DIGIRATI

Auditoria eletrônica de web sites de segunda geração (Isee1)

DRAFT

Digirati Informática, Serviços e Telecomunicações Ltda.
Rua do Mercado, 34 sala 1401. Tel.: (21) 2233-5950. CEP 20010-120. Rio de Janeiro / RJ.
www.digirati.com.br . info@digirati.com.br

Introdução

Na auditoria de web sites de primeira geração, a identificação de um usuário é, basicamente, feita através do endereço IP. Além disso, existe uma dependência negativa entre a contagem de páginas solicitadas e a tecnologia de conteúdo empregada pelo auditado. Essas características são reconhecidas como um limite do modelo.

O principal objetivo da auditoria de segunda geração é a identificação avançada de usuário e o fim da influência da tecnologia de conteúdo sobre a métrica de páginas solicitadas. A tecnologia de auditoria de segunda geração criada e desenvolvida pelo IMD (o Instituto de Midia Digital é uma divisão da empresa de tecnologia Digirati especializada em auditoria), com a colaboração e apoio do mercado de mídia interativa, chama-se Isee1 (“I see one”).

Um segundo, e também muito importante, objetivo perseguido pelo Isee1 é ser uma plataforma de auditoria para audiência e campanha publicitária.

O Isee1 utiliza a tecnologia de cookie. Todavia, essa é apenas uma das peças da solução, até porque toda a arquitetura tem abertura para aperfeiçoamentos. Logo, a expressão “identificação por cookies” deve ser evitada. Dá-se preferência à identificação Isee1 ou por Isee1.

Todo o procedimento é descrito de forma incremental. Uma vez que existem inúmeros detalhes envolvidos no processo, cada cenário e seus papéis serão descritos isoladamente.

Requisitos (princípios de projeto)

Toda a proposta atende aos requisitos abaixo:

- **Baixo overhead durante a navegação do usuário.**
  Objetivo: proporcionar um acesso rápido ao usuário.
  Conseqüência: minimizar os artefatos instalados nos web servers e transferir a carga de processamento para o processo de contabilização numérica.

- **Alta estabilidade dos web servers.**
  Objetivo: evitar transtornos para os auditados.
  Conseqüência: os artefatos nos web servers devem ser simples, responsáveis por tarefas elementares; a complexidade deverá ser transferida para o processo de contabilização numérica.

- **O Isee1ID do internauta deve ser comum a todos os auditados.**
  Objetivo: permitir responder perguntas do tipo “quantos visitantes únicos visitaram todos os auditados em um mês e quantos desses, em porcentagem, visitaram determinado auditado?”, além de permitir integração entre parceiros.
  Conseqüência: mais complexidade.

- **Ser implementável por todos os auditados.**
  Objetivo: não privilegiar parte dos auditados, todavia o auditado terá de se submeter a um conjunto de regras.
  Conseqüência: a solução deve ser escalável.

Modelo para auditoria de audiência

Todos os arquivos que devem contar como uma página solicitada devem receber um pequeno código, doravante chamado de **tag.**

Essa tag será responsável por realizar o request ao cgi way.cgi (Who Are You?); esse retornará uma “imagem vazia” (imagem de um pixel transparente) e, se for necessário, enviará um cookie (Isee1Cookie).

Uma vez que a tag é inserida apenas nos arquivos a serem contados como Page Impression (ver tópico sobre métricas), os padrões internacionais são atendidos sem a dura exigência de reestruturar o web site renomeando arquivos e editando links.

Os web servers que são responsáveis pela hospedagem / execução do way.cgi são chamados de **Isee1 servers**. Os Isee1 servers são de propriedade / responsabilidade do IMD. Com isso é garantida a segurança dos arquivos de log, pois somente o IMD tem acesso a eles. Não há acréscimo do uso do link do auditado, nem necessidade de investimentos em hardware, software ou contratação de profissionais especializados.

Todas as métricas de audiência passam a ser obtidas através dos arquivos de log dos Isee1 servers. Os demais arquivos de log não têm serventia para o Isee1.

