Perfis de Usuários de Web Sites por Mineração de Uso para Avaliação de Usabilidade

Rafael Crespo, Luis A. Rivera, Ausberto Castro

rafaelbpcrespo@gmail.com, {rivera, ascv}@uenf.br

Laborátorio de Ciências Matemáticas – LCMAT
Universidade Estadual do Norte Fluminense – UENF
Av. Alberto Lamego, 2000; CEP 28015-620, Campos dos Goytacazes – RJ – Brasil

Resumo: *Hoje em dia a quantidade de pessoas acessando a internet tem crescido exponencialmente e isso acontece por diversos motivos, entre eles a facilidade de acesso à Internet, a difusão dos dispositivos móveis e o crescimento das redes sociais. A consequência do aumento do número de usuários é o crescimento do fluxo de dados, pois quanto mais pessoas acessam a internet, muito mais dados são gerados. Como a quantidade de informações é extremamente grande, tornou-se inviável realizar o tratamento ou a análise delas de forma manual, então surgiu o conceito de mineração web, que visa extrair informações a respeito dos dados gerados pelo acesso à web. O objetivo deste trabalho é mostrar as diferentes etapas da mineração do uso da Web, utilizando o log de acesso a um determinado* site.

Palavras chave: Mineração de dados, mineração web, padrões de usuário, perfil de Usuário.

Abstract: *Nowadays the number of people accessing the internet has grown exponentially and this happens for several reasons, including the ease of access to the Internet, the spread of mobile devices and the growth of social networks. The result of the increase in the number of users is the growth of the data flow, because the more people access the internet, much more data is generated. As the amount of information is extremely large, it has become feasible to perform the treatment or analysis of them manually, then came the concept of web mining, which aims to extract information about the data generated by web access. The objective of this work is to show the different stages of mining Web usage, using the log access to a particular sit*e.

Keywords: Data mining, Web mining, user pattern, user profile.

1 Introdução

Sem dúvida alguma, a Internet é hoje a maior fonte de informação atualizada para qualquer pessoa e qualquer assunto de interesse. Pode-se buscar qualquer tipo informação utilizando técnicas de buscas simples, em diversos sites. Em estas buscas, um usuário pode entrar em páginas de diferentes conteúdos, apresentação e organização. Algumas destas páginas são inspiradoras, confiáveis, agradáveis, e outras não. O objetivo do usuário que navega no site é encontrar informações convincentes para seu interesse.

Um determinado site pode estar sendo acessado por um grande número de usuários em um determinado intervalo de tempo, cada um com diferentes objetivos e comportamentos. Então, é de responsabilidade do administrador do site implementar mecanismos que tornem essas páginas do agrado e satisfação da maioria dos usuários. Uma parte fundamental desses mecanismos está relacionada com conceitos de usabilidade que lida com o ditado: "*com tantas opções, as pessoas não vão querer perder tempo 'quebrando a cabeça' para buscar informações ou produtos dentro deste site*".

A escolha de um site por parte de um usuário para busca de informações de interesse esta diretamente relacionada com a organização e usabilidade nele implementados.

Considerando que a usabilidade é um atributo muito importante na escolha de um site por parte do usuário, é necessário encontrar técnicas que auxiliem aos web designers a entender de que forma eles podem atender melhor às necessidades dos usuários. Uma destas técnicas é a mineração web nos documentos ocultos produzidos e armazenados nos servidores. Os perfis dos usuários podem variar no tempo de modo que as informações obtidas pela mineração web periodicamente são elementos fundamentais para uma atualização periódica das páginas web.

Segundo Liu (2007), **web mining** é o *uso das técnicas de data mining para descobrir e extrair automaticamente informações relevantes dos documentos e serviços ligados à internet*. O uso de técnicas de data mining, com algumas modificações devido a que os dados na web não são estruturados, permite extrair as informações relevantes dos documentos e serviços ligados à internet e suas aplicações com o intuito de generalizar, validar e interpretar os padrões de navegação visando à melhora de usabilidade.

Apesar de ter como base a mineração de dados, a Mineração Web desenvolveu suas formas próprias de extração devido à grande variedade e quantidade de informações que são dinamicamente geradas. Por isto, a Mineração Web hoje alcançou significativa importância. Isto se deve a vários fatores: o crescimento significativo de dados e informações, a grande variedade de tipos de dados, a heterogeneidade das informações, a quantidade imensa de fontes para uma mesma informação.

A literatura tem definido três setores de estudo (Jeria, 2007) e (Srivastava et al, 2000): 1) a *Mineração Web de Conteúdo*, que foca na mineração dos conteúdos exibidos nas páginas da web, e bem similar a uma mineração de texto aplicada na web; 2) a *Mineração Web de Estrutura*, que é voltada para o estudo das estruturas de hiperlinks da web; 3) a *Mineração de Uso da Web*, que é a análise dos

dados gerados através da utilização e navegação dos usuários.

A Mineração de Uso da Web (WUM: *web usage mining*) tem chamado a atenção principalmente das empresas de e-commerce, pois tem sido vista como uma grande ferramenta para "conhecer" seus clientes, que neste caso são os usuários que navegam nos sites em busca de produtos. Essas empresas visam investir em pesquisa e desenvolvimento de métodos de mineração e reverter os dados obtidos a seu favor, direcionando para uma estratégia de marketing mais eficaz.

A grande quantidade de informação que continua a crescer significativamente por causa do desenvolvimento constante da internet em diferentes plataformas e das bases de dados dinâmicas, fez com que novos recursos de armazenagem e processamento de dados começassem a surgir, pois tornou-se impossível realizar essas tarefas em um volume tão grande de dados sem o auxílio da tecnologia, tais como a Mineração Web. Na WUM os dados analisados são os registros deixados em arquivos logs: de acesso, de erro ou de *proxy*. Estes arquivos são gerados dinamicamente durante a navegação do usuário.

Todas as empresas que vendem produtos ou serviços dependem de seu nível de aceitação e de seus clientes, portanto uma boa relação entre a empresa e o cliente é de extrema importância para que este se torne fiel à empresa. O problema da fidelização é que com a globalização e o comércio eletrônico, veio uma maior dificuldade de entender e "analisar" o comportamento do cliente, saber o que ele procura, qual abordagem utilizar, o que oferecer, pois se tornou um cliente a distância sem um contato direto entre vendedor e cliente, e a partir disto que a Mineração Web passou a ser uma ferramenta essencial.

A eficiência de análise de informações, chamadas perfis, depende da quantidade e da qualidade das informações registradas como perfis. Com esse enfoque é necessário considerar aspectos de captura dos perfis por onde os usuários navegarem.

O objetivo deste trabalho é analisar perfis de usuários de um site através dos logs de acesso do servidor web, aplicando a WUM, para, a partir do resultado gerado, produzir evidências para serem analisadas utilizando os conceitos e atributos de Usabilidade, e chegar à sugestão de possíveis melhorias de navegação que possam ser aplicadas neste site. Desta forma, a navegação do usuário será facilitada, pois terá melhores mecanismos de encontrar as informações que está buscando.

