e Lognormal representam bem a chegada de sessões para conteúdo educacional (*eTeach*), uma Pareto de cauda longa se mostrou um bom modelo para chegadas das sessões para um dos servidores de conteúdo de entretenimento (ISP/Áudio) e a Exponencial foi encontrada para os outros dois (TV/UOL e Radio/UOL).

- Aplicando o estudo na implementação de estratégias de cache, foi observado que alguns objetos são acessados raramente incentivando a utilização da popularidade das mídias como parâmetro para a definição de que objetos armazenar em *cache*. O acesso a seguimentos dos objetos se dá de forma diferente aos vídeos de acordo com sua popularidade, tipo e duração.

Variáveis de interatividade como posição inicialmente acessada de um vídeo, tempos de atividade e inatividade (ON/OFF) dos usuários nas sessões, freqüência das interações, distância dos saltos, entre outras, são estudadas e algumas caracterizadas e modeladas através de uma distribuição parametrizada que se aproxime bem da distribuição empírica da métrica. A duração dos objetos de mídia é considerada durante a caracterização e escolha dos modelos de distribuições, ou seja, a busca por distribuições é feita baseada no acesso aos objetos de forma separada para mídias mais curtas e de mais longa duração.

3.1.4 Caracterização de acessos a Vídeos sob a Web

Em [1], é feita a caracterização do acesso de usuários a vídeos na *World Wide Web* (WWW). O estudo de vídeo sob a web (VOW-*Vídeo on Web*) foi realizado na Universidade de Tecnologia de Lulea, na Suécia.

O sistema conta com alta largura de banda, ou seja, alta capacidade de transmissão. Isto permite que os usuários realizem a escolha de suas interações sem interferência do fator banda. Em muitos sistemas o fato de uma interação do usuário acarretar em uma latência até que a requisição seja atendida, devido ao fato dos recursos de rede não serem suficientes, pode interferir na escolha do usuário, que pensará mais antes de realizar uma interação. Os links disponíveis no sistema vão de 2 (dois) a 34 (trinta e quatro) megabits por segundo. O sistema de software usado é denominado mStar.

Para identificar estratégias de cache em proxies, sistemas de arquivo multimídia e servidores VOW a serem utilizados de modo a beneficiar o funcionamento do sistema, procurou-se responder as seguintes questões:

- O acesso aos vídeos possui localidade temporal?
- Com que freqüência os vídeos são acessados em comparação com documentos HTML?
- Existem padrões no acesso dos usuários aos vídeos?

A coleta de dados foi feita de agosto de 1997 a março de 1998. Foram medidos: tempo entre chegadas de pedidos, padrões de navegação no vídeo, localidade temporal dos acessos e tendências de tamanho dos arquivos.

Entre os resultados obtidos foi constatado um tempo entre chegadas de pedidos médio de 411 segundos. Segundo os autores, isto mostra que os acessos deste tipo são tão freqüentes quanto a documentos HTML.

Outra observação foi que 45% dos usuários param a exibição do vídeo antes dele terminar, ou seja, não assistem o vídeo até o final. Uma alta localidade temporal também foi percebida. Isto quer dizer que as interações possuem forte relação com o intante do vídeo que está sendo vizualizado.

O tamanho médio dos arquivos contidos no servidor era de 110 *MegaBytes*. A duração média dos arquivos era de 77 minutos.

A quantidade total de bytes de arquivos no servidor era de 15,7 *GigaBytes*. O servidor possuía arquivos de áudio e vídeo separados, mas para análise todos

foram tratados juntamente. Os arquivos mais comuns são os de 125 *MegaBytes*, e o tamanho médio dos arquivos é de 121 *MegaBytes*. Os arquivos eram várias ordens de grandeza maiores que os arquivos comuns presentes na *Web*.

A duração dos arquivos varia de 10 minutos a 2 horas. Os arquivos mais populares possuem de 90 a 100 minutos.

A taxa de bits, calculada dividindo o tamanho do vídeo por sua duração, na maioria dos casos era de 150 a 250 kbps, valores relativamente baixos. Os motivos do uso de tais valores são, garantir banda para acesso de países com acesso de baixa largura de banda e o uso do esquema de compressão H.261 que produz estas baixas taxas. O uso destas taxas possibilita também uma economia maior de espaço em disco.

As conclusões foram que, com estes vídeos com taxa mais baixa, foi observada alta localidade temporal nas interações e que o tipo de filme afeta o padrão de acesso do usuário (vídeo educacional tende a não ser continuamente visualizado até o fim pelo usuário, como ocorre nos vídeos de entretenimento em geral). Um alto grau de localidade temporal também foi detectado, sendo assim, o uso de cache possibilita uma melhora no desempenho do sistema. O acesso por título do vídeo não segue uma distribuição Zipf como outros documentos da *Web*. Houve sobrecarga de acesso a algumas poucas máquinas por estas possuírem alguns dos filmes mais populares.

3.2 Caracterização e modelagem

3.2.1 Caracterização e Modelagem de usuários de servidores de áudio e vídeo para entretenimento e educação

Um estudo complementar ao de [7] é feito em [17], onde é avaliado o impacto do comportamento interativo dos usuários de sistemas multimídia, tanto para entretenimento quanto para educação, no uso de protocolos de compartilhamento de banda. É proposto um modelo para geração de carga sintética baseado nos dados reais dos servidores de mídia de entretenimento (UOL) e de educação (eTeach e MANIC). O modelo é usado para avaliar e propor protocolos e compartilhamento de banda para usuários interativos.

3.2.2 Caracterização e Modelagem de usuários de um sistema para ensino a distância

O trabalho de [20] realiza um estudo detalhado do comportamento dos usuários do servidor de mídia educacional MANIC. Diferente do sistema MANIC avaliado em [15], onde era disponibilizado apenas conteúdo em áudio e slides, aqui o sistema já contava com vídeos sincronizados com os slides em HTML.

O principal objetivo do trabalho foi caracterizar variáveis que representassem o comportamento dos usuários e propor um modelo para geração de carga sintética.

As principais métricas analisadas, visando a construção de um modelo de comportamento foram:

- Tempo em *play* (ou tempo em ON): Período em que o vídeo é tocado continuamente;
- Tempo de OFF: composição de uma ou mais interações dos tipos *Pause, Index, FF e RW* seqüenciais;
- Posição inicial acessada: primeira posição da aula acessada pelo usuário no início da sessão;
- Tamanho dos saltos: Quantidade de vídeo avançada ou retrocedida em tempo através de uma interação FF, RW ou index;
- Tempo de permanência em um slide: tempo em que um slide é asssitido continuamente, independente do ponto do slide onde a exibição iniciou ou terminou.

Foram obtidas as distribuições de probabilidades que mais se adequam a representar cada uma das métricas descritas acima.

Os principais resultados com relação à análise do sistema foram:

- Há um forte correlação entre a posição do aluno na aula e a duração da requisição bem como entre a posição do aluno e o tamanho do salto;
- Quando ocorre uma mudança no conteúdo da aula, existe uma maior probabilidade de ocorrer uma interação do aluno;
- A distribuição Lognormal foi a que melhor modelou o tempo em OFF, *Pause, Index, FF e RW* e o tamanho dos saltos para trás. O tempo em ON e o tamanho dos saltos para frente podem ser aproximados pelas distribuições Gamma, Weibull e Lognormal;

- Em mais de 58% das sessões o primeiro slide acessado foi o primeiro da aula, os demais acessos iniciais se distribuem uniformemente entre as outras posições da aula;
- O tempo de permanência em um slide pode ser modelado por uma distribuição Exponencial, com média de 138 segundos.

Baseados no estudo do comportamento dos usuários do sistema, os autores propuseram um modelo para geração de carga sintética baseado em uma cadeia de Markov oculta (*hidden Markov model*).

3.2.3 Caracterização hierárquica para carga em servidores multimídia ao vivo

Em [23] é feita uma caracterização de forma hierárquica das variáveis aleatórias presentes em um sistema multimídia, que transmite ao vivo um popular "*Reality TV Show*" no Brasil através da internet. Este sistema se difere dos analisados nos outros trabalhos descritos por disponibilizar um conteúdo de entretenimento ao vivo, onde a evolução dos objetos sendo transmitidos é dirigida pelo servidor e não pelos usuários.

