Asociación Argentina
de Mecánica Computacional

Mecánica Computacional Vol XXIX, págs. 9407-9423 (artículo completo)
Eduardo Dvorkin, Marcela Goldschmit, Mario Storti (Eds.)
Buenos Aires, Argentina, 15-18 Noviembre 2010

METODOLOGIA PARA O DESENVOLVIMENTO DE UM WEB SITE ADAPTATIVO UTILIZANDO TÉCNICAS DE MINERAÇÃO DO USO DA WEB E REGRAS DE ASSOCIAÇÃO

Pedro H. de Oliveira e Silva[a], **Gray F. Moita**[a] **e Thiago M. Rodrigues Dias**[a]

[a]*Centro Federal de Educação Tecnológica de Minas Gerais Av. Amazonas, 7675, 30.510-000 Belo Horizonte, MG, Brasil, http://www.cefetmg.br/*

Palavras Chave: Mineração do Uso da Web, Web Sites Adaptativos, Arquitetura Orientada a Serviços, Serviços Web.

Resumo. Este artigo apresenta uma metodologia para o desenvolvimento de Web Sites Adaptativos: sites que automaticamente melhoraram a sua organização e apresentação a partir da aprendizagem de padrões de acesso dos seus visitantes. Para tal desenvolvimento é utilizado técnicas de mineração do uso da web e regras de associação. Para facilitar o desenvolvimento de tal metodologia foi definido a utilização do conceito de Arquitetura Orientada a Serviços (SOA - Service Oriented Architecture). Esta arquitetura se preocupa com a construção independente de serviços, negócios alinhados que podem ser combinados em significantes processos de negócio de alto nível e soluções dentro do contexto do empreendimento. Todavia, de forma a avaliar e validar a metodologia, um protótipo de site adaptativo, focado em uma instituição de ensino (Centro Federal de Educação Tecnológica de Minas Gerais - CEFET-MG) vem sendo desenvolvido, no sentido de fornecer a apresentação do material em diferentes mídias adaptando-se ao perfil / estilo do usuário. Acredita-se que a utilização destas técnicas poderão ser utilizadas para examinar os logs de acesso dos usuários, descobrindo padrões de acesso automaticamente, a fim de tornar o web site adaptativo.

1 INTRODUÇÃO

O desenvolvimento e a organização do conteúdo de um web site não tem sido uma tarefa trivial. Normalmente um web site sintetiza diversas informações distribuídas entre as páginas, imagens e hiperlinks. Mesmo que o volume de acesso do site não seja grande, geralmente os seus usuários são diversificados. Cada visitante que acessa um site pode ter um objetivo diferente, pode estar procurando por informações diferentes. É muito provável que a maioria dos usuários da Internet tenham visitado sites que poderiam ter a informação, o produto ou o serviço que ele estava procurando, mas não conseguiu encontrá-los e se dirigiu para um outro site. Ao mesmo tempo que projetar um web site é uma tarefa complicada, é muito difícil identificar as características do público alvo. Assim, construir uma interface que atenda os requisitos de todos os usuários de um site não é uma tarefa fácil. Se o projetista tivesse uma maneira de conhecer o seu público, grande parte dos problemas de interação entre usuário e interface poderiam ser resolvidos. Assim, a fim de auxiliar na tarefa de conhecer o público que um site possui, várias técnicas estão disponíveis como a mineração de dados na web (WebMining).

Independente das características dos usuários de web sites, a sua principal necessidade consiste em encontrar a informação desejada de modo fácil e rápido. Ainda que seja possível identificar o comportamento de todos os usuários em um site, tornasse difícil disponibilizar informações de forma clara e simples para todos. Para isto, um site adaptativo, que se ajusta automaticamente a cada usuário de acordo com seus padrões de comportamento, é muito útil.

Sites adaptativos são desenvolvidos com base em técnicas que auxiliam o projetista na tarefa de personalizar páginas web e por este motivo, são chamadas de técnicas de personalização. Algumas possíveis conseqüências da aplicação dessas técnicas são: aumento da média de páginas visitadas por sessão, os usuários podem demorar mais tempo visitando o site; pode haver um aumento da taxa de retenção (número de visitantes que retornam ao site) e da taxa de conversão (visitas que se convergem em possíveis compras) (Greening (2000)).

Nossa abordagem básica é analisar os logs de acesso de web, afim de encontrar grupos de usuários que frequentemente acessam páginas semelhantes, assumindo que estas páginas representam temas coerente na mente dos usuários. Para tal análise utilizaremos os conceitos de algoritmos de aprendizagem e clustering. Os algoritmos de aprendizagem, possibilita encontrarmos uma descrição conceitual de uma classe de objetos de uma coleção de dados com base em exemplos da classe (exemplos positivos) e exemplos de objetos não pertencentes a classe (exemplos negativos). Essencialmente, o algoritmo deve determinar o que os objetos têm em comum que os distingue de outros objetos na coleção. Contudo, o conceito de algoritmos de aprendizagem são algoritmos de aprendizado supervisionado: eles exigem que suas entradas sejam classificadas antecipadamente. No nosso domínio, tudo o que temos é uma coleção de objetos (páginas Web) e alguns dados sobre padrões de uso. Algoritmos de agrupamento (clustering), por outro lado, não são supervisionados: eles recebem uma coleção de objetos como entrada e produzem uma divisão desta coleção - uma classificação onde cada objeto se encontra exatamente dentro de sua classe, ou "cluster".

Assim este artigo visa realizar um estudo sobre a adoção de características e vantagens de técnicas de mineração do uso da web e regras de associação, utilizando o conceito de Arquitetura Orientada a Serviços (SOA - Service Oriented Architecture) para o desenvolvimento de web sites adaptativos, apontando os benefícios que este tipo de metodologia pode trazer para o desenvolvimento de sistemas adaptativos, gerando automaticamente melhorias e sugestões a partir de logs do servidor web, melhorando a organização e apresentação de web sites.

2 WEB SITES ADAPTATIVOS

Adaptação, em informática, significa definir um conjunto de parâmetros para atender às exigências de um usuário específico; ajustar para o uso pessoal. Em Batista (2008) é apontado alguns fatores que justificam a opção pela personalização:

Quando um web site é voltado a um público com perfil diversificado; estilos cognitivos variados; abrangendo iniciantes até experts em interação humano-computador; de crianças até idosos; desde os que possuem maior conhecimento acerca do conteúdo abordado, até aqueles que estão começando a se interessar pelo assunto; em suma, quando existem diversos 'modelos de usuários', torna-se viável promover a adaptação de conteúdo, navegação e apresentação.