Na Seção 2 são bordados fundamentos de usabilidade de paginas web e os tipos de mineração de uso. Na Seção 3 é formulado o modelo de mineração de WUM para uso de usabilidade de páginas. Na Seção 4 é realizada a descoberta de padrões com mineração dos dados de log. Na Seção 5, os padrões são analisados desde o ponto de vista de usabilidade. Finalizando, na Seção 6 serão apresentadas conclusões e formulações dos trabalhos futuros.

2 Mineração de Uso para Avaliação de Usabilidade de Páginas

A análise de padrões ocultos, gerados pelos comportamentos dos usuários ao navegar as páginas web, permite estabelecer conceitos de melhoria das páginas de forma a satisfazer as demandas dos usuários que buscam alguma informação. A análise de arquivos logs é um dos métodos de avaliação de usabilidade, porém, existem outros tais como, heurísticos, ensaio de interação, questionários, relatos de incidentes críticos dos usuários (Campos, 2012). Neste trabalho foi utilizado a análise de logs por mineração, estabelecendo os conceitos e atributos de usabilidade e mineração de uso, de acordo com Carmona et al. (2012).

2.1. Usabilidade de web sites

Os conceitos e atributos de usabilidade foram desenvolvidos na interação humano-computador como trabalhos de pesquisa interdisciplinar (Shneiderman et al, 2009), utilizando métodos de psicologia experimental nas ferramentas de Ciência da Computação, e outras áreas de aprendizado e expertos preocupados em fatores humanos e ergonómicos. Neste caso, foram considerados os critérios de usabilidade de páginas web.

Nas páginas web, a preocupação é com uma grande quantidade de usuários e a heterogeneidade de seus perfis. Quando um usuário busca uma informação ou produto no site, o foco está naquilo que esta sendo procurado, fazendo uma análise superficial das informações contidas nas páginas até encontrar algo relacionado com o procurado. Outras vezes, o usuário passa para outras páginas e volta algumas vezes à mesma página, sinalizando que a informação de seu interesse se encontra nessa página.

Nielsen (1993) define os cinco atributos da *usabilidade*, como sendo os seguintes:

a) *Aprendizagem*: atributo relacionado com a facilidade de aprendizagem do produto pelo usuário no primeiro contato, de forma que lhe seja fácil de reiniciar a operação.

b) *Eficiência*: o usuário deve obter bons resultados interagindo com esta ferramenta.

c) *Memorabilidade*: para usuários casuais, o produto deve ser fácil de lembrar como operar.

d) *Minimização de erros*: baixa taxa de erros. Se tivesse, deve ser facilmente contornado.

e) *Satisfação*: A ferramenta deve ser prazerosa para o usuário.

Quando o termo usabilidade começou a ser abordado segundo uma visão da Tecnologia da Informação e Interação Homem-Computador, uma nova definição foi dada pela ISSO 9241 (1998) e ela dizia que a usabilidade era *a medida na qual um produto pode ser usado por usuários específicos para alcançar objetivos com eficácia, eficiência e satisfação dentro de um contexto de uso*.

2.2. Mineração de Uso da Web

De acordo com Jeria (2007), a WUM trata da aplicação de técnicas de mineração para descobrir padrões de uso da informação Web com o objetivo de entender como é que os usuários comuns utilizam a página e de satisfazer suas necessidades. Para este tipo de aplicações, a principal fonte de informações são os *arquivos log* dos servidores Web.

Enquanto a Mineração de Conteúdo e a Mineração de Estrutura utilizam os dados reais ou primários da Web, a Mineração de Uso lida com os dados secundários, que são gerados a partir da interação do usuário com a Web. Os dados de uso da Web incluem informações de logs de servidores web, logs de servidores proxy, logs de browsers, perfis de usuário, cookies, seções ou transações de usuários, pasta de favoritos, consultas do usuário, cliques de mouse e qualquer outro dado gerado pela interação do usuário com a Web. O objetivo principal é capturar, modelar e analisar o padrão de comportamento e os perfis dos usuários que interagem com o sistema. Os padrões descobertos geralmente são recursos frequentemente acessados por grupos de usuários com interesses em comum.

O processo de mineração de uso da Web pode ser classificado segundo duas abordagens. Uma delas mapeia os dados de uso do servidor Web em tabelas relacionais antes das técnicas adaptadas de mineração de dados serem aplicadas. A outra utiliza os dados de logs diretamente utilizando técnicas especiais de pré-processamento. Assim como no *Knowledge Discovery Data* (KDD), que é um processo de descobrimento de conhecimento a partir de um grande volume de dados, a limpeza e pré-processamento dos dados, aqui, é uma parte crucial do processo, pois a qualidade desses dados vai determinar a eficiência dos algoritmos de mineração.

2.3. Estrutura de um website

Para propósito de análise e exemplos, sem perda de generalidade, é utilizado um exemplo de estrutura de uma página web que é estabelecida no processo de design do site uma estrutura de possíveis formas de navegação entre as diferentes páginas identificadas por seus nomes e seus links. A Figura 1 ilustra um exemplo representado por uma árvore, onde os nós são as páginas visitadas e identificados com nomes a, b, …, j. As arestas são os links. Assim, da página a podem ser requisitadas as páginas b, c e d, da página b são acessadas as páginas e e f, e assim sucessivamente.

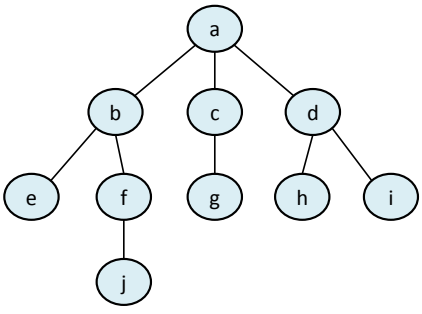


Figura 1: Estrutura de relacionamento entre as páginas do site

2.4. Trabalhos relacionados

O tema de mineração de web é bastante explorado, bem como seu uso na avaliação de usabilidade de paginas web, mesmo assim, a comunidade académica continua pesquisando vários aspectos de eficiência em diferentes métodos de operação e caracterização dos dados incompletos para a mineração e a interpretação dos perfis.

Cho et al. (2002) desenvolveram um trabalho voltado para a área de aplicação de recomendações personalizadas, onde foi desenvolvido um sistema de recomendações. Inicialmente as características do usuário são recolhidas pelo rastreamento de cliques (Mineração de Uso), em seguida e para evitar recomendações ruins que afastarão clientes, o sistema seleciona aqueles que geralmente compram produtos recomendados pelo site usando uma árvore de indução, e para finalizar, medidas são elaboradas para escolher produtos mais eficientes entre os produtos candidatos.

Já em Carmona et al (2012), a WUM é aplicada com a finalidade de ajudar ao webmaster da empresa a melhorar o design do site. No trabalho a técnica de regra de associação é utilizada. Depois de realizar todo o estudo, chegaram à conclusão de que era importante prestar atenção nos acessos que eram gerados através de referências de outros sites, pois os usuários visitam um número muito baixo de páginas, onde a maioria dos acessos feitos ao site eram oriundos de usuários que utilizavam o Internet Explorer como navegador. Graças a este estudo o webmaster tem agora um caminho para seguir e executar o seu trabalho seguindo informações concretas e uma das suas ações deveria ser se preocupar em fazer um layout mais optimizado para o IE.

