SISTEMA AVALIADOR DE SITES E-COMMERCE PARA AUXÍLIO À REALIZAÇÃO DE COMPRA *ONLINE* COM SEGURANÇA

Natascha Sava Hun[1]; Renata Corrêa Pimentel[2]

Resumo

Com o decorrer dos anos, o computador pessoal e a banda larga têm se tornado cada vez mais acessíveis à população, atingindo diferentes classes sociais e, conseqüentemente, proporcionando aumento do comércio eletrônico, ou seja, compra e venda através da Internet. Além deste tipo de comércio (também conhecido como e-commerce), um novo grupo aumenta gradativamente: o grupo dos e-consumidores. Estes "consumidores eletrônicos", ao realizar suas compras pela Internet, desejam naturalmente efetuar suas transações sem nenhum tipo de risco no que se refere tanto a segurança das transações quanto da confiabilidade do lojista, pelo fato de ocorrerem remotamente (ao contrário de uma loja física). Através de técnicas de raspagem de dados da Web (também conhecidas como WebScraping), o intuito deste projeto é proporcionar ao e-consumidor uma consulta às informações do lojista para avaliar sua credibilidade, porém de forma centralizada e ágil: um website para avaliação de lojistas virtuais a partir de informações de outros websites-fonte, considerados atualmente como referências em registro de opiniões de usuários que já tiveram a experiência de compra na loja em questão consultada e registram publicamente na Internet suas satisfações ou reclamações; além da verificação do uso de criptografia pelo lojista e de utilização de serviços de auditoria. Desta forma, todas as informações até então encontradas em consulta a diversos sítios da rede estarão concentradas e disponíveis em um único local, para consulta rápida e centralizada, auxiliando os usuários da Internet na tomada de decisão de compra em determinado site e-commerce, bastando somente informar seu endereço virtual antes de realizar a compra.

Palavras-chave: Comércio eletrônico, e-Consumidor, Raspagem de Dados

Abstract

Over the years, the personal computer and broadband are becoming increasingly accessible to the public, reaching different social class and therefore providing increased electronic commerce, ie buying and selling via the Internet. Besides this type of commerce (also known as e-commerce), a new group gradually increases: the group of e-consumers. These "electronics consumers" when making their purchases over the Internet, naturally want to make transactions without any risk in terms of both transaction security and reliability of the shopkeeper, because they occur remotely (as opposed to a physical store). Currently, for this kind of customer to verify the credibility of the retailer before finalizing a purchase online, is possible consult various registration post-sale sites or audit and verify postage stamps adoption of transaction security for the retailer, however queries various sources may be needed, which requires time to the consumer. Through techniques of data web scraping (also known as WebScraping), the aim of this project is to give consumers information's consultation of

[1] Graduada em Engenharia da Computação pela Faculdade de Engenharia de Sorocaba-FACEN**S** e analista Marketing e-Commerce do gGrupo SBF, e-mail: natascha.sh@gmail.com.

[2] Mestre em Engenharia da Computação pela Universidade Estadual de Campinas-UNICAMP, professora da Faculdade de Engenharia de Sorocaba-FACEN**S,** e-mail: renata@facens.br.

the merchant, but in a centralized and flexible: a website for evaluation of virtual store from of other websites-source, currently regarded as references in recorded reviews from users who have had the experience of shopping at the store in question referred, other than testing the use of encryption and the use of audit services. Thus, all information previously found in several places will be concentrated and available in a single location for quick and centralized search, helping users of the Internet in making purchasing decisions in a e-commerce website, just simply writting your virtual address.
Keywords: e-commerce, e-consumer, WebScraping

1 Introdução

O comércio eletrônico teve início juntamente com a popularização da Internet, em meados de 1995, período em que a empresa Netscape Communications lança seu navegador *Web* e o protocolo *Secure Socket Layer* (SSL), que proporcionou maior segurança das transações virtuais ao criptografar as informações que trafegavam na rede, sendo utilizado até hoje. Como exemplo de primeiras empresas que notaram neste meio a oportunidade de comercializar seus produtos com menores custos e vinte e quatro horas por dia, é possível citar a Amazon.com e eBay Inc., ambos fundados em 1995. No Brasil, o site Submarino.com.br é pioneiro em comércio eletrônico, realizando suas vendas 100% eletrônicas (sem loja física), desde 1999.

Constituir ou manter uma loja virtual é o foco da maioria dos atuais lojistas físicos ou ainda de empreendedores que encontram na Internet uma maneira mais prática e de menor custo de comercializar seus produtos quando comparada à montagem de um comércio físico, em razão das vantagens que a Internet proporciona (eliminação de custos com estrutura física e vendedores; custo reduzido de divulgação e propaganda, etc.). Segundo o presidente da Câmara Brasileira de Comércio Eletrônico Manuel Matos, em entrevista à revista Istoé Dinheiro (MELO, 2011), o *e-commerce* no Brasil está em expansão e aos poucos se popularizando. Consequentemente, um novo grupo aumenta gradativamente: o dos "e-consumidores", ou seja, os clientes das lojas virtuais.

Infelizmente, na rede mundial de computadores há usuários mal intencionados que agem criminalmente a partir de seus conhecimentos mais aprofundados em computação ou, no âmbito do comércio eletrônico, podem prejudicar outros usuários sem mesmo possuírem conhecimento técnicos em questão, como por exemplo, simplesmente não enviar ao cliente um produto adquirido já pago. Essas questões geram em alguns usuários da Internet, principalmente iniciantes, o receio em realizar ou não uma compra virtual. O medo de ser lesado de alguma forma pode inibir muitos usuários

a realizá-la, mesmo que o risco possa ser baixo ou até mesmo nulo dependendo do lojista e do computador utilizado. Atualmente, estes consumidores necessitariam consultar diferentes sites especializados em pós-venda virtual quando desejarem se certificar da credibilidade do lojista antes de efetuar a compra, demandando tempo ao consumidor. Essas consultas são realizadas principalmente por novos e-consumidores, que não possuem ainda o hábito de comprar pela Internet, ou por usuários que já possuem o hábito, porém desejam efetuar uma compra em *websites* de menor popularidade, sendo necessário averiguar sua confiabilidade.

