# Reconhecimento de padrões de comportamento individual baseado no histórico de navegação em um *Web Site*

**Luiz Fernando Rust da Costa Carmo, Danielle Costa**

Núcleo de Computação Eletrônica – Universidade Federal do Rio de Janeiro (UFRJ)
Caixa Postal 23.24 – 20.010-974 – Rio de Janeiro – RJ – Brasil

rust@nce.ufrj.br, daniellecosta@posgrad.nce.ufrj.br

***Abstract****. This paper investigates the use of a trust evaluation process for access control and user authentication in Web applications. Trust is evaluated by means of a mechanism based on user behavioral analysis, and depends on different factors as: an appropriate Web environment for collecting/storage user behavior information and the attribution of a trust measure in function of the specific user behavior. To quantify, and consequently to establish a trust measure, several pattern recognition techniques of behavior has been investigated and evaluated through an empiric experimental process.*

***Resumo.*** *Este artigo investiga o uso do conceito de confiança para controle de acesso e autenticação de usuários em aplicações Web. A evolução da confiança é estabelecida através da avaliação do comportamento do usuário, e depende de fatores como: um ambiente Web apropriado à coleta e armazenamento de informações do comportamento do usuário, e a atribuição de uma medida de confiança a este comportamento. Para quantificar, e conseqüentemente, estabelecer uma medida de confiança, várias técnicas de reconhecimento de padrões de comportamento foram investigadas e avaliadas empiricamente através de um processo experimental.*

## 1. Introdução

Grande parte do desenvolvimento de aplicações para a Internet hoje gira em torno da *Web,* no entanto, à medida que a *Web* cresce, também crescem os problemas e conseqüente preocupação quanto à segurança dos aplicativos.

Em Carmo et al. [2007] é apresentado uma proposta de mecanismo de segurança para aplicações *Web* baseado em confiança. A idéia básica é desenvolver um sistema contínuo de avaliação comportamental do usuário, onde confiança e restrições de acesso podem ser deduzidas automaticamente. O conceito de confiança apresentado é definido informalmente como uma medida de quão certo o provedor de uma aplicação está a respeito da identidade de um usuário. A medida de confiança por sua vez provém de uma avaliação comportamental baseada na análise de trilhas de navegação, superposta a uma assinatura contextual histórica.

O presente artigo fundamenta-se nesta proposta e tem como objetivo o estabelecimento de novos conhecimentos úteis para a avaliação comportamental e conseqüente estabelecimento da confiança.

Sendo assim, este trabalho propõe a construção de um *web site* experimental como gerador de subsídios para a avaliação comportamental e a investigação de novas técnicas passíveis de serem empregadas ao cálculo da confiança.

Assume-se que através da observação sistemática ou da alteração de algum elemento do experimento pode-se obter as evidências de um comportamento. Este comportamento refere-se à maneira de navegar do indivíduo (trilha de navegação). Estas trilhas são as instâncias comportamentais que uma vez identificadas farão parte de um histórico comportamental do mesmo indivíduo.

Através destes históricos pode-se trabalhar com a confiança de forma analítica e a partir dos dados de interação obter um indicador de confiança para um determinado usuário. Técnicas utilizadas para o reconhecimento de padrões de comportamento são avaliadas como propostas para quantificar a confiança.

Este artigo está organizado da seguinte forma: Na seção 2 são apresentados os trabalhos sugestivos para o desenvolvimento deste artigo. Na seção 3 são apresentados os conceitos necessários a compreensão das abordagens propostas para quantificar um comportamento e estabelecer a confiança. Na seção 4 é apresentada uma proposta para avaliar um comportamento na *Web*. Na seção 5 são expostos os resultados da avaliação experimental. Finalmente, na seção 6 são apresentadas as limitações e dificuldades encontradas e trabalhos futuros a serem realizados.

## 2. Trabalhos Relacionados

Sistemas de segurança, que envolvem a análise do comportamento contextual, são aqueles que procuram adquirir experiência a partir do perfil de utilização de serviços pelo usuário. São exemplos de comportamento contextual, comandos Unix e informações de navegação na *Web.* Assim, no domínio de detecção de intrusão, Lane e Brodley [1999], apresentaram um trabalho que aborda a análise comportamental contextual como conhecimento para caracterizar o comportamento de um indivíduo, sistema ou rede em termos de seqüências de dados temporais discretos. O foco está no modelo *Intance Based-Learning* (IBL). Os autores desenvolveram um protótipo de sistema de detecção de intrusão por anomalia que emprega um *framework* IBL para classificar comportamentos de usuários como normais e anormais. Os dados de entrada do sistema são linhas de comando Unix que são passadas através de um analisador, o qual reduz o dado a um formato interno, e faz a seleção das características. O resultado é comparado com o histórico de características do usuário via medida de similaridade.