A organização hierárquica das métricas é composta por três camadas. A primeira camada é focada nos clientes. O interesse dos clientes, a concorrência no acesso ao sistema, o intervalo entre os acessos e a freqüência de acesso são algumas das características avaliadas. Na segunda camada (camada de sessão), são estudadas características como duração das sessões, tempo de inatividade entre as sessões, número de sessões abertas, entre outras. O último nível da hierarquia (camada de transferências), foca no comportamento dos usuários em cada sessão, como tempos de atividade e inatividade na sessão, utilização da banda nas transferências, entre outras.

Baseado na caracterização das variáveis estudadas, um modelo para geração de carga sintética foi construído e parametrizado. A ferramenta GISMO (*Generator of Internet Streaming Media Objects and workloads*) foi utilizada para geração da carga segundo o modelo construído.

3.2.4 Modelagem e caracterização de usuários do sistema VoD McIVER

No trabalho apresentado em [4], o sistema analisado para estudar as características de interatividade de clientes acessando um servidor multimídia foi o McIVER. O objetivo central do trabalho realizado foi criar um modelo e caracterizar as interações dos usuários com o sistema McIVER acessando um conteúdo educacional.

As distribuições encontradas para representar as variáveis aleatórias usadas na caractrização do comportamento dos usuários foram: a Exponencial para a duração das sessões; Exponencial e Lognormal (com melhor ajuste da Lognormal) para intervalos entre interações onde nenhum vídeo estava sendo tocado; e a distribuição Lognormal para durações de pausa, número de visualizações, play, *fast forward* e *rewind*.

O comportamento dos usuários é modelado, neste trabalho, usando distribuições Exponenciais para as métricas, apesar de ter sido observado que a distribuição Lognormal é um melhor ajuste para a grande maioria das métricas estudadas, isto foi feito para possibilitar uma análise mais simples através do uso de Cadeias de Markov.

Capítulo 4

Caracterização do Comportamento dos Usuários

NESTE capítulo são apresentados a caractetização das variáveis do sistema, estatísticas sobre o sistema e a metodologia para análise dos dados obtidos do sistema. Esta metodologia é composta das seguintes etapas:

1. Filtragem dos arquivos de log;
2. Definição das variáveis do sistema a serem analisadas;
3. Cálculo de parâmetros de diversas distribuições de probabilidade para cada uma das variáveis do sistema;
4. Escolha de uma ou mais distribuições que mais se adequam para representar as variáveis do sistema.

É feita também uma comparação com resultados de trabalhos relacionados existentes na literatura.

4.1 Metodologia

A metodologia consiste de quatro etapas conforme citado na introdução deste capítulo. Na primeira etapa, logs foram descartados segundo critérios que serão apresentados na seção 4.1.1.

A definição das variáveis ou métricas do sistema a serem analisadas encontra-se na seção 4.2.4. Na terceira, etapa algumas das variáveis do sistema serão selecionadas

e um conjunto de distribuições de probabilidade da literatura serão parametrizadas usando técnicas que serão descritas na seção 4.1.2. Após a parametrização, avaliaremos qual a distribuição mais adequada para representar uma determinada variável. Os métodos usados para esta análise estão descritos na seção 4.1.3.

Para analisar os logs e obter as estatísticas desejadas foram utilizados scripts na linguagem Perl. Estes scripts realizam a leitura dos arquivos de log e geram arquivos de amostras e resultados com as informações necessárias para a análise a ser realizada.

Em algumas ocasiões os scripts em perl possuem interação com scripts em outras linguagens como Matlab e Dataplot. Para aumentar a confiabilidade dos resultados, implementações equivalentes nas duas linguagens foram comparadas ou usadas em conjunto. Uma descrição detalhada da implementação e funcionamento dos scripts pode ser vista no apêndice A.

4.1.1 Filtragem dos arquivos de log

Antes de analisar os arquivos de log foi necessário realizar uma filtragem, para que fossem analisados apenas logs de sessões que sejam de nosso interesse. O primeiro passo foi descartar arquivos corrompidos e arquivos que possuíam algumas falhas no conteúdo, como por exemplo, falhas ocasionadas na escrita dos registros. Além disso, os logs são gerados em sessões abertas tanto por tutores que trabalham no projeto, quanto pelos alunos que cursam as disciplinas. Neste trabalho serão analisados apenas os logs dos alunos. Isto foi feito porque o comportamento de um aluno aprendendo um conteúdo através do sistema é diferente de um tutor conferindo tópicos, tirando dúvidas de um aluno, revisando uma aula, etc. Nosso interesse portanto está em como um usuário se comporta aprendendo diversos conteúdos através do sistema.

Outra filtragem também realizada foi a utilização apenas de arquivos de log das sessões em que o índice dos slides foi carregado corretamente. Se o índice dos slides é carregado com sucesso, o usuário pode navegar pela aula livremente através dos marcadores que levam a slides específicos que estão sincronizados com a vídeo-aula. Por outro lado, quando o índice não é carregado com sucesso, este tipo de interação não está disponível para o usuário, podendo este somente realizar operações como *fast forward/rewind*, saltos através da barra de progresso, *pause* e *stop*. Ou seja, neste último caso, o comportamento do usuário é diferente do primeiro, uma vez que em um dos casos só lhe é permitido navegar sobre o vídeo enquanto que em outro também é possível navegar pelo índice podendo pular diretamente para qualquer

tópico da aula.

No decorrer da análise foi observada uma falha no tempo registrado de algumas ações, em alguns arquivos. Este problema estava interferindo nas medidas e gerando problemas como tempos negativos para algumas métricas, tempos que ultrapassavam um limite físico imposto pela duração dos vídeos entre outros. Para resolver o problema, tentamos identificar qual a porcentagem dos arquivos estava invalidada pelo problema, como o número destes arquivos invalidados representava menos de 30 % (trinta por cento) dos logs que foram selecionados para análise, estes logs foram descartados, uma vez que a massa de logs continuou suficientemente grande.

Com os arquivos já selecionados, na primeira análise realizada a respeito da duração das sessões, observamos um elevado número de sessões de curta duração. Mais de 45% das sessões duram menos que 5 minutos, como pode ser claramente visto em um pico inicial que se destaca na Figura 4.1. Acreditamos que 5 minutos seja um tempo muito pequeno para que um aluno se concentre na aula durante uma sessão. Com isso, o comportamento dos alunos nestas sessões, pode fugir do padrão de características presentes em sessões em que o aluno assiste às aulas calmamente e interessado no que está sendo exibido. Sessões onde o aluno entrou, viu do que se tratava a aula e não assistiu o material com interesse de aprender, não são de nosso interesse e podem prejudicar a análise que faremos buscando avaliar o comportamento dos alunos em aprendizado e encontrar padrões neste comportamento. Portanto, toda análise presente neste capítulo será baseada somente em sessões com duração acima de 5 minutos. Apesar desta escolha, serão feitas algumas análises comparativas incluindo as sessões abaixo de 5 minutos.

Para determinar se o uso de 5 minutos como limite para separação das sessões para análise é satisfatório, foram feitas comparações de estatísticas encontradas utilizando este valor e outros valores limites como 20, 30, 40 e 50 minutos, valores entre a média de duração dos vídeos (cerca de 50 minutos) e de duração das sessões (cerca de 20 minutos quando todas as sessões são observadas).

Após a filtragem descrita, realizada para uma melhor análise do comportamento dos usuários do sistema, 1452 registros de sessões, gravados nos dois semestres letivos do ano de 2005 do projeto CEDERJ, foram separados para análise.

4.1.2 Métodos usados para parametrização de distribuições

Para cada uma das métricas selecionadas, uma parametrização utilizando o método *Maximum Likehood Estimate* ([21, 13]) foi feita para cinco distribuições bem conhecidas, são elas: Exponencial, Gamma, Normal, Lognormal e Weibull [21]. O

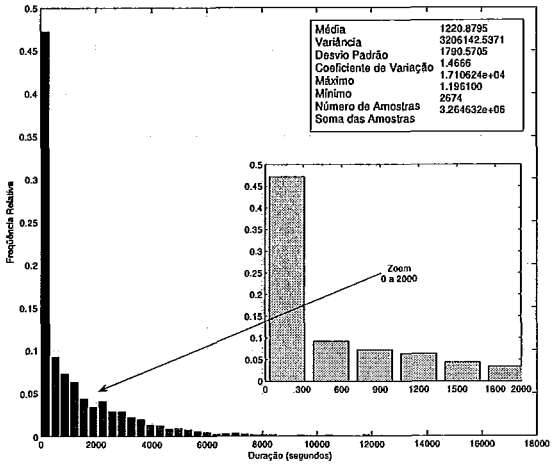


Figura 4.1: Duração das sessões abertas pelos usuários.

