Análise de Comportamento de Consumidores por Agrupamento de Sessões para Avaliar o Consumo de Recursos Computacionais e de Comunicação

Eduardo V. Franco
Wilson V. Ruggiero
{evfranco, wilson}@larc.usp.br

Resumo

O presente trabalho apresenta uma nova metodologia, baseada no trabalho de Menasce et al. (1999), para análise de negócio relativo à avaliação da importância de grupos de consumidores para o modelo de negócio para aplicações de negócio eletrônico para a *World Wide Web* (WWW). Esta avaliação é feita através do agrupamento dos consumidores em grupos e da medição da receita gerada e do consumo de recursos por cada uma dos grupos. Este trabalho também descreve a modelagem, a implementação e a validação da eficácia de uma ferramenta para a medição do consumo de recursos computacionais de cada um desses grupos de consumidores.

Abstract

This work presents a new methodology, based on the work of Menasce et al (1999) to build business analysis for the evaluation of the importance of individual customer groups to the business model of a web-based electronic business application. This can be accomplished by breaking the customer base into groups and measuring the monetary income and the resource consumption for each group. This work also describes a software tool modeling to measure th

1. Introdução

O comércio eletrônico atual, o comportamento das empresas que vendem produtos ou prestam serviços pela WWW e o comportamento dos consumidores, só podem ser entendidos se estudados em conjunto com a evolução das tecnologias que permitiram a existência da Internet e com a evolução do processo de adoção destas tecnologias por empresas e consumidores.

Em 1969 o Departamento Norte-Americano de Defesa com o propósito de compartilhar recursos computacionais, criou a ARPANET, uma rede de transmissão de pacotes entre três universidades americanas. Dois anos depois, em 1971, foram criados os primeiros protocolos para utilização desta rede recém criada: o FTP e o Telnet. Porém, foi somente em 1972 que surgiu o E-Mail, a primeira *Killer Application*, ou seja, a forma de utilização que popularizou a Internet além dos círculos de pesquisa voltados a seu estudo. O IP e o TCP, protocolos de rede e transporte utilizados até hoje, só surgiram mais tarde em 1980.

Em 1989, já com 100.000 dispositivos conectados, a Internet já era uma rede de escala global, porém predominantemente acadêmica, até que neste ano a WWW foi criada por Tim Berners-Lee através da especificação do formato HTML e do protocolo HTTP. A partir desta invenção, a Internet foi popularizada em todo o mundo principalmente após o lançamento em 1994 do primeiro Navegador WWW comercial: o MOSAIC da NCSA que permitiu que pessoas fora do meio acadêmico acessassem a WWW.

Percebendo a adoção cada vez maior do uso da Internet pela população em geral, diversos empreendedores passaram a utilizar a Internet como canal de venda de produtos e meio de prestação de serviços, criando desta forma o negócio eletrônico e as empresas ponto-com. Depois de passada a empolgação inicial, o mercado de comércio eletrônico se estabilizou e acabou dividido entre as empresas tradicionais, que adotaram a Internet como mais um canal para atuação e as empresas surgidas na WWW. Nesta nova era de comércio eletrônico, a estratégia de negócio de todas as empresas passou a ser guiada pela inovação tecnológica somada às estratégias de negócio tradicionais.

Com a maior competição entre as empresas, o consumo de recursos computacionais passou a ser um fator primordial para a rentabilidade de um modelo de negócio eletrônico. Para que a diminuição do consumo de recursos computacionais seja possível sem que o faturamento seja afetado, é necessário equilibrar a quantidade da informação passada a cada consumidor com o seu potencial de retorno para o negócio. A avaliação do potencial de retorno dos consumidores para um negócio faz parte da disciplina de marketing e é feita através da classificação dos consumidores em grupos através de um ou mais critérios sócio-econômicos e comportamentais. Para avaliar o potencial de retorno de cada grupo de consumidores, as empresas tradicionais valem-se da observação do comportamento dos consumidores nos pontos de venda, de pesquisas qualitativas e quantitativas, e de aplicação de técnicas de *data-mining* no seu repositório de informação sobre vendas. Já as empresas de negócio eletrônico não podem observar diretamente os consumidores e por isso precisam de métodos e ferramentas para avaliar o comportamento dos consumidores em seu *web-site*.

A análise de comportamento de consumidores em *web-sites* tem sido estudada desde a popularização da WWW para melhorar a eficiência dos servidores WWW e para automatizar a elaboração de mapas de navegação para aplicações WWW. A partir das mesmas informações armazenadas nos *logs* de servidores WWW que permitem conhecer o comportamento dos consumidores, também é possível calcular o consumo de recursos neste servidor e, quando conhecido o comportamento da aplicação, estimar o consumo de recursos dos outros servidores que dão suporte à aplicação.

Para tornar um modelo de negócio eletrônico mais rentável é necessário equilibrar o faturamento e o consumo de recursos através da diminuição do consumo de recursos sem que isto afete o faturamento. Para que isto seja possível, é necessário dividir os consumidores em diversos grupos de consumidores, aferir o consumo e o faturamento de cada um dos grupos e adotar estratégias de marketing e de priorização de serviços diferenciada para cada uma dos grupos de consumidores do negócio eletrônico.