Zhang et al. (2007) descreve um conjunto de ferramentas que exploram dados de uso da web, as quais identificam padrões de navegação na internet. A partir dos dados minerados pelas ferramentas, os padrões identificados são usados para alimentar a recomendação personalizada de produtos para vendas online. O objetivo principal foi mostrar que ao utilizar a rede neural de Kohonen treinada para trabalhar offline, o problema de escalabilidade, que assombra esses sistemas, seria resolvido.

Duan e Liu (2012) projetaram uma ferramenta de mineração de uso da web (WebLog Mining Tool) utilizando o método de Padrões Sequênciais. A ferramenta tem como entrada de dados os arquivos logs da web, e como saída são entregues os padrões que foram identificados no trecho de log analisado. Web Log Mining Tool é dividida em três fases: pré-processamento, mineração de padrões e visualização de padrões. O objetivo do trabalho é que os profissionais especializados em usabilidade possam encontrar problemas no site observando o resultado gerado pela ferramenta desenvolvida.

3 Modelado de um WUM

Neste trabalho é seguido o modelo tradicional de WUM, tal como ilustrada pela Figura 2, composto por quatro fases: Coleta de dados, pré-processamento, mineração, e análise.

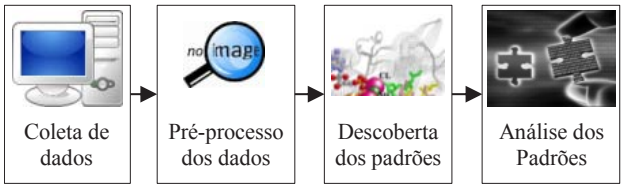


Figura 2: Processo de identificação de perfis em WUM

Na *coleta de dados*, os arquivos log da web registram as informações das atividades dos visitantes quando eles realizam alguma requisição do servidor web. Estes arquivos podem ser armazenados em três lugares diferentes, sendo eles: servidores web, servidores de proxy e nos browsers dos clientes.

Pré-processamento, segundo Sanjay et al (2010), é o processo de preparação dos dados coletados, e geralmente consome mais tempo porque exige maior poder computacional que as outras etapas. Isso acontece porque as informações disponíveis na web são extremamente heterogêneas e desestruturadas. Nesta etapa é feita a toda a limpeza e o tratamento dos dados coletados e acontece uma série de manipulações das informações para que sejam minuciosamente tratadas e consequentemente repassadas para a fase seguinte com a melhor qualidade possível.

Descoberta do padrão é a fase onde são utilizadas as técnicas de mineração para descobrir os padrões de comportamento existentes nos dados que foram pré-processados. Entre as técnicas utilizadas e que possuem maior relevância temos: Regras de Associação, Classificação, Clustering e Padrões de Sequência.

Na *análise de padrão* é aplicada as estatísticas, e os padrões encontrados são processados e filtrados para gerar um modelo de usuários agregados que poderão ser utilizados como entrada de dados para ferramentas de visualização e análise na web, geração de relatórios ou mecanismos de recomendação. Nesta fase os resultados obtidos a partir da aplicação dos algoritmos de mineração são analisados e logo transformados ou convertidos em conhecimento, para a tomada de decisões, medidas de correção, melhorias, etc.

3.1. Coleta de dados

A coleta dos dados do servidor web, no formato ELF é feito por acesso remoto via SSH, e então copiado para uma máquina local para evitar outro tipo de processamento no servidor.

Nos servidores web podem-se encontrar tipos de log e informações diferentes. Os mais comuns são: o log de erro que ocorrem no servidor, e log de acesso da navegação do usuário. O log de acesso do servidor web é a informação usada para aplicações de WUM. Essas informações, registradas no formato *Common Log* e *Extended Log Format (ELF)*, retratam o comportamento dos visitantes que navegam nas páginas do site hospedado nesse servidor. Na Tabela 1, é ilustrado um exemplo de ELF que contem IP, data de requisição, bytes transferidos, URL (caminho de destino), *Referrer* URL (caminho de origem). Neste caso, não registra as páginas visitadas novamente, por estarem salvas no cache dos navegadores, localmente, ou nos servidores *proxy*.

Tabela 1: Exemplo de Log no ELF

IP Adrress	123.456.78.9
Userid	-
Time	[25/Apr/1998:03:04:41 -0500]
Method/URL/Protocol	“GET B.html http/1.0”
Status	200
Size	2050
Referrer	A:html
Agent	Mozilla/3.04 (Win95,I)

Outros dados de interesse para um melhor entendimento dos perfis dos usuários podem ser obtidos dos browsers dos clientes e servidores de proxy. Os dados do browser do cliente podem informar os padrões de navegação de um mesmo usuário, evitando confundir com os padrões de outros usuários do mesmo IP. A inconveniência é que por se tratar de uma máquina do usuário, geralmente diferente ao servidor, requer-se da permissão do usuário para ter acesso às informações da máquina, fato que na prática é quase impossível. Os servidores proxy, que atuam como intermediário de cache entre navegador do cliente e servidores web, guardam, também, dados que podem ser úteis para caracterizar o comportamento de grupos de usuários anónimos que compartilham o mesmo servidor proxy.

3.2. Pré-processamento

O formato dos dados deve permitir que o processamento seja fácil e eficiente. Para isto os dados coletados são filtrados de impurezas, identificados os usuários, as sessões e estabelecido os caminhos de navegação. A Figura 3 ilustra a sequência seguida nesta fase.

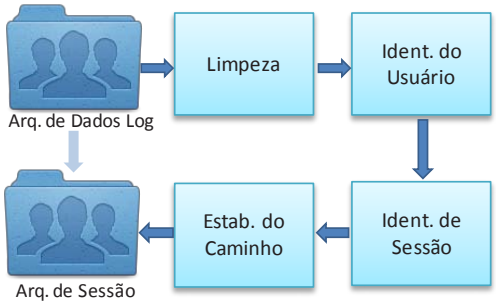


Figura 3: Etapas de pré-processamento

3.2.1. Limpeza dos dados

Os dados do servidor estão em formato texto, portanto deve ser colocado em formato apropriado para a identificação de suas partes. São selecionados só os dados importantes para nosso propósito, como data, hora, IP, código de resposta, URL acessada, URL requerente, browser, informação do sistema operacional.

O arquivo log de acesso (*access log*) registra toda requisição direta ou indireta feita ao servidor. Por exemplo, quando um site que tem imagens é acessado diretamente pelo usuário (acesso à página inicial), é feita a requisição das imagens necessárias e do favicon (ícone da página). No arquivo log de acesso se encontram todos os objetos de requisições, incluindo nomes de imagens com sufixos GIF, JPEG, JPG, PNG, BMP, ICO, como folhas e estilo (CSS) e Java Script (JS). Também contêm

dados relacionados com robôs (*web robots* ou *spiders*). O robô é uma ferramenta que, de tempo em tempo, realiza acesso aos sites com o intuito de coletar os conteúdos. A requisição realizada por acesso de robôs é identificada por conter o texto "robots.txt". Outra forma é através de listagens de IP's robôs.