MLE permite calcular os parâmetros de uma distribuição conhecida, baseado na probabilidade máxima de ocorrência das amostras empíricas encontradas. Os parâmetros calculados serão usados para a geração das curvas destas distribuições, a fim de compará-las com a curva de distribuição dos dados empíricos. Para métricas com variabilidade mais alta (coeficiente de variação maior que 1), a distribuição Hiperexponencial também será parametrizada e usada na comparação. Para parametrizar a Hiperexponencial uma implementação de algoritmos EM (*Expectation Maximization*) interativos foi utilizada, o programa usado foi o EMpht [14]. Este programa permite parametrizar, dentre outras distribuições compostas de estágios (ou fases), a Hiperexponencial com vários números de estágios.

A Tabela 4.1 apresenta a nomenclatura que será utilizada nos parâmetros das distribuições. No caso da hiperexponencial, os vetores α de probabilidades α_i e λ de taxas λ_i das exponenciais que compõem a hiperexponencial, para i variando de 1 ao número de estágios, serão apresentados.

Distribuição	Parâmetro 1	Parâmetro 2
Gamma	γ = *shape*	k = escala
Weibull	γ = *shape*	k = escala
Exponencial	μ = média	
Lognormal	Θ = posição	k = escala
Normal	Θ = posição	k = escala

Tabela 4.1: Descrição dos parâmetros utilizados.

4.1.3 Métodos usados para escolha de distribuições

Para escolher que distribuição, dentre as opções estudadas, deve ser usada para representar cada métrica, uma análise bastante criteriosa foi feita. A primeira análise é gráfica e tem como objetivos avaliar a cauda da distribuição empírica e verificar qual distribuição parametrizada possui a curva que mais se aproxima da curva da distribuição empírica. Os gráficos da distribuição cumulativa complementar de probabilidade (CCDF) empírica e parametrizadas são plotados com o eixo $Y(1 - F(x))$ em escala logarítmica para evidenciar a cauda das distribuições. A importância de se analisar o comportamento da cauda das distribuições está no impacto desta característica na escalabilidade do sistema. Um exemplo são os casos onde a grande ocorrência de amostras de valores altos para variáveis que representam a geração contínua de carga para o sistema reduzem a escalabilidade do sistema, ou seja, quando mais tempo os usuários passam ativos durante a visualizacão das aulas menor o número de usuários que o sistema é capaz de suportar simultaneamente. O corpo da distribuição possui um interesse maior quando o interesse está em analisar características como tempo entre chegadas de requisições. Os resultados da análise sobre estes dois pontos de vista impactam de formas diferentes na implementação e sucesso de técnicas das compartilhamento de recursos.

De forma a complementar a análise gráfica uma análise quantitativa será realizada. Para cada métrica o erro quadrático médio (MSE) de cada distribuição parametrizada em relação à distribuição empírica é calculado. O MSE nada mais é do que o cálculo da média das diferenças entre os pontos de duas curvas, elevadas ao quadrado. A fórmula para realização deste cálculo é a seguinte:

$$MSE = E\left[(X' - X)^2\right]$$

Nesta equação X' representa os pontos da curva empírica e X os pontos da curva da distribuição parametrizada. Quanto menor for o MSE, mais próxima uma curva está da outra, em média.

Outro método utilizado é o teste de Kolmogorov-Smirnov [21]. Este teste aceita a hipótese H_0 de os dados empíricos serem gerados a partir da distribuição parametrizada que está sendo testada, se a estatística D_{max} (que chamaremos de KSSTAT) calculada no teste for menor ou igual a um valor crítico, proveniente de uma tabela de valores pré-definida para o teste. O valor de D_{max} corresponde à distância máxima encontrada entre as curvas que estão sendo avaliadas (empírica e parametrizada). A sensibilidade do teste é definida pelo nível de significância α, ou seja, a probabilidade de o teste rejeitar a hipótese nula erroneamente. Neste trabalho

será usado um nível de significância igual a 0,10. Quando o teste rejeita a hipótese nula não se pode afirmar que os dados empíricos vêm da distribuição parametrizada. Além disso, quanto menor o valor de D_{max}, mais próxima a distribuição empírica está da parametrizada. Portanto, caso a distribuição sendo testada seja rejeitada pelo teste, usaremos do teste o parâmetro KSSTAT para identificar a distribuição cuja curva apresenta a menor distância máxima em relação à curva empírica.

Para finalizar, o teste gráfico QQPlot [13] é usado para mostrar se a distribuição parametrizada escolhida realmente é um bom modelo para os dados sendo analisados. Este mostra se dois conjuntos de amostras vêm de uma mesma população, ou seja, possuem distribuições provenientes de uma mesma "família". Inicialmente são calculados os quantiles usando dois conjuntos de amostras: o conjunto de amostras empíricas e o de amostras geradas segundo uma distribuição parametrizada. Um quantile é a fração de amostras menores que um dado valor. No QQPlot são representados os valores dos quantiles obtidos para ambos os conjuntos de amostras. Se os dois conjuntos de amostras são provenientes de uma mesma distribuição, os pontos do gráfico devem formar, aproximadamente, uma reta com inclinação de 45 graus.

Para identificar que número de estágios são necessários para que a Hiperexponencial se ajuste bem aos dados, serão feitas para cada métrica, quatro parametrizações com 2, 4, 6 e 8 estágios. Serão calculados o MSE e será aplicado o teste de duas amostras de Kolmogorov-Smirnov entre cada parametrização e a curva empírica . O teste de duas amostras de Kolmogorov-Smirnov é assim chamado quando o teste é aplicado em duas amostras e não entre uma amostra real e uma distribuição parametrizada. O número de estágios que alcançar os melhores resultados nos testes será utilizado na Hiperexponencial.

Com esses métodos, que permitem a análise visual da cauda das distribuições (gráficos da distribuição complementar), calculam a distância quadrática média entre as curvas da distribuição (MSE), computam os pontos mais distantes entre elas (teste de Kolmogorov-Smirnov), e permitem avaliar se dois conjuntos de amostras pertencem a uma mesma distribuição (QQplot), acreditamos que podemos obter resultados bastantes confiáveis. Uma análise tão criteriosa quanto essa não é observada, do nosso conhecimento, em nenhum dos trabalhos relacionados presentes na literatura.

4.2 Estatísticas e Resultados

Nesta seção serão apresentados os resultados obtidos da análise do comportamento dos usuários do sistema através dos arquivos de log disponíveis.

Os resultados serão classificados em três categorias para uma melhor organização. O primeiro nível da hierarquia, que chamaremos "Interesse dos usuários" apresenta informações sobre o acesso dos alunos às aulas disponibilizadas, algumas métricas como o número de sessões abertas, qual a aula e a disciplina com maior número de acessos, quais os vídeos mais acessados baseado em sua duração, etc. Num segundo nível (Sessão) serão apresentados dados sobre cada uma das sessões abertas como duração, tempo de atividade e inatividade do usuário, número de requisições geradas ao servidor, número médio de interações e características das interações. Por fim, o nível de maior importância para nosso trabalho é o chamado de "Interatividade do cliente", neste estão os tempos de duração de cada ação do usuário, o tempo em que o mesmo fica em play contínuo, o tamanho dos saltos realizados, posição inicial acessada, entre outros. É neste terceiro nível da hierarquia que o trabalho estará mais concentrado e detalhado, pois as variáveis avaliadas serão usadas em um modelo para geração de carga sintética [24].

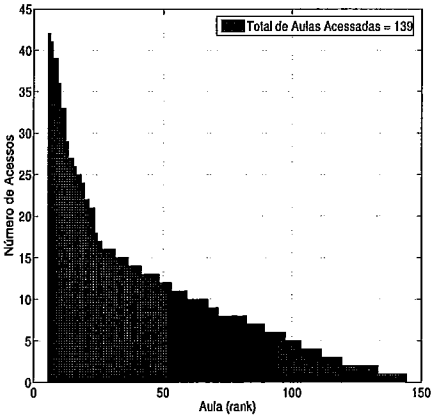
Na seção 4.2.5, para cada métrica serão apresentados resultados comparativos do pólo de Piraí com os pólos de Volta Redonda e Três Rios. Analisaremos o pólo de Piraí em separado para verificar se há alguma mudança no comportamento dos usuários entre estes pólos. Isto foi feito porque este pólo contém mais alunos que moram na cidade do Rio de Janeiro, ou seja, o fato de terem que viajar até Piraí para assistir às aulas pode gerar uma mudança no comportamento além de estes poderem dar preferência ao uso de DVD's ao invés de irem ao pólo.