Uma alternativa que promove melhor assistência à heterogeneidade de perfis de usuários é o desenvolvimento de Web Sites Adaptativos (Brusilovsky (2004)). Segundo Koch (2000), os Sistemas Web Adaptativos potencializam a abordagem centrada no usuário: o sistema adapta os aspectos visíveis de acordo com o 'modelo do usuário' (construído a partir de dados do usuário), gerando uma interface que disponibiliza a "informação apropriada, com layout adequado para cada usuário".

Segundo Brusilovsky et al. (2004), "a Web Adaptativa tem atraído atenção considerável devido ao seu potencial para fornecer aplicações e serviços personalizados para os cidadãos da sociedade do conhecimento".

Os web sites adaptativos promovem, automaticamente, sua organização e apresentação de acordo com os padrões de acesso do usuário. As páginas são mais acessíveis, há possibilidade de destacar links interessantes, conectar páginas relacionadas e promover agrupamento de documentos similares (Perkowitz e Etzioni (2009)).

Um Web Site pode ser adaptativo de duas formas básicas, a personalização e a Otimização. A personalização é a adaptação da interface do site devido as necessidades dos seus visitantes individuais, com base em informações sobre esses indivíduos. A otimização o processo que define uma estrutura melhor para o site baseado nas interações de todos os seus visitantes. Em vez de fazer alterações para cada indivíduo, o site aprende a partir de inúmeros visitantes, tornando sua utilização mais fácil para todos, incluindo aqueles que nunca usaram antes.

Sites adaptativos observam as atividades dos usuários, os seus erros e aprendem sobre os perfis de usuário, sobre os seus padrões de acesso e problemas com a organização do conteúdo de um site. A personalização de web sites é uma estratégia para aproveitar as informações deixadas pelo usuário com o objetivo de tornar o site mais próximo das necessidades do seu público.

A seguir, apresenta-se uma taxonomia discutida em Ruas et al. (2001) que classifica as técnicas de personalização existentes e define todo o processo de personalização de um site. O processo de criação de um site adaptativo envolve quatro fases Figura 1:

- definição dos objetivos: o administrador do site deve definir as metas a serem alcançadas com a personalização;
- observação: a coleta dos dados de acesso ao site;
- transformação: gera as regras de adaptação a partir dos dados coletados na fase anterior;
- aplicação: a utilização das regras na sua estrutura.

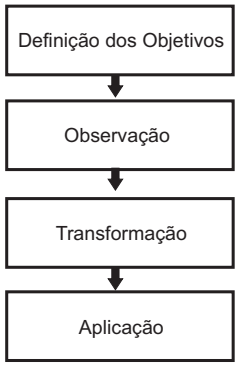

Figura 1: Processo de Personalização de um Web Site (Ruas et al. (2001)).

3 MINERAÇÃO DE DADOS NA WEB

Nos últimos anos tem crescido a aplicação da mineração de dados em um ambiente altamente dinâmico a Internet. Esta nova aplicação tem sido denominada de mineração de dados na WEB e o crescimento de sua importância pode ser atribuído a alguns fatores como: a inexistência de padrões, a falta de estruturação e a heterogeneidade; a Internet apresentou um crescimento acentuado e como conseqüência um aumento significativo no acúmulo de informações e as mesmas mudam muito rapidamente. Neste contexto observa-se a existência de um mundo de dados altamente dinâmico, e com extrema riqueza, sendo assim, torna-se um meio de grande interesse para aplicação da mineração de dados para um melhor aproveitamento das informações através da descoberta e extração de conhecimento.

A *WEB* (Zaiane (2000)) pode ser definida como uma fonte de matéria-prima, amplamente distribuída, altamente heterogênea, semi estruturada ou não estruturada, interconectada, evolutiva, repositório de informações de hipertexto/hipermídia. Este meio apesar da abundância e diversidade de informações apresenta problemas para o seu uso devido a dinamicidade e diversificação de estruturas que a caracterizam.

Quando se enfoca a mineração de informações no ambiente da Internet, utiliza se a expressão mineração de dados na WEB ou WEB MINING, que segundo Cook (2000) pode ser definida da seguinte forma:

Mineração de dados na WEB pode ser definida como a descoberta e análise de informação útil originada na WEB.

Sendo assim, mineração de dados na WEB configura-se em um processo não trivial de mineração ambientado na Internet.

A mineração de dados na Web tem como objetivo obter o conhecimento ou descobrir informações uteis através da **estrutura da Web (Hyperlink)**, o **conteúdo da página**, e **dados do uso**. Apesar mineração da *Web* utilizar diversas técnicas de mineração de dados, não é meramente uma aplicação de mineração de dados tradicionais, devido à heterogeneidade e a natureza semi-estruturada ou não dos dados da *Web*. Muitas tarefas de mineração e novos algoritmos foram inventadas na década passada. Com base nos principais tipos de dados utilizados no processo de mineração, as tarefas de mineração da Web podem ser classificados em três subáreas conforme a figura 2: mineração da estrutura da *Web*, mineração do o conteúdo da *Web* e a mineração do uso da *Web*.

Devido à riqueza e diversidade de informações na *Web*, há um grande número de tarefas de mineração *web*, neste trabalho vamos concentrar-se na mineração do uso da *Web*.

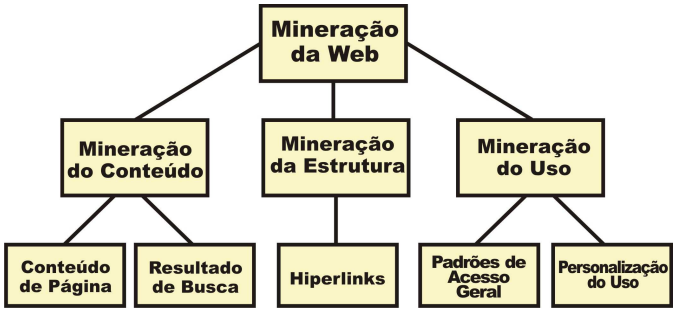

Figura 2: Taxonomia da mineração da WEB (Zaiane (2000)).