O objetivo deste trabalho é descrever uma técnica de análise do consumo de recursos de um *web-site* através da contabilização do consumo de recursos por módulo do *web-site* e por grupo de consumidores. A técnica proposta é baseada na técnica proposta por Menasce et al. (1999), que consiste na classificação das requisições feitas a uma determinada aplicação WWW em sessões e a partir disso o cálculo da probabilidade de um consumidor navegar de um determinado módulo desta aplicação para outro, porém, alterado de forma a contabilizar diversas informações que são filtradas e descartadas na análise original, mas são importantes para a análise de consumo de recursos computacionais.
O artigo está organizado da seguinte forma: No capítulo 2 são descritas as tecnologias que tornaram

o negócio eletrônico através da WWW possível, no capítulo 3 são discutidas técnicas de modelagem de comportamento de consumidores, no capítulo 4 são listadas técnicas de identificação de sessões. No capítulo 5 a técnica proposta e a ferramenta implementada são descritas. O sexto capítulo descreve os critérios utilizados para a simulação dos dados que são analisados no capítulo seguinte. O capítulo 8 descreve as conclusões do trabalho e o último capítulo contém as referências.

2. Tecnologia para Negócios através da World Wide Web

Os negócios através da WWW surgiram quando alguns empreendedores notaram que a WWW e a Internet podiam ser utilizadas para outras atividades além das atividades acadêmicas e de comunicação eletrônica. Estes empreendedores perceberam que a WWW poderia ser mais um canal de interação com o consumidor para modelos de negócio tradicionais e que também poderia servir como suporte para diversos novos modelos de negócio, baseados na facilidade de comunicação provida pela WWW.
Porém, para permitir a realização de negócios pela WWW, os primeiros empreendedores tiveram que superar duas limitações técnicas impostas pelo formato HTML e pelo protocolo HTTP. A primeira limitação era a forma estática de publicações de informações utilizada até então na WWW visto que o formato HTML não permitia uma interação rica o suficiente com o consumidor para permitir a realização de negócios. Tal limitação foi superada com a geração de páginas HTML de forma dinâmica por aplicações de negócio eletrônico e através sa integração de servidores WWW com bancos de dados.

A segunda limitação a ser superada foi a ausência de controle de estado no protocolo HTTP, que dificultava a identificação inequívoca do consumidor que por sua vez limitava a interação dos consumidores com as aplicações de negócio eletrônico. Tal limitação foi solucionada com a inclusão de informações de controle de estado nas comunicações entre clientes e servidores WWW batizadas como *cookies*. O advento dos *cookies* permitiu a criação de sessões explicitas de comunicação entre servidores e clientes WWW que permitiu uma interação mais rica e segura entre negócio e consumidor.

2.1 Servidores WWW e Logs

Um servidor WWW funciona da seguinte forma: para cada requisição feita, o servidor gera uma resposta e transmite através da mesma conexão em que foi feita a requisição. Dados sobre a conexão, requisição e reposta são então armazenadas no log do servidor WWW.

Uma transação de negócio feita através de uma aplicação de negócio eletrônico é realizada da mesma forma que as demais navegações realizadas na WWW. Desta forma, para realizar uma transação, um consumidor navega por diversos módulos do negócio eletrônico exibidos na forma de páginas HTML onde são obtidas informações sobre o produto/serviço ou são passadas informações para permitir a realização da transação. Por este motivo, assim como as informações de navegação de usuário são armazenadas em *logs* de servidores WWW, aplicações de negócio eletrônico também geram os mesmos *logs* na forma de arquivos texto. Para exemplificar, na tabela 1 e na figura 1, estão as descrições dos campos e a relação entre campo do protocolo http e informação armazenada no log:

#	Nome	Descrição
1	date	**Data da resposta da requisição.**
2	time	**Hora da resposta da requisição.**
3	c-ip	**IP da origem da requisição**
4	cs-username	**Nome do usuário que fez a requisição**
5	s-ip	**IP do servidor que atendeu a requisição**
6	cs-method	**Método utilizado na requisição**
7	cs-uri-stem	**URI do objeto requisitado**
8	cs-uri-query	**Parâmetros passados ao objeto requisitado**
9	sc-status	**Status da resposta**
10	sc-bytes	**Tamanho em bytes da resposta**
11	cs-bytes	**Tamanho em bytes da requisição**
12	time-taken	**Tempo gasto para processar a requisição**
13	cs-version	**Versão do protocolo**
14	cs(User-Agent)	**User-Agent da origem da requisição**
15	cs(Cookie)	**Cookies enviados com a requisição**
16	**cs(Referrer)**	**Objeto que originou a requisição**