As linhas do arquivo log possuem informações irrelevantes como ".cio", ".css", ".js", ".png", ".jpg", ".jpeg", ".tff", ".cgi", ".bmp", "robots", devem ser removidos, tal como realizado por *shell script* do arquivo_log colocando os dados limpos no arquivo dados_importantes.txt:

>cat arquivo_log | grep –v "\.ico\|\.css\|\.js\|\.png\|\.jpg\|\.jpeg\|\.tff\|\.cgi\|\.bmp\|\robots" >> dados_importados.txt

3.2.2. Identificação do usuário

Existem alguns métodos para identificar os usuários que acessam ao site. A forma mais simples é associar a um usuário cada requisição de um IP, porém a simplicidade de este método faz que não seja confiável. O servidor proxy, segundo Microsoft[8] que é o computador que funciona como intermediário entre o navegador da web e a internet, armazena cópias das páginas da Web que são acessadas com mais frequência, faz que diferentes usuários de navegadores possam acessar o site através do mesmo servidor Proxy. Registrando o IP do servidor proxy no log do servidor para diferentes usuários. Cooley (2000) sugere o uso de uma combinação de atributos IP e *User Agent* registrados nos logs. Se duas requisições tem mesmo IP e mesmo agente (informação de navegador e sistema operacional), então se trata de um mesmo usuário, caso contrário se trata de dois usuários distintos.

Como exemplo de ilustração simplificado, consideremos uma linha de um Log composto da seguinte forma:

177.43.182.126 - - [04/nov/2014:17:24:21 +0000] ´´GET b HTTP/1.1´´ 200 1371´´a´´ ´´Mozilla/5.0 (Windows NT 6.1; WOW64) Applewebkit/537.36 (KHTML, like Gecko) Chrome/38.0.2125.111 Safari/537.3´´

Para simplificar as tabelas, considere-se as strings α e □, com as informações de agentes de log (extraídos do arquivo log):

α = ´´Mozilla/5.0 (Windows NT 6.1;WOW64) Applewebkit/537.36 (KHTML, like Gecko) Chrome/38.0.2125.111 Safari/537.3´´

□ = ´´Firefox (Windows NT 6.1;WOW64) Applewebkit/537.36 (KHTML, like Gecko) Chrome/38.0.2125.111 Safari/537.3´´

Com isso, é simplificado o log limpo de 17 linhas como apresenta a Tabela 2:

Na Tabela 2, observa-se que um segmento, mais da metade das requisições tem o mesmo IP=188.3.182.126, mas de dois tipos de navegadores agentes α e □, fato que indica que se trata de usuários diferentes.

[8] http://windows.microsoft.com/pt-br/windows-vista/what-is-a-proxy-server.

3.2.3. Identificação de sessão

As sessões são os conjuntos de páginas acessadas em um site durante um período determinado vindas de um mesmo IP. Nesta etapa não é feita a diferenciação de usuários e sim a de sessões de navegação, ou seja, quando um mesmo usuário visita mesmo site em momentos diferentes sem interessar o que matéria busca.

Tabela 2: Exemplo de Log de entrada simplificada

#	IP	DateTime	URLd	URLr	Ag	ID
01	177.43.182.126	04/11/2014:17:23:21	a	-	α	1
02	177.43.182.126	04/11/2014:17:24:21	b	a	α	1
03	177.43.182.126	04/11/2014:17:24:25	d	b	α	1
04	177.43.182.126	04/11/2014:17:24:28	c	d	α	1
05	177.43.182.126	04/11/2014:17:25:28	b	c	α	1
06	177.43.182.126	04/11/2014:17:26:29	d	b	α	1
07	177.43.182.127	04/11/2014:17:23:21	a	-	α	2
08	177.43.182.127	04/11/2014:17:24:18	b	a	α	2
09	177.43.182.127	04/11/2014:17:24:26	c	b	α	2
10	177.43.182.128	04/11/2014:17:23:24	a	-	α	3
11	177.43.182.128	04/11/2014:17:24:26	c	a	α	3
12	177.43.182.128	04/11/2014:17:25:25	e	c	α	3
13	177.43.182.126	04/11/2014:17:23:24	d	-	□	4
14	177.43.182.126	04/11/2014:17:24:35	g	d	□	4
15	177.43.182.126	04/11/2014:17:25:21	b	g	□	4
16	177.43.182.126	04/11/2014:17:26:17	a	b	□	4
17	177.43.182.126	04/11/2014:17:27:10	c	a	□	4
…	…	…	..	..	..	..

Em um arquivo log é comum a existência de informações de acesso de um mesmo usuário, porém em situações diferentes, com datas e objetivos diferentes. Por exemplo, se um usuário acessa o mesmo site, em diferentes dias, por tópicos de interesse diferentes por vez, são considerados como acessos distintos, em momentos distintos e com objetivos distintos. Para se ter uma divisão mais estruturada desses casos, se considera que cada vez que um usuário realiza um acesso ao site, ele tem uma sessão de navegação.

O objetivo da identificação de sessão é dividir o acesso às páginas de cada usuário em sessões individuais. Uma maneira de fazer isto, proposta por Cooley et al (1999), é considerando a diferencia horária entre as requisições, e se ultrapassar um certo limite de tempo (*timeout*), assume-se que o usuário está iniciando uma nova sessão de navegação. Páginas comerciais geralmente utilizando 30 minutos como timeout, mas pode variar dependendo com o site.

3.2.4. Complemento de caminho de navegação

"*Path completion*" é a etapa da preparação dos dados que permita completar o caminho percorrido pelo usuário durante a navegação de um site em uma sessão. Por tanto, um arquivo limpo de log com *n* sessões espera representar *n* estruturas de caminhos. Na Tabela 2, que tem como mínimo 4 sessões, se espera 4 caminhos.

Uma sessão de usuário nem sempre está completa, no sentido de encontrar todas as informações importantes da navegação, porque que nem todo evento realizado pelo usuário é registrado nos logs de acesso. Tal é o caso do evento *voltar* (página anterior) que é registrada em cache do browser, relacionada às páginas anteriormente visitadas com o objetivo de evitar maior consumo de banda e tempo nas novas requisições feitas ao servidor

para páginas. Estas informações ausentes no arquivo log, porém acessíveis ao usuário, são conhecidas como *Missing Reference* (Mitharam, 2012), sendo lacunas de navegação, devem ser preenchidas pelo método de path completion.

