*Avaliação do acesso a periódicos eletrônicos na web pela análise do arquivo de log de acesso**

Guilherme Ataíde Dias
Bacharel em Ciência da Computação – Universidade Federal da Paraíba – 1990. Mestre em Administração – Central Connecticut State University, USA –1996. Doutorando em Ciência da Informação – Universidade de São Paulo, ECA/USP.
E-mail: guilhermeataide@aol.com

Resumo

Este artigo apresenta uma abordagem sobre a avaliação do acesso a periódicos eletrônicos disponibilizados na World Wide Web por meio da análise do arquivo de log de acesso. O arquivo de log de acesso da revista ***Informação & Sociedade: Estudos*** *é processado e apresentado como um exemplo de aplicação do uso de uma ferramenta automatizada de análise para arquivo de* ***log*** *de acesso. As características inerentes à análise do arquivo de* ***log*** *de acesso são apresentadas e discutidas.*

Palavras-chave

Periódicos eletrônicos; Avaliação de acesso; Arquivo de log de acesso.

Evaluating the access of electronic periodicals at the Web through the analysis of the access log file

Abstract

This article presents an approach for the evaluation of the access to electronic journals made available in the World Wide Web through the analysis of the access log file. The access log file of the journal Informação & Sociedade: Estudos is processed and presented as an example of the use of an automated tool for log file analysis. The inherent features on the analysis of the access log file are presented and discussed.

Keywords

Electronic periodicals; Access evaluation; Access log file.

INTRODUÇÃO

O processo de disponibilização de um periódico eletrônico na *World Wide Web* é um empreendimento composto de várias etapas. A partir do momento em que estas etapas estejam completas, torna-se necessário avaliar de alguma maneira o acesso ao respectivo periódico eletrônico pelos seus usuários. Pode-se medir, por exemplo, dentre uma variedade de opções, os seguintes itens: o acesso à página de entrada (*home page*), acessos aos resumos dos artigos, acesso ao texto completo dos artigos (*download* de artigos). A importância em estudar o acesso a periódicos eletrônicos traduz-se de várias maneiras: é importante ter conhecimento dos conteúdos acessados pelos usuários como forma de identificar as suas necessidades e atendê-los de forma adequada; no caso de uma biblioteca disponibilizar o acesso a periódicos eletrônicos pagos, é fundamental ter uma estatística de acessos a estes periódicos como forma de justificar o investimento feito em assinaturas.

Os dados compilados neste artigo contemplam o acesso ao texto completo dos artigos disponibilizados no periódico eletrônico *Informação & Sociedade: Estudos*, uma publicação semestral do Curso de Mestrado em Ciência da Informação da Universidade Federal da Paraíba (CMCI/UFPB). O periódico em questão passou a ter uma versão eletrônica a partir do primeiro semestre de 2000 e pode ser acessado através da URL http://www.informacaoesociedade.ufpb.br. A versão impressa de forma tradicional em papel continua sendo editada.

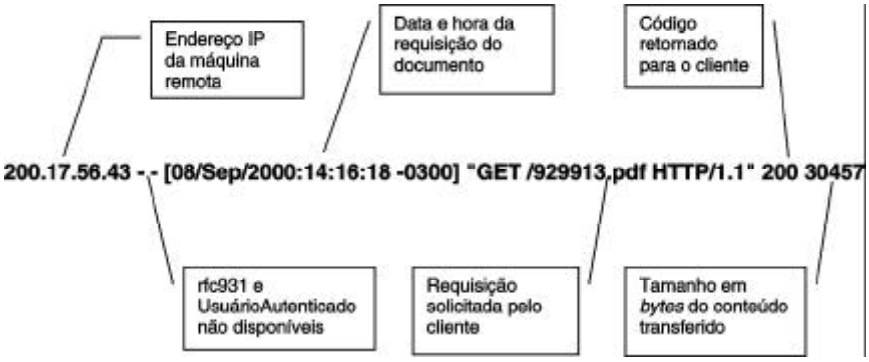
DADOS SOBRE USO DOS PERIÓDICOS ELETRÔNICOS

Por ser uma atividade relativamente nova, a análise do acesso a periódicos eletrônicos apresenta alguns pontos passíveis de análise e discussão, pois estes pontos de uma forma geral precisam ser compreendidos e normatizados. Segundo Luther (2000), um problema fundamental sobre o qual não se chegou ainda a um entendimento seria o de

* Este artigo faz parte da tese de doutorado em ciência da informação do autor, que tem como orientadora Dinah Población.

como gerar dados de modo que os mesmos pudessem ser comparados e utilizados.