A Internet é uma rede rica em informação e serviços, porém muitas vezes dispersos em diversas fontes. A coleta de informações na rede mundial para auxílio na tomada de decisão pode exigir tempo e paciência para alcançar a conclusão certa a partir dos dados levantados. Atualmente, desconhece-se haver um meio rápido e centralizado de extrair informações sobre uma loja virtual, sem ter que realizar consulta a diversos sites para auxiliar na decisão de compra em determinada loja. Portanto, comprar pela Internet significa ter alguns cuidados extras os quais poderiam ser desconsiderados quando se realiza uma compra em loja física, pelo simples fato de não se ter na *Web* o mesmo contato com o produto, loja e vendedor. O risco de não receber o atendimento esperado é maior, assim como a dificuldade de exigir seus direitos em caso de eventuais problemas com a compra. Com opiniões de consumidores anteriores e averiguação da adoção de medidas de segurança, a decisão em comprar ou não comprar poderá ser tomada com maior certeza, e este é o foco deste sistema: de forma rápida, prática e centralizada fornecer informações relevantes a decisão de compra em *e-commerce*.

Este sistema possibilita ao usuário, de forma simples e prática (sem cadastros ou *logins*), que seja informado somente o endereço da loja virtual e, a partir de então, através de técnicas de raspagem de dados da *Web* (também denominadas como *WebScraping*) a realização automática de varredura em *websites*-fonte pré-determinados focados em registro de opiniões de usuários que já realizaram compra(s) na respectiva loja virtual, além da verificação de adoção de auditoria de segurança e uso de protocolo criptográfico nas transações, exibindo os resultados encontrados ao usuário. A abrangência deste sistema é em âmbito nacional, tanto para as lojas virtuais quanto para as fontes de informação.

Para que seja possível a realização precavida deste projeto quanto às questões legais, no que se refere ao uso de informações provenientes de terceiros, além da certificação da liberação constante nas páginas de "Termos e Condições" das fontes,

este projeto deverá ter ainda respaldo em comunicar via e-mail o acesso aos dados em seus sites da *Web*. Embora a Internet seja capaz de tornar qualquer informação pública mundialmente e rapidamente, há ainda preservação da informação e da fonte conforme o dado que está sendo veiculado na rede, protegido por direitos autorais e leis de propriedade intelectual. É preciso cuidado no uso das informações, principalmente se vierem a distorcer a autoria dos fatos publicados conforme o destino dado às informações coletadas. Comumente, essa questão é clara e objetiva em página normalmente denominada "Termos e Condições de Uso" (sigla em Inglês "TOS") disponibilizada pela maioria dos grandes sites da *Web*, onde a utilização de qualquer técnica de extração, raspagem, mineração ou indexação é declarada como autorizada ou proibida. Mesmo que haja a aprovação do uso de alguma técnica, é preciso certificar-se para qual fim há liberação de uso dos dados para que não ocorra qualquer surpresa futura, como solicitações de remoção do site da *Web*, bloqueio de IP ou ainda questões judiciais. (TURLAND, 2010).

Além dos cuidados com direitos autorais, é preciso ética no volume e periodicidade da extração de dados, para evitar que de alguma forma a fonte de informação seja prejudicada por um possível acesso excessivo, por exemplo. O ideal é a extração somente dos dados desejados, para facilitar a avaliação dos mesmos evitando possíveis erros na análise, consultas desnecessárias ao site-fonte ou ainda bloqueio pela origem em razão de quantidade abusiva de consultas, mesmo estando liberado nos Termos e Condições de Uso. Determinar quais "*tags*" (ou identificadores) contêm a informação pode auxiliar em uma busca direta, evitando que todas as marcações da linguagem da página sejam consultadas em busca do dado desejado. Além da freqüência de consulta ao site-fonte, aplica-se também bom senso na quantidade de páginas acessadas ou baixadas, evitando, por exemplo, *download* ou acesso à arquivos desnecessários. Uma boa prática é o armazenamento do conteúdo ou das páginas em caso de repetidos acessos em curto espaço de tempo, agilizando a consulta e evitando inclusive falta de informações em momentos em que o site-fonte esteja indisponível, já que os dados foram há pouco tempo extraídos e salvos. (HEMENWAY; CALISHAIN, 2004).

As seções seguintes visam explanar as técnicas de extração de dados e, de forma mais detalhada, a técnica de *WebScraping* utilizada neste projeto, assim como o desenvolvimento, características específicas observadas, resultados obtidos, publicação e aperfeiçoamentos futuros.

2 Raspagem de Dados

A rede mundial de computadores permite que todas as pessoas usuárias não somente tenham acesso à informações globais, mas também à publicação dessas informações. A Internet tornou-se um meio prático, barato, acessível e rápido de expor opiniões em âmbito mundial, como por exemplo, demonstrar satisfação ou insatisfação sobre algo ou alguém de maneira totalmente pública e disponível. Desta forma, qualquer pessoa pode acessar as satisfações ou insatisfações do mesmo serviço ou produto, podendo determinar ou não sua decisão final de adquirir o mesmo, a partir das informações encontradas na Internet. Assim como essas informações estão disponíveis para leitura por qualquer usuário, podem ser extraídas por um software, algoritmo ou outra página da Web que desejar utilizá-las. Isso significa que, a partir do momento que uma informação é lançada na Internet, estará disponível para qualquer usufruto, desde que sempre observados os direitos legais do autor, se houver, e ética no destino das informações.