O software utilizado pelo IMD para realizar o processamento dos dados (audiência e campanha) é o Digirati Auditor.

Modelo para auditoria de campanha

A principal diferença entre a auditoria de audiência e a de campanha está na tag. Essa, aqui, é agregada à peça publicitária e assume duas responsabilidades: realizar o request ao way.cgi para informar o Ad Impresion (ver tópico sobre métricas) e o request para informar o Ad Click.

Os demais recursos são compartilhados pelas duas auditorias: way.cgi, Isee1 servers, Isee1Cookie, Isee1ID. Apenas adaptados aos objetivos de cada auditoria.

ISee1Cookie / way.cgi

O Isee1Cookie terá o domínio .isee1.net, o que garantirá a globalidade do valor entre todos os auditados.

O way.cgi retorna um Isee1Cookie acompanhado de um redirect (HTTP result 302) para o way.cgi sempre que o mesmo não exista, independente do motivo (não aceitar ou ser a primeira vez), ou o Isee1Cookie possuir um Isee1ID inválido. Se o Isee1Cookie existir, o mesmo será gravado no arquivo de log na respectiva entrada da conexão em questão.

Para evitar que o User-Agent, que não aceita cookies, fique permanentemente recebendo redirects, o way.cgi adiciona à lista de parâmetros passados “redirected=1”. Com isso, na segunda vez será possível saber que o User-Agent não aceita cookie e o way.cgi não envia um novo redirect.

O redirect para User-Agents sem identificação agrega à solução duas vantagens: o primeiro acesso do User-Agent que aceita cookie não é perdido; ganha-se uma forma simples de identificar os acessos que não aceitam cookies e quando o usuário “nasceu” (acabou de ganhar um Isee1ID e aceitou).

Os acessos que não aceitam cookie aparecem com um result 200 e Isee1ID ausente (o “redirected=1” também estará na lista de parâmetros, mas é redundante para verificar se o User-Agent aceita cookie). O usuário que nasceu é identificado por um 200, Isee1ID válido e “redirected=1” na lista de parâmetros.

O Isee1Cookie tem a data de expiração em 31 de dezembro de 9999 (sexta - friday) às 23:59:59h GMT, path=/ e domain=.isee1.net.

Para evitar os sistemas de cache, o way.cgi adiciona no cabeçalho do HTTP response as instruções: “Pragma: no-cache\nCache-Control: max-age=0\nExpires: Thu, 01 Jan 1970 00:00:00 GMT\n”.

Sob controle do auditado há dois parâmetros: “redirect” e “StringID”. As respectivas semânticas serão descritas em seus próprios tópicos; nesse será explicada apenas a sua funcionalidade.

O parâmetro “redirect” faz com que, ao invés de ser enviada uma imagem vazia, seja feito um redirect para a URI indicada (valor do parâmetro). Esse parâmetro, se houver, tem de ser, OBRIGATORIAMENTE, o último da lista (permite que o auditado utilize qualquer URI, sem restrições), estar escrito em caixa baixa (por motivos de

otimização) e precedido de um outro parâmetro (por motivos também de otimização). Esse parâmetro só é utilizado pela auditoria de campanha.

Há um comportamento especial a mais: quando o parâmetro “redirect” está presente, o redirecionamento com “redirected”, descrito acima, não ocorre. Isso é devido ao fato desse parâmetro ser utilizado para registrar o click do usuário. Se o usuário já tiver sido identificado anteriormente, ele não o será agora, logo faz-se um ganho de performance.

Na corrente versão, redirect só está disponível para auditoria de campanha (por motivos de desempenho). A URI deve ser absoluta e estar da forma de escapes.

O parâmetro “StringID” registra uma string como identificador de um canal ou banner. Não há distinção de caixa, tanto no nome do parâmetro (“StringID”) quanto no valor. Esse parâmetro é utilizado por ambas as auditorias.