FIGURA 1
Linha de um arquivo de log associado à Revista Informação & Sociedade: Estudos



Tomando-se como exemplo os periódicos científicos eletrônicos brasileiros na área da ciência da informação, apenas o periódico *Ciência da Informação,* hospedado no SciELO, apresenta de forma aberta para a comunidade de usuários a opção de consultar estatísticas. Os relatórios de utilização disponíveis são os de acessos da revista, acessos aos fascículos e acessos aos artigos. De acordo com o sugerido em Luther (2000), pode-se inferir que seria interessante estabelecer uma padronização na geração dos dados associados aos diferentes periódicos eletrônicos. Qualquer tentativa de comparação do acesso aos *sites* destas revistas ficaria comprometida sem uma prévia padronização dos dados estatísticos a serem gerados. Uma maneira de contornar a inexistência de uma padronização de dados a serem processados e comparados pode ser feita com a utilização dos arquivos de *log* gerados pelos servidores *web*. Estes arquivos de *log* podem eventualmente apresentar diferenças no formato em virtude da diversidade de servidores *web* existentes, contudo é possível configurar os diferentes servidores web para gerar um arquivo de *log* de acesso de acordo com uma especificação única e amplamente conhecida. Esta especificação é o *Common Logfile Format (CLF)* tal qual detalhado pelo World Wide Web Consortium (W3C). O World Wide Web Consortium pode ser acessado através da URL http://www.w3.org.

Neste trabalho, considera-se que um *log* é um registro com informações relativas à ocorrência de determinados eventos. A seguir, são apresentados dois exemplos práticos do que vem a ser um *log*:

1. sempre que determinado artigo disponibilizado em um periódico eletrônico é consultado (evento), informações relativas a esta consulta (*log*) são armazenadas em um arquivo;

2. em uma biblioteca, quando um usuário solicita o empréstimo de um livro (evento), as informações relativas a este empréstimo são armazenadas em linhas de uma ficha (*log*) e posteriormente armazenadas em um arquivo.

A seguir apresenta-se, de forma detalhada, como cada entrada no arquivo de *log* de acessos do servidor *web* deve estar estruturada de acordo com o *Common Logfile Format*:

MáquinaRemota rfc931 UsuárioAutenticado [data] "requisição" status bytes

Cada campo do *Common Logfile Format* armazena as seguintes informações:

MáquinaRemota: o nome da máquina remota ou o endereço IP no caso de o nome da máquina não estar disponível;

rfc931: nome do usuário remoto, se a informação não estiver disponível, um sinal de menos (–) será colocado no campo;

UsuárioAutenticado: no caso de o documento requisitado ser protegido por uma senha de acesso, então este campo conterá o nome do usuário autenticado (Laurie,1999); se a informação não estiver disponível, um sinal de menos (–) será colocado no campo;

[data]: data e hora de requisição do documento;

"*requisição*": a linha da requisição exatamente como solicitada pelo cliente;

status: código de três dígitos retornado para o cliente indicando o *status* da requisição;