Para coletar informações disponíveis na *Web*, há diversas técnicas e terminologias de extração de dados conforme a maneira que se deseja extrair, todas abrangidas pelo termo "*Data Scraping*". As técnicas podem variar de acordo com o uso ou não de programas de auxílio na coleta de dados, manutenção da correspondência das informações se houver alguma alteração no site fonte de dados, localização do dado na página em linguagens de marcação, imagens, animações, *Really Simple Syndication* (RSS), por exemplo. Segundo Hemenway & Calishain (HEMENWAY; CALISHAIN, 2003), se a mesma informação estiver disponível de várias formas, o caminho ideal será sempre o mais simples e de formato mais estruturado, além de se considerar também em qual meio há maior possibilidade de estar sempre atualizado e disponível para raspagem de dados, evitando falhas na extração. Os dados podem inclusive estar disponíveis em *WebServices*, *Application Programming Interface* (API) ou microformatos, e sempre devem ser considerados e consultados se são disponibilizadas pelo site fonte, de modo a facilitar a coleta de dados. *WebServices* e API's garantem informações atualizadas e, principalmente, são um meio de permissão de uso de dados pelo autor.

Um dos termos para extração de dados é "*Web Mining*", variante do termo "*Data Mining*", ou Mineração de Dados: representa mineração de dados, porém focado na *Web* como fonte. Abrange desde a busca do site-fonte a partir de termos-chave, análise e

interpretação dos dados, a até a descoberta de padrões entre sites da *Web* através de algoritmos com conceitos de inteligência artificial. (AMO, 2003).

É possível ainda referir-se à extração de dados como "*Web Crawling*", "*Web Wrapping*", "*Web Spidering*", "*Screen Scraping*" ou "*Web Scraping*". Os termos são resultantes dos programas agentes utilizados para realização completa da coleta e análise de dados, desde a busca da fonte até localização da informação desejada na página *Web*.

Crawlers ou *Spiders* são denominações dadas aos programas responsáveis em buscar na *Web* as páginas que contêm o conteúdo almejado, gerando a denominação *Web Crawling* ou *Web Spidering*, podendo portanto serem alvos diretos de uma ação por parte do site-fonte se vierem a agir de forma nociva ao mesmo, como coleta de dezenas de páginas por segundo. Sites de busca possuem programas *Spiders*, por exemplo, com a finalidade de indexar páginas da *Web* para composição de seus resultados de pesquisa, mas também com ética na freqüência de acesso e até mesmo com a possibilidade do *Webmaster* permitir ou não a ação do *Spider* em determinadas páginas, através de um arquivo tipo texto nomeado como "robots.txt" no diretório raiz do site, com finalidade de informar à *Spiders* quais limites devem ser seguidos.

Após a localização da(s) página(s) com conteúdo necessário com auxílio de um *Crawler*, estas são repassadas aos programas *Wrappers* (*Web Wrapping*) que, segundo Lage et al. (2002), podem ser denominados como: "(...)*Web Wrappers*, programas que extraem dados não estruturados a partir de páginas *Web* e os armazena em formatos adequados, tais como XML (...)" (traduzido pela autora). Este tipo de programa tem a finalidade de encontrar a informação desejada em meio a uma página da *Web* e retorná-la de forma estruturada. Utilizando um *Wrapper* ideal, há retorno sem nenhuma informação irrelevante, além da coleta informações em diferentes formatos e em diferentes arquivos, não especificamente para uma única e determinada página. Existem, inclusive, programas reparadores de *Wrappers*, para garantir que caso haja alguma modificação na página, ainda sejam capazes de localizar a informação desejada.

Screen Scraping, como nome já diz, refere-se à captura de informações a partir de uma saída em vídeo. O programa responsável pela captura chama-se *Screen Scraper* e comumente utiliza mecanismo de *Optical Character Recognition* (OCR), responsável por converter informações de imagens em texto editável ou extração em controles gráficos de aplicações *Graphical User Interface* (GUI).

O termo *WebScraping* envolve a coleta de dados sem distinguir e determinar padrões semânticos como ocorre em *Web Mining*, focando apenas em obter os dados. Conhecimentos de como trabalha o protocolo HTTP são necessários para que seja possível a recuperação de documentos que contém a informação a ser extraída, principalmente quando houver necessidade de requisições GET ou POST. Para facilitar a varredura na linguagem de marcação em busca do dado requerido e eliminando o conteúdo indesejado, recursos como *XML Path Language* (XPath) e Expressões Regulares (Regex) são utilizados. XPath compõe expressões que contêm o caminho a nós em documentos de marcação; porém, em algumas situações, o uso de XPath pode não retornar resultados ou retorná-los não como esperado, conforme a estrutura do documento. Esse problema pode corrigido com o uso de expressões regulares. Regex permitem definir uma sintaxe para validação dos dados obtidos, mas requer acompanhamento a longo prazo caso haja modificações na página-fonte, já que expressões regulares não foram criadas exclusivamente para esta finalidade. (TURLAND, 2010).

Para este projeto, optou-se em aplicar os conceitos de *WebScraping*. Portanto, o desenvolvimento baseou-se na utilização de conceitos e funcionalidades do protocolo HTTP descritos na RFC2616 (LAFON, 2011), além de recursos como XPath e Regex citados anteriormente.