4.2.1 Estatísticas do Sistema

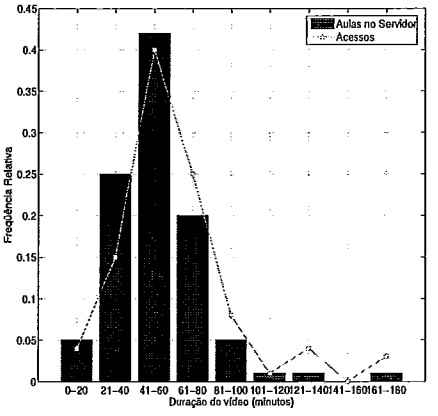
Foram armazenados 60 GigaBytes de aulas no servidor, composto por um total de 146 aulas. O tamanho das aulas varia de 93 MegaBytes a 1 GigaByte. O tamanho médio das aulas armazenadas é de 415 MegaBytes.

Os vídeos possuem duração desde 11:36 minutos a 2:52:21 horas, com média de aproximadamente 50 minutos. A taxa de bits média dos vídeos é de 1,1 Mbps, com vídeos codificados de 0,56 Mbps a 1,33 Mbps, sendo a grande maioria (principalmente os mais recentes) em 1,1 Mbps.

O curso disponibilizado contou, no ano de 2005, com oito disciplinas. As disci-



(a) por ranking



(b) pela duração dos vídeos

Figura 4.2: Acessos às aulas disponibilizadas.

plinas, identificadas por um código formado por "ead"[ano][número da disciplina], possuem o número de aulas mostrado na Tabela 4.2.

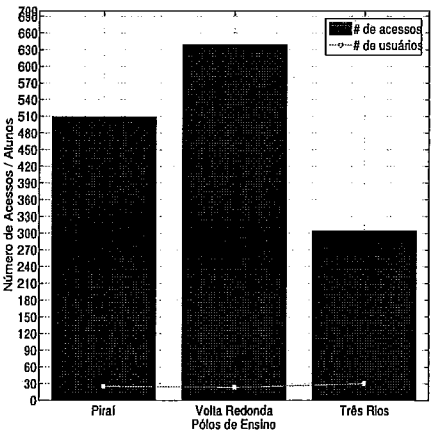
Disciplina	Código	Número de Aulas
Introdução à informática	ead05001	13
Construção de página web	ead05002	20
Fundamentos de algoritmos para computação	ead05004	24
Projeto e desenvolvimento de algoritmos	ead05005	15
Álgebra linear	ead05006	15
Estrutura de dados	ead05007	36
Matemática para computação	ead05008	9
Fundamentos de programação	ead05009	14

Tabela 4.2: Aulas das disciplinas disponíveis

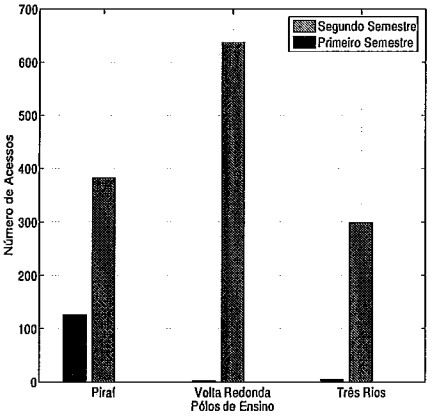
4.2.2 Interesse do Usuário

A questão de maior interesse aqui é que objetos e partes destes objetos são os mais acessados pelos usuários. Na Figura 4.2 temos em (a) a popularidade das aulas baseado em um *ranking* decrescente do número de acessos, em (b) que aulas são mais assistidas pelos alunos de acordo com sua duração.

Apenas 139 das 146 aulas foram acessadas, ou seja, 7 aulas não foram acessadas nem uma única vez. O gráfico 4.2(b) exibe também a percentagem de aulas para



(a) por pólo



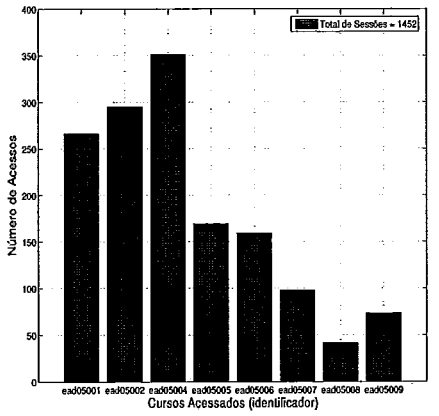
(b) para cada semestre de 2005

Figura 4.3: Sessões abertas nos pólos de ensino.

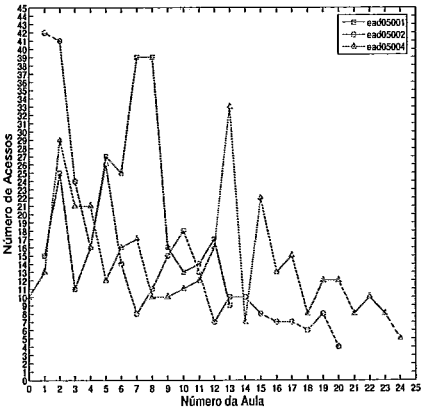
cada faixa de duração presentes no servidor. Mais de 40% das aulas possuem duração de 40 a 60 minutos, estas são também as aulas com maior percentagem dos acessos. As aulas mais curtas e as de longa duração (maiores que 80 minutos), são menos numerosas no servidor e possuem uma percentagem menor de acessos.

Em [10], os vídeos mais populares do servidor BIBS são os de 50 a 55 minutos de duração, que receberam 70% das requisições feitas ao servidor. No eTeach, cerca de 25% a 30% dos acessos foram para vídeos de curta duração (até 5 minutos), mais de 20% dos acessos foram para vídeos de 30 a 35 minutos, sendo que os vídeos de até 5 minutos são bem mais numerosos no servidor que os de duração de 30-35 minutos. Os vídeos mais curtos são também os mais populares em [6], onde 78% dos vídeos assistidos foram acessados apenas uma vez.

O número de alunos matriculados em cada um dos três pólos de ensino é de aproximadamente 30 alunos. Apesar de o pólo de Volta Redonda contar com o menor número de alunos (25 matriculados), foi o pólo com o maior número de acessos às aulas do curso, seguido por Piraí e depois por Três Rios (ver Figura 4.3(a)). Em todos os três pólos de ensino, o número de acessos aos servidores foi consideravelmente menor no primeiro semestre de curso, podemos ver no gráfico 4.3(b) que apenas Piraí recebeu um número razoável de acessos neste período (maior que 100), nos outros pólos quase todos os acessos foram feitos no segundo semestre de curso. Alguns problemas técnicos foram encontrados neste período, principalmente por este ser o primeiro período de curso, reduzindo o número de arquivos de logs. Além disso, novas turmas ingressaram no curso no segundo período, fazendo com



(a) para cada curso



(b) para cada aulas dos cursos mais acessados

Figura 4.4: Acessos às aulas de cada curso.

que o número de acessos aumentasse significativamente.

As disciplinas introdutórias do curso são as mais acessadas. O comportamento evidenciado na Figura 4.4(a) mostra as disciplinas "Introdução à informática", "Construção de página web" e "Fundamentos de algoritmos para computação" com um número de acessos cerca de duas vezes maior (em média) que a quarta disciplina mais acessada "Projeto e desenvolvimento de algoritmos". As quatro disciplinas citadas pertencem ao primeiro semestre de curso e mesmo com a grande maioria de acessos tendo sido realizada no segundo semestre de curso, as disciplinas introdutórias são as mais estudadas pelos alunos. A já mencionada entrada de alunos no segundo período talvez seja a explicação para este fato, além de haver reprovações de alguns alunos, fazendo com que acessem aulas das disciplinas em que foram reprovados.