3.1 Mineração do Uso da Web

Com o contínuo crescimento e proliferação do comércio eletrônico, serviços *Web* e Sistemas de Informação baseados na *Web*, os volumes de dados e de **Clickstream** (Sequência de Cliques) do usuário coletados por Organizações Baseadas na *Web* em suas Operações diárias atingiu proporções astronômicas.

A análise estes dados podem ajudar essas organizações determinar o de tempo de vida de clientes, desenhar caminhos estratégicos de marketing através de produtos e serviços, avaliar a eficácia das campanhas promocionais, otimizar a funcionalidade dos aplicativos baseados na *Web*, fornecer um conteúdo mais personalizado aos visitantes, e encontrar a estrutura lógica mais eficaz para o seu espaço *Web*. Este tipo de análise envolve a descoberta automática de padrões e relações significativas de uma grande coleção de dados semi-estruturados, muitas vezes armazenados em aplicações *Web* e *logs* de acesso ao servidor, bem como nas respectivas fontes de dados operacionais.

A mineração do uso da Web refere-se à descoberta e análise automática de padrões em dados de páginas visitadas e associadas coletados ou gerados como resultado de interações do usuário com os recursos da *Web* em um ou mais sites da *Web* (R. Cooley e Srivastava (1997), Mobasher (2006), J. Srivastava e Tan (2000)). O objetivo é capturar, modelar e analisar os padrões comportamentais e perfis de usuários que interagem com um *Web Site*. Os padrões descobertos são geralmente representados como conjuntos de páginas, objetos ou recursos que são acessados com freqüência por grupos de utilizadores com necessidades ou interesses comuns.

Na sequência do processo padrão de mineração de dados (U. M. Fayyad e Smyth (1996)), o uso global do processo de mineração da *Web* pode ser dividido em três fases inter-dependentes: coleta dos dados e pré-processamento, descoberta de padrões e análise de padrões. Na fase de pré-processamento, os dados de *clickstream* são limpos e particionados em um conjunto de transações do usuário que representam as atividades de cada usuário durante diferentes visitas ao site. Outras fontes de conhecimento, tais como o conteúdo do site ou da estrutura, bem como o conhecimento semântico das ontologias de domínio do site (tais como catálogos de produtos ou hierarquias conceito), também pode ser usado em pré-processamento de dados ou de melhorar a transação do usuário. Na fase de descoberta de padrões, estatísticas, banco de dados e operações de máquinas de aprendizagem são realizados para obtenção de padrões escondidos refletindo o comportamento típico dos usuários, bem como resumos estatísticos sobre os recursos da *Web*, sessões e usuários. Na fase final do processo, os padrões descobertos e as estatísticas são futuramente processados, filtrados, possivelmente resultando em modelos de usuários agregados que pode ser utilizado como entrada em aplicações tais como sistemas de recomendação, ferramentas de visualização e análise *Web* e ferramentas de geração de relatórios.

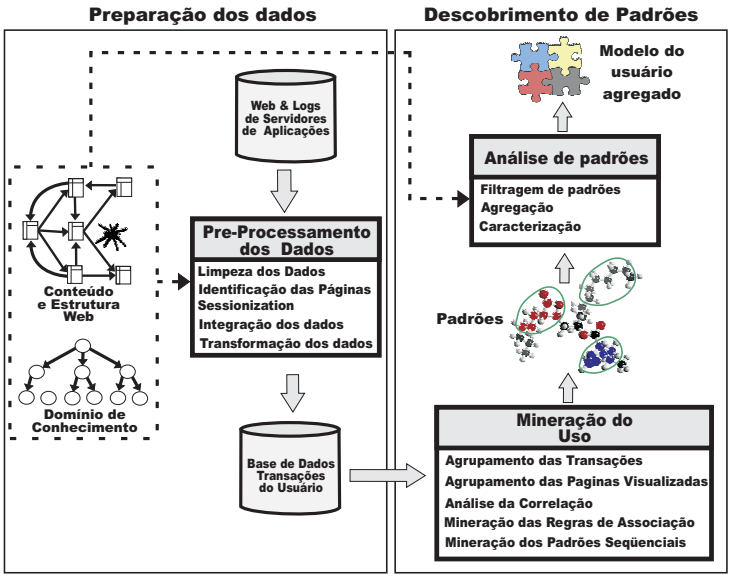

Figura 3: Processo de Mineração do uso da Web Zaiane (2000).

O processo global é representado na figura 3.

No restante desta seção, apresentamos uma análise detalhada da mineração do uso da *Web* como um processo, e discutir os conceitos relevantes e técnicas comumente utilizadas em todas as diversas fases mencionadas acima.

3.2 Mineração de Regras de Associação

As Regras de associação são uma importante classe para encontrar relacionamentos ou padrões frequentes entre conjuntos de dados. Mineração de regras de associação é uma tarefa fundamental de mineração de dados. É talvez o modelo mais importante inventado e amplamente estudado pela comunidade de banco de dados e mineração de dados. Desde que foi introduzido por R. Agrawal e Swami (1993), tem atraído muita atenção. Muitos algoritmos eficientes, extensões e aplicações têm sido relatados.

Como exemplo de uma aplicação clássica de regras de associação pode-se citar um carrinho de compras em um supermercado, onde o objetivo desta aplicação é descobrir quais itens existentes no carrinho de compras dos seus clientes estão associados. Uma regra de associação é por exemplo:

$$Carne \rightarrow Cerveja\,[suporte = 10\%confianca = 80\%]$$

A regra extraída no exemplo citado diz que 10% dos clientes compram carne e cerveja, e 80% de quem compra carne também compra cerveja. Suporte e confiança são duas medidas de força de regra, que serão definidos no decorrer do trabalho.

Este modelo de mineração é de fato muito geral e que pode ser usado em muitas aplicações. Por exemplo, no contexto da Web e documentos de texto, ele pode ser usado para encontrar relações de co-ocorrência de palavras e padrões de uso da Web.

3.2.1 Conceitos Básico sobre Regras de Associação

O problema de mineração de regras de associação pode ser declarado do seguinte modo (R. Agrawal e Swami (1993)): seja $L = i_1, i_2, \ldots, i_n$ um conjunto literais chamados de itens.