Em outro trabalho apresentado por Platzer [2004], é desenvolvido a ferramenta *SimOffice* que utiliza conceitos de confiança para controle de acesso em *web services.* A idéia principal foi a criação um sistema auto-gerenciável onde as permissões são configuradas automaticamente. Para esta finalidade, foram implementados algoritmos que simulam o julgamento humano sobre um outro indivíduo. Com a utilização correta do sistema o usuário ganha créditos podendo ter acesso a novos serviços. Do contrário, com a utilização incorreta do sistema, a confiança depositada por ele no usuário diminui e o mesmo pode perder acesso a serviços antes disponíveis.

Véras e Ruggiero [2005] propõem um processo de autenticação contínua baseada em uma métrica de confiança para aplicações seguras na *Web*. A proposta de autenticação contínua permite em função da monitoração do comportamento do usuário e da

integração com uma aplicação, verificar o valor do seu indicador de confiança, e dessa forma, manter ou revogar a sua autenticação quando certos limites de confiança ou desconfiança são ultrapassados. Neste caso, a confiança é uma medida probabilística quantificada com valores entre [*0,1*], onde *zero* apresenta desconfiança completa e *um* a confiança total.

Uma das diferenças principais destas abordagens em relação a este trabalho é a forma de avaliar o comportamento do usuário. A presente proposta avalia o comportamento do usuário em função do seu passado (base de dados de utilização) com a finalidade de aumentar a confiança na identificação do usuário. Também não há a preocupação em se identificar um comportamento como normal ou anormal.

No domínio da mineração de dados, Onoda [2006] propõe o estudo do comportamento de navegação dos usuários, integrando *Web Usage Mining* com conceitos da análise de comportamento fundamentada no behaviorismo de Skinner. A partir desta análise, várias hipóteses sobre o comportamento dos usuários em um *web site* são levantadas. A conscientização dos donos da empresas sobre essas hipóteses permite o desenvolvimento de ações mais convenientes ao *site.*

Em relação à adaptação de *web sites* El-Ramly e Stroulia [2004] propõem o uso de técnicas mineração de padrão de comportamento. Após a navegação, o comportamento é um indicador dos interesses dos usuários, e os registros do *log* do servidor *Web* podem ser minerados para inferir sobre o que os usuários estão interessados. Assim, um *web site* pode ser reorganizado para tornar o conteúdo mais interessante, ou as recomendações podem ser dinamicamente geradas para ajudar novos usuários a encontrar a informação de seu interesse mais rapidamente.

Deshpande e Karypis [2004], Anderson et al. [2002] analisam o comportamento da navegação de usuários na *Web* para fazer predições a respeito da navegação futura e personalizar as páginas para cada usuário. Modelos de Markov são usados para analisar os comportamentos.

Diferentemente dos trabalhos anteriores, o mecanismo proposto neste trabalho permite que a coleta das informações de comportamento seja direcionada e sem a participação direta do usuário no processo, ou seja, ele não fornece diretamente os dados que serão usados para compor o seu histórico de navegação. Neste caso somente os dados considerados relevantes para a avaliação são capturados. Esse tipo de abordagem dispensa a mineração dos *logs* ou pré-processamento das informações.

## 3. Reconhecimento de Padrões de Comportamento

Normalmente, aplicações para o reconhecimento de assinaturas ou faces utilizam abordagens, tais como: Cadeias de Markov, Distância de Edição e Norma de Frobenius, [Schimke et al., 2004; Cheng et al., 1992]. Trabalhos utilizando Distância de Levenshtein também são vistos no domínio da detecção de intrusão [Unterleitner, 2006].

Esta seção descreve os principais conceitos sobre essas abordagens o qual este artigo se propõe a investigar para o problema de reconhecer um padrão de usabilidade de um *web site* (maneira como o usuário navega no *web site*).