Segundo Fielding (FIELDING et al., 1999), HTTP é um protocolo em nível de aplicação para sistemas de informação de hipermídia colaborativos e distribuídos. Para identificar ou referenciar um recurso disponível na Internet, utiliza-se seu endereço, conhecido como *Universal Resource Identifier* (URI): seqüência compacta de caracteres que identifica um recurso físico ou abstrato, permitindo uma aplicação analisar os componentes comuns de uma referência URI sem conhecer os requisitos do esquema específico de cada possível identificador. (BERNERS-LEE; FIELDING; MASINTER, 2005). Conforme a RFC2616, uma requisição de um cliente para um servidor inclui, dentro da primeira linha da mensagem, o método a ser aplicado ao recurso identificado pela URI, o identificador do recurso e a versão do protocolo em uso. Atualmente, são oito métodos ao todo especificados na RFC2616, sendo os métodos HEAD, GET e POST utilizados pela técnica de *WebScraping*. A resposta de uma requisição HTTP é composta por diversos segmentos e, um dos segmentos da resposta é denominado "*Status Code*", importante para indicar a disponibilidade do site em que se deseja coletar informação.

A linguagem XPath, segundo a W3C - *World Wide Web Consortium* (CLARK; DEROSE, 1999) - *XML Path Language* tem a finalidade de endereçar partes de um documento XML, oferecendo também recursos básicos para manipulação de string, números e valores booleanos através de uma sintaxe compactada (não-XML). XPath possui esse nome por utilizar notação de "caminho" (*path*), como uma *Uniform Resource Locator* (URL), para navegar através da estrutura hierárquica de um documento XML. XPath é projetado para localizar e processar os itens dentro de Documentos XML ou HTML formatados corretamente (HEMENWAY; CALISHAIN, 2003), utilizando uma expressão composta por barras ("/") para localizar hierarquicamente as tags (identificações) e, conseqüentemente, seus conteúdos. Através de XPath, é possível construir expressões numéricas, expressões de igualdade, relacionais e boolenas a partir de operadores matemáticos e comparativos. É possível ainda utilizar funções pré-definidas disponíveis para cadeias de caracteres, números, valores booleanos e manipulação de nós ou movimentação entre nós-pai e nós-filho. (CASANOVA, 2003).

Expressões Regulares, conforme Jargas (JARGAS, 2009), é uma maneira versátil de busca de texto por permitir a procura de conteúdo em que não o se conhece exatamente, mas sabe-se suas possíveis variações; procura de texto em posições específicas em uma linha; ou ainda procura de conteúdo que atenda determinados padrões. De acordo com Turland (TURLAND, 2010), em alguns casos os documentos de marcação estarão mal formados, não permitindo que seja aplicável a utilização de extensão XML, como o XPath citado acima, ou os resultados podem conter menos ou diferentes dados do que se esperava. Casos em que se queira verificar se os dados extraídos por extensão XML são realmente os desejados. Estas tarefas podem ser realizadas com as funções básicas de manipulação e operadores de comparação de texto, mas na maioria dos casos a aplicação seria confusa e não confiável. Expressões Regulares fornecem uma sintaxe que consiste de meta caracteres onde padrões de strings são expressos de maneira flexível e concisa.

3 Desenvolvimento

Para o desenvolvimento deste projeto, utilizou-se a linguagem PHP (*Hypertext PreProcessor*), por ser uma linguagem livre atualmente considerada uma das cinco linguagens mais utilizadas (TIOBE SOFTWARE, s.d.). e que possui funções e classes

capazes de realizar as técnicas de *WebScraping*. A interface desenvolvida em *eXtensible Hypertext Markup Language* (XHTML), por permitir validação de código pela W3C e ser atualmente interpretado inclusive por navegadores em versões anteriores; e *Cascade StyleSheet* (CSS) para organização dos estilos de formatação aplicados. Para armazenamento das informações, utilizou-se banco de dados MySql, por ser também livre como a linguagem PHP.

Este projeto possui âmbito nacional e, portanto, as fontes de informação foram pré-determinadas a partir de pesquisa na Internet à sites registradores de satisfação de compra e de auditoria. O único resultado exceção a essa pesquisa de fontes refere-se à adoção de protocolo criptográfico (SSL), o qual é verificado na página do lojista.

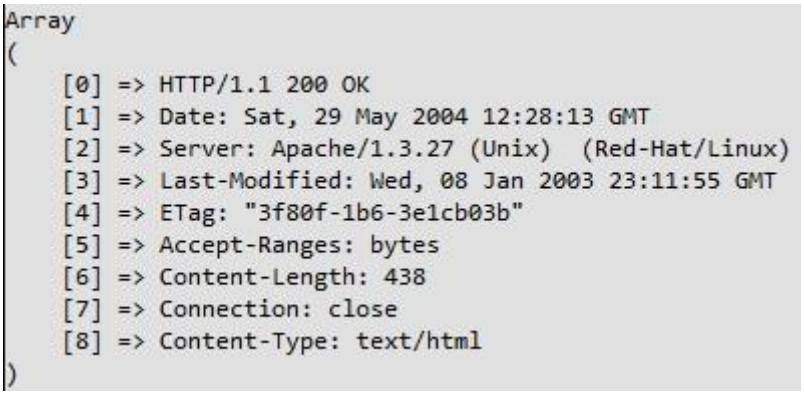
Conforme citado anteriormente sobre ética no consumo das informações de terceiros, o intuito deste projeto é não extrair informações com alta frequência pela razão da improbabilidade da situação da loja ser modificada em curto espaço de tempo nos sites-fonte. Salvar os resultados em banco de dados evitará consultar diversas vezes informações coletadas há pouco tempo, além de agilizar a exibição do resultado ao usuário quando for consultado em um banco de dados ao invés de uma página da *Web*. Também na hipótese do site-fonte estar indisponível, ter informações registradas garantem uma resposta ao usuário (se forem recentes).