O formato do valor de StringID deve ser o mesmo de um identificador aceito pelas linguagens Pascal ou C/C++, ou seja, iniciar com uma letra ou underscore (“_”) e seguir de letras, underscores e números. Não pode conter espaços. Ver abaixo a definição formal segundo a sintaxe descrita no RFC 2616.

```
UPALPHA    = <"A".."Z">
LOALPHA    = <"a".."z">
ALPHA      = UPALPHA | LOALPHA
DIGIT      = <"0".."9">
IDENTIFIER = (“_” | ALPHA) *(“_” | ALPHA | DIGIT)
```

Tag de audiência

Nos arquivos com código html, o tag nada mais é do que o fragmento de código abaixo (o comentário é opcional) acrescentado no início do arquivo.

```
<!-- Tag para chamada do CGI do Isee1 -->
<img src="http://nome.Auditado.isee1.net/?StringID=canal"
width="1" height="1" border="0">
```

O endereço DNS dos Isee1 servers são no formato: nome.Auditado.isee1.net. “nome” é um nome atribuído pelo IMD (típicamente “a0” para auditoria de audiência); “Auditado” é o nome do auditado. Isso permitirá que o Isee1ID seja global a todos os auditados, pois haverá um Isee1Cookie para o domínio .isee1.net.

O parâmetro StringID é opcional e serve para identificar o canal. Se o mesmo for omitido, será considerado o default (o canal principal do auditado). Os parâmetros

width="1" height="1" têm valores diferentes de zero para evitar problemas com alguns browsers.

É recomendável que a tag seja colocada no início do documento. Isso faz com que o browser carregue rapidamente a mesma e garante que, caso o usuário tente mudar de página antes de encerrar a carga efetiva da página corrente, seu acesso seja registrado.

Os arquivos de conteúdo que se enquadrarem na lista abaixo não devem conter a tag de chamada ao way.cgi:

- **não geram conteúdo html;**

No caso particular de frames (frameset), os arquivos compostos por frames devem incluir uma chamada especial a um arquivo HTML com uma tag de chamada ao way.cgi.

- **popups;**
- **arquivos chamados através de iframe;**
- **arquivos chamados através de ilayer;**
- **arquivos carregados através de chamadas frameset.**

Segue abaixo um exemplo das chamadas acrescentadas nos arquivos de frames.

```
<!-- Arquivo com frame -->
<HTML>
<HEAD>
<TITLE>IMD</TITLE>
</HEAD>

<!-- Incluir as duas linhas abaixo -->
<frameset rows="0,*" border=0>
  <frame src="Isee1.html" name="Isee1">
<!-- fim da inclusão -->

  <FRAMESET ROWS="36,*,45" BORDER="0">
    <FRAME SRC="home_ms.frm" NAME="MenuSuperior">
    <FRAMESET COLS="128,471,*">
      <FRAME SRC="home_ml.frm" NAME="MenuLateral">
      <FRAME SRC="home_ct.frm" NAME="Corpo">
    </FRAMESET>
    <FRAME SRC="rodape.frm" NAME="Rodape">
```

```
<!-- Incluir a linha abaixo -->
</frameset>
<!-- fim da inclusão -->

  <NOFRAMES>
  <BODY>
  <P>Seu browser não aceita frames</P>
  </BODY>
  </NOFRAMES>

  </FRAMESET>

</HTML>
<!-- Fim do arquivo com frame -->


<!-- Arquivo Isee1.html -->
<html><body>
<img src="http://nome.Auditado.isee1.net/?StringID=canal"
width="1" height="1" border="0">
</body></html>
<!-- Fim arquivo Isee1.html -->
```

Tag de campanha

A tag de campanha é um pouco mais elaborada em relação à de audiência, porque a mesma deve registrar o Ad Impression e o Ad Click, métricas originadas de acessos distintos.

Primeiro é adicionada uma tag semelhante a de audiência à peça publicitária. Se a peça inteira consistir de um único banner, serão adicionadas duas tags, a do banner e a de auditoria. Se a peça tiver um arquivo próprio a tag de auditoria será adicionada lá.