### 3.1. Cadeias de Markov

Um Processo de Markov é um processo estocástico onde as distribuições de probabilidade para o seu desenvolvimento futuro, dependem somente do estado presente, não levando em consideração como o processo chegou a tal estado. Os processos markovianos são modelados formalmente pelos modelos de Markov, que são sistemas de transições de estados, onde os estados são representados em termos de seus vetores probabilísticos, que podem variar no espaço temporal (discreto ou contínuo), e as transições entre estados são probabilidades que dependem apenas do estado corrente. Se o espaço de estados é discreto (enumerável), então o modelo de Markov é denominado de Cadeias de Markov.

Suponha um sistema em $n$ estados possíveis. Para cada $i = 1,2,...,n$ $j = 1,2,...,n$, seja $t_{ij}$ a probabilidade de que se o sistema está no estado $j$ em um determinado período de observação então ele estará no estado $i$ no próximo período de observação; $t_{ij}$ é chamado de uma probabilidade de transição. Além disso, $t_{ij}$ se aplica a todos os períodos de tempo, isto é, não muda com o tempo. Como $t_{ij}$ é uma probabilidade, temos que ter $0 \leq t_{ij} \leq 1$ $(1 \leq i, j \leq n)$ [Kolman, 1998; Rabiner, 1989; Winston ,1994].

As propriedades desses modelos são estudadas em termos das propriedades das matrizes de transições de estados que são utilizadas na sua descrição e representadas através de um grafo [Haykin, 2001].

### 3.2. Distância de Levenshtein

A Distância de Levenshtein também conhecida como Distância de Edição é a medida de semelhança ou diferença entre duas cadeias de caracteres. Corresponde à transformação ou edição da primeira cadeia de caracteres na segunda através de uma série de operações de edição, de forma individual, sobre cada um dos caracteres da cadeia.

A distância entre duas cadeias $x$ e $y$ é dada pelo número mínimo de operações necessárias para transformar $x$ em $y$. Entende-se por operações a inserção, remoção e a substituição de um caractere [Navarro 2001]. Por exemplo, a distância de Levenshtein entre as palavras banana e laranja é *3*, já que o custo de cada operação é *1* e não há maneira de transformar a palavra na outra com menos de três edições. 1) banana; 2) lanana (substituição de '*b*' por '*l*'); 3) larana (substituição de '*n*' por '*r*'); 4) laranja (inserção de '*j*'). Um algoritmo de programação dinâmica é usado freqüentemente para calcular a Distância Levenshtein.

### 3.3. Distância de Frobenius

A medida de distância entre matrizes pode ser determinada através de uma norma. Uma norma de matriz de uso comum na álgebra linear é a norma de Frobenius que é dada pela expressão:

$$||A||_f = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}$$

onde $a_{ij}$ representa o elemento da $i$ - ésima linha e $j$ - ésima coluna da matriz $A$. Equivale ao tradicional cálculo da norma de vetores, se $A$ é interpretado como um vetor [Golub and Loan, 1996].

## 4. Proposta de Avaliação Comportamental

Nesta seção é abordado o problema de avaliar o comportamento de um usuário em uma aplicação *Web* de forma a caracterizá-lo e diferenciá-lo em termos de uma seqüência de dados discretos. A avaliação depende de fatores como, a determinação de um ambiente *Web* apropriado, à coleta e armazenamento de informações do comportamento do usuário, e a atribuição de uma medida de confiança a este comportamento.

### 4.1. Coleta da Assinatura Comportamental em aplicações *Web*

O problema da coleta de assinatura segundo Carmo et al. [2007] pode ser formulado como uma tarefa de aprendizagem para caracterizar o comportamento típico de um usuário (assinatura) em termos de seqüência de dados discretos. Para tanto, é necessário definir o modelo da aprendizagem, bem como o formato representacional dos dados de entrada.

Os autores consideram um modelo simplificado de *Instance Based-Learning* (IBL). Neste modelo, um conceito é representado implicitamente por um conjunto de instâncias que exemplificam este conceito (dicionário de instâncias). Cada instância comportamental é diretamente classificada de acordo com o usuário gerador. Desta forma, a assinatura comportamental é representada pelo conjunto de instâncias comportamentais de um usuário específico, sendo gerada para cada macro-estado (contingente de páginas onde os comportamentos ocorrem).