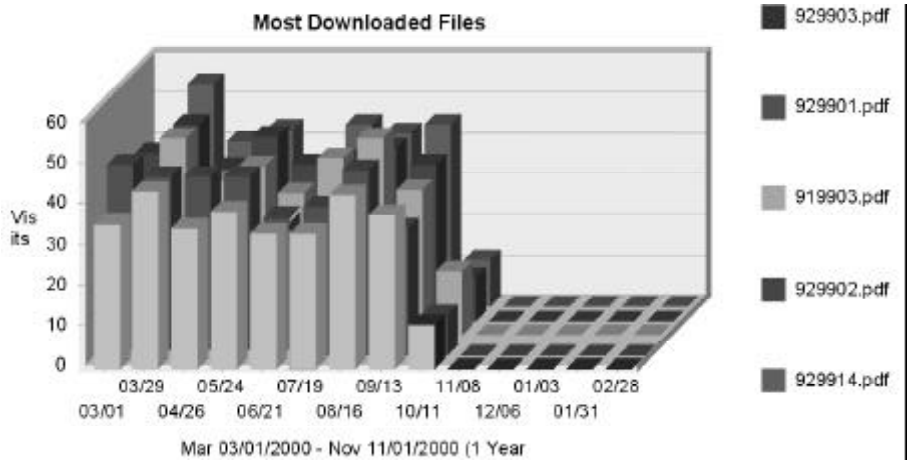
bytes: o tamanho em *bytes* do conteúdo transferido.

A linha apresentada a seguir (figura 1), extraída do arquivo de *log* do servidor *web Apache,* onde está hospedado o periódico eletrônico *Informação & Sociedade: Estudos*, permite-nos apresentar na prática a forma de um arquivo de *log* gerado de acordo com o *Common Logfile Format* .

ACESSO A PERIÓDICOS ELETRÔNICOS EM SITES DA WEB

A partir do exposto, serão mostrados alguns tipos de relatórios que podem ser obtidos no acesso de *sites* da *web* mediante análise dos arquivos de *log* de acesso. Em um primeiro momento, os arquivos de *log* de acessos podem apresentar-se como a solução ideal para a análise do acesso a *sites* da *web*. Contudo, é importante mencionar que os arquivos de *log* de acesso nos oferecem recursos para que sejam realizadas análises apenas de cunho estritamente quantitativo, facilitando a identificação de questões relativas a "o quê", "quando" e "por quem".

FIGURA 2
Revista Informação & Sociedade: Estudos
Arquivos mais baixados entre 1º de março de 2000 a 1º de novembro de 2000



De acordo com Haigh (1998), os dados contidos em um arquivo de *log* de acessos podem ser processados para gerar relatórios, tais como:

– total de arquivos e *kbytes* servidos com sucesso;

– número distinto de endereços IP servidos e número de requisições associadas a cada endereço;

– número de requisições feitas por sufixos de domínios;

– número de requisições para arquivos específicos ou diretórios;

– totalizações e médias por períodos específicos de tempo (horas, dias, semanas, meses, anos);

– URLs visitadas anteriormente pelo usuário (informação indisponível no caso de utilização do Commom Logfile Format).

Para obterem-se relatórios extraídos a partir de um arquivo de *log* de acesso, é recomendável fazer-se uso de uma ferramenta automatizada adequada para tal fim. É possível, contudo, analisar-se o conteúdo de um arquivo de *log* de acesso de forma manual, porém tal procedimento não é recomendável, pois o tamanho do referido arquivo freqüentemente possui milhares de linhas, se for considerado que o *site* hospedeiro do periódico eletrônico estudado seja mediamente visitado. Uma listagem contendo diversas ferramentas automatizadas para a análise de *log* de acesso pode ser encontrada na URL http://www.w3.org/WCA/loganalysis-tools.html.

A figura 2 e as tabelas 1 e 2, a seguir, são exemplos de relatórios obtidos a partir da análise de um arquivo de *log* de acesso. O arquivo de *log* de acesso utilizado foi obtido do *site* da *web* onde está hospedado o periódico eletrônico *Informação & Sociedade: Estudos*. O gráfico e as tabelas mostram o número de vezes que os arquivos associados a artigos mais acessados do periódico foram "baixados" (*downloaded*) no período entre 1º de março de 2000 e 1º de novembro de 2000. A ferramenta utilizada para auxiliar a geração dos relatórios apresentados foi o *software Webtrends Log Analyzer*. Mais informações sobre esta ferramenta podem ser encontradas pela URL http://www.webtrends.com.