O conjunto de caminhos (CC) é o caminho de acessos de todos os usuários (UID) obtidos de sessões de usuários (SU) dados pela fórmula:

$$CC = (UID, (URL_1, DH_1, CR_1), \dots, (URL_k, DH_k, CR_k))$$

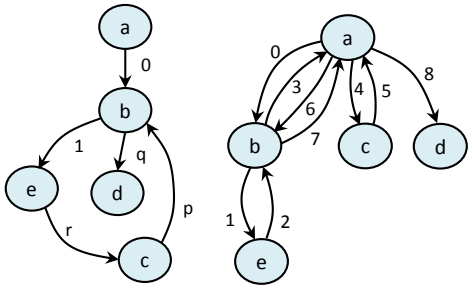
com comprimento de referencia CR, URL corrente, e data-hora DH.

Segundo Chitra et al (2010) e Li et al (2008), primeiramente devem ser identificados os caminhos para cada sessão de usuário, colocando-os em linhas contiguas e computando comprimento de referencia para cada linha de log limpo. A combinação de caminhos é efetuada se duas páginas consecutivas são a mesma em um caminho. Se qualquer um dos URL especificados no URL-referente é diferente ao URL da linha anterior, então tal URL de URL-referente da linha atual é inserido nessa sessão, com visas a completar o caminho. O seguinte passo é determinar o comprimento de referência de novas páginas anexadas no completar do cominho, e modificar o comprimento de caminho de aqueles adjacentes. Considerando que as paginas assumidas são normalmente consideradas como páginas auxiliares, o comprimento é determinado pelo comprimento médio de referências de páginas auxiliares. O comprimento de referencia de páginas adjacentes também é reajustado. Considera-se que uma sessão de usuário é dividida em grupos significativos de referências de páginas como uma transação. Essas transações podem ser identificadas pelo método de "diferença direta máxima" (*maximal forward reference*), comprimento de referência, e tempo que dura a criação do arquivo de transação, como em (Li et al, 2008).

O resultado dessa operação é um arquivo completo de todas as sessões de todos os usuários que tiveram acesso às páginas do website em um intervalo de tempo estabelecido. Este arquivo é considerado como o arquivo log de acesso web completo (*complete web access log* - CWAL).

Comprimento de referência é o tempo gasto pelo usuário para exibir uma página específica. Teoricamente, seria a diferença de tempo de acesso da linha atual e próxima linha, mas que também envolve tamanho de página transferida em relação à taxa de transferência (Li et al, 2008).

Na Tabela 2, com identificação de sessão, é evidente a ausência de algumas páginas intermediárias na navegação, pois analisando na Figura 2 observa-se, por exemplo, que não há um link direto entre as páginas E e C, B e D, nem C e B. Sem tratar o caminho completo seria como ilustra a Figura 4(a), que é um caminhamento não real, porque não existem caminhos *r*, *p* e *q* na tabela de log. Enquanto, a Figura 4(b) ilustra o caminhamento tal como a sequencia de 9 ações que realizaria um usuário a partir da página A. Esse caminhamento é conseguido inserindo as arestas 2, 3 e 4 ao invés da aresta r, arestas 5 e 6 ao invés de p arestas 7 e 8 ao invés da aresta q.



(a) (b)

Figura 4: Caminhos de navegação da primeira sessão: (a) encontrada no log; (b) após path completion

4. Descoberta de padrões

Nesta fase as informações são analisadas no intuito de extrair conhecimento delas, como descobrir padrões comportamentais e de navegação dos usuários que passaram pela website a través de técnicas de mineração.

As técnicas de maior relevância em mineração dos CWALs são: regras de associação, classificação, agrupamento (clustering) e padrões de sequência.

A *classificação*, em mineração de dados, consiste em classificar os dados em relação às categorias previamente estabelecidas. As categorias são definidas de acordo com a similaridade de alguma característica entre os dados. Geralmente são utilizadas árvores de decisão, regressão e redes neurais. O *clustering* busca estabelecer grupos de elementos relativamente homogêneos (similares) entre si e diferentes respeito aos elementos dos outros grupos. O critério de similaridade para agrupamento de grupos é realizado em relação aos descritores representativos, chamados atributos, dos elementos. Foi utilizado com sucesso nas primeiras vezes em grandes documentos HTML por Cutting et al (1992). Clustering agrupa os elementos em número previamente estabelecido de grupos, enquanto a classificação estabelece o número de grupos no próprio processo.

As r*egras de associação* (Hipp et al, 2000), é uma proposição probabilística de estados de dados da forma "Se X então Y", sendo X□Y=□, expressando que una transição T contiver X então provavelmente também conterá Y, baseado com regras de supermercados "o cliente que compra o produto X também comprará o produto Y com probabilidade de p%" (Pal et al, 2002). Esta técnica é utilizada para conhecer as rotas de visitas seguidas pelos usuários das páginas web, para poder assistir a estruturação das páginas no servidor. Enquanto *padrões sequenciais* (Ezeife et al, 2005) é um processo de mineração de dados sequenciais para descobrir as relações de correlação que existem em conjunto de lista ordenada de eventos, neste caso, por ordem cronológica dos acontecimentos.

Para a descoberta de padrões, neste trabalho, é utilizada a aplicação de mineração WEKA[9] (*Waikato Enviroment for*

[9]http://www.cs.waikato.ac.nz/ml/weka/.

Knowledge Analysis), que é uma ferramenta desenvolvida em Java pela Universidade de Waikato, Nova Zelândia, em 1999. Atualmente, esta disponível na versão 3.8.

WEKA usa de entrada de formato ARFF (*Attrubute-Relation File Format*), forma organizada contendo domínio do atributo, valores que os atributos poderão ter e a classe. Para isso, o arquivo CWAL é transformado em ARFF. Um arquivo ARFF tem duas partes: cabeçalho que é responsável por listar todos os atributos usados e os respectivos valores; lista dos dados propriamente ditos, sendo exibidos na mesma ordem em que os atributos foram listados e separados por vírgula. A seguir um exemplo do arquivo ARFF.

```
@RELATION Sessoes
@ATTRIBUTE sessão {1, 2, 3, 4, 5, 6}
@ATTRUBUTE pagina0 {a, b, c, d, e, f, g, h, i, j, x}
@ATTRUBUTE pagina1 {a, b, c, d, e, f, g, h, i, j, x}
@ATTRUBUTE pagina2 {a, b, c, d, e, f, g, h, i, j, x}
@ATTRUBUTE pagina3 {a, b, c, d, e, f, g, h, i, j, x}
@ATTRUBUTE pagina4 {a, b, c, d, e, f, g, h, i, j, x}
@ATTRUBUTE pagina5 {a, b, c, d, e, f, g, h, i, j, x}
@ATTRUBUTE pagina6 {a, b, c, d, e, f, g, h, i, j, x}
@ATTRUBUTE pagina7 {a, b, c, d, e, f, g, h, i, j, x}
@ATTRUBUTE pagina8 {a, b, c, d, e, f, g, h, i, j, x}
@ATTRUBUTE pagina9 {a, b, c, d, e, f, g, h, i, j, x}
@DATA
1, a, b, e, b, a, c, a, b, a, d
2, a, d, a, c, x, x, x, x, x, x
3, a, c, q, x, x, x, x, x, x, x
4, a, d, i, d, a, b, f, x, x, x
5, d, h, d, i, d, h, d, g, c, g
6, a, b, f, j, x, x, x, x, x, x
```

WEKA tem opções de mineração por classificação, clustering e regras associativas, para serem utilizadas com os dados gerados no arquivo ARFF.