Tabela 1 - Informações armazenadas em um log de servidor WWW

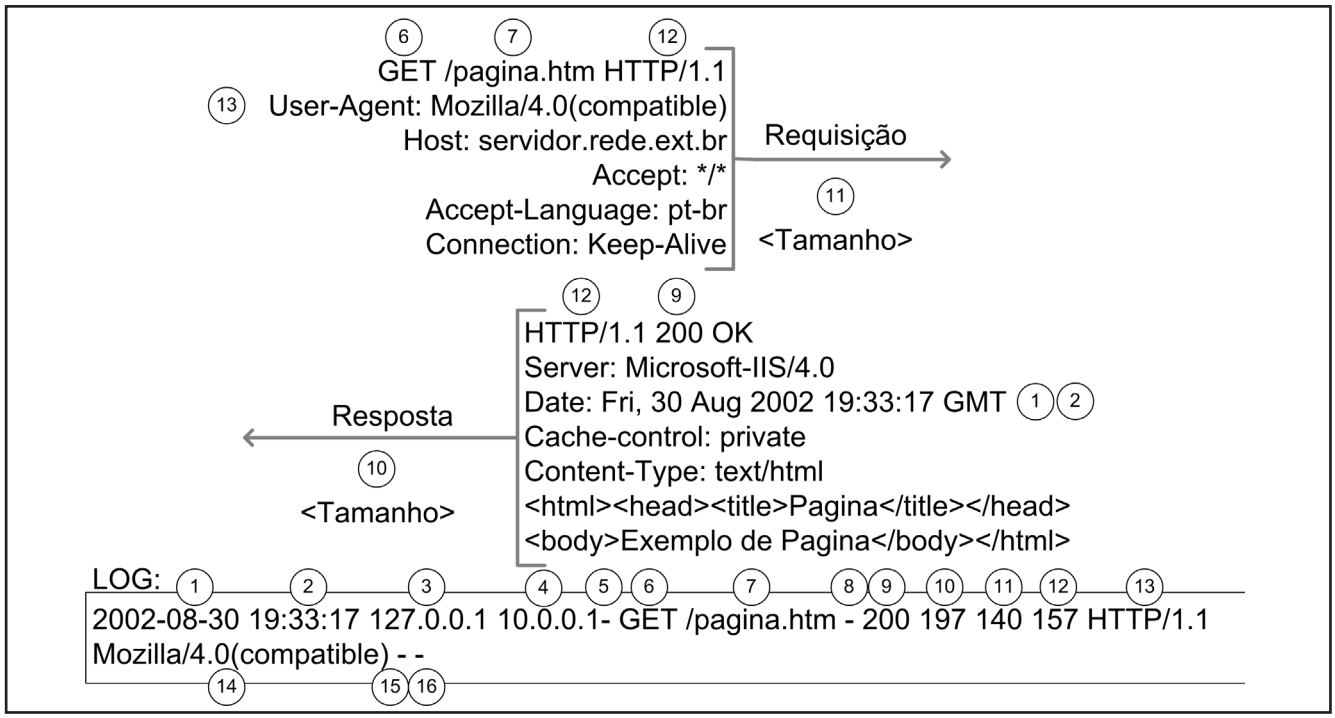

Figura 1 - Dados armazenados no log do servidor WWW

3. Análise e Modelagem do Comportamento dos Consumidores

Em uma aplicação de negócio eletrônico, cada passo de negócio seguido por um consumidor é realizado através do envio de informações pelo consumidor, do processamento das informações enviadas por um módulo do negócio eletrônico e, por último, através da montagem de uma página HTML contendo orientações para o consumidor ou um novo pedido de informações complementares através de um formulário.

O comportamento do consumidor é estudado através da análise de quais passos de negócio foram seguidos, em qual seqüência e qual o passo final de negócio atingido. Como cada um dos passos de negócio é implementado através de um módulo, é possível analisar o comportamento do consumidor através do conhecimento de quais módulos foram acessados e em qual momento cada uma desses módulos foi acessado. Esta informação pode ser obtida através da monitoração ativa da aplicação ou através do processamento das informações armazenadas nos *logs* dos servidores WWW.

As mesmas informações que permitem analisar o comportamento do consumidor também permitem medir o consumo de recursos de cada requisição através da análise da quantidade de informação transmitida e do tempo necessário para o processamento da requisição do consumidor.

A modelagem do comportamento de consumidores consiste em estudar a forma que os consumidores interagem com aplicações de negócio eletrônico. Como toda interação através da WWW se dá através de servidores WWW, a melhor forma de se estudar estas interações é através da análise das informações trocadas entre o consumidor e o servidor WWW. Esta análise pode ser feita através da adaptação da aplicação de negócio eletrônico para armazenar as informações necessárias ou através da análise das informações contidas nos *logs* de servidores WWW.

Diversos trabalhos [2, 3, 4, 5, 6, 7] propuseram formas e ferramentas distintas para a análise destas informações. Esta distinção é devida ao fato de cada um dos trabalhos pretender analisar um determinado fator do comportamento dos consumidores, e por isso, apenas as informações pertinentes à análise são extraídas e armazenadas.

A principal referência deste trabalho [1] descreve uma metodologia de caracterização da carga gerada em um *web-site* de negócio eletrônico para um conjunto de consumidores chamada de *Customer Behavior Model Graph (CBMG)* . Esta caracterização é feita através da elaboração de uma matriz de probabilidade de transição entre cada módulo de um *web-site* a partir do processamento das informações contidas em logs de servidores WWW.

A solução proposta por [1] é baseada na construção de uma matriz de probabilidades de transição entre os diversos módulos de um web-site e na caracterização da carga de ambientes transacionais para

o correto dimensionamento destes ambientes. Por este motivo, a metodologia proposta despreza muita informação que é necessária para a correta medição do consumo de recursos e não se preocupa em definir formalmente a metodologia de agrupamento de consumidores em grupos de consumidores para a realização da análise.

4. Identificação de Sessões

Para corretamente medir o consumo de recursos, é necessário identificar nos dados extraídos dos *logs* do servidor WWW quais requisições pertencem a uma mesma navegação. Este conjunto é definido como sessão.

Após a filtragem, as requisições devem ser individualizadas por usuários e agrupadas seqüencialmente em sessões. Existem três formas de se identificar sessões: Explícita, Seqüencial e Implícita.

Uma requisição possui uma sessão explicita quando algum dos elementos contidos nas requisições da lista de acessos permite identificar univocamente uma sessão. Este elemento geralmente é um código identificador de sessão armazenado em um *cookie* ou passado como parâmetro.