PROBLEMAS NA UTILIZAÇÃO DE DADOS EM ARQUIVOS DE LOG DE ACESSO

Serão analisados agora, de maneira detalhada, alguns problemas associados ao uso dos dados contidos nos arquivos de *log* de acesso. Para isso, é necessário que sejam apresentadas algumas definições. Os termos "*hit*" e "sessão de usuário (*user session*)" serão explicados:

– Hit: toda troca de dados realizada entre um cliente e um servidor *web*. Exemplo: um usuário solicita, através de seu navegador, (*browser*) uma página HTML (.html). Neste caso, tem-se um *hit*. Supondo que este usuário acesse uma outra página com três imagens associadas à mesma, tem-se então quatro *hits*, um *hit* para a página HTML e mais três para os arquivos de imagem associados.

– Sessão de usuário (user session): uma sessão de usuário pode ser definida por meio da delimitação de um período de tempo em que ocorrem solicitações ao servidor *web* provenientes de um mesmo endereço IP. Uma sessão é considerada encerrada após determinado período de inatividade, por exemplo, de 30 minutos. As ferramentas usadas para analisar arquivos de *log* de acesso utilizam variações desta seqüência apresentada para determinar uma sessão de usuário.

Com certa freqüência, encontra-se publicado, nos periódicos especializados, o número de *hits* que determinado *site* obteve em um período de tempo. Esta informação é publicada com o intuito de quantificar o número de acessos a determinado *site*. A medição do número de acessos de um *site*, baseada no número de *hits*, não fornece um indicador confiável, pois, de acordo com o exposto na definição de *hit*, uma página consultada uma só vez pode gerar mais *hits* do que uma página que seja consultada várias vezes, mas que gere uma quantidade menor de *hits*. Conseqüentemente, não se recomenda utilizar o número de *hits* como medida para analisar o acesso a periódicos eletrônicos disponibilizados em *sites* da *web*.

O processo de contagem e identificação de sessões de usuários não é preciso, pois não se pode associar com total segurança um endereço IP a um único usuário. No caso de o usuário estar utilizando uma estação de trabalho com endereço IP estático, esta estação, mesmo que esteja sendo utilizada por uma dezena de usuários diferentes, vai apresentar sempre o mesmo endereço IP,

TABELA 1

Revista Informação & Sociedade: Estudos – arquivos mais baixados entre 1º de março de 2000 a 1º de novembro de 2000

Arquivos mais baixados

	Arquivo	Número de *Downloads*	% Sobre o Número de *Downloads*	Visitas
1	929914.pdf	652	5.43%	354
2	929902.pdf	557	4.63%	347
3	919903.pdf	695	5.78%	342
4	929901.pdf	609	5.07%	333
5	929903.pdf	496	4.13%	314
6	919901.pdf	442	3.68%	307
7	919905.pdf	495	4.12%	286
8	919908.pdf	470	3.91%	283
9	929905.pdf	431	3.58%	280
10	929908.pdf	363	3.02%	270
11	929915.pdf	314	2.61%	249
12	929907.pdf	282	2.34%	248
13	929916.pdf	283	2.35%	244
14	929922.pdf	288	2.39%	243
15	919909.pdf	291	2.42%	242
16	929904.pdf	282	2.34%	241
17	929911.pdf	302	2.51%	239
18	929906.pdf	277	2.3%	237
19	919904.pdf	292	2.43%	236
20	929920.pdf	265	2.2%	235
Totais		**8,086**	**67.24%**	**N/A**

TABELA 2

Revista Informação & Sociedade: Estudos – títulos dos artigos mais baixados entre 1º de março de 2000 a 1º de novembro de 2000