Seja D um conjunto de operações (transações), no qual cada transação T é um conjunto de itens tal que $T \subseteq L$. Associado com cada transação está um atributo que a identifica unicamente, chamado TID. Uma transação T contém X, sendo X um conjunto de itens em L, se $X \subseteq T$. Uma regra de associação é uma implicação do tipo $X \rightarrow Y$, onde $X \subset L, Y \subset L$ e $X \cap Y = \oslash$. A regra $X \rightarrow Y$ é válida no conjunto de transações D com o grau de confiança c, se $c\%$ das transações em D que contêm X também contêm Y. A regra $X \rightarrow Y$ tem suporte s em D se $s\%$ das transações em D contêm $X \cup Y$. Um conjunto X contendo k itens é chamado de um conjunto-de-k-itens. O conjunto de itens X, que aparece à esquerda do operador de implicação, é denominado antecedente(ou precedente) da regra; por sua vez, o conjunto de itens Y, que aparece à direita do operador, é denominado de consequente.

O problema na mineração de regras de associação, dado um conjunto de transações, está em gerar todas as regras que tenham suporte e confiança maiores do que os valores mínimos definidos pelo usuário, os quais, geralmente, são referidos como $minsup$ e $minconf$, respectivamente. Se o suporte de um conjunto de itens for maior ou igual ao mínimo estabelecido ($sup(X) \geq minsup$), diz-se que é frequente. O suporte de uma regra $X \rightarrow Y$ é dado por $sup(XY)$ e a sua confiança é $sup(XY)/sup(X)$ (Agrawal et al. (1996)). Dentro do conceito de que uma regra trata-se de uma afirmação sobre uma distribuição probabilística, o suporte pode ser descrito como a probabilidade de que uma transação qualquer satisfaça tanto X como Y, ao passo que a confiança é a probabilidade de que uma transação satisfaça Y, dado que ela satisfaz X.

Um grande número de algoritmos de mineração de regras de associação, têm sido relatados na literatura, onde cada um com diferentes benefícios na mineração. Os conjuntos de regras resultantes são, no entanto, todos iguais com base na definição de regras de associação. Isto é, dado um conjunto de dados de transações T, um $minsup$ e uma $minconf$, o conjunto de regras de associação existentes em T é unicamente determinado. Qualquer algoritmo deve encontrar o mesmo conjunto de regras, embora sua eficiência computacional e os requisitos de memória pode ser diferente. Este trabalho propõe a utilização do algoritmo Apriori proposto por Agrawal e Srikant (1994), que será descrito abaixo.

3.2.2 O Algoritimo Apriori

O algoritmo *Apriori* foi proposto em 1994 pela equipe de pesquisa do Projeto QUEST da IBM que originou o software *Intelligent Miner*. Trata-se de um algoritmo que resolve o *problema da mineração de itemsets frequentes*, isto é, dados um banco de dados de transações *D* e um um nível mínimo de suporte β, o algoritmo encontra todos os itemsets frequentes com relação a *D* e β.

O algoritmo Apriori funciona em duas etapas:

1. encontrar todos os conjuntos de itens freqüentes, isto é, com suporte acima do suporte mínimo estabelecido. Por ser a etapa mais onerosa em termos de uso de CPU e de E/S, esta recebe a maior atenção no projeto de algoritmos de mineração como o caso do Apriori;

2. para gerar as regras de associação utilizando os conjuntos de itens freqüentes, devem-se selecionar apenas as regras que possuam o grau de confiança mínimo, correspondente a $minconf$, o que pode ser implementado da seguinte forma: para cada conjunto de item freqüente l, encontram-se todos os subconjuntos não vazios de l; para cada subconjunto

```
Procedure Apriori
    L1 = {frequent 1 - itemsets};
    for (k=2; Lk-1 ≠ 0; K++ ) do
        Ck = apriori_gen(Lk-1);
        forall transactions t ∈ Ct do
            Ct = subset(Ck, t);
            forall candidates c ∈ Ct do
                c.count++;
            od
        od
        Lk = {c ∈ Ck | c.count ≥ minsup};
    od
    Answer = Uk Lk;
end
```

Figura 4: Algoritmo APRIORI. (Liu (2007))

a, gera-se uma regra na forma $a \rightarrow (l - a)$, se o suporte de l dividido pelo suporte de a for, no mínimo, igual ao $minconf$.

Esse algoritmo faz diversas passagens sobre a base de transações para encontrar todos os conjuntos de itens freqüentes; em cada um desses passos, primeiro, gera conjuntos de itens candidatos e, depois, percorre a base de dados para determinar se os candidatos satisfazem o suporte mínimo estabelecido. Na primeira passagem, o suporte para cada item individual (conjuntos-de-1-item) é contado e todos aqueles que satisfazem o suporte mínimo são selecionados, constituindo-se os conjuntos-de-1-item freqüentes. Na segunda iteração, conjuntos-de-2-itens candidatos são gerados pela junção dos conjuntos-de-1-item e seus suportes são determinados pela pesquisa no banco de dados, sendo, assim, encontrados os conjuntos-de-2-itens freqüentes. Assim, o algoritmo, apresentado na Figura 4, prossegue iterativamente, até que o conjunto-de-k-itens freqüentes encontrado seja um conjunto vazio.

Esse algoritmo usa o princípio de que cada subconjunto de um conjunto de itens frequente também deve ser frequente. Tal constatação é utilizada para reduzir o número de candidatos a serem comparados com cada uma das transações no banco de dados. Todos os candidatos gerados que contenham algum subconjunto que não seja freqüente são eliminados.

4 ARQUITETURA ORIENTADA A SERVIÇOS

Arquitetura de software é uma descrição de um sistema de software em termos de seus componentes principais, as relações deles, e a informação que passam entre eles. Em essência, arquitetura é um plano para construir sistemas que satisfazem para exigências bem-definidas e, através de extensão, sistemas que possuem características precisaram satisfazer as exigências agora e no futuro.

O propósito fundamental de arquitetura de software é ajudar administrar a complexidade de sistemas de software e as modificações que sistemas inevitavelmente sofram em resposta a mudanças externas no negócio, organizacional, e ambientes técnicos.

Segundo ANSI/IEEE (2000), uma arquitetura de software trata basicamente de como os componentes fundamentais de um sistema se relacionam intrinsecamente e extrinsecamente (ANSI/IEEE (2000)). Enquanto arquitetura de software tradicional é enfocada na construção de aplicações de software, SOA é enfocado na construção de soluções com um empreendimento ou âmbito inter-organizacional, baseado nas interações entre consumidores com necessidades (freqüentemente processos empresariais) e provedores com capacidades (serviços).