Inicialmente, os dados do usuário deste *website* são salvos em banco de dados, registrando seu IP, data e hora do acesso realizado. Ao receber a URL informada pelo usuário, o sistema extrai somente o endereço raiz da loja, descartando qualquer *query string* contida na URL por não ter finalidade neste projeto, caso o usuário informe endereços compostos pelo domínio principal seguido de passagem de parâmetros.

Com somente o domínio principal da loja, é verificado se corresponde a um endereço de domínio válido, através da utilização de expressão regular. A linguagem PHP possui duas bibliotecas para interpretação de expressões regulares, POSIX e PCRE, sendo a biblioteca PCRE utilizada neste projeto para a verificação e busca através de regex, por possui funções mais poderosas e rápidas. (JARGAS, 2009). Se a URL não for considerada válida, o usuário é notificado a informar o endereço correto do lojista. Caso contrário, com uma URL validada, antes da verificação da reputação da loja, seu endereço virtual é consultado se realmente existe na *Web*, através do resultado da resposta HTTP.

Em PHP, uma maneira de consultar o cabeçalho de resposta enviado pelo servidor após uma requisição HTTP é através da função *get_headers()*, a qual retorna um *array*

que pode ser visualizado na figura 1, composto por dados como código de resposta (*status code*) de três dígitos, conforme documentado na RFC2616, como versão do servidor, data da modificação, entre outros.



```
Array
(
    [0] => HTTP/1.1 200 OK
    [1] => Date: Sat, 29 May 2004 12:28:13 GMT
    [2] => Server: Apache/1.3.27 (Unix)  (Red-Hat/Linux)
    [3] => Last-Modified: Wed, 08 Jan 2003 23:11:55 GMT
    [4] => ETag: "3f80f-1b6-3e1cb03b"
    [5] => Accept-Ranges: bytes
    [6] => Content-Length: 438
    [7] => Connection: close
    [8] => Content-Type: text/html
)
```

Figura 1 - Resposta à Requisição HTTP conforme RFC2616

*Status Code (*traduzido de Fielding et al., 1999):

- 1xx: Informativa - Pedido recebido, processo contínuo;
- 2xx: Sucesso - A ação foi recebida com sucesso, compreendida e aceita;
- 3xx: Redirecionamento - Outras ações devem ser tomadas a fim de completar o pedido;
- 4xx: Erro no Cliente - O pedido contém sintaxe inválida ou não pode ser completado;
- 5xx: Erro no Servidor - O servidor não completou um pedido aparentemente válido.

(traduzido pela autora). (FIELDING et al., tools.ietf.org, 1999)

O retorno "2xx" da requisição HTTP à URL do lojista informado indica que a página existe e que o processo de raspagem de dados poderá ser iniciado. Se não for este o retorno da requisição, o usuário é notificado que site está indisponível.

Conforme citado anteriormente, para evitar excessivos acessos às fontes para extração de informação que possivelmente tenha sido coletada em tempo não hábil para atualizações, os resultados obtidos são armazenados em banco de dados, o que permite também agilidade na busca das informações caso estejam já salvas. Desta forma, o endereço virtual do lojista em questão é consultado à tabela de lojas e, caso encontrado, a data da última busca é verificada na tabela de histórico do lojista. Se a credibilidade da loja foi rastreada no dia em questão, as informações coletadas das fontes salvas em uma tabela de resultados no banco de dados são exibidas ao usuário.

A comunicação entre a linguagem PHP e o banco de dados MySql foi realizada através da API *PHP Data Objects* (PDO), unificando a chamada a métodos através de orientação a objetos com o *Design Pattern Query Object*, onde as instruções SQL são criadas por meio de objetos, permitindo expressar vários tipos de instruções por meio de conjunto de métodos e posteriormente transformar a informação em comandos SQL.

(DALL'OGLIO, 2009). A criação do banco de dados, tabelas e relacionamentos foi realizada por meio da ferramenta PHPMyAdmin - ferramenta desenvolvida em linguagem PHP para administração de banco de dados MySql através de um navegador.

Porém, quando o lojista é consultado pela primeira vez ou a última consulta não ocorreu no dia em questão, a raspagem de dados é realizada. A forma como é realizada varia de acordo com disponibilidade das informações em cada fonte.

O site é composto por fontes registradoras de opinião e fontes auditoras, todas pré-determinadas. Abaixo, as fontes consultadas e suas respectivas descrições publicadas na Internet:

- **www.ebit.com.br** -"*A e-bit Informação é uma empresa com informações do comércio eletrônico fundada em 1999, pioneira na realização de pesquisas sobre hábitos e tendências de e-commerce no Brasil. A e-bit possui um sistema de avaliação que reúne informações sobre comércio eletrônico coletadas junto a consumidores após realizarem compras em aproximadamente 2.000 lojas virtuais. Por isso, os associados da e-bit que acessam a página lojas virtuais têm à sua disposição uma lista completa de lojas divididas por categorias de produtos. As lojas conveniadas ao bitConsumidor, sistema de pesquisas da e-bit onde o cliente relata sua experiência de compra no momento em que a conclui, são classificadas por meio de medalhas de bronze, prata, ouro ou diamante, conforme a opinião de seus próprios clientes.* ".[3]
- **www.reclameaqui.com.br** - "*Reclame Aqui! é o espaço do consumidor na Internet. Aqui você pode exercer sua cidadania expressando sua reclamação quanto a atendimento, compra, venda, produtos e serviços. Sem qualquer custo, a reclamação é publicada e um aviso é encaminhado via e-mail à parte reclamada, caso a empresa tenha seu Serviço de Atendimento ao Cliente Cadastrado no Reclame Aqui. As empresas poderão responder a qualquer momento, publicando assim a resposta à reclamação do cidadão, bastando apenas estarem cadastradas no site. As reclamações cadastradas no Reclame Aqui irão gerar um ranking sempre atualizado das empresas conforme critérios de número de reclamações, tempo de resposta, ausência de resposta, índice de Solução, Número de Avaliações, nota do reclamante e índice de voltar a fazer negocio com a empresa considerados a partir do momento da publicação e da respostas das mesmas. O sistema é*