Títulos mais baixados

	Arquivo	Nome do Artigo	Nome do Autor
1	929914.pdf	OS DESTINOS DA CIÊNCIA DA INFORMAÇÃO: entre o cristal e a chama	Aldo de Albuquerque Barreto
2	929902.pdf	O PAPEL DOS SERVIÇOS DE INFORMAÇÃO NA TRANSFERÊNCIA DE CONHECIMENTO ENTRE UNIVERSIDADE E INDÚSTRIA: uma análise nacional	Fátima Portela Cyrne
3	919903.pdf	CONHECIMENTO PARA O DESENVOLVIMENTO: reflexões para o profissional da informação	Vânia Maria Rodrigues Hermes de Araújo Iea Maria Freire
4	929901.pdf	WORLD WIDE WEB: aspectos teóricos dos mecanismos de busca	Fernanda Nahuz
5	929903.pdf	A TRANSFERÊNCIA DA INFORMAÇÃO NA EDUCAÇÃO UNIVERSITÁRIA: implicações do uso da oralidade, da escrita e outras tecnologias: metodologia e instrumentos.	Henriette Ferreira Gomes
6	919901.pdf	INTELIGÊNCIA COMPETITIVA: uma abordagem sobre a coleta de informações publicadas	Marilia Damiani Costa Ivanise Alves da Silva
7	919905.pdf	REESTRUTURAÇÃO DE INFORMAÇÃO & SOCIEDADE: ESTUDOS: periódico do Curso de Mestrado em Ciência da Informação da Universidade Federal da Paraíba	Joana Coeli Ribeiro Garcia Maria das Graças Targino
8	919908.pdf	INFOMAÇÃO PARA MUDANÇA SOCIAL	Gilvando Sá Leitão Rios Emeide Nóbrega Duarte Denise Gomes Pereira de Melo
9	929905.pdf	PÓS GRADUAÇÃO PARA BIBLIOTECÁRIOS: educação em permanência	Jomima Marques de Oliveira
10	929908.pdf	PRÁTICAS DE INFORMAÇÃO NO ENSINO DE BIBLIOTECONOMIA	Maria Nilza Barbosa Rosa
11	929915.pdf	A POLÍTICA GOVERNAMENTAL PARA A PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO NO BRASIL	Johana W. Smit
12	929907.pdf	INFORMAÇÃO E VIOLAÇÃO DA CIDADANIA NOS PRESÍDIOS	Josinaldo José Fernandes Malaquias
13	929916.pdf	GLOBALIZAÇÃO E MÉTODO IMPRESSIONISTA	Maria Auxiliadora de Carvalho
14	929922.pdf	CARACTERÍSTICAS DAS DISSERTAÇÕES DE MESTRADO PRODUZIDAS NO CURSO DE MESTRADO EM CIÊNCIA DA INFORMAÇÃO DA UFPB	Marlene de Oliveira
15	919909.pdf	LEITURA EM SALA DE AULA: avatares do desejo ou recalcamento?	Mirian de Albuquerque Aquino
16	929904.pdf	O MOVIMENTO DOS SEM TERRA (MST) COMO ESPAÇO INFORMACIONAL: análise das práticas informacionais desenvolvidas por coordenadores e líderes	Hilda Coutinho Barbosa
17	929911.pdf	TRATAMENTO DO TEXTO JORNALÍSTICO ESCRITO À LUZ DA ANÁLISE DOCUMENTÁRIA: o caso do resumo	Rildeci Medeiros
18	929906.pdf	POLÍTICA DO SILÊNCIO: o fluxo informacional no sistema de arquivos do estado de Pernambuco	Josemar Henrique de Melo
19	919904.pdf	BIBLIOTECA E AÇÃO CULTURAL: apontamentos conceituais a partir da experiência na Universidade Federal de São Carlos	Márcio de Assumpção Pereira da Silva Lígia Maria Silva e Souza Lourdes de Souza Moraes
20	929920.pdf	O BIBLIOTECÁRIO E O MERCADO DA DOCUMENTAÇÃO POPULAR: CPDOs	Luciana Moreira Carvalho

gerando, desta forma, só uma sessão de usuário. De maneira contrária, no caso de se utilizar uma estação que trabalha com endereços IP dinâmicos, tem-se uma variedade de sessões de usuários, quando, na realidade, existe a possibilidade de essas sessões estarem associadas a um só usuário. Portanto, contar e identificar sessões de usuários fornece, apenas, uma estatística aproximada do número de usuários distintos e do número de vezes que os respectivos acessaram o *site* da *web*, hospedeiro do periódico eletrônico.