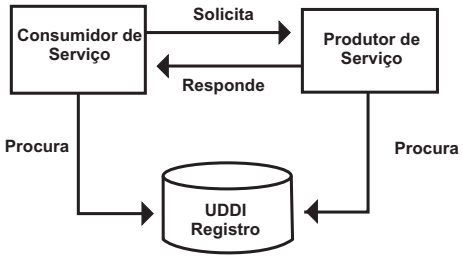

Figura 5: Paradigma Básico para Arquitetura Orientada a Serviços.

SOA é um estilo arquitetônico para construir soluções de empreendimento baseado em serviços. Mais especificamente, SOA se preocupa com a construção independente de serviços negócio-alinhados que podem ser combinados em significantes, processos de negócio de alto-nível e soluções dentro do contexto do empreendimento.

A arquitetura orientada a serviços é uma abordagem da tecnologia da informação ou uma estratégia em que as aplicações fazem o uso dos serviços disponíveis em uma rede, tais como a internet. O conceito de SOA permite encontrar uma solução relativamente barata e com um custo benefício maior quando se refere a sistemas que precisam conversar entre si e processos que demandam maior flexibilidade e agilidade para atender as revoluções do mercado. A arquitetura orientada a serviço representa um novo modelo de evolução para a construção aplicações distribuídas Erl (2004).

A arquitetura orientada a serviços é atualmente reconhecida como um tipo de desenvolvimento significativo, especialmente para sistemas de aplicações de negócios. Ela permite maior confiabilidade, integridade de mensagem, integridade transacional e segurança de mensagem.

A Figura 5 mostra o conceito de SOA que resume as três principais entidades em uma solução típica da arquitetura orientada a serviços:

- Consumidor de serviço
- Produtor de serviço
- Diretório de Serviço

Há então dois pontos importantes no processo de Arquitetura Orientada a Serviços, o consumidor, que consome e requisita os resultados ao provedor, ou seja, o outro lado da questão executa o serviço e responde às necessidades [Figura 5].

O importante é que para ser considerado arquitetura SOA, um componente deve ser um serviço, o que implica em: granularidade grossa; relevância e alto nível da transação executada; baixo acoplamento; interface bem definida e detalhes de implementação bem encapsulados (Sampaio (2006)).

5 VIABILIDADE PARA O DESENVOLVIMENTO DE SITES ADAPTATIVOS UTILIZANDO TÉCNICAS DE WEB MINING

O desenvolvimento de diferentes tipos web sites adaptativos têm sido muito explorado recentemente. É bastante comum web sites permitirem aos seus utilizadores personalizar o local para os mesmos. A maioria dos "portais", por exemplo, permitem personalizações manual, tais como listas dos links favoritos, imagens, cotações de ações do interesse do utilizador, relatório de condições climáticas locais, entre outras funções. Um site também pode, por exemplo, manter o controle das páginas que foram alteradas desde a última visita de um determinado usuário.

Alguns sites também permitem que seus usuários descrevam os seus interesses e a partir disso apresenta informações relevantes para esses interesses. Sites mais sofisticados realizam previsões de caminhos: supõe que o usuário quer ir em um determinado lugar e direciona-o de imediato (ou, pelo menos, fornece um link). A previsão de caminho pode ser feito on-line, prevendo o objetivo do usuário com base em seu caminho, até então, ou pode ser feito offline, estaticamente calculado com base em modelos de usuário. Em Joachims T. (1997) é proposto o sistema WebWatcher que aprende a prever quais links os usuários irão acessar em uma página específica, em função de um modelo de seus interesses. O WebWatcher destaca graficamente o link, que um determinado usuário provavelmente irá acessar e o repete na parte superior da página quando a mesma é apresentada. Já em Fink J. (96) é proposto o sistema AVANTI que centra-se na personalização dinâmica baseada nos gostos e necessidades dos usuários. Com base no que sabe sobre o usuário, o AVANTI tenta prever tanto o objetivo final do usuário e seu provável próximo passo., destacando os links que levam diretamente às páginas que ele pensa que um usuário vai querer acessar, baseado no interesses do usuário. Os dois sistemas apresentados dependem, em parte, de informações fornecidas pelos usuários quando eles entrarem no site. Em Yan T. (1999) é proposto uma abordagem de personalização automática. É realizado técnicas de mineração de dados na web, mais especificadamente uma clusterização sobre os arquivos de log do site, a fim de identificar as categorias de usuários. Essas categorias podem ser utilizadas para classificar novos visitantes do site, para personalizar o site enquanto eles navegam. Por exemplo, links que são muito acessados por outros membros da categoria podem ser destacados.

A fim de tornar possível a mineração uma tarefa mais fácil de ser implementada poderia ser utilizado a proposta feita por Dias (2008), que propõem a utilização de uma nova arquitetura para desenvolvimento de software, SOA (Arquiteturas Orientadas a Serviços), que tem como principal objetivo o reuso intenso dos seus componentes (serviços) para que, em médio prazo, a tarefa do desenvolvimento de uma aplicação seja primordialmente a tarefa da composição e coordenação dos serviços já implementados, aumentando o reuso e diminuindo o dispêndio de recursos.

A proposta é que estas técnicas sejam aplicadas na construção de sites adaptativos proporcionando maior qualidade destes sistemas e fazer uso das técnicas de mineração de dados na web juntamente com os benefícios que a Arquitetura Orientada a Serviços se propõe.

6 METODOLOGIA PARA O DESENVOLVIMENTO DE SISTEMAS WEB ADAPTATIVOS UTILIZANDO MINERAÇÃO DO USO DA WEB

Ao desenvolver um web site, o desenvolvedor meticulosamente cria a aparência do site, a estrutura das informação e determina os tipos de interações que devem estar disponíveis. Diante disso, Acreditamos que deveria ser feita uma distinção severa entre as mudanças estratégicas (longo prazo) e as técnicas de adaptação (curto prazo) ao implementar um web site adaptativo, uma vez que devemos evitar danos à estrutura do site. Este cenário e pesquisas de outros trabalhos nesta área nos levou a formular as seguintes considerações para o desenvolvimento de um web site adaptativo:

- Evitar que os usuários preencham questionários sobre interesses: Usuários da internet não interessam por trabalho extra (por exemplo, preenchimento de questionários), especialmente se não tem uma recompensa clara, e podem optar por sair e não participar.