Quando não é possível identificar uma sessão de forma explicita, e o campo "cs(Referrer)" esta presente nos dados armazenados, é possível identificar as sessões de forma seqüencial, pois o *referrer* é um dos campos obrigatórios do cabeçalho do protocolo HTTP e transmite a URI do objeto que gerou a requisição. Ou seja, quando é requisitado um módulo Y após o usuário ter navegado por um link contido na página HTML gerada pelo módulo X, a URI do módulo X é enviada no campo *referrer* da requisição de Y. Essa referência também ocorre no caso das requisições secundárias, que são requisições resultantes da requisição principal para obter os demais elementos que são utilizados na montagem da página HTML como, por exemplo, as imagens. Esses elementos não fazem parte do caminho de navegação, mas são importantes para identificar o total de recursos consumidos e o tempo gasto.

Quando não é possível identificar uma sessão de forma explicita ou seqüencial, sempre é possível identificar a sessão de forma implícita. Esta deve ser a última técnica de agrupamento de sessão a ser utilizada quando nenhum outro método possa ser utilizado já que é extremamente susceptível a erros de identificação causados pela utilização de *proxies*. A sessão implícita é obtida através do agrupamento das requisições por IP e o estabelecimento de um tempo máximo entre requisições (*Time-out*). Desta forma duas requisições existentes na lista de acesso pertencem a uma mesma sessão caso ambas tenha sido feitas a partir de um mesmo endereço IP e o tempo entre elas seja menor que um limite de tempo pré-estabelecido (*Time-out*).

5. Medição do Consumo de Recursos por Grupos de consumidores

Para permitir a extração de informações mais completas quanto ao consumo de recursos de cada uma dos grupos de consumidores, além da obtenção da matriz de probabilidade como é feito na análise CBMG, é necessário também contabilizar informações referentes à quantidade de informação recebida, à quantidade de informação transmitida e o tempo de processamento de cada transação. Por este motivo, boa parte das informações desprezadas pela análise CBMG deve ser contabilizada e, para cada nó do gráfico CMBG, as seguintes informações também precisam ser extraídas:

• O tamanho total do documento HTML principal que implementa o passo de negócio (CI),

• O tempo de transmissão do documento HTML (CT)

Desta forma, o consumo de recursos em tempo CT e o consumo de recurso em informação CI podem ser calculados em função do tempo total de transmissão do documento principal (T), do número de requisições secundárias (N), do tempo médio das requisições secundárias (TN), da quantidade de dados transmitidos na requisição principal (I), e da quantidade média de dados transmitidos nas requisições secundárias (IN):

$CT = T + N * TN$ e
$CI = I + N * IN$

O passo seguinte após a identificação das sessões é identificar à qual grupo o consumidor que fez a requisição pertence. O processo de identificação do grupo, ao qual uma sessão pertence, está intimamente ligado ao processo de agrupamento de consumidores em grupos e por isso não é discutido neste artigo.

5.1. Ferramenta para Análise do Consumo de Recursos de Grupos de consumidores

A ferramenta proposta neste trabalho foi implementada para medir o consumo de recursos através da realização de uma análise CBMG modificada. Nesta análise, as seguintes informações foram extraídas:

- O tamanho total dos documentos HTML principais que implementam os passos do negócio,
- A quantidade média de elementos extras carregados,
- O tempo médio para se carregar os elementos extras,
- O tamanho médio dos elementos extras.

Além da extração destas informações, a ferramenta também calcula a variância destes valores.

Para modelar a solução foi escolhida a técnica de modelagem orientada a objeto, pois essa técnica permite transportar elementos do domínio problema diretamente para o domínio da solução para posteriormente selecionar quais características (atributos) dos objetos são necessárias para o correto entendimento do problema e sua solução. Esta técnica também permite que sejam facilmente encontrados as classes e os métodos de cada classe, para resolver de forma atômica os diversos passos necessários para a solução do problema. O banco de dados necessário para o armazenamento dos *logs* pré-analisados e para o armazenamento das análises e das sessões foi modelado utilizando-se modelos entidade-relacionamento obtidos a partir do diagrama de classes.

Os arquivos de *logs* foram modelados na forma da classe "LogFile". Sobre estes objetos são efetuadas as principais operações de análise através dos seus métodos de análise e de extração de sessão. Os objetos da classe "LogFile" são persistentes

Cada arquivo de *log* é formado por diversas entradas, modeladas pela classe "EntradaDeLog". A classe "Entrada de Log" é utilizada somente na fase de análise dos arquivos de Log e por isso esta classe não é persistente.

As informações importantes de uma entrada de log, que são necessárias para diversas fases da análise, foram modeladas através da classe "Chamada". Esta classe é persistente e contém todas as informações da "Entrada de Log" referentes à requisição do usuário que podem ser usadas para identificar o seu comportamento.