FIGURA3
Página armazenada na cache do browser do próprio usuário

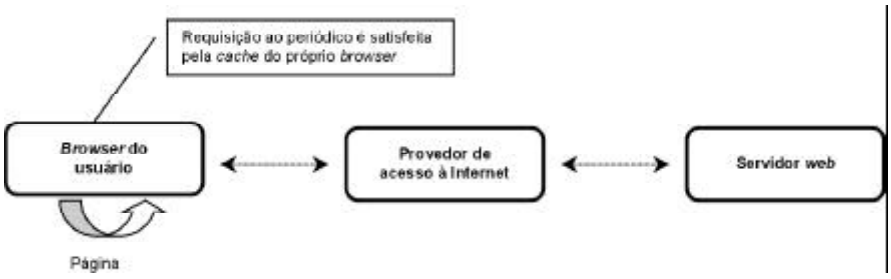
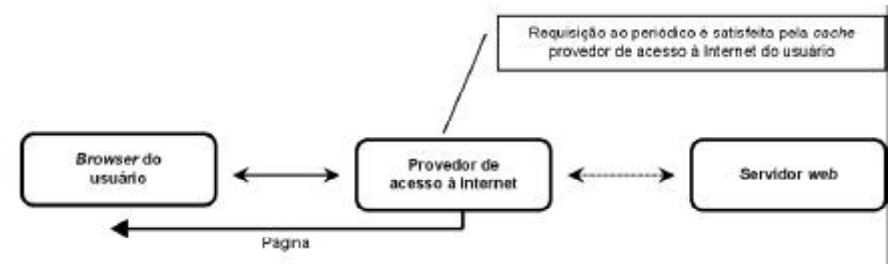


FIGURA4
Página armazenada na cache do browser do próprio usuário



Um evento que também pode afetar o processo de determinar uma sessão, bem como aumentar o número de *hits* no servidor *web*, seria a visita de um *software* do tipo *robot* (espécie de navegador automático) que faz a varredura completa de um *site*. Da mesma forma que um usuário comum, um *robot* também tem suas atividades registradas no arquivo de *log* de acesso. Esse tipo de *software* está normalmente associado a *sites* que disponibilizam ferramentas de busca. Alguns *softwares* para a análise de arquivos de *log* permitem que seja isolado o uso gerado por *robots* (Haigh, 1998), contribuindo, portanto, para reduzir a incidência de erro quando da análise dos arquivos de *log* de acesso.

Para complementar esta explanação sobre os problemas oriundos da utilização de dados armazenados nos arquivos de *log* de acesso, será explicado o conceito de uma *cache* de dados. No jargão da ciência da computação, a *cache* é entendida como uma área onde dados são armazenados de forma temporária. A função primordial de uma *cache* de dados é permitir que usuários tenham acesso à informação de maneira otimizada.

A utilização da *cache* de dados permite que as informações solicitadas pelos usuários sejam recuperadas de maneira mais veloz. Mas, em compensação, pode reduzir a significância das informações contidas nos arquivos de *log* de acesso, pois um usuário pode recuperar um determinado artigo e esta ação pode não ficar registrada no arquivo de *log* de acesso do servidor *web*. No caso de um usuário requisitar um artigo de periódico através de uma URL digitada em seu *browser*, pode-se verificar algumas situações distintas*:

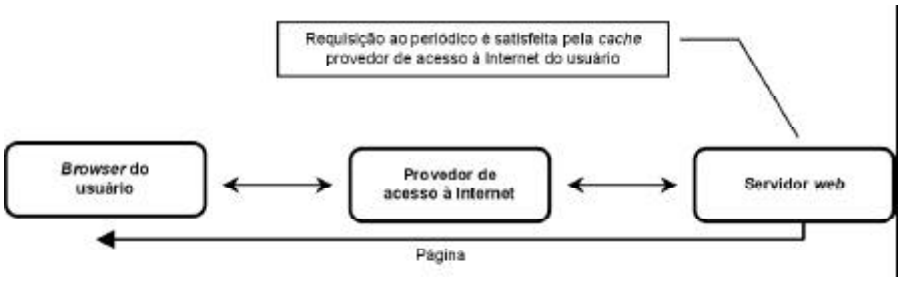
1. A página associada ao artigo de periódico já está armazenada na *cache* local do próprio *browser*. Todo o processo é realizado na própria estação de trabalho do usuário. Conseqüência: o arquivo de *log* de acesso associado ao servidor *web*, ao qual o artigo foi requisitado, nunca vai registrar tal solicitação, pois a requisição já foi satisfeita pela própria *cache* local do *browser* (situação representada na figura 3);