[3] Fonte: http://www.ebit.com.br/ebit/html/quem_somos.asp

totalmente automatizado, não havendo interferência de operador na geração dos dados de ranking.".[4]

- **www.buscape.com.br** - "*O site BuscaPé é uma ferramenta de mídia feita para você encontrar a melhor compra através de informações, incluindo comparação, que auxiliarão na sua decisão de compra consciente. Disponibilizamos, através de ferramentas de busca inteligente (engines e algoritmos), uma lista das ofertas com informações que auxiliam na hora de fechar sua aquisição. É importante dizer que o site BuscaPé não é uma loja. Nossa tecnologia, e ferramentas, permite que você não perca tempo procurando em várias lojas (físicas ou na internet) preços ou informações que sobre produtos e serviços que agrupamos e organizamos para você. Todos os dias, milhões de ofertas capturadas nas lojas são catalogadas, organizadas, ajustadas e apresentadas a você, sempre, em prol de uma compra consciente.* ".[5]
- **www.confiometro.com.br** - "*O Confiômetro é um espaço onde você, Consumidor, pode gratuitamente expressar sua opinião em relação ao atendimento, compra e venda de produtos e serviços. Qualquer pessoa pode deixar sua opinião, basta apenas se cadastrar no Confiômetro. A opinião será enviada via e-mail à empresa, que poderá respondê-la a qualquer momento. Na parte dos rankings, exibimos e ordenamos as empresas de forma a mostrar as que têm maior índice de solução, mais rápidas em suas respostas, entre outros critérios. Dessa forma, o consumidor poderá verificar rapidamente como está a reputação das empresas. O banco de dados do Confiômetro é confidencial. Seus dados são utilizados apenas para possibilitar seu acesso e de empresas, e para que possamos nos comunicar com ambos. A divulgação das opiniões na mídia e encaminhamento a órgãos e autoridades que possam colaborar na defesa dos interesses coletivos, será feita a critério do Confiômetro. A responsabilidade das mensagens e informações publicadas é dos Consumidores, e o Confiômetro enviará apenas a essência do conteúdo das mensagens, guardando a origem dessas em seu banco de dados.*".[6]
- **www.procon.sp.gov.br** - "*A Fundação de Proteção e Defesa do Consumidor – Procon-SP, tem como objetivo principal equilibrar e harmonizar as relações entre consumidores e fornecedores. Sua missão é planejar, coordenar e executar a*

[4] Fonte: http://www.reclameaqui.com.br/como_funciona/ajuda/?id=1
[5] Fonte: http://www.buscape.com.br/sobre-o-buscape
[6] Fonte: http://www.confiometro.com.br/como-funciona.html

política estadual de proteção e defesa do consumidor em São Paulo. Constatando que uma questão de consumo está ou pode prejudicar muitos consumidores, o Procon-SP pode propor ações coletivas aos órgãos competentes. também atua auxiliando o Poder Judiciário em suas decisões e o Poder Legislativo acompanhando e oferecendo propostas a projetos de lei, sempre que o assunto envolver relações de consumo.".[7]

- **www.siteblindado.com.br** - "*Líder em segurança para e-commerce no Brasil, a Site Blindado S/A analisa vulnerabilidades em servidores (IPs) e aplicações (URLs), e fornece a empresas de todos os portes uma solução entregue como serviço SaaS (Software as a Service), capaz de monitorar operações não só no Brasil, mas no mundo.*".[8]
- **www.lojacertificada.com.br** - " *A Loja Certificada, emprega vários critérios de consulta e avaliação das empresas cadastradas. Basicamente, o selo de LojaCertificada garante ao consumidor que a loja virtual é de uma empresa que existe legalmente, possuíndo autorizações do governo federal e estadual para operar. Que a empresa tem um histórico confiável no mercado e não se trata de uma loja falsa. Também é analisada a segurança da ferramenta de e-commerce, que garanta que os dados do consumidor estejam protegidos de forma segura.*".[9]

Além disso, este projeto verifica também indício de uso de protocolo SSL nas transações e se o lojista faz parte do movimento Internet Segura:

- **www.internetsegura.org** - "*O lançamento do site e da campanha publicitária, no dia 5 de abril de 2005, marca o início da fase externa do Movimento Internet Segura (MIS), dentro de seu objetivo de levar informações que permitam aos usuários da internet no Brasil uma navegação mais segura e com maior confiança pela rede. Os membros do Movimento incluem bancos, bandeiras de cartões de crédito, lojas de comércio eletrônico, fabricantes de softwares e equipamentos de segurança e portais de acesso, além de outras organizações*".[10]

Antes de qualquer varredura, o mesmo procedimento adotado para verificação da URL do lojista é aplicado às fontes de informação: as fontes somente são consultadas mediante à resposta positiva do *status code* à requisição HTTP.