Esta abordagem requer o armazenamento de um conjunto completo de instâncias comportamentais por assinatura que, em uma aplicação *Web,* são constituídas pelas seqüências de páginas visitadas pelo usuário durante a sua interação com o ambiente (trilhas de navegação). A assinatura por sua vez pode ser visualizada como parte de um histórico de comportamento individual (figura 1).

| | | | | |
|---|---|---|---|---|
| 24 | 200.150.38.198 | 2006-12-06 | 10:01:00 | Principal#PopNacional#RockNacional#Detalhes#ConcluirVenda# |
| 24 | 200.150.38.198 | 2006-12-04 | 10:07:00 | Principal#RockNacional#Detalhes#ConcluirVenda# |
| 24 | 200.150.38.198 | 2006-12-07 | 13:18:00 | Principal#RockNacional#Detalhes#ConcluirVenda# |
| 24 | 201.58.109.144 | 2006-12-08 | 15:27:00 | Principal#RockInternacional#Detalhes#ConcluirVenda# |
| 24 | 200.165.79.37 | 2006-12-05 | 18:45:00 | Principal#RockNacional#Detalhes#ConcluirVenda# |
| 24 | 200.150.34.232 | 2007-02-05 | 19:26:00 | Principal#RockNacional#Detalhes#ConcluirVenda# |
| 24 | 200.150.34.232 | 2007-02-05 | 19:27:00 | Principal#RockNacional#Detalhes#ConcluirVenda# |
| 24 | 200.150.34.232 | 2007-02-05 | 19:26:00 | Principal#RockNacional#Detalhes#ConcluirVenda# |
| 24 | 200.150.34.232 | 2007-02-05 | 19:25:00 | Principal#Lancamentos#Detalhes#ConcluirVenda# |
| 24 | 200.150.34.232 | 2007-02-05 | 19:27:00 | Principal#RockInternacional#Detalhes#ConcluirVenda# |

**Figura 1: Exemplo de assinatura comportamental do usuário 24**

A idéia essencial deste artigo é de que as instâncias sejam coletadas por meio de um experimento com diferentes usuários (extraindo requisitos para a implementação de *web sites*), de forma que os comportamentos sejam analisados em caráter de diferenciação comportamental (avaliando quais são as melhores técnicas para cálculo da confiança associada).

### 4.2. Aspectos da construção de *Web Sites*

Segundo Nielsen [1999], na Internet não existe uma padronização, mas certas convenções podem ser utilizadas e respeitadas. Sendo assim, esta subseção aborda aspectos envolvidos na construção de *web sites* estáticos direcionados a avaliação comportamental.

Tratando-se da avaliação dos comportamentos, a dificuldade está em definir uma coleção de páginas *Web* onde os mesmos possam ocorrer. O problema está em selecionar um contingente que não seja totalmente indutivo, o que dificulta o processo de diferenciação comportamental, ou que seja aleatório demais, o que pode levar a uma não padronização do comportamento de um mesmo indivíduo.

Segundo Nielsen [2000] pode-se controlar onde o usuário vai navegar, contudo na *Web*, os usuários controlam fundamentalmente a navegação pelas páginas. É possível forçar os usuários por caminhos definidos e evitar que estabeleçam *links* com determinadas páginas, mas os *sites* que usam esta estratégia são considerados rígidos. Os *sites* que são projetados visando a liberdade de movimento são vistos como uma melhor opção. No entanto, a estratégia de inserir elementos de navegação global nos "subsites", com essa finalidade, pode levar a problemas de acúmulo navegacional ou de má condução.

Métodos de filtragem ou de coleta de dados direcionada podem ser empregados para reduzir o acúmulo navegacional. Assim, eliminam-se as páginas as quais não são interessantes à avaliação comportamental, como por exemplo, páginas de interesse comum aos usuários que não refletem seu comportamento diferenciado (páginas como "*Login"* ou "Cadastro").

Considerando que o interesse do usuário em visitar determinada página é um indicativo de seu comportamento existe também a necessidade de se definir onde começa e termina uma instância comportamental. Nielsen [2000] afirma que a estrutura de um *web site* deve ser determinada pelas tarefas que os usuários desejam realizar. Um *web site* estruturado desta forma favorece a delimitação de uma instância que pode iniciar com a página principal do *web site* e terminar quando o usuário alcança a página definida como uma página de conclusão da tarefa.

Geralmente os *logs* armazenados em servidores *Web* são as principais fontes de informações sobre os usuários. No entanto, os dados armazenados no *log* não são propícios para a identificação de um usuário com precisão. Na abordagem mais conhecida assume-se que cada endereço IP é um usuário. Contudo, um usuário pode apresentar mais de um endereço IP (figura 1), o que dificulta a reunião de um conjunto de instâncias comportamentais relacionadas ao mesmo indivíduo. A utilização de *Cookies*, que dependem do consentimento do usuário, e de mecanismos de *login* são abordagens conhecidas por serem mais eficazes.