5. Análise dos padrões

Análise requer dos indicadores estatísticos dos padrões encontrados para tomar decisões e medidas de correção, melhorias, gerar um modelo de usuários agregados que poderão ser utilizados como entrada de dados para ferramentas de visualização e análise de Web, geração de relatórios ou mecanismos de recomendação. A ferramenta fornece resultados estatísticos de acordo os perfis desejados. Neste caso, estados de inicio de sessões; acessos condicionados; páginas de navegação breve; códigos de retorno.

5.1. Estados de inicio de sessões

O objetivo é encontrar quais são as páginas que mais são utilizadas como porta de inicio para o site pelos usuários. Esta informação é importante para fornecer ao usuário do site um acesso rápido. Neste caso, os dados utilizados são apenas as primeiras páginas no inicio de cada sessão. A Figura 5 mostra, na interface de WEKA, as frequências das páginas a e c que corresponde a primeira e terceira página. Indica que a página a esta bem como pagina inicial, mas a página c é uma opção secundaria.

5.2. Acessos condicionados

São os acessos frequentes de páginas que se relacionam entre si, ou seja, que os usuários nas sessões acessam o link X e o link Y, modelo relacionado com regra de associação da forma X➔Y. O termo de frequência de pares de links acessados por usuários fornece informações úteis sobre a disposição de links relacionados em sites. Este enfoque surge da disposição dos produtos dos supermercados para os clientes, como a típica associação {cerveja}➔{fraldas}, que significa que o cliente que pega cervejas para levar, e se vê fraldas perto, aproveita levar para os filhinho em casa.

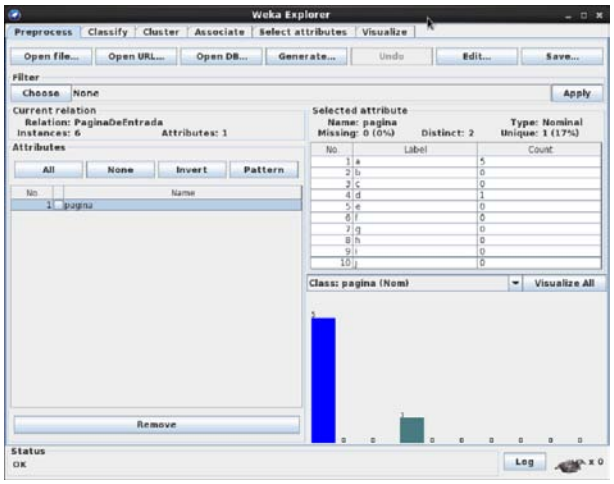


Figura 5: Frequência de páginas de entrada do site

Para se encontrar as regras de associações entre os itens (neste caso referencia às páginas), pode ser utilizado o método *a priori* (Agrawal, 2015), para encontrar conjunto de itens que possuam suporte de ocorrências (*Supo*) acima de um limite mínimo (dado como entrada). O valor de *Supo* é a possibilidade de frequência ocorrer X➔Y em relação do total de transações (ou sessões) T:

$$Supo = \frac{freq(X,Y)}{T}.$$

As regras selecionadas devem possuir um grau de confiança mínima (*Conf*) para serem utilizadas:

$$Conf = \frac{freq(X,Y)}{freq(X)}.$$

O método *a priori*, abordados em detalhe por Agrawal et al (1994) e Agrawal (2015), como ilustra o Código 1, consiste em determinar o conjunto L(k) de itens frequentes de tamanho k (conjunto com k elementos) que atende ao suporte estabelecido, a partir do conjunto C(k) de k itens candidatos. Os padrões que não são frequentes são eliminados.

Código 1: Algoritmo Apriori.

```
1 Apriori (Transaction T, MinimumSupport minSup)
2     k = 1;
3     L1 = {All frequente 1-itemsets};
4     for (k=2; Lk-1□□; k++)
5         Ck = aprioriGen (Lk-1); // News candidats
6         forall t □ T
7             Ct = subSet(Ck, t); // candidate em t
8             forall c □ Ct
```

```
9                c.count++;
10          Lk = {c □ Ck / c.count ≥ minSup};
11      Answer = U_k Lk
```

O função aprioriGen gera conjunto de itens de candidatos, enquanto subSet extrai regras associadas em si.

A seguir é ilustrado conteúdo do arquivo ARFF para análise de esse perfil por WEKA.

```
@Relation acessoSessao
@ATTRIBUTE a {true}
@ATTRIBUTE b {true}
@ATTRIBUTE c {true}
@ATTRIBUTE d {true}
@ATTRIBUTE e {true}
@ATTRIBUTE f {true}
@ATTRIBUTE g {true}
@ATTRIBUTE h {true}
@ATTRIBUTE i {true}
@ATTRIBUTE j {true}

@DATA
true, true, ?, ?, ?, true, ?, ?, ?, true
true, true, ?, ?, ?, true, ?, ?, ?, true
true, true, ?, ?, ?, true, ?, ?, ?, true
true, true, ?, ?, ?, true, ?, ?, ?, true
true, true, ?, ?, ?, true, ?, ?, ?, true
true, true, ?, ?, ?, true, ?, ?, ?, true
true, true, ?, ?, ?, true, ?, ?, ?, true
true, true, ?, ?, ?, true, ?, ?, ?, true
true, ?, true, ?, ?, ?, true, ?, ?, ?
true, true, ?, ?, ?, true, ?, ?, ?, true
true, true, ?, ?, ?, true, ?, ?, ?, true
```

Neste arquivo é indicado por true se uma página foi acessada em cada uma das sessões de navegação. Por exemplo, na primeira linha de @DATA, nas respetivas posições, são indicadas com true as páginas acessadas a, b, d, f, j. A Figura 6 mostra a execução de regras de associação pelo algoritmo *a priori* de WEKA com suporte mínimo minSup = 0.2 e confiança 0.5. Observa-se na primeira iteração encontra L(1) ={a, b, c, f, j} de 5 elementos, sendo que o item a foi acessada nas 13 sessões, por tanto seu fator de suporte é 1; já o item c foi acessada apenas em 3 sessões, portanto tem um fator de suporte 0,23. Na segunda iteração é achado 7 relações, L(2) = {{a,b}, {a,c}, {a,f}, {a,j}, {b,f}, {b,j}, {f,j}}. O processo é executado até que k tal que L(k) não possua subconjunto com o suporte válido. Neste caso até k=5 que L(5) = { }, porque L(4) = {{a,b,f,j}}.

Foram encontradas 51 regras possíveis, tal como mostra a Figura 6 (continuação - direita), as quais deverão ser analisadas em função do parâmetro de confiança conf (0.5). Observando a regra 5, conclui-se que todas as sessões que visitaram a página b também visitaram a página f, com grau de confiança 1, que esta acordo da jerarquia ilustrada pela Figura 2, por tanto a estrutura da página esta bem. Similar caso ocorre com a regra 9, com as páginas f e j.