A navegação de um usuário foi modelada através da classe transição, que representa um passo da navegação de um usuário no sistema. As informações do destino da transição sempre são conhecidas e modeladas através de objetos da classe "Chamada". Já as informações do módulo de origem dependem das informações armazenadas em cada entrada do log e também da fase e passo atual da análise. Por isso a origem é modelada através da classe "Página" que pode ser uma chamada, um módulo do sistema,

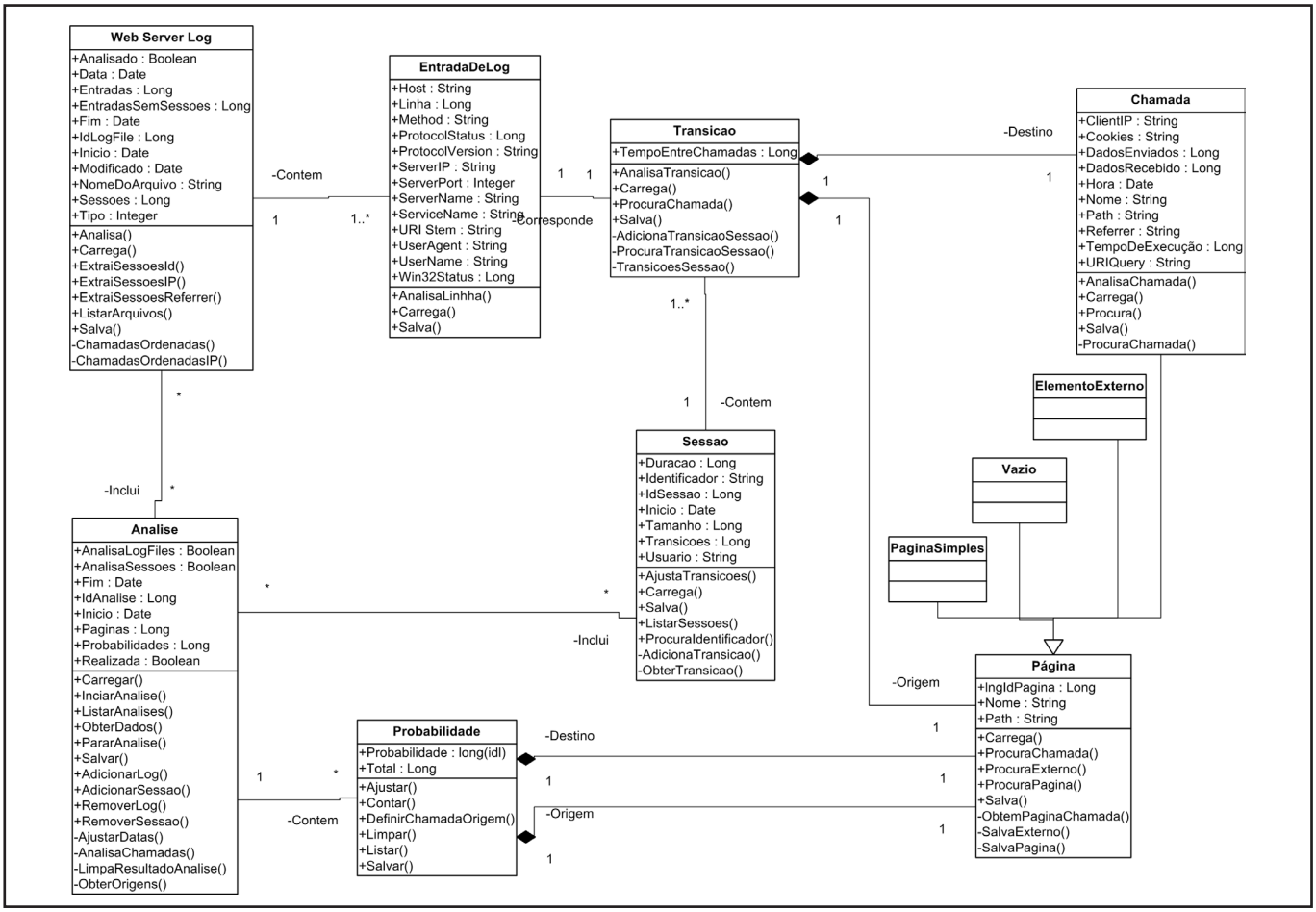

Figura 2 - Diagrama de classes da solução

um elemento externo ou ser desconhecida.

Cada um dos objetos da classe transição pode ou não ter sua sessão identificada. Esta sessão é modelada através da classe "Sessão". Esta classe é a responsável por identificar a relação entre os usuários e as chamadas.

Uma análise é efetuada sobre um conjunto de arquivos de *logs* e apenas as chamadas pertencentes a determinadas sessões são consideradas para análise. Estas análises são modeladas através da classe "Análise". Cada análise gera como resultado uma matriz de probabilidades.

A probabilidade de navegação entre dois módulos quaisquer do sistema para uma determinada análise é modelada através da classe "Probabilidade". Desta forma, a matriz de probabilidades de uma determinada análise é formada pelo conjunto de objetos da classe "Probabilidade" relacionados a uma determinada análise.

6. Aplicação da Ferramenta sobre Dados Simulados

O primeiro objetivo deste teste foi verificar a eficácia da ferramenta em identificar corretamente a sessão a qual pertence cada requisição e uma vez identificada a qual grupo pertence cada uma das sessões analisar o consumo de recursos para cada grupo.

Para preparar a massa de testes, primeiro foi especificada uma aplicação simulada de negócio eletrônico com os seguintes parâmetros:

Na (Número de módulos): 6,

Ng (Quantidade de grupos de usuários): 3,

Ns(g) (número de sessões por grupo): Ns(A) = 900, NS(B)= 1000 e NS(C) = 800.