Shahabi et al. [1997] discutem algumas linguagens utilizadas para programar agentes de software para capturar informações de navegação em *web sites*. Problemas relacionados à utilização de opções disponíveis nos navegadores são apontados como fatores limitantes a captura dos comportamentos de navegação dos usuários.

No entanto, os eventos de abertura das páginas podem ser monitorados. Esta estratégia possibilita a identificação das páginas quando o usuário utiliza os botões do navegador ou mantém várias páginas abertas, diferentemente dos trabalhos que utilizam a técnica

de *Clickstream* onde o comportamento dos usuários é monitorado *click a click* [Brainerd and Becker, 2001; Hu and Zhong, 2005].

Para fins de avaliação comportamental foi desenvolvido uma aplicação de um *web site* experimental de forma a possibilitar a captura dos comportamentos. Para conceder maior mobilidade aos usuários, aspectos ergonômicos foram observados para possibilitar que um comportamento fosse explicitado com liberdade. O ambiente foi estruturado inicialmente no formato de um grafo fortemente conexo, assim é possível a partir de qualquer página se chegar a qualquer outra. Já para as páginas do *site*, foi utilizado um *layout* considerado como padrão para o desenvolvimento de interfaces, o que atribui às páginas uma característica familiar às interfaces de ambientes reais.

Neste *web site*, onde diferentes usuários podem navegar, a coleta dos dados de navegação é feita através de agentes de software, de forma direcionada e sem a participação direta do usuário. Os agentes monitoram cada abertura de página capturando somente os dados relevantes para a avaliação.

### 4.3. Cálculo da Confiança Comportamental

A coleta das assinaturas dos usuários possibilita estabelecer uma medida de confiança a uma instância comportamental em função de sua assinatura. O cálculo da confiança (*Trust*) é expresso pelo produto dos seguintes fatores:

- Similaridade comparativa (*SComp*): similitude entre a instância atual e o conjunto de instâncias que compõe a assinatura. Espelha o quanto este comportamento se aproxima dos demais previamente capturados.
- Intra-similaridade (*SIntra*): independente da amostra atual de instância comportamental. Representa se um usuário possui um comportamento bem formado (quando as instâncias pertencentes à assinatura são repetidas ou levemente diferentes).
- Inter-similaridade (*SInter*): traduz a qualidade da assinatura de um usuário em função do conjunto completo de assinaturas (de diferentes usuários).

Diferentemente da heurística apresentada por Carmo et al. [2007] para quantificar os fatores de confiança, este artigo expõe a aplicação de abordagens utilizadas em reconhecimento de padrões para calcular a confiança comportamental de forma simplificada.

### 4.4. Cadeias de Markov aplicadas à métrica *SComp*

A condição inicial considera que as páginas do *web site* são os estados da cadeia de Markov e os *links* as transições que envolvem os estados. Assim, uma trilha de navegação composta de *3* páginas é interpretada como os estados *1*, *2* e *3* da cadeia. Seja $a_{ij}$ a probabilidade de transição do estado *i* para o estado *j*, então a matriz *3* x *3* é a seguinte:

$$A = \{ a_{ij} \} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

A probabilidade de transição de estados $a_{ij}$ refere-se ao número de vezes que houve transição do estado *i* para o estado *j*, dividido pelo número de ocorrências do estado *i* no histórico de trilhas. O valor de *SComp* para a instância atual é o produto das probabilidades de transição dos estados dados pela matriz de transição. Imagine que o usuário realize a seguinte trilha mais recente: *1 # 2 # 3*, calculando-se *SComp*:

$$a_{12} \times a_{23} = 0.3 \times 0.2 = 0.06$$

### 4.5. Distância de Levenshtein aplicado à métrica *SIntra*

O cálculo das distâncias entre as trilhas do histórico do usuário realiza basicamente os seguintes procedimentos:

- aplica-se o algoritmo para o cálculo da Distância de Levenshtein entre todas as trilhas que compõem o histórico de um usuário; o número de distâncias é uma combinação

$$C_{n,2} = n!/(n-2)!\,n!$$

- realiza-se a normalização pelo valor máximo das distâncias;
- calcula-se a média $\overline{\mathbf{x}}$ das distâncias: (soma das distâncias normalizadas)/(quantidade de trilhas do histórico)

Quanto menor a distância maior a similaridade entre as trilha. Logo,

$$SIntra = 1 - \overline{X}$$

### 4.6. Distância de Frobenius aplicado à métrica *SInter*

Cada histórico de navegação é representado por uma matriz de transição de probabilidade de Markov. *SInter* é a média das distâncias de Frobenius que calcula a diferença entre a matriz do usuário e as demais matrizes existentes. Quanto mais diferenciado for o histórico do usuário maior o valor da distância de Frobenius.