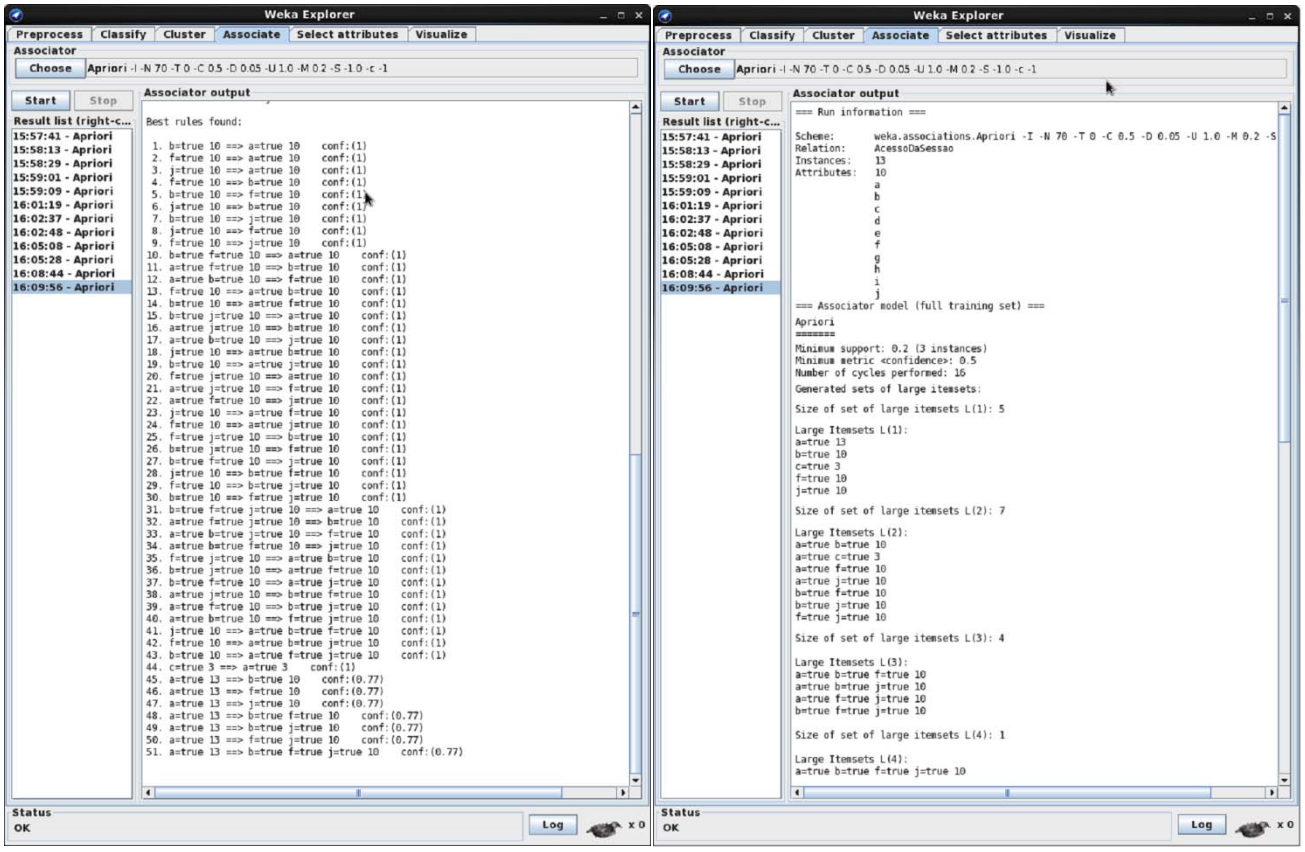


Figura 6: Operação por regras de associação em WEKA

Em relação ao comportamento dos usuários, tem algo que não esta funcionando tal como prevista na estrutura da Figura 2, pois não esta seguindo o fluxo previsto, fato que pode ser relacionado a dois atributos de usabilidade como

Aprendizagem e *Memorabilidade*. No primeiro caso, indica que possivelmente os usuários estejam enfrentando dificuldades de navegação. O segundo indica que provavelmente, o usuário não esta lembrando qual era o caminho que seguiu da última vez em que acessou. Por tanto, o designer deve investigar o motivo de não seguir a sequencia do fluxo comum, tal como o link da página seguinte não esta na posição intuitiva ou visível, ou texto não é significativo.

5.3. Outros perfis

Análise de código de retorno

O código de retorno 500 significa que aconteceu um erro no servidor. O responsável pelo site deve corrigir na direção correta na página. A Figura 7 (esquerda) mostra o código de retorno 404, que indica *page not found*, gerando um impacto negativo de interesse do usuário pelo uso da página. A análise de código de retorno é importante para que, neste caso, o atributo de usabilidade *Minimização de Erros* seja melhorado.

Páginas com pouco tempo de navegação

Páginas com pequeno tempo de navegação pode indicar que possivelmente a página não deve ser elemento de ligação ou se o conteúdo dela não esta prendendo a atenção do usuário. Para identificar a página basta subtrair horário de requisição de acesso da pagina seguinte. A regra 7, da Figura 7 (direita), com grau de conf=1, mostra que a página a tem duração menos que 1, o que faz supor que esta é apenas uma página de ligação; enquanto a regra 1 mostra que todos os acessos à página c ficaram mais de 1 minuto navegando por ela, portanto permite supor que esta página não é uma simples página de ligação, ela deve conter algum conteúdo que prende a atenção do usuário. Pode ser considerada que a página a é de conteúdo relevante para o site, e não apenas uma página de ligação.