[7] Fonte: http://www.procon.sp.gov.br/pdf/acs_orientando_defendendo_consumidor.pdf)
[8] Fonte: https://www.siteblindado.com/pt/institucional/site-blindado-sa/
[9] Fonte: https://www.lojacertificada.com.br/porque-e-seguro.asp
[10] Fonte: http://www.internetsegura.org/institucional/institucional01.asp

A extração da informação varia conforme recursos disponibilizados para cada fonte. Para os sites eBit, Reclame Aqui, Confiômetro, Procon e Site Blindado, foi possível a utilização de *query string*, pela passagem do nome ou da URL do lojista como parâmetro:

- **www.ebit.com.br -** http://www.ebit.com.br/*nomedaloja*
- **www.reclameaqui** - http://www.reclameaqui.com.br/compare/*id-nomedaloja*/
- **www.confiometro.com.br** - http://www.confiometro.com.br/*nomedaloja* e http://www.confiometro.com.br/buscar.html?nome=*nomedaloja*
- **www.procon.sp.gov.br** - - http://www.procon.sp.gov.br/reclamacoes.asp?ano=*ano*&pesquisa=*nomedaloja*
- **www.siteblindado.com.br** - - http://selo.siteblindado.com.br/verificar?url=*urldaloja*

As páginas retornadas pelas URL acima contêm a informação desejada e foram coletadas através da linguagem PHP a partir da utilização da biblioteca *libcurl*, que permite conectar e comunicar com vários tipos diferentes de servidores com vários tipos diferentes de protocolos. (br.php.net).

A partir da página-fonte buscada, é necessária a extração da informação desejada contida em algum local da página coletada. Essa extração é possível com biblioteca *DOMXPath* do PHP para manipulação da linguagem XPath já citada, a qual descreve o caminho dentro da linguagem de marcação (no caso *HyperText Markup Language* - HTML) em que se encontra o dado buscado. Para que o dado seja encontrado a partir desse caminho, é fundamental que não haja erros na linguagem de marcação que possam vir a confundir a localização, como por exemplo *id's* (identificação de *tag*) duplicados. XPath percorre a linguagem utilizando uma expressão composta por barras ("/") para localizar hierarquicamente as *tags* (identificações) e, conseqüentemente, seus conteúdos. O conteúdo entre barras é composto pelo nome do tipo da *tag*, sendo a utilização de uma única barra a representação de um caminho absoluto para o determinado elemento. Para realizar uma busca no documento inteiro, a expressão XPath deverá ser compostas por duas barras iniciais ("//"), então todos os elementos no documento que se encaixam no critério serão selecionados, mesmo que eles estejam em níveis diferentes da árvore XML. (PENATTI, s.d.).

Abaixo, exemplo do caminho XPath para o conteúdo desejado da URL da fonte Site Blindado:

//div[6]/div/div/div[2]/p[2]/span[2]

A informação desejada, no caso acima a data da auditoria, consta dentro de *tags* conforme a hierarquia descrita no exemplo, por fim segunda *tag* <p> e segunda *tag* .

A fonte Buscapé é a única que disponibiliza API's para desenvolvedores, não sendo necessária a utilização de técnicas de *WebScraping*, somente cadastro na área de desenvolvedores para que seja adquirido um *id*, a ser adicionado como um dos parâmetros da URL durante a requisição de dados.

A API Buscapé utilizada neste projeto é denominada "**Busca por Loja**", a qual retorna informações de uma loja virtual cliente Buscapé através de seu nome como parâmetro. Todos os serviços API do Buscapé utilizam a tecnologia REST no tratamento de requisições: Transferência de Estado Representacional *(Representational State Transfer*), ou somente REST, é uma técnica de engenharia de software para sistemas distribuídos, que descreve uma interface *Web* simples que utiliza XML, HTTP, JSON ou texto puro, sem abstrações adicionais dos protocolos baseados em padrões de troca de mensagem como o SOAP. (developer.buscape.com). Desta forma, é possível construir facilmente uma URL para ser executada em navegador, linha de comando ou código.

O formato padrão de resposta é *eXtensible Markup Language* (XML), mas há a opção de utilizar resposta em formato *JavaScript Object Notation* (JSON). Neste projeto, optou-se pelo formato XML, possível de ser lido com a biblioteca PHP *SimpleXML*, a qual permite facilmente percorrer os nós de um arquivo XML.

O site auditor Loja Certificada foi o único em que a extração de dados somente é possível através de método POST. Não há a utilização de *query string* e nem disponibilização de API's ou *WebServices*. Para consultar se determinado lojista utiliza os serviços desta fonte, o endereço virtual do mesmo é informado na página inicial do site lojacertificada.com.br e, através do método post, os dados da loja são exibidos se possuir o serviço contratado. Este mesmo procedimento de post foi realizado também a partir da biblioteca PHP *libcurl*, alterando-se somente os parâmetros repassados à mesma para indicar o método de envio a ser aplicado, além de indicar o nome do campo do formulário (verificado no código da página da fonte) e conteúdo deste campo (URL do lojista). A partir do resultado, o mesmo procedimento de XPath foi aplicado para verificar se a resposta é afirmativa para utilização (ou não) deste serviço de auditoria.

A constatação do uso de SSL/https e de adesão ao movimento Internet Segura foi possível pela aplicação de expressões regulares à página do lojista em busca das informações correspondentes, retornando verdadeiro ou falso quando encontrado o termo desejado.