2. A página associada ao artigo de periódico não está armazenada na *cache* local do próprio *browser*, mas na *cache* do provedor de acesso à *internet* do usuário, em virtude de outro usuário já ter requisitado o mesmo artigo de periódico, anteriormente. ***Conseqüência***: o arquivo de *log* de dados associado ao servidor *web*, ao qual o artigo foi requisitado, nunca vai registrar tal solicitação, pois a requisição já foi satisfeita pela *cache* do provedor de acesso à Internet do usuário (situação representada na figura 4);

* Situações de caráter ilustrativo. É possível a existência de outras variantes.

3. A página associada ao artigo de periódico não está armazenada na *cache* local do próprio *browser*, nem na *cache* do provedor de acesso à Internet, mas diretamente no servidor *web*, associado à URL digitada pelo usuário no *browser*. Conseqüência: o arquivo de *log* de acesso associado ao servidor *web* ao qual o artigo foi requisitado vai ter a solicitação registrada (situação representada na figura 5).

FIGURA5
Página é retornada pelo servidor web associado à URL digitada pelo usuário



CONCLUSÃO

Após estas considerações relativas ao uso dos arquivos de *log* de dados, torna-se evidente a necessidade de ter-se bastante cuidado sempre que for necessário gerar análises baseadas nestas informações. Parafraseando Goldberg (2001), é possível concluir que atribuir sentido para informações sem nenhum sentido é pior do que não ter nenhuma informação.

Um assunto importante que deve ser considerado diz respeito à escolha de uma ferramenta automatizada para a análise do *log* de acesso. No momento da escolha de uma ferramenta, é fundamental levantar alguns questionamentos como os seguintes: a ferramenta leva em consideração as linhas do *log* geradas por *robots*? Como esta ferramenta determina uma sessão de usuário? A correta escolha de uma ferramenta automatizada para a análise de *log* é decisiva para ter-se uma idéia próxima da realidade das dinâmicas de acesso a um periódico eletrônico.

Foi visto, no decorrer deste texto, que, mediante análise dos arquivos de *log* de acesso, não se pode ter um perfil completamente preciso do acesso a periódicos eletrônicos hospedados em *sites* da *web*, mas apenas um modelo aproximado do que acontece na realidade, pois esta é uma abordagem quantitativa que não fornece subsídios para endereçar questões de caráter qualitativo, tais como a opinião dos usuários com relação ao conteúdo do *site*, satisfação, usabilidade e os motivos que o levaram a acessar o *site*. Um perfil mais realista, não só de acesso, mas também de uso, deve ser elaborado em conjunto com outras técnicas tais como entrevistas, preenchimento de questionários e grupos focais.

Relato de Experiência aceito para publicação em 07-09-2001.

REFERÊNCIAS BIBLIOGRÁFICAS

LAURIE, Ben, LAURIE, Peter. Apache*: the definitive guide. 2. ed. [S. l.]: Sebastopol: O´Reilly, 1999. 369 p.*

LUTHER, Judy. White paper on electronic journal usage statistics. Journal of Electronic Publishing, *v. 6, n. 3, mar. 2000. Disponível em: <http://www.press.umich.edu/jep/06-03/luther.html> Acesso em : 12 maio 2001.*

GOLDBERG, Jeff. *Why web usage statistics are (worse than) meaningless*. Disponível em: < http://www.cranfield.ac.uk/docs/stats/ >. Acesso em: 23 maio 2001.

*HAIGH, Susan, MEGARITY, Janette. Measuring web site usage: log file analysis. N*etwork Notes, *n.* 57, *ago. 1998. Disponível em: <http://www.nlc-bnc.ca/9/1/p1-256-e.html. Acesso em: 20 maio 2001.*