- Fazer com que o site seja mais fácil de se usar para todos usuários, incluindo usuários que estão o acessando pela primeira vez, usuários casuais, etc.

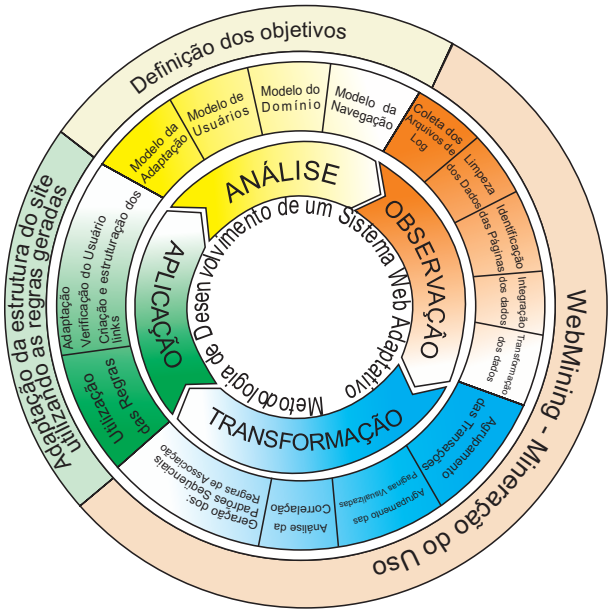

Figura 6: Metodologia para o desenvolvimento de sistemas web adaptativos.

- Proteger o design original do site de mudanças destrutivas: Limitar as transformações de modo que as mesmas não modifiquem a estrutura do site. Podemos acrescentar links, mas não removê-los, criar páginas, mas não destruí-las, adicionar novas estruturas, mas não embaralhar as já existentes.

- Manter o desenvolvedor no controle. O desenvolvedor precisa de se manter no controle do site, tanto para obter confiança nas técnicas de adaptação automática e evitar possíveis mudanças indesejáveis.

Diante do exposto e da identificação das necessidades dos sistemas web adaptáveis podemos definir uma metodologia para que estes sistemas possam ser desenvolvidos atendendo a todas as necessidades para se obter o resultado esperado no processo de adaptação de um site.

Considerando que o processo de desenvolvimento de interface não é uma atividade top-down (Batista (2008)), propõe-se um modelo que descreve uma seqüência iterativa de etapas, de forma a facilitar o desenvolvimento de um sistema web adaptativo.

No modelo apresentado na Figura 6, pode se observar todas as etapas que irão compor o desenvolvimento do sistema, que permite ao projetista obter uma visão global do processo de desenvolvimento e realizar um trabalho estruturado, sistemático e organizado. A análise inicia a seqüência de etapas; ao percorrer o sentido horário, faz-se a transição para as seguintes etapas: observação, transformação e aplicação.

6.1 Análise

Enquanto no desenvolvimento de sistemas não adaptativos, dá início ao processo a partir da etapa levantamento de dados, iniciamos sua atividade definindo e analisando os objetos referente a adaptação do site. Ou seja, a etapa do levantamento de dados já foi realizada para construir os modelos do usuário, do domínio, da navegação e da adaptação. Cabe então nesta etapa definir os objetivos a serem alcançados com a aplicação dessas estratégias. Somente com esses objetivos definidos é que as estratégias poderão ser projetadas. O administrador de um

site pode ter diversos motivos para decidir aplicar estratégias de personalização em um web site. Pode-se classificar esses motivos da seguinte forma:

1. *Funcionais*: O administrador do site quer facilitar a visita do usuário. A estrutura do site deve ser rearranjada de forma que o usuário encontre o que estiver procurando ou que acesse seções que o interessam e com facilidade.

2. *Comerciais*: Neste caso, se o objetivo é fomentar a venda de um determinado produto. O administrador espera que depois da personalização aumente o número de vezes que uma ação ocorra, por exemplo, pagar o produto *"A"*.

6.2 Observação

Considerando um web site, a necessidade de torná-lo adaptativo e os objetivos a serem alcançados com a personalização, o próximo passo do processo deve ser a coleta de dados. Nesta fase, o objetivo é coletar os dados relacionados aos usuários do site e, opcionalmente, dados relacionados ao negócio do site. Deve-se ressaltar que não é possível realizar uma análise da eficiência da estrutura do site, com o fim de alterar esta estrutura, se não existem dados sobre interação dos usuários com o mesmo. Neste ponto, pode-se considerar as seguintes perspectivas (Perkowitz e Etzioni (2009)):

1. *Limpeza dos Dados*: Limpeza de dados é geralmente especifica de cada site, e envolve tarefas tais como, a remoção de referências a objetos estranhos embutido que pode não ser importante para fins de análise, incluindo referências a arquivos de estilo, gráficos ou arquivos de som.

2. *Identificação dos usuários*: Define se os dados coletados serão utilizados para identificar o usuário.

   - Usuário é identificado: nesse caso, é possível guardar as informações de cada usuário, e a cada vez que ele retornar ao site, estas informações podem ser aproveitadas. É possível analisar o comportamento dos usuários ao longo de um período de tempo e perceber quais partes do site estão sendo descobertas e em qual ordem.
   - Usuário não é identificado: não é necessário para tornar um site adaptativo que os seus usuários sejam identificados. Quando a forma de aquisição dos dados é através do log do servidor web, pode ser que apenas a identificação das sessões seja suficiente. Essa situação pode ser comparada às análises realizadas com os dados de compra de um supermercado, onde a identidade do comprador não é considerada quando se aplica as técnicas de mineração de dados.

3. *Fonte de dados*: Trata da natureza dos dados coletados. Especifica quais tipos de dados serão utilizados na próxima fase, a de transformação.

   - Dados de acesso dos usuários: para descobrir detalhes do comportamento dos usuários do site, é suficiente conhecer as páginas acessadas em cada sessão e a ordem em que elas foram acessadas. As preferências dos usuários podem ser inferidas através das páginas do site que ele acessou.