Após todos os resultados extraídos, os mesmos são salvos em tabela de resultados no banco de dados para que em uma próxima consulta ao lojista em curto espaço de tempo não seja necessária novamente a raspagem de dados (atualmente, utiliza-se o intervalo de um dia, ou seja, se não houver consulta em banco de dados realizada no dia atual, efetua a técnica de *WebScraping*; caso contrário, seleciona do banco de dados). Caso alguma fonte não retorne resultados, mensagem de fonte indisponível é exibida ao usuário, erro registrado na tabela de erros do banco de dados e e-mail de notificação enviado à desenvolvedora para possível averiguação do ocorrido.

5 Considerações Finais

Quando se utiliza raspagem de dados da *Web*, ou seja, resultados provenientes de sites de terceiros, resultados satisfatórios ao usuário só são possíveis tratando e verificando todas as possibilidades de erro que podem ocorrer, como por exemplo a indisponibilidade de alguma fonte no momento da consulta ou ainda na mudança da composição de sua *query string*. Alteração do endereço virtual ou do caminho da informação na página fonte é um risco/vulnerabilidade que corre qualquer site que utiliza práticas de *WebScraping*. Para isso, todo tratamento é necessário, para que não haja erros visíveis ao usuário.

Como melhorias futuras, pretende-se aperfeiçoar a averiguação do uso de protocolo criptográfico SSL nas transações, com o intuito de exibir ao usuário a validade do certificado utilizado pelo lojista e exibição parcial dos resultados por lojista conforme são coletados, ao invés da exibição total somente após toda extração de dados, para tornar a experiência do usuário mais rápida ao demonstrar carregamento dos resultados.

O site descrito neste projeto foi nomeado "*Guru da Compra*" e sitiado no endereço *www.gurudacompra.com.br*. O intuito é que haja acompanhamento de possíveis novas fontes de informação que possam ser agregadas, oferecendo portanto mais dados para que o usuário possa determinar sua decisão de compra fundada em

mais indícios de credibilidade do lojista. O contato da desenvolvedora estará disponível no site para usuários que queiram também sugerir outras fontes de informação que sejam atualmente desconhecidas pela autora, ou ainda reportar algum resultado exibido que julgarem ser inconsistente.

5. Referências

AMO. **Curso de Data Mining**. 2003. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Uberlândia-UFU, Uberlância, 2003. Disponível em: <http://www.deamo.prof.ufu.br/arquivos/Aula17.pdf>. Acesso em: 17 abr. 2011.

BERNERS-LEE, T.; FIELDING, R. T.; MASINTER, L. **Uniform Resource Identifier (URI):** Generic Syntax. Internet RFC 3986, 2005. Disponível em: < http://tools.ietf.org/html/rfc3986>. Acesso em: 08 mai. 2011.

CASANOVA, M. A. **Consultas em XML – X*Path***. Rio de Janeiro: PUC-Rio, 2003. Disponível em: <http://www.inf.puc-rio.br/~casanova/INF2328-Topicos-WebBD/modulo0-XML/modulo2a-xml-consultas-xpath.pdf>. Acesso em: 11 mai. 2011.

CLARK, J.; DEROSE, S. **XML *Path Language (XPath)***. Boston: MIT, 1999. Disponível em: <http://www.w3.org/TR/xpath>. Acesso em: 09 mai. 2011.

DALL'OGLIO, P. **PHP - Programando com Orientação a Objetos**. São Paulo: Novatec, 2009.

FIELDING, R. et al., ***Hypertext Transfer Protocol - HTTP/1.1***. RFC 2616, 1999. Disponível em: <ttp://tools.ietf.org/html/rfc2616>. Acesso em: 08 mai. 2011.

PALMIERI, J. et al. **Collecting hidden web pages for data extraction**. ACM WIDM, 2002. Disponível em: <ttp://tools.ietf.org/html/rfc2616>. Acesso em: 08 mai. 2011.

LAGE, J. P. et al. Collecting hidden weeb pages for data extraction. In: **Proceedings of the 4th international workshop on Web information and data management**, Nova York, ACM, 2002. p. 69-75. Disponível em: < http://dl.acm.org/citation.cfm?id=584946>. Acesso em: 08 mai. 2011.

HEMENWAY, K.; CALISHAIN, T. **Spidering Hacks:** 100 Industrial-Strength Tips & Tools. Sebastopol: O'Reilly Media, 2004. 404p. Disponível em: < http://www.newsmth.net/bbsanc.php?path=%2Fgroups%2Fcomp.faq%2FPerl%2Fsmth_cd%2Febook%2FM.1154839056.A0&ap=386>. Acesso em: 08 mai. 2011.

JARGAS, A. M. **Expressões Regulares -** Uma Abordagem Divertida. São Paulo: Novatec, 3 ed., 2003.

LAFON,Y. **HTTP - *Hypertext Transfer Protocol***. World Wide Web Consortium, 2011. Disponível em: <W3C: http://www.w3.org/Protocols/>. Acesso em: 08 mai. 2011.

MELO, J. Comércio Eletrônico em Expansão. São Paulo, **Isto é Dinheiro,** 31 mai. 2011. Disponível em: <

http://www.istoedinheiro.com.br/noticias/58520_COMERCIO+ELETRONICO+EM+EXPANSAO+NO+BRASIL>. Acesso em: 11 nun. 2011.

PENATTI, O. A. B. ***XML - Tutorial XPath***. Macoratti.net, (s.d.). Disponível em: < http://www.macoratti.net/vb_xpath.htm >. Acesso em: 09 mai. 2011.

TURLAND, M. ***php|architec'ts Guide to Web Scraping***. Victoria (Canadá): Nanobooks, 1 ed. 2010. 192p.

TIOBE SOFTWARE. *TIOBE Programming Community Index for September 2011*. Holanda, TIOBE Software BV, (s.d.). Disponível em: <http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html >. Acesso em: 23 set. 2011.