Em seguida, o comportamento de cada grupo foi especificado de acordo com as probabilidades de transição nas tabelas 2, 3 e 4:

Ga	A	B	C	D	E	F	O
A	0	1	0	0	0	0	0
B	1/2	0	1/2	0	0	0	0
C	0	1/3	0	1/3	1/3	0	0
D	0	0	1/3	0	0	1/3	1/3
E	0	0	1/3	0	0	1/3	1/3
F	0	0	0	0	0	0	1

Tabela 2 - Probabilidade de Transição para o Grupo A

Ga	A	B	C	D	E	F	O
A	0	1	0	0	0	0	0
B	1/2	0	1/2	0	0	0	0
C	0	1/2	0	1/2	0	0	0
D	0	0	1/2	0	0	1/2	0
E	0	0	1/3	0	0	1/3	1/3
F	0	0	0	0	0	0	1

Tabela 3 - Probabilidade de Transição para o Grupo B

Ga	A	B	C	D	E	F	O
A	0	1	0	0	0	0	0
B	0	0	1	0	0	0	0
C	0	0	0	1/3	2/3	0	0
D	0	0	1/3	0	0	2/3	0
E	0	0	2/3	0	0	0	1/3
F	0	0	0	0	0	0	1

Tabela 4 - Probabilidade de Transição para o Grupo C

Para garantir que cada grupo apresentasse um consumo diferenciado de recursos e faturamento, foi definida na tabela 5 a quantidade média de dados transmitidos e recebidos por grupo, assim como o valor médio de transação. Estes valores, e os valores presentes nas tabelas 2, 3 e 4, foram inspirados no comportamento de 3 grupos típicos de consumidores: Consumidores Eventuais (A), Consumidores Focados (B) e Consumidores Fiéis(C).

Grupo	Dados Enviados	Dados Recebidos	Valor Médio de Transação
A	1.000 bytes	10.000 bytes	100,00
B	2.000 bytes	20.000 bytes	200,00
C	4.000 bytes	40.000 bytes	300,00

Tabela 5 - Consumo de Recursos

No cálculo do tempo médio de requisição, foi considerado que todos os grupos apresentam a mesma velocidade de transmissão de dados e por isso os tempos de requisição são sempre proporcionais aos valores sorteados para a quantidade de dados enviados e recebidos.

Por último, para determinar se a realização ou não de uma transação, deve-se verificar se a sessão passa pelo módulo F já que pelo modelo utilizado, o módulo F gera a página de confirmação de pagamento.

A partir da massa de testes definida anteriormente, uma aplicação computacional desenvolvida em Visual Basic simulou o comportamento dos consumidores dentro da aplicação conforme definido na massa de testes. Para aproximar a simulação da realidade, foi utilizada uma variação de 10% sobre a média pretendida em todos os casos de sorteio de valores.

O resultado da simulação foi um conjunto de requisições de módulos da aplicação eletrônica que

por sua vez foi inserido diretamente no repositório da ferramenta desenvolvida como aconteceria nos casos onde o sistema analisado estivesse sendo monitorado através de formas de monitoração dinâmicas ou ativas.

Para evitar que erros na identificação das sessões afetassem a validação da ferramenta ou a validação da análise, foi definido que todas as requisições teriam o identificador da sessão explicito nos parâmetros passados pelo cliente durante a requisição (uri-query) assim como acontece em diversas aplicações existentes atualmente. Além disso, para simplificar a identificação de qual grupo contém cada uma das sessões, um identificador de grupo também foi adicionado aos parâmetros transmitidos do cliente para a aplicação. Desta forma, toda requisição possui no campo uri-query a seguinte informação: '?NumGrupo=<Identificacão do Grupo>&Sessao=<Número da Sessão>'.

Sobre a massa de testes criada foram feitas 4 análises: Uma análise contemplando todas as sessões existentes na massa de testes utilizada e uma análise individualizada para cada uma das sessões. Estas análises foram realizadas em duas etapas.

A primeira etapa, comum a todas as análises, consistia na identificação e agrupamento de sessões. Devido a um numero explícito de sessão estar presente em cada chamada conforme definido na simulação, foi utilizado método de identificação de sessões explicitas.

Depois de realizada a primeira etapa foi então realizada a primeira análise de probabilidade de transição utilizando como entrada todas as requisições simuladas. O resultado obtido foi:

P	A	B	C	D	E	F	O
A	0	1	0	0	0	0	0
B	0.46	0	0.54	0	0	0	0
C	0	0.34	0	0.41	0.25	0	0
D	0	0	0.43	0	0	0.50	0.07
E	0	0	0.53	0	0	0.12	0.35
F	0	0	0	0	0	0	1

Tabela 6 - Probabilidade Agregada de Transição Medida

Como resultado da análise também foi produzida a tabela 7 com o consumo de recursos:

Módulo	Enviados(Kb)	Recebidos(Kb)	Tempo(ms)
A	1868 ± 791	18701 ± 7898	564 ± 306
B	1840 ± 798	18416 ± 8047	556 ± 313
C	2217 ± 1005	22122 ± 10001	665 ± 379
D	2168 ± 992	21660 ± 9809	653 ± 369
E	2928 ± 1502	29286 ± 14996	879 ± 546
F	2191 ± 901	22013 ± 9073	676 ± 370
Média	2056 ± 914	20568 ± 9142	621 ± 351

Tabela 7 - Consumo Total de Recursos Medido

Depois de realizada a analise completa, foi criada uma nova analise e nela inseridas todas as sessões cujo identificador indicasse pertencer ao grupo A. No total foram inclusas 900 sessões conforme esperado. Foi então realizada a análise de probabilidade de transição e o resultado obtido foi:

P(A)	A	B	C	D	E	F	O
A	0	1	0	0	0	0	0
B	0.52	0	0.48	0	0	0	0
C	0	0.32	0	0.32	0.36	0	0
D	0	0	0.35	0	0	0.31	0.34
E	0	0	0.30	0	0	0.33	0.37
F	0	0	0	0	0	0	1

Tabela 8 - Probabilidade de Transição Medida para o Grupo A

Módulo	Enviados(Kb)	Recebidos(Kb)	Tempo(ms)
A	997 ± 114	9975 ± 1148	303 ± 104
B	1002 ± 116	9994 ± 1153	299 ± 104
C	1002 ± 116	9997 ± 1143001	300 ± 102
D	1005 ± 111	10001 ± 1156	304 ± 105
E	1000 ± 119	10055 ± 1149	305 ± 100
F	989 ± 116	1002 ± 1163	304 ± 108
Média	999 ± 115	9995 ± 1150	302 ± 103

Tabela 9 - Consumo de Recursos Medido para o Grupo A

A mesma análise realizada para o Grupo A também foi feita para os grupos B (1000 sessões) e C (800 Sessões):

P(A)	A	B	C	D	E	F	O
A	0	1	0	0	0	0	0
B	0.50	0	0.50	0	0	0	0
C	0	0.50	0	0.50	0	0	0
D	0	0	0.48	0	0	0.52	0
E	0	0	0	0	0	0	0
F	0	0	0	0	0	0	1

Tabela 10 - Probabilidade de Transição Medida para o Grupo B

Módulo	Enviados(Kb)	Recebidos(Kb)	Tempo(ms)
A	1998 ± 228	20038 ± 2304	607 ± 211
B	2001 ± 230	20007 ± 2298	605 ± 206
C	1998 ± 231	19972 ± 2323001	603 ± 206
D	1997 ± 232	20058 ± 2327	606 ± 204
E	-	-	-
F	2009 ± 227	20137 ± 2286	619 ± 215
Média	2000 ± 230	20020 ± 2308	606 ± 207

Tabela 11 - Consumo de Recursos Medido para o Grupo B

P(A)	A	B	C	D	E	F	O
A	0	1	0	0	0	0	0
B	0	0	1	0	0	0	0
C	0	0	0	0.32	0.68	0	0
D	0	0	0.32	0	0	0.68	0
E	0	0	0.67	0	0	0	0.33
F	0	0	0	0	0	0	1

Tabela 12 - Probabilidade de Transição Medida para o Grupo C

Módulo	Enviados(Kb)	Recebidos(Kb)	Tempo(ms)
A	4013 ± 473	39983 ± 4593	1181 ± 399
B	4000 ± 471	40304 ± 4707	1126 ± 403
C	4012 ± 461	39942 ± 4607	1196 ± 405
D	4002 ± 453	39677 ± 4551	1190 ± 399
E	4025 ± 456	40236 ± 4639	1206 ± 409
F	3965 ± 460	39896 ± 4554	1232 ± 418
Média	4009 ± 462	40033 ±4616	1202 ± 405

Tabela 13 - Consumo de Recursos Medido para o Grupo C

A diferença entre a linha E da tabela 10 e a linha E da tabela 3 é devido ao fato de não haver nenhuma probabilidade de transição para o módulo E. Por isso não houve nenhuma transição a partir do módulo E.

7. Análise dos dados simulados

Foram seguidos os passos:
• Aferir o faturamento para cada um dos grupos,
• Calcular o número total de transações realizadas por grupo,
• Medir o consumo de recursos de um dos grupos.

Para aferir o faturamento médio para cada grupo foram levados em conta dois fatores: A probabilidade de realização de uma transação que, conforme a definição da massa de testes, é igual à probabilidade de uma sessão passar pelo módulo F e o valor médio da transação que também é dado pela definição da massa de testes.

Para calcular a probabilidade de uma sessão passar pelo módulo F foi utilizada a fórmula para o cálculo do número médio de vistas (Vn) proposta no artigo de Menasce et al. (1999). Para calcular Vn deve-se resolver um sistema de equações lineares obtido através da transformação da matriz de probabilidade ($p_{k,n}$), onde:

$V_1 = 1$ e $Vn = \sum_{k=1}^{p} V_k * p_{k,n}$

Calculando-se $V_f = V_5$ para todos os grupos obtivemos:

Grupo	Probabilidade de Transação P(t)
A	48,16 %
B	100,00 %
C	50,00 %
A+B+C	66,91 %

Tabela 14 - Probabilidades de Transação Medidas para o módulo F

Multiplicada a probabilidade de realização de transação Pt(G) de cada grupo pelo valor definido na massa de testes como valor médio de cada transação do grupo Vm(G), foi aferido o valor médio esperado de transação por sessão para cada grupo Vs(G):

$$Vs(G) = Pt(G) * Vm(G)$$

Obtendo como Resultado:

Grupo	Valor Médio
A	48,16
B	200,00
C	150,00
A+B+C	198,18

O valor médio faturado por sessão para todos os grupos Vs(A+B+C) foi calculado através da média ponderada do valor da transação em relação ao número de sessões que realizaram transação de um determinado grupo G: (Pt(G) * Ns(G)):

$$Vs(A+B+C) = \frac{Vs(A) * Pt(A) * NS(A) + Pt(B) * VS(B) * NS(B) + VS(C) * Pt(C) * NS(C)}{Pt(A) * NS(A) + Pt(B) * NS(B) + Pt(C) * NS(C)}$$