- Dados pessoais dos usuários: se informações pessoais dos usuários do site estiverem disponíveis - como profissão, cidade/região/país onde mora, estado civil - outros tipos de análises podem ser realizadas. O site pode descobrir com estas análises que o perfil dos usuários do site é, por exemplo, homem, executivo e casado. Com essa informação, pode destacar os links para as páginas relacionadas ao Mercado de Executivo.
- Dados do mercado: utilizar dados de um data warehouse, como por exemplo os dados das vendas realizadas pela empresa que não foram feitas via Internet (Gomory et al. (2000)). Informações de quais produtos são vendidos juntos podem ser obtidas neste data warehouse e aproveitadas nas vendas via Internet.

4. *Forma de aquisição dos dados de acesso*: Especifica a forma como os dados são coletados, definindo se este processo é ou não transparente ao usuário.

- Não transparente: o usuário, voluntariamente, pode fornecer dados para o site. Por exemplo, o site pode fazer perguntas ao usuário para saber qual o seu objetivo naquela sessão e saber se ele foi ou não alcançado. Assim, pode criar um histórico de acessos, baseado nos sucessos e insucessos dos usuários.
- Transparente: o site pode coletar as informações de acesso do usuário sem que ele precise responder qualquer tipo de questionário. Num arquivo de log estão todas as requisições feitas ao servidor web com o endereço IP da máquina que fez as requisições e a data/horário do acesso. É possível identificar sessões e identificar qual a ordem de acesso das páginas de um site, durante uma sessão. Com estas informações, já é possível realizar alguns tipos de análises.

6.3 Transformação

Depois que já foram coletados dados suficientes, é preciso descobrir padrões de acesso (ou de comportamento), os perfis dos usuários, tendências, para que a estrutura do site possa ser melhorada. Os dados coletados na fase de observação precisam ser trabalhados, relacionados e bem entendidos para que informações válidas sejam produzidas.

É nesse contexto que entra em cena a transformação, cujo objetivo é obter informações a partir de dados. Nesta fase, várias decisões caracterizam a estratégia:

1. *Técnicas de análise*: Esta técnica identifica quais técnicas de análise serão utilizadas para explorar os dados.

- Padrões Informativos: quando não se tem um problema específico, uma forma de se obter informações de uma grande massa de dados é utilizando técnicas para descobrir padrões informativos. Os algoritmos de agrupamento são bons exemplos para este tipo de padrão. Eles identificam características comuns a um conjunto de entradas de forma a particionar a base de dados de acordo com os seus atributos. Uma possível aplicação nesse caso, seria descobrir os perfis de usuários que acessam um site, com base nas páginas visitadas e nas palavras pesquisadas, ou mesmo com as suas características pessoais.
- Padrões de previsão: dada uma entrada de dados e os seus atributos, padrões de previsão têm o objetivo de fazer uma estimativa razoável de um atributo desconhecido

com base em outros atributos conhecidos. Redes neurais e regressão linear vêm sendo amplamente utilizadas para determinação de padrões de previsão e também podem ser usados para gerar as regras de transformação do site.

2. *Nível de automatização*: Define se a fase de transformação estimula ou não, para que o administrador do site participe nas deduções sobre a personalização.

   - Não automatizado: uma abordagem interessante é que o sistema estimule a interação do administrador do site em momentos em que as heurísticas aplicadas não conseguem decidir qual caminho seguir. Por exemplo, um sistema pode identificar que determinados grupos de páginas são acessadas simultaneamente, e sugere então ao administrador do site que seja criada um página no site com links para essas páginas. Nesse caso, o administrador do site ficaria encarregado de descobrir a que assunto essas páginas estão relacionadas e dar um título para a nova página do site.
   - Automatizado: se a técnica de transformação dos dados não depender da interferência humana, o processo exige uma menor interação do administrador do site que, em alguns casos, pode ser a melhor opção

6.4 Aplicação

Durante a fase de transformação, obteve-se conhecimento sobre os padrões de acesso ao site e o que fazer para melhorá-lo. Na fase de aplicação, deve-se aplicar este conhecimento adquirido na estrutura do site. Neste ponto, pode-se considerar as seguintes perspectivas:

1. Escopo das aplicações: Define qual o alcance das transformações.

   - Global: nesse caso, o site aprende com os dados dos usuários e faz alterações permanentes na estrutura do site. Quando a fase de transformação, por exemplo, sugere a criação de uma nova página, a aplicação da técnica então altera definitivamente a estrutura do site.
   - Local: o site é dinâmico e se adapta à sessão/usuário corrente. Quando a técnica utilizada apenas inclui links não permanentes na página que está sendo acessada, esta alteração não é definitiva. A cada sessão, a estrutura do site pode sofrer transformações diferentes, que serão desfeitas assim que aquela sessão terminar.

2. Automatização: Define se as mudanças feitas no site necessitam da autorização do administrador do site.

   - As transformações são aplicadas automaticamente: nesta abordagem, o site pode interagir diretamente com o usuário, alterando as páginas pelas quais está navegando, sem que haja a permissão de um administrador do site.
   - As transformações são apenas sugeridas: o site pode apenas sugerir mudanças na estrutura para o administrador do site e ele as aplica se achar conveniente. Esta abordagem geralmente leva a transformações definitivas na estrutura do site.

3. Recurso de formatação: Identifica quais recursos de formatação das páginas são utilizados na aplicação das mudanças.

Figura 7: Arquitetura de um Web Site Adaptativo.

- Proprietário: já foram desenvolvidos alguns tipos especiais de linguagens de marcação, destinado a sites adaptativos. Uma página pode ser construída com tags especiais no HTML e as adaptações são feitas nas regiões delimitadas por essas tags. Por exemplo, essa tag pode ter argumentos que indicam que um link deve ser destacado, caso a página seja visualizada por um determinado usuário, num determinado contexto.
- Público: o HTML ou XML padrão são utilizados e os recursos de adaptação são feitos utilizando as tags já existentes.

4. Amplitude das transformações: Identifica quais regiões do site serão afetadas durante a fase de aplicação.

5.

6. As regiões "adaptáveis"são fixas, pré-determinadas: o site pode, por exemplo, ter um frame superior onde sempre aparecem links relacionados com a página corrente ou os links que as pessoas seguem depois que acessam a página corrente. Esta é uma forma simples de aplicar o conhecimento adquirido na fase de transformação.

7. As adaptações podem ocorrer em qualquer lugar do site: nesse caso, nenhuma região do site é destinada somente a adaptações. Essas podem ocorrer em qualquer parte da estrutura do site.