Em cada uma das três disciplinas mais acessadas, as aulas mais assistidas são também as primeiras de cada curso. No gráfico da Figura 4.4(b) fazemos uma comparação entre os acessos dos alunos para cada aula destas três disciplinas. Alguns picos de acesso podem ser vistos em algumas aulas de numeração intermediária, as últimas aulas são consideravelmente menos vistas. Este fato é semelhante ao comportamento dos alunos de cursos presenciais onde a freqüência gradativamente diminui com o avanço do semestre.

A Figura 4.5 é um histograma do número de slides existentes em cada aula, são apresentados dados apenas das aulas que receberam acessos dos alunos. A maior

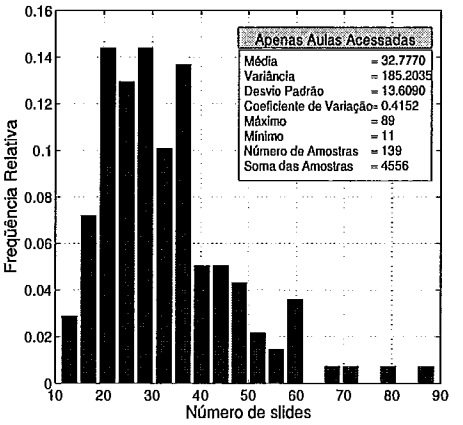


Figura 4.5: Número de slides das aulas.

parte das aulas tem de 20 a 40 slides. As aulas possuem em média 32 slides. Outros dados sobre o número de slides por aula são mostrados nesta mesma figura.

Com os resultados encontrados nesta seção, vimos que há um maior número de acessos dos alunos aos conteúdos introdutórios. Este comportamento se dá por diversos fatores e se assemelha ao que costuma ser observado em cursos presenciais.

4.2.3 Sessão

A primeira variável a ser analisada é a duração da sessão, que corresponde ao tempo entre o início da visualização de uma aula e o término desta visualização. Do histograma da Figura 4.6 podemos concluir que a grande maioria das sessões são de até 1:30 horas aproximadamente, a média é de um pouco mais de 36 minutos e existem algumas raras ocorrências de sessões de até aproximadamente 4 horas.

Em [15] a duração média das sessões foi de 1731 segundos (cerca de 28 minutos), com 2662 segundos de desvio padrão, a média das sessões com áudio foi 2827,22 segundos (47 minutos) e das sessões sem áudio 304,52 segundos (5 minutos). As sessões analisadas em [20] tiveram duração média de aproximadamente 2278 segundos (38 minutos) com coeficiente de variação 0,856. Já as sessões analisadas em [6] são na maioria de curta duração (duração média de 2,2 minutos), 85% das sessões, segundo o autor, são menores que 5 minutos. A duração média das sessões analisadas em [4] foi de 1095,1 (18 minutos) segundos com desvio padrão 999,63 segundos, sendo que as sessões de menor duração foram também as mais populares.

Os alunos ficam em média cerca de 51% da sessão ativos, ou seja, assistindo o vídeo e cerca de 49% do tempo inativos (ver figura 4.7(a)). As interações dos alunos

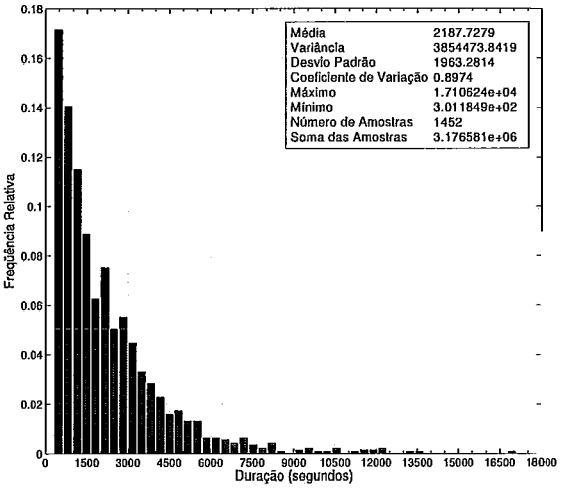
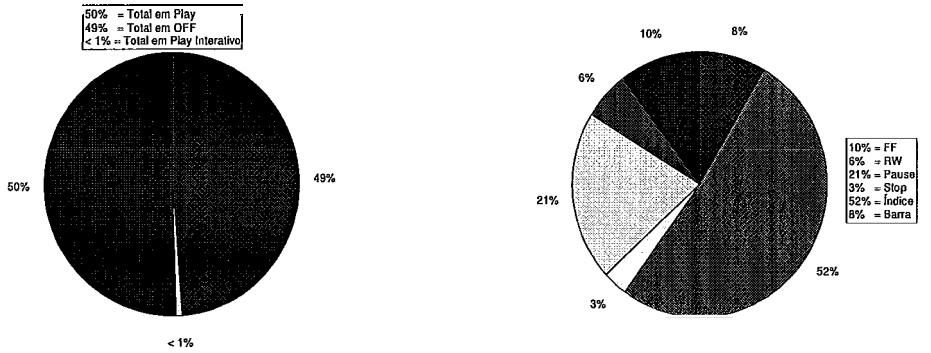
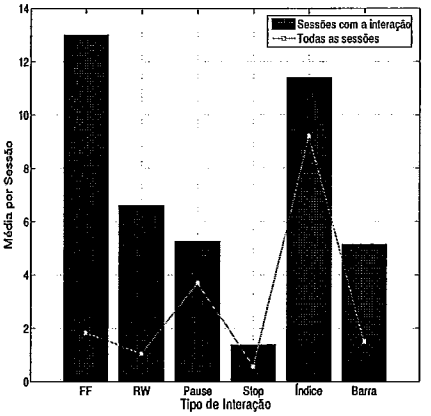


Figura 4.6: Duração das sessões.

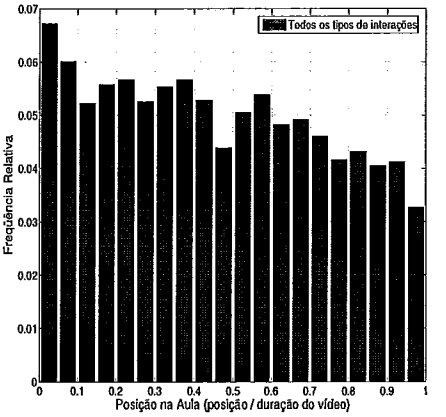


(a) fração de atividade e inatividade (b) ocorrência das interações

Figura 4.7: Comportamento dos usuários em uma sessão.



(a) ocorrências das interações



(b) posição onde ocorrem as interações

Figura 4.8: Interações em uma sessão.

com o vídeo através do índice e as interações de *pause* são, de acordo com a Figura 4.7(b), as interações mais comuns, sendo que a percentagem de acesso ao índice é consideravelmente maior que qualquer outra interação. Os avanços através do botão *fast forward* são mais freqüentes que os retrocessos através de *fast rewind*. As outras interações (stop e movimentação da barra de progresso) são menos comuns em uma sessão.

Em [10] também foi observada uma pequena quantidade de interações *fast forward* e *rewind*, uma vez que o sistema permite a utilização de marcadores que apontam mais claramente para partes de conteúdo. Interações de pausa são também mais freqüentes do que as do tipo *fast forward* e *rewind* em [4].

A Figura 4.8(a) apresenta dados sobre a ocorrência das interações, onde são mostradas as médias de ocorrência de cada tipo de interação em uma sessão condicionada ao subconjunto de sessões em que o tipo de interação em questão ocorreu. Também são mostradas as médias descondicionadas, ou seja, considerando o número de interações (de cada tipo) igual a zero nas sessões em que estas não ocorreram. É possível perceber a diferença na média quando um certo tipo de interação não costuma ocorrer em um grande número de sessões.

Para ter uma idéia de em que posições do vídeo estas interações costuma acontecer com mais freqüência, um gráfico com o histograma de ocorrência de interações baseado na posição em que ocorreu é mostrado na Figura 4.8(b). Por este histograma podemos ver que as interações se distribuem pelas posições do vídeo de forma quase uniforme, sendo que o início da aula parece ser um ponto de interação um

pouco mais freqüente e no fim as interações aparecerem um pouco menos.