O último passo foi a medição do consumo de recursos por sessão de cada grupo. Para isso foi necessário extrair o consumo médio de recursos por módulo para cada grupo das tabelas 9, 11 e 13 e multiplicar estes valores pelo número de requisições por sessão medida para cada grupo:

Grupo	Requisições	Dados Enviados	Dados Recebidos	Tempo
A	10,51	999 ± 115	9995 ± 1150	302 ± 103
B	16,59	2000 ± 230	20020 ± 2308	606 ± 207
C	7,01	4009 ± 462	40033 ± 4616	1202 ± 405
A+B+C	11,73	2056 ± 914	20568 ± 9142	621 ± 351

Tabela 16 - Consumo médio de recursos por requisição por sessão

Foi calculado o consumo de recursos por sessão através da multiplicação do número de requisições por sessão pela quantidade de recursos consumidos por requisição. A quantidade de dados recebidos e a quantidade de dados enviados foram então combinados em uma única medida chamada de dados trafegados:

Grupo	Dados Trafegados	Tempo
A	115547 ± 18806	3174 ± 1082
B	365312 ± 59486	10054 ± 3434
C	308734 ± 50323	8426 ± 2859
A+B+C	265380 ± 166828	7284 ± 4117

Tabela 17 - Consumo médio de recursos por sessão

Para medir a importância do grupo em relação ao total, foi calculado o total de dados trafegados, tempo consumido e faturamento de um grupo em relação ao total. O faturamento foi calculado levando-se em conta o valor médio das transações do grupo e a probabilidade de transação calculada anteriormente.

Grupo	Sessões	Dados Trafegados	Tempo	Faturamento
A	33,33%	14,52 %	14,54 %	11,93 %
B	37,04%	51,00 %	51,16 %	55,04 %
C	29,63%	34,48 %	34,30 %	33,03 %

Tabela 18 - Faturamento e Consumo de Rercursos por Grupo

Por último, ao se comparar os grupos com a média global, obteve-se a tabela 19 com o consumo de recursos e faturamento relativo de cada sessão do grupo em relação ao consumo e faturamento médio para cada sessão independente de grupo:

Grupo	Dados Trafegados	Tempo	Faturamento
A	- 56,46 %	- 56,43 %	- 75,70 %
B	+ 37,66 %	+ 38,02 %	+ 0,92 %
C	+ 16,37 %	+ 15,67 %	-24, 31 %

Tabela 19 - Comparação da Sessão média de cada grupo com a Sessão média

Os dados obtidos confirmaram a validade da metodologia de análise, pois permitiram identificar que o grupo B era o mais rentável apesar do tamanho médio de suas sessões ser maior que o tamanho das sessões dos outros grupos. Também permitiu identificar que o grupo A era o menos rentável de todos devido a uma baixa probabilidade de realização de transação. O grupo C conforme esperado foi caracterizado pelo consumo e faturamento intermediários.

8. Conclusões

Este trabalho propôs e implementou uma ferramenta para fazer análises de desempenho voltadas ao consumo de recursos computacionais.

A ferramenta foi validada, pois todos os valores intermediários produzidos pela análise foram próximos aos valores esperados que constavam da massa de testes projetada, visto que a discrepância encontrada na validação dos dados analisados foi sempre pequena.

A análise dos resultados produzidos pela ferramenta permitiu também validar a metodologia de análise proposta. Isto foi possível, pois o resultado da análise coincide com o comportamento esperado de cada um dos grupos e também os resultados da análise permitiriam sugerir alterações no modelo de negócio e na aplicação de negócio eletrônico para melhorar o desempenho da aplicação já que o grupo de consumidores com maior retorno apresentava um consumo de recursos alto se comparado aos demais grupos.

9. Referências

[01] MENASCÉ, D., ALMEIDA, V., FONSECA, R., MENDES, M. A Methodology for Workload Characterization of E-commerce Sites. Proceedings of the ACM Conference in Eletronic Commerce, p.119-129, 1999.

[02] ARLITT, M., WILLIAMSON, C. Web Server Workload Characterization. Proceedings of the 1996 SIGMETRICS Conference on Measurement of Computer Systems, 1996.

[03] BORGES, J., LEVENE, M. Data mining of user navigation patterns. Lecture Notes in Artificial Intelligence, p.92-111, 2000.

[04] CHANG, H., ZHANG, F. Research and development in Web usage mining system-key issues and proposed solutions: a survey. Proceedings of the International Conference on Machine Learning and Cybernetics, v.02, p.986-990, 2002.

[05] CHEN, Z., FOWLER, R., FU, A., WANG, C. Linear and Sublinear Time Algorithms for Mining Frequent Traversal Path Patterns From Very Large Web Logs. Proceedings of the Seventh International Database Engineering and Applications Symposium, 2003.

[06] CHI, E., ROSIEN, A., HEER, J. LumberJack: Intelligent Discovery and Analysis of Web User Traffic Composition. The Proceedings of the ACMSIGKDD Workshop on Web Mining for Usage Patterns and User Profiles, 2002.

[07] PALIOURAS, G., PAPATHEODOROU, C., KARKALETSIS, V., SPYROPOULOS, C.D., TZITZIRAS, P. From Web Usage Statistics to Web Usage Analysis. Proceedings of the IEEE Conference on Systems Man and Cybernetics, p.159-164, 1999.