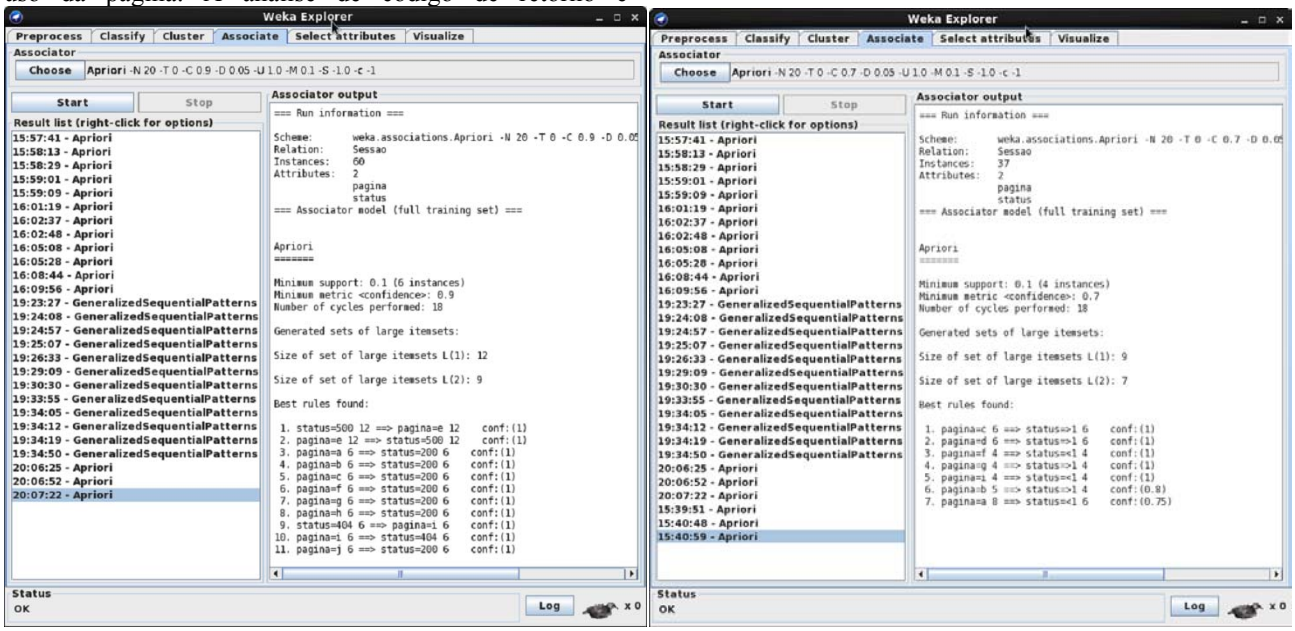


Figura 7: Análise: código de retorno (esquerda) e páginas com pouco tempo de navegação (direita)

6. Conclusões e trabalhos futuros

A mineração de dados na web permite que páginas possam ser enriquecidas e atualizadas constantemente, guiados pelos indicadores dos comportamentos dos usuários de maneira que cada vez seja melhorado o atributo de aceitação.

Nestes processos, utilizar os logs dos servidores permite verificar a usabilidade das páginas neles hospedadas, para melhorar os objetivos atingidos para os quais foram desenvolvidas. Os resultados da mineração podem ser analisados e interpretados como um conhecimento que auxiliasse o responsável pelo site, seja ele um web designer ou um desenvolvedor, em correções de problemas e também em alterações que tornassem a navegação dos usuários mais simples, fluida e agradável.

O processo seguido, neste trabalho, desde a coleta de dados até o *path completion*, é fundamental para conseguir uma boa informação e obter um conhecimento confiável na análise dos perfis. Embora este processo demande técnicas elaboradas, em particular para o *path completion*, com a ajuda de informações de construções de páginas e dados ocultos no pr*oxy*, podem ser melhorados. Na seguinte etapa, descoberta de padrões, as ferramentas de WEKA facilita processar os dados completos dos logs para uma análise dos perfis dos usuários desde o ponto de vista de atender atributos para os conceitos de usabilidade.

Entre os trabalhos futuros que podem ser pesquisados e desenvolvidos, é fazer uma análise do perfil de navegação observando como parâmetros o tempo total das sessões registradas nos logs do servidor. Também se pode avaliar a possibilidade de implementar a realização do pré-processamento de dados, através de um programa próprio, na ferramenta WEKA como possível adaptação, para os logs oriundos dos logs.

Referências bibliográficas

Agrawal, R.; Srikant, R. (1994), Fast Algorithms for Mining Association Rules. ACM: *Proceedings of the 20th VLDB Conference,* Santiago Chile, pp 487-499.

Agrawal, R. (2015). Data Mining, The Textbook. Springer, pags 734.

Campos, M. M. (2012), Avaliação de Usabilidade de Sites Web. Revista Caminhos, On-Line, v. 5, p. 189-203.

Carmona, C. J. et al. (2012), Web usage mining to improve the design of an e-commerce website: Orolivesur.com. *Expert Syst. Appl.*, v. 39, n. 12, p. 11243-11249.

Cooley, R. W. (2000), Web usage mining: Discovery and application of Interesting Patterns from Web data. Tese (Doutorado), Faculty of The Graduate School, University of Minnesota.

Cooley, R.; Mobasher, B.; Srivastava, J. (1999), Data preparation for mining world wide web browsing patterns. *Knowl. Inf. Syst*., v. 1, n. 1, p. 5-32.

Chitraa, V.; Davamani, A.S. (2010) An Efficient Paht Completion Technique for Web Log Mining. *IEEE International Conference on Computational Intelligence and Computing Research.*[s.p].

Cho, Y. H.; Kim, J. K.; Kim, S. H. A personalized recommender system based on web usage mining and decision tree induction. *Expert Syst. Appl*., v. 23, n. 3, p. 329-342.

Cutting, D. R. et al. (1992) A cluster based approach to browsing large document collections. *Proceedings of the Fifteenth International Conference on Research and Development in Information Retrieaval*.

Duan, J.L.; Liu, S.X. (2012), Application on web mining for web usability analysis. *ICMLC. IEEE*, 2012. p. 1981-1985.

Ezeife, C.I.; Lu, Y..(2005) Mining Web Log Sequential Patterns with Position Coded Pre-Order Linked WAP-Tree. *Springer Science: Data Mining and Knowledge Dsicovery,* 10,pp 5-38.

Hipp, J.; Güntzer, U.; Nakhaeizadeh, Gh.(2000) Algorithms for Association Rule Mining – A General Survey and Comparation.*ACM SIGKDD Exploration,*Vol. 2, Issue. 1,pp 58-64.

Jeria, V. H. E. (2007). Minería Web de Uso y Perfiles de Usuario: Aplicaciones con Lógica Difusa. Tese (Doutorado), Departamento de Ciencia de la Computación e Inteligencia Artificial, Universidad de Granada, 2007.

Li, Y.; Feng, B.; Mao, Q. (2008) Research on Path Completion Technique in Web Usage Mining.*IEEE International Symposium on Computer Science and Computationjal Technology,* pp 554-559.

Liu, B. Web Data Mining. [S.l.]: Springer, 2007.

Mitharam, M. D. (2012), Preprocessing in web usage mining. *International Journal of Scientific & Engineering Research*, v. 3.

Nielsen, J. (1993), Usability Engineering, Academic Press.

Pal, S. K.; Talwar, V.; Mira, P. (2002) Web Mining in Soft Computing Framework: Relevant, State of the Art and Future Directions. *IEEE Transaction on Neural Networks,* v. 13, p. 1163-1177.

Raiyani, Sh.; Jain, Sh. (2012) Efficient Preprocessing Technique Using Web log Mining. *IEEE International Journal of Advancements in Research & Technology.*[s.p].

Sanjay, M. (2010), An Efective and Complete Preprocessing for Web Usage Mining.

Shneideman, B.; Plaisant, C. (2009), Designing the User Interface: trategies for Efefctive Human-Computer Interaction, 5th Edition, Pearson, 606 pags.

Srivastava, J. et al. (2000) Web usage mining: discovery and applications of usage patterns from web data. *SIGKDD Explor. Newsl., ACM*, New York, USA, v. 1, n. 2, p. 12-23.

Zhang, X.; *Edwards*, J.; Harding, J. (2007), Personalised online sales using web usage data mining. *Comput. Ind.*, Elsevier Science Publishers, v. 58, n. 8-9, p. 772-782.