Nesta seção vimos que as sessões de curta e média duração (de duração até cerca de uma hora e meia) ocorrem com maior freqüência que sessões de longa duração (com duração entre 1:30 hrs e mais de 4 horas). A duração média de uma sessão (36 minutos aproximadamente) é próxima a duração média de uma aula que é de aproximadamente 50 minutos. Os alunos passam mais tempo assistindo o vídeo do que parados em algum ponto dele. Interações através do índice são mais comuns que todas as outras interações. A facilidade em atingir a parte desejada da aula através deste tipo de interação deve ser a causa desta preferência por parte dos usuários.

4.2.4 Interatividade dos alunos

Nesta seção serão apresentadas e analisadas todas as características relacionadas à interatividade dos alunos. Estudaremos diversas variáveis e a relação entre elas, e tentaremos representar cada variável com um único modelo de distribuição visando a construção de um modelo de usuário para geração de carga sintética. As principais variáveis que serão analisadas aqui são:

- Tempo em ON (*Play* e *Play Interativo*)
- Tempo em OFF (*stop* e *pause*
- Tamanho dos saltos (para frente e para trás)
- Posição inicial acessada
- Tempo de permanência em um slide

Para entender melhor cada uma destas variáveis, que foram fundamentais na construção do modelo de usuário para geração de carga sintética, a Figura 4.9 apresenta um esquema exemplificando a evolução de uma sessão. Cada uma das variáveis pode ser facilmente percebida de acordo com a descrição que é apresentada a seguir.

O tempo em ON corresponde ao tempo em que o usuário se encontra ativo na sessão solicitando continuamente blocos de vídeo ao servidor. O tempo em ON foi dividido em duas categorias que chamaremos "Play"e "Play interativo". Esta divisão se deve ao fato de observarmos a ocorrência de diversos intervalos em *play* de curta duração entre seqüências de saltos consecutivos dos usuários, estes pequenos tempos entre saltos correspondem ao *play* interativo e os tempos em play restantes

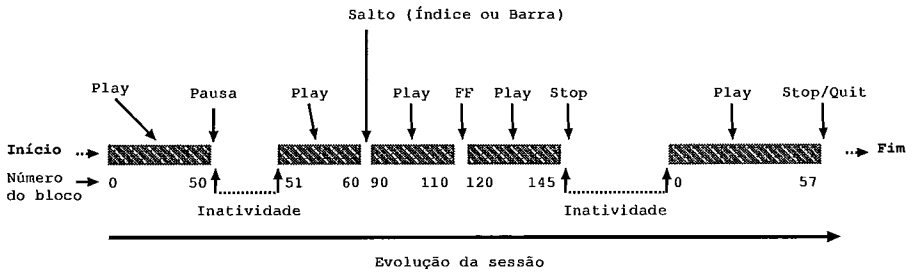


Figura 4.9: Exemplo de evolução de uma sessão (variáveis de interatividade).

são o que chamamos tempo em *play*. Maiores detalhes sobre a divisão da métrica tempo em ON nas duas métricas mencionadas acima serão vistos mais adiante.

O tempo em OFF, em contraposição ao tempo em ON, corresponde aos intervalos de tempo em que o usuário permanece inativo em uma sessão, não realizando pedidos ao servidor. O tempo em OFF é composto por intervalos de tempo em *pause* e *stop*, que são tempos gastos pelo usuário com o vídeo paralisado ou parado, respectivamente. As amostras de *pause* e *stop* serão também analisadas separadamente.

No sistema o usuário pode realizar diversos tipos de saltos para frente ou para trás no vídeo. Analisaremos então o tamanho (distância) dos saltos realizados pelo usuário em ambos os sentidos. A posição inicial acessada é uma estimativa de que posição do vídeo o usuário (aluno) acessa assim que inicia a sessão, ou seja, a primeira parte da aula escolhida por ele para iniciar o aprendizado. Serão estudados também os tempos gastos pelo aluno em cada slide, que serão chamados de tempo de permanência em slides.

Para algumas métricas foram observadas certa correlação com outra variável do sistema. A posição onde se inicia a contabilização da métrica pode limitar o valor que a métrica pode assumir. Por exemplo, se um usuário está na posição x de um vídeo de tamanho y, o tempo que este usuário pode permanecer em *play* contínuo está limitado ao tempo residual $(y - x)$ do vídeo. Portanto, para cada métrica serão plotados gráficos das posições correspondentes a cada amostra versus o valor da amostra, afim de avaliar esta correlação, quando esta existir.

Tempo em ON

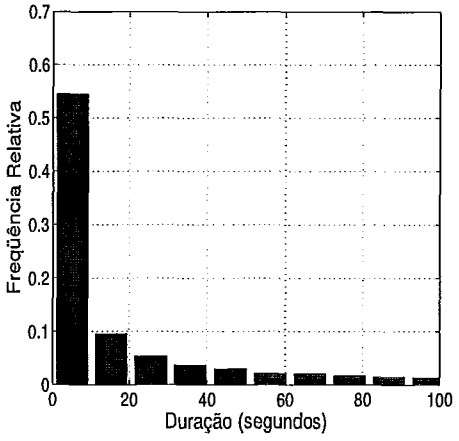
No decorrer da análise do tempo em ON, algumas características importantes foram observadas. Durante as sessões, grande parte dos usuários, muito freqüentemente, realizavam vários saltos consecutivos, sejam eles saltos pela barra de pro-

gresso, pelo índice ou usando os botões *fast forward* e *fast rewind*. Entre estes saltos, em grande parte das vezes, o tempo em ON é relativamente pequeno. Estas seqüências de saltos em intervalos curtos de tempo, sugerem que o usuário está procurando pelo ponto da aula que é de seu interesse. Como o usuário nestas situações não está assistindo o vídeo continuamente concentrado no conteúdo, resolvemos separar estes intervalos em ON da situação onde o vídeo está sendo tocado e visto pelo usuário em aprendizado.

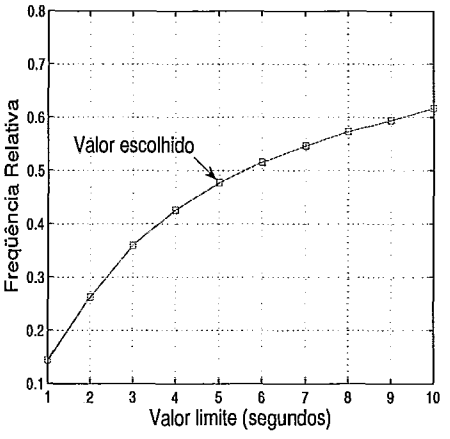
Uma outra situação que chamou a atenção foi que, quando o usuário pausa (*pause*) ou pára (*stop*) a visualização, o cliente não permite que o usuário realize saltos, permitindo somente que ele pressione o botão *play* para retomar a reprodução do vídeo. Caso o usuário deseje neste ponto realizar um salto, é necessário que antes pressione o botão *play* para em seguida poder realizar o salto. Neste caso, se o intervalo entre o uso do botão *play* e o salto for muito curto, pode indicar que o aluno não estava interessado na parte da aula que seria tocada e sim foi forçado a pressionar o botão *play* para poder realizar o salto.

A melhor forma de analisar estas situações cuidadosamente foi separá-las em uma categoria própria que chamaremos de "play interativo", diferente dos momentos em "play normal". Sendo assim, a métrica tempo em *play* interativo será composta de intervalos em *play* de curta duração entre saltos consecutivos ou entre uma pausa seguida de um salto. O tempo em *play* corresponderá aos tempos que não se enquadram nesta situação, fazendo com que estes tempos retratem melhor os momentos em que o usuário está assistindo realmente o vídeo escolhido.

Para que esta divisão entre as duas categorias possa ser feita, faz-se necessária a escolha de um valor limite (*threshold*) para os tempos entre saltos e entre pause/salto. Um histograma de todas as amostras de ON encontradas, truncado em 100 segundos, é mostrado na Figura 4.10(a), onde mais da metade das amostras é menor do que 10 segundos. Para escolher um valor de *threshold* foram analisadas as amostras de ON que estão entre saltos ou entre pause/salto. Um gráfico com a fração de amostras menores que vários valores limites é apresentado na Figura 4.10(b). Para o valor limite de 5 segundos, tem-se que a fração de amostras selecionada seria quase a metade. Por este valor ser razoável para que o aluno observe uma pequena parte do vídeo, não se interesse e siga procurando pela parte de seu interesse, este será o valor escolhido como limite entre as duas classes mencionadas.



(a) Histograma de todas as amostras de *play*.