Essa primeira tag tem a mesma forma da de audiência e é a mesma para ambas as formas citadas acima, e tem o objetivo de viabilizar a contagem de Ad Impression. A tag (o comentário é opcional):

```
<!-- Tag para chamada do CGI do Isee1 -->
<img src="http://nome.Auditado.isee1.net/?StringID=banner"
width="1" height="1" border="0">
```

O item “nome” no endereço DNS dos Isee1 server é um nome atribuído pelo IMD, típicamente “ad” para auditoria de campanha. “Auditado” é o nome do auditado, na auditoria de campanha é o cliente dono da campanha, em geral uma agência.

O parâmetro StringID, aqui, tem um significado bem diferente e utilizado de forma mais rígida. O mesmo, aqui, *não é opcional* e deve indicar o país onde o veículo publicitário está, o veículo propriamente dito e um identificador do banner (atribuído pelo IMD). Conforme a descrição formal abaixo.

```
COUNTRY    = ALPHA ALPHA
VEHICLE    = 1*ALPHA
IDBANNER   = 1*DIGIT
BANNER     = COUNTRY “_” veiculo “_” IDBANNER
```

Exemplo para o veículo “onon”, hospedando, respectivamente, o banner “128” na Agentina e o banner “127” no Brasil: “ar_onon_128” e “br_onon_127”. Os dois alphas que designam o país é fornecido pelo IMD. Basicamente são os mesmos adotado para a rede de DNS.

A segunda e última tag a ser adiciona é responsável pela contagem do Ad Click. A mesma é apenas uma reescrita do link para onde o usuário será encaminhado quando clicar na campanha. Suponhamos que a URI de destino seja “http://www.axe.com.br”, reescrito para a agência “Agg” fica:

```
http://ad.agg.isee1.net/?StringID=ar_onon_128&redirect=http:%2F%2Fwww.axe.com.br%2F
```

Segue uma pequena listagem de escapes para consultas rápidas.

Caractere	Código		Caractere	Código
“/”	%2F		“?”	%3F
“=”	%3D		“&”	%26

Isee1ID

O Isee1ID é uma string com três campos, todos codificados em base 16 para se ter máxima velocidade de conversão por parte da máquina e legibilidade imediata por humanos. O Isee1ID é em caixa baixa (só utiliza os caracteres “0” ... “9” e “a” ... “f”).

O primeiro é um campo de 16 bits (4 caracteres) representando a tecnologia empregada para a geração do Isee1ID. Esse documento só define o número 0. Esse campo é fundamental para a expansão futura da tecnologia Isee1.

Um exemplo de implementação futura que fará uso desse primeiro campo é um plug-in para ser instalado nos browsers com o intuito de emular um cookie permanente na máquina do usuário. O que simplificaria todo o processo, ainda que nem todos os usuários tivessem o plug-in. Esse poderia ser anexado aos kits de acesso distribuídos pelos filiados.

O segundo campo é um identificador de unicidade de 128 bits (32 caracteres). Esse identificador pode repetir se o primeiro campo for diferente, e é obtido com a aplicação da função de one-way hash MD5 sobre um conjunto de dimensões de unicidade. O conjunto utilizado é determinado pelo primeiro campo do Isee1ID.

O conjunto de dimensões deve ser único no usuário; não é necessária a unicidade no tempo e/ou no espaço. Para o primeiro campo igual a 0, as dimensões são o endereço IP e a porta do User-Agent e a data (dia, mês, ano, hora, minuto e segundo) do request feito ao way.cgi.

A função one-way hash garante o “anonimato” do User-Agent. O objetivo é diferenciar os usuários uns dos outros e não, identificar.

O último campo é um crc (POSIX.2 checksum) 32 bits (8 caracteres) dos campos anteriores. Esse permitirá validar o Isee1ID.

Na criação do Isee1ID a data é posta em formato “aaaa-mm-dd:hh:mm:ss” (ex: “2001-06-15:13:01:50”) usando caracteres ASCII e hora em GMT para compatibilidade entre plataformas distintas.