7 ARQUITETURA PROPOSTA DO WEB SITE ADAPTATIVOS

Diante do exposto e da identificação das etapas compostas no desenvolvimento de um web site adaptativo podemos definir uma arquitetura de alto nível, como mostra a Figura 7.

O alto nível de arquitetura de web site adaptativo é mostrado na Figura 7:

O primeiro passo de um web site adaptativo é agrupar seus visitantes através de um padrão de mineração de dados, o próximo passo é selecionar as transformações para o site de cada usuário ou grupo. As transformações selecionadas para cada usuário ou grupo são armazenadas no servidor web para serem utilizadas. Quando um usuário visita o site, no futuro, o servidor verifica qual grupo o mesmo pertence e realiza sua transformação correspondente. A página web solicitada é enviada a partir do servidor, transformada e em seguida enviada para o visitante.

Com esta implementação as características e vantagens descritas anteriormente poderão ser absorvidas pelos sistemas e fornecendo a capacidade de integração com diferentes fontes de dados (logs de servidor) permitindo assim a adaptação do web site.

8 CONCLUSÃO

Este trabalho teve como objetivo apresentar um estudo exploratório sobre as vantagens de se adotar técnicas de mineração do uso da web e regras de associação, juntamente com a utilização do conceito de Arquitetura Orientada a Serviços e uma possível arquitetura para estes sistemas

para sua implementação. A adaptação de web site apresenta-se como uma aplicação e área de pesquisa com excelentes perspectivas de futuro devido à variedade de informações e perfis de usuários disponíveis na internet. São inúmeras as vantagens obtidas com a aplicação de técnicas de mineração do uso da web, regras de associação e SOA conforme foram apresentadas. A aplicação destas vantagens em sistemas web adaptáveis poderá proporcionar inúmeros benefícios tanto para os web sites como para seus usuários, trazendo qualidade tanto na usabilidade, como na interação dos seus visitantes, visando a excelência nos seus resultados e atendendo a natureza dinâmica e heterogênea da Web. Por fim, acredita-se que a metodologia e a arquitetura proposta trazem contribuições para facilitar a atividade projetual dos profissionais nas áreas do Design e de Tecnologia da Informação e Comunicação a desenvolverem sistemas web adaptáveis.

REFERÊNCIAS

Agrawal R., Mannila H., Srikant R., Toivonen H., e Verkamo A.I. Fast discovery of association rules. advances in knowledge discovery and data mining. *AAAI/MIT Press*, 1996.

Agrawal R. e Srikant R. Fast algorithms for mining association rules. pagina 487 499. Proc. 20th Int. Conf. Very Large Data Bases, VLDB, Morgan Kaufmann, 1994.

ANSI/IEEE. Recommended practice for architectural description of software - intensive systems, 2000. ANSI/IEEE Std 147.

Batista C.R. *Modelo e Diretrizes para o processo de design de interface web adaptativa*. Tesis de Doutorado, Programa de Pós graduação em Engenharia e Gestão do Conhecimento, UFSC, Florianópolis, 2008.

Brusilovsky P. *Adaptive Navigation Support: From Adaptive Hypermedia to the Adaptive Web and Beyond*, volume 2, pagina 7 23. PsychNology Journal, 2004.

Brusilovsky P., Karagiannidis C., e Sampson D. Layered evaluation of adaptive learning systems. In *International Journal of Continuing Engineering Education and Lifelong Learning*, volume 14, pagina 402 421. 2004.

Cook Diane J. e Holder L.B. Graph-based data mining. volume 15, páginas 32–41. IEEE Intelligent Systems, Los Alamitos, 2000.

Dias T.M.R. *Uma arquitetura orientada a serviço para sistemas de mineração de dados na Web*. Dissertação mestrado, CEFET-MG, Belo Horizonte, 2008.

Erl T. *Service-oriented architecture: A field guide to integrating XML and Web Services*. 2004.

Fink J. Kobsa A. N.A. User-oriented adaptivity and adaptability in the avanti project. *Designing for the Web: Empirical Studies, Microsoft Usability Group*, 96.

Gomory S., Hoch R., e Leea. E-commerce intelligente measuring, analyzing and reporting on merchandising effectiveness of online stores. Relatório Técnico, IBM - Institute for Advanced Comerce, 2000.

Greening D.R. *Data Mining on the Web. Web Techniques*, volume 5. 2000.

J. Srivastava R. Cooley M.D. e Tan P. *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*, volume 1. SIGKDD Explorations, 2000.

Joachims T. Freitag D. M.T. Webwatcher: A tour guide for theworldwideweb. páginas 770–775. Proc. IJCAI, Nagoya, Japan, 1997.

Koch N.P. *Software Engineering for Adaptive Hypermedia Systems: Reference Model, Modeling Techniques and Development Process.* Tesis de Doutorado, Ludwig Maximilians University Munich., Munich, 2000.

Liu B. *Web Data Mining Exploring Hyperlinks, Contents, and Usage Data*. Springer, 2007.

Mobasher B. *Encyclopedia of Data Warehousing and Mining*. Idea Croup, 2006.

Perkowitz M. e Etzioni O. Adaptive web sites. 2009.

R. Agrawal T.I. e Swami A.N. Mining association rules between sets of items in large databases. pagina 207 216. Proceedings of the 1993 ACM SIGMOD international conference on Management of data (SIGMODŠ93), SIGMOD '93, 1993.

R. Cooley B.M. e Srivastava J. Web mining: Information and pattern discovery on the world wide web. pagina 558Ű567. Proc. of the 9th IEEE Intl. Conf. on Tools With Artificial Intelligence (ICTAIŠ97), 1997.

Ruas F., Ribeiro F., e Freitas L. Técnica para a construção de web sites adaptativos. *Technical report, Departamento de Ciência da Computação, Universidade Federal de Minas Gerais*, 2001.

Sampaio C. *SOA e Web Services em Java, Brasport.* 2006.

U. M. Fayyad G.P.S. e Smyth P. Data mining to knowledge discovery: An overview. in advances in knowledge discovery and data mining. pagina 1Ű34. AAAI/MIT Press, 1996.

Yan T. Jacobsen H. G.M.H.D.U. From user access patterns to dynamic hypertext linking. Proc. 5th Internat. WWW Conference, Paris, France, 1999.

Zaiane O.R. Web mining: Concepts, practices and research. *In Simpósio Brasileiro de Banco de Dados, Tutorial, XV SBBD*, 2000.