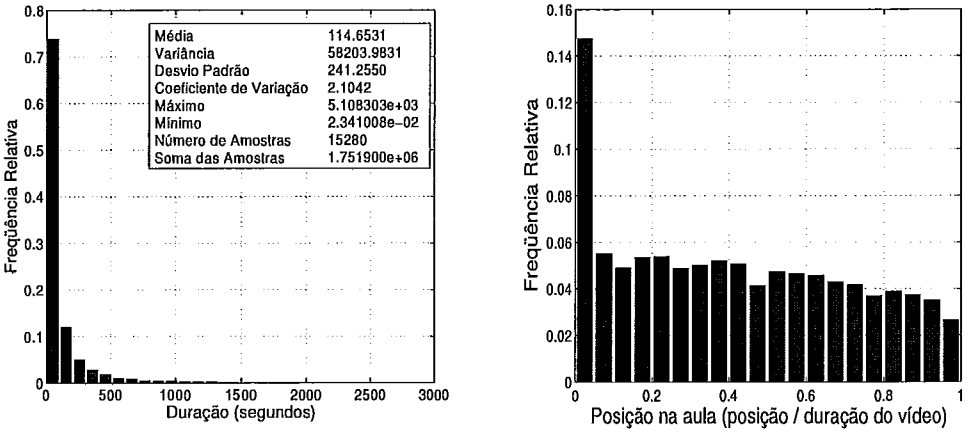
(b) Fração de amostras entre saltos e pause/salto menores que o valor limite (*threshold*) indicado no eixo x.

Figura 4.10: Escolha do *threshold*.

Play

A maior parte dos tempos em *play* é de curta duração, 84% dos tempos são menores que 180 segundos (3 minutos). As Figuras 4.11(a) e 4.11(b) mostram um resumo sobre as amostras encontradas e onde ocorreram na aula acessada. O início da aula é o local de maior ocorrência dos tempos em *play*, o resto das amostras se distribui bem entre as outras posições do vídeo. A predominância de valores pequenos indicam uma alta interatividade dos alunos com o vídeo, característica também percebida através do alto número de interações por sessão apresentados na Seção 4.2.3. Este alto grau de interatividade pode estar relacionado ao fato de o conteúdo das aulas ser altamente interativo, com animações, exemplos, exercícios, etc. Além disso, o aluno é encorajado a estudar pelo livro texto e pode ficar alternando entre a leitura do livro e a verificação do ponto da aula no vídeo correspondente ao trecho do livro. Finalmente, pode acontecer de os alunos não terem tempo de assistir todas as aulas e assistirem apenas partes que acreditem ser de maior relevância.

O tempo médio em play encontrado foi de 114,65 segundos (1,9 minutos), com coeficiente de variação 2,1. A Figura 4.12(a) apresenta uma comparação entre os valores médios e de desvio padrão das amostras encontrados, caso sejam usados filtros de sessão nos valores de 20, 30, 40 e 50 minutos, ou ainda se nenhum filtro for usado. No eixo das abcissas do gráfico da Figura 4.12(a) $s > x$ representa que foram consideradas sessões maiores que x minutos, por exemplo, se nenhum filtro



(a) Histograma

(b) Posição onde ocorreram

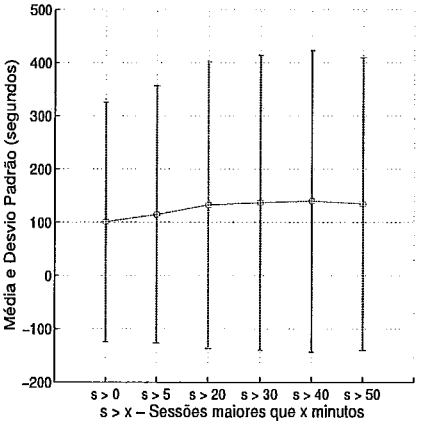
Figura 4.11: Tempo em *play*.

foi usado temos $s > 0$. O filtro de 5 minutos parece ter sido uma boa escolha, uma vez que a média do tempo em *play* não fica muito maior ou menor que os valores obtidos para outros filtros.

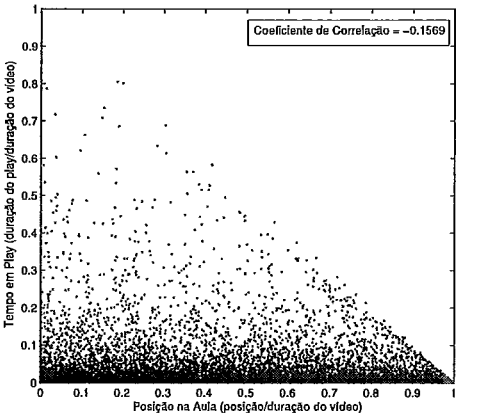
Como já era previsto, existe uma certa correlação entre o tempo do usuário em *play* com a posição em que o play foi iniciado. Na Figura 4.12(b) os pontos do gráfico se concentram sobre e abaixo de uma linha que indica o limite em tempo residual que o usuário tinha disponível a partir da posição que estava no vídeo. O sentido da linha de referência reflete um coeficiente de correlação negativo de -0.1569, que indica que as variáveis estão inversamente correlacionadas. Esta mesma relação foi mostrada em [20], onde o coeficiente de correlação encontrado foi de -0,1968.

Para fins de comparação temos que estes valores mostram que nossos usuários são muito mais interativos que os de outros sistemas analisados em outros trabalhos. Os valores da métrica tempo em play encontrados, mesmo tendo sido filtradas as sessões menores que 5 minutos e tendo sido criada uma métrica complementar que absorve as amostras de play menores que 5 segundos entre interações (Play interativo), foram consideravelmente menores que em outros trabalhos.

Analisando o tempo em play ponderado pela duração do vídeo que estava sendo assistido, para ter uma idéia da fração que estes tempos representam dentro de um vídeo, temos que o tempo em play é em média cerca de 3,5% de um vídeo. Estes valores indicam que o usuário passa mais tempo interagindo com o vídeo e menos tempo assistindo a aula continuamente que nos sistemas estudados anteriormente.



(a) Comparação com outros filtros de sessão

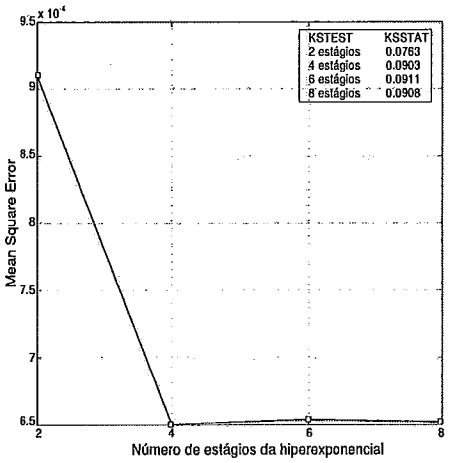


(b) Análise da correlação com a posição do vídeo

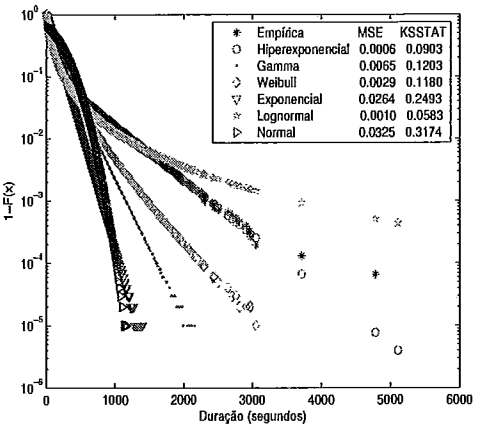
Figura 4.12: Filtros de sessão e gráfico de correlação (Tempo em *play*).

Procederemos agora com o fitting das amostras de tempo em play com as cinco distribuições já citadas, para identificar qual delas é o melhor modelo para representar o tempo em play. Como o coeficiente de variação das amostras é relativamente alto, faremos também a parametrização da Hiperexponencial.