O way.cgi, para validar o Isee1ID só verifica o crc e não a versão do Isee1ID, isso garante compatibilidade imediada com novas formas de geração do Isee1ID. O responsável por conhecer as versões existentes do Isee1ID é o Digirati Auditor.

Métricas de audiência

As métricas de audiência são computadas somente sobre entradas válidas dos arquivos de log dos Isee1 servers.

User. Quantidade de usuários identificados pelo Isee1. Equivalente à métrica Visitor (by Unique Cookie) do IAB e Unique User do IFABC.

Time-spent. Tempo médio gasto por um usuário. Equivalente à métrica Average Time (Per Visitor by Unique Cookie) do IAB e Unique User Duration do IFABC.

Page Impression. Quantidade de páginas vistas. Ou, a quantidade de entradas válidas nos arquivos de log dos Isee1 servers. Equivalente à métrica Page Request do IAB e Page Impression do IFABC.

Alcance (web site). Porcentagem de Users que acessaram o canal, no mês, em relação aos Users que acessaram o web site a que pertence o canal.

Alcance (auditados). Porcentagem de Users que acessaram, no mês, em relação ao Users de todos os auditados pelo IMD.

O início da contagem do time-spent é o primeiro pedido ao web site; o término da sessão é o último acesso feito. O tempo que o usuário fica na última página não será contabilizado por não existir no log. Intervalos de acesso superiores a 30 minutos caracterizam uma nova sessão. Ou seja, a contagem de tempo anterior ao grande intervalo já havia sido encerrada e será iniciada uma nova contagem.

O User e Time-spent são métricas atômica. User é computado no intervalo de uma hora, um dia e um mês. Time-spent é computado apenas no intervalo de um mês e um dia. Esse último intervalo é utilizado apenas para controle, e não é divulgado por apresentar razoáveis distorções devido ao acesso contínuo ao web site. Essas distorções são eliminadas no intervalo de um mês por ser uma faixa maior de tempo.

Considerar apenas os usuários identificados acrescenta um imenso valor ao relatório de auditoria. A métrica Time-Spent não sofrerá perturbações quanto a NAT (Network Address Translator) ou DHCP (Dynamic Host Configuration Protocol) e a taxa Page Impression por Users (utilizada pelas agências) será extremamente realista. É conhecido que, pelo menos 80% dos User-Agents, aceitam cookie.

As entradas dos arquivos de log que atenderem a todos os itens listados abaixo são consideradas válidas:

- **Ser do método GET;**
- **Ter status 200;**
- **Referenciar o way.cgi;**
- **Não ter endereço IP do User-Agent com origem do IMD ou interno ao auditado;**
- **Possuir o Isee1ID válido;**
- **O campo reference deve indicar um domínio sobre o qual o auditado é proprietário da audiência;**
- **Ter um User-Agent de consulta (browsers);**

- **Referenciar uma URL diferente da anteriormente acessada pelo usuário se não for o primeiro acesso (evitar contagens por refreshs).**

Métricas de campanha

As métricas de campanha são computadas somente sobre entradas válidas dos arquivos de log dos Isee1 servers.

Ad Impression. Quantidade de peças publicitárias vistas. Ou, a quantidade de entradas válidas nos arquivos de log dos Isee1 servers de campanha para Ad Impression. Equivalente à métrica Ad Request do IAB e Ad Impression do IFABC. Como há usuários que não são identificados, Ad Impression é apresentado por usuários identificados e não identificados.

Ad Click. Quantidade de Ad Impression clicado por um usuário. Ou, a quantidade de entradas válidas nos arquivos de log dos Isee1 servers de campanha para Ad Click. Equivalente à métrica Click do IAB e Ad Click do IFABC. Como há usuários que não são identificados, Ad Click é apresentado por usuários identificados e não identificados.

Click Rate. Porcentagem de Ad Clicks sobre Ad Impressions. Equivalente à métrica Click Rate do IAB. O IFABC não denomina essa métrica.

User. Quantidade de usuários identificados pelo Isee1. Equivalente à métrica Visitor (by Unique Cookie) do IAB e Unique User do IFABC. Essa métrica é apresentada tanto para Ad Impression quanto para Ad Click isoladamente.

As entradas dos arquivos de log que atenderem a todos os itens listados abaixo são consideradas válidas para Ad Impression:

- **Ser do método GET;**
- **Ter status 200;**
- **Referenciar o way.cgi;**
- **Não ter endereço IP do User-Agent com origem do IMD ou interno ao auditado;**

As entradas dos arquivos de log que atenderem a todos os itens listados abaixo são consideradas válidas para Ad Click:

- **Ser do método GET;**

- **Ter status 302;**
- **Referenciar o way.cgi;**
- **Não ter endereço IP do User-Agent com origem do IMD ou interno ao auditado;**

Extras

Os itens dessa seção são extras, no sentido do aperfeiçoamento da plataforma Isee1, sem alteração de qualquer semântica descrita anteriormente.

Para evitar que os mecanismos de busca indexem o domínio isee1.net é disponibilizado o arquivo "/robots.txt" em todos os subdomínios. O conteúdo do mesmo é listado abaixo.

```
# Tells Scanning Robots Where They Are And Are Not Welcome
#
# User-agent:  can also specify by name; "*" is for
everyone
# Disallow:    disallow if this matches first part of
requested path
#
# For now disallow all we can modify this as needed to
allow certain crawlers.
#

User-agent: *
Disallow: /
```

A plataforma P3P versão 1.0 já se encontra em estado "Candidate Recommendation" pelo W3C. A especificação já é suportada pelo Microsoft Internet Explorer 6.0 e espera-se a adesão de outros desenvolvedores. O Isee1 já suporta o P3P.

Pauta do comitê gestor das normas do IMD

Avaliar a margem de erro do uso, pelo IMD, da base de dados do site http://ip-to-country.com/database/ em uma análise demográfica no processo de Auditoria. O erro pode não ser os meros 2% indicados pelo ip-to-contry devido a regionalização da America Latina.

O Comitê entende que esses números não podem entrar no relatório comercial, IAB Compliant, por não ter o endosso Internacional. Todavia, em um relatório interno,

nos moldes do relatório técnico, poderia conter essa informação bem como sua margem de erro estimada.

Referências

- **W3C - World Wide Web Consortium (http://www.w3c.org):**
  Hypertext Transfer Protocol -- HTTP/1.1, RFC 2616;
  Platform for Privacy Preferences (P3P) Project
  (http://www.w3.org/P3P/);
  Extended Log File Format (http://www.w3.org/TR/WD-logfile);
  Session Identification URI (http://www.w3.org/TR/WD-session-id).

- **MD5, Livro de criptografia - APPLIED CRYPTOGRAPHY – Protocols, Algorithms, and Source Code in C, Schneier, Bruce, USA, John Wiley & Sons Inc., 1996, ISBN 0-471-11709-9.**

- **Livro de cookies. RFC de cookies - COOKIES, St. Laurent, Simon, São Paulo, Ed.Berkeley, 1999, ISBN 85-7251-503-8.**

- **The MD5 Message-Digest Algorithm, RFC 1321.**

- **Base 64, ver RFC 1421.**

- **Persistent Client State HTTP Cookies,**
  http://home.netscape.com/newsref/std/cookie_spec.html.

- **HTTP State Management Mechanism, RFC 2109.**

- **IFABC - International Federation of Audit Bureaux of Circulations (http://www.ifabc.org):**
  Web Measurement Standards (http://www.ifabc.org/web/index.html).

- **IAB - Interactive Advertising Bureau / Internet Advertising Bureau (http://www.iab.net):**
  THE MEDIA MEASUREMENT TASK FORCE
  (http://www.iab.net/advertise/content/mmtf3.html);
  IAB's Glossary of Interactive Advertising Terms
  (http://www.iab.net/main/glossary1.htm).