O gráfico usado para escolha do número de estágios da Hiperexponencial é mostrado na Figura 4.13(a). Através do gráfico podemos observar que quatro estágios são suficientes para serem usados com a Hiperexponencial, acima desse número não parece haver melhora muito significativa entre as diferenças do modelo e as amostras. Passando então à primeira etapa para a escolha da melhor distribuição parametrizada, observando o gráfico com as CCDF's (ver Figura 4.13(b)), a Hiperexponencial parece ser a distribuição mais fiel aos dados. Os valores de MSE e a estatística KSSTAT confirmam esta distribuição como uma boa opção. Apesar de a distribuição Lognormal apresentar valores de teste baixos para esta métrica, a Hiperexponencial parece ser um melhor modelo na maioria dos testes, esta afirmação pode ser melhor observada comparando o gráfico de quantiles gerado para ambas as distribuições apresentado na Figura 4.14. Os parâmetros encontrados para a Hiperexponencial são $\alpha = [2.591410e-01, 1.230875e-01, 2.331885e-01, 3.845831e-01]$ e $\lambda = [5.155328e-02, 2.024318e-03, 4.065414e-02, 8.925916e-03]$.



(a) MSE e KSSTAT para vários estágios da hiperexponencial



(b) Gráfico de CCDF com eixo y em escala logarítmica.

Figura 4.13: *Fitting* das distribuições para o tempo em *play*.

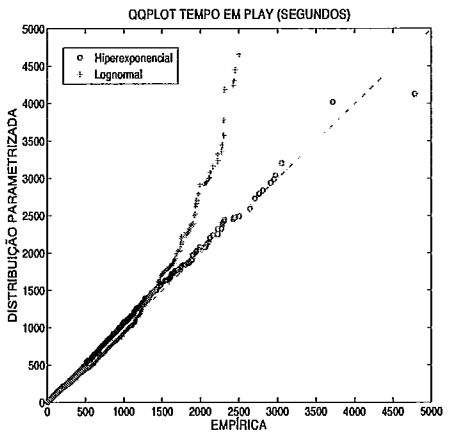


Figura 4.14: QQPlot de amostras geradas pela hiperexponencial e lognormal com as amostras reais do tempo em *play*

Play Interativo

A métrica play interativo foi criada para representar tempos em play de situações onde acreditamos que o usuário não está assistindo o vídeo continuamente com interesse no conteúdo apresentado. A Figura 4.15 apresenta um esquema que descreve a ocorrência de intervalos play entre saltos. Uma vez separadas as amostras para a categoria de play interativo, procederemos com a análise da métrica. Na Figura 4.16(a) temos o histograma que resume o comportamento das amostras. Segundo o

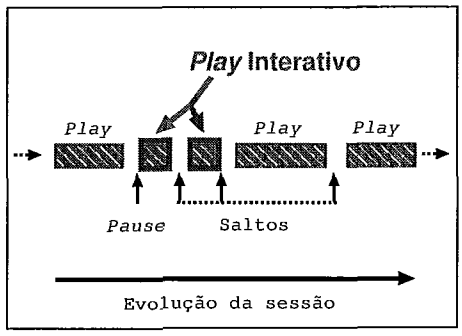
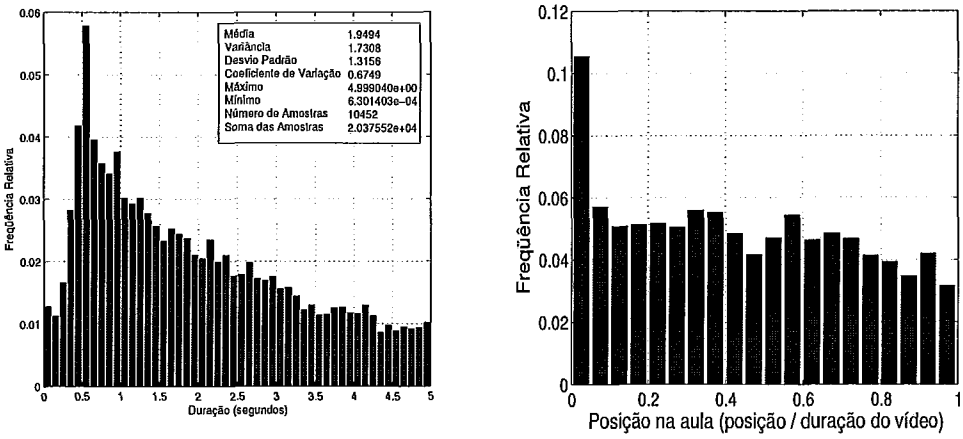


Figura 4.15: Ocorrência de play's entre saltos e pausa/salto.



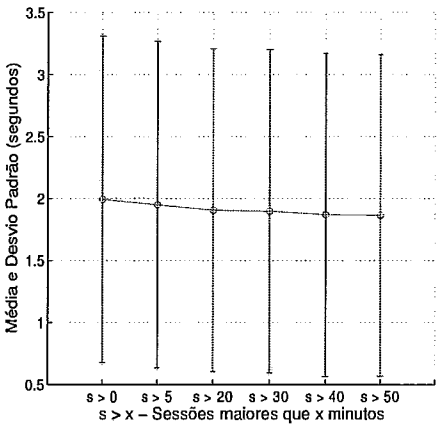
(a) Histograma

(b) Posição onde ocorreram

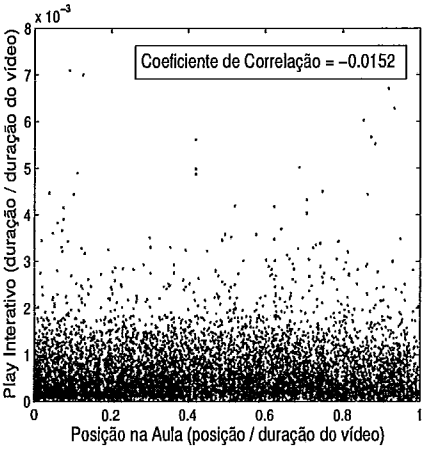
Figura 4.16: Tempo em *play* interativo.

gráfico podemos verificar um pequeno pico na freqüência de amostras com duração de cerca de 0,5 a 1 segundos, que é um tempo razoável para que o usuário tenha tempo de realizar uma nova interação em busca da posição de seu interesse. Pelo gráfico da Figura 4.16(b), como aconteceu com os tempos em *play*, boa parte das amostras ocorrem no início da aula, as outras são também bem distribuídas entre as outras posições.

Analisando as amostras obtivemos um tempo médio para a métrica de 1,95 segundos, com coeficiente de variação de 0,67. Comparando os valores encontrados com os obtidos com outros filtros de sessão (Figura 4.17(a)), mais uma vez não há uma variação significativa. A falta de correlação entre o *play* interativo e a posição no vídeo (coeficiente de correlação igual a -0.0152) é lógica, já que as amostras são muito pequenas para sofrer influência do tempo residual do vídeo, como acontece com o tempo em *play* (Figura 4.17(b)).



(a) Comparação com outros filtros de sessão



(b) Análise da correlação com a posição do vídeo

Figura 4.17: Filtros de sessão e gráfico de correlação (*play* interativo).

Para esta métrica não utilizaremos a distribuição Hiperexponencial, pois a variabilidade das amostras é relativamente baixa e uma das cinco distribuições iniciais pode capturar bem o comportamento das amostras. Por esta métrica ser composta de amostras truncadas (*threshold* igual a 5 segundos), precisaremos tratar o modelo para que as amostras que serão geradas por ele não ultrapassem o threshold. Para isso, inicialmente identificaremos quais são as distribuições mais adequadas, exatamente como explicamos na seção 4.1.3, em seguida serão geradas amostras aleatórias da distribuição parametrizada escolhida, estas amostras serão truncadas ou reestimadas, de forma que nunca ultrapassem o valor limite. Truncar uma amostra significa atribuir a ela o valor limite (5) quando esta for maior que este limite. Por outro lado, reestimar as amostras é simplesmente resortear o valor de uma amostra, caso esta seja maior que o valor limite, até que seja gerada uma amostra dentro do intervalo esperado. Para escolher qual das opções é mais adequada, é calculada a estatística KSSTAT (para o teste de Kolmogorov-Smirnov de duas amostras) entre as amostras reais e as geradas truncadas ou reestimadas.

Passando então à etapa de escolha da melhor distribuição, observando a Figura 4.18(a), não fica muito claro que curva mais se aproxima da CCDF empírica. Entretanto os valores de MSE e KSSTAT apontam para as distribuições Gamma e Weibull. O truncamento e a reestimação de valores gerados aleatoriamente segundo estas distribuições apontam a distribuição Weibull (com $\gamma = 1,4532$ e $k = 2,1446$) truncada como a melhor opção, como pode ser visto nos valores da KSSTAT e no comportamento do gráfico QQplot (Figura 4.18(b)).