Os seguintes procedimentos são adotados:

- calcula-se a distância de Frobenius entre a matriz do usuário e as demais matrizes;
- realiza-se a normalização pelo valor máximo das distâncias;
- calcula-se a média das distâncias: (soma das distâncias normalizadas)/(quantidade de distâncias).

Considere o exemplo de duas matrizes X e Y tais que:

$$X = \begin{bmatrix} 0.5 & 0.5 \\ 1 & 0 \end{bmatrix} \quad Y = \begin{bmatrix} 0.4 & 0.6 \\ 0 & 1 \end{bmatrix}$$

O valor da distância de Frobenius para as matrizes é:

$$|X - Y|_f = \sqrt{\sum_{i=1}^{2} \sum_{=1}^{2} |x_{ij} - y_{ij}|^2} = |0.5-0.4| + |0.5-0.6| + |1-0| + |0-1| = 2.2$$

## 5. Avaliação Experimental

Esta seção inicialmente descreve o desenvolvimento do experimento e discute a sua forma de implementação, o processo de composição do histórico comportamental de cada usuário e a determinação do valor de confiança. Em seguida, os respectivos resultados a partir da base de dados coletada são analisados.

### 5.1. Metodologia

A primeira etapa compreendeu a definição do contexto do experimento. Definiu-se que os comportamentos ocorreriam no contexto de uma loja virtual para que os usuários pudessem ser diferenciados por sua preferência de compra. O experimento simula uma loja virtual de vendas de cds pela Internet.

A etapa seguinte foi a implementação do *web site* da loja desenvolvido em HTML (Hypertext Markup Language) e PHP (Hipertext Preprocessor). Nesta etapa também foi implementado em *JavaSript* o mecanismo de coleta dados de navegação através dos agentes de software. A cada sessão os agentes monitoram a abertura das páginas coletando e armazenando as trilhas de navegação em um *Cookie* temporário. Após autenticação via *login* e senha, as informações como IP, data e hora de acesso e a trilha são incorporados ao banco de dados de utilização do usuário.

Dois testes *on-line* com experimentos foram realizados em períodos diferentes por um grupo de pessoas que receberam as instruções de acesso à loja. Foram gerados pelos participantes 21 históricos e 179 instâncias comportamentais.

Na primeira fase do teste, foi solicitada aos participantes a tarefa de efetuar compras diárias de qualquer produto de sua preferência. O objetivo desta fase foi verificar a eficácia do mecanismo de coleta de dados e a criação do histórico individual de cada participante. Foram identificadas falhas e dificuldades envolvendo a utilização do *web site.*

Na segunda fase foi solicitada aos participantes a tarefa de efetuar compras sucessivas de qualquer produto de sua preferência. O objetivo deste segundo teste foi coletar uma maior quantidade de dados para a criação dos históricos e verificar a eficácia das alterações realizadas no *web site*.

Na etapa final as técnicas para reconhecimento de padrões foram aplicadas e um indicativo de confiança foi estabelecido para cada usuário.

### 5.2. Resultados e Análise

Um indicador de confiança, calculado por meio da proposta sugerida neste artigo, foi estabelecido para cada usuário. A tabela 1 apresenta os resultados do cálculo de confiança realizado considerando a base de dados gerada pelo experimento. Os valores considerados para o *Trust* são os resultados da aplicação da Distância de Levenshtein para *SComp* e *SIntra* e Frobenius para *SInter*.

**Tabela1: Resultados do *Trust* = *SComp x SIntra x SInter***

| Participantes | SComp-Markov | SComp-Levenshtein | SIntra-Levenshtein | SInter-Frobenius | Trust |
|---|---|---|---|---|---|
| 19 | 4,0000E-05 | 0,5500 | 0,4250 | 0,4485 | 0,104836875 |
| 22 | 1,0000E-08 | 0,5185 | 0,5000 | 0,4789 | 0,124154825 |
| 23 | 1,0000E-08 | 0,0000 | 0,3929 | 0,4060 | 0 |
| 24 | 1,1110E-01 | 0,5000 | 0,6806 | 0,4993 | 0,16991179 |
| 26 | 1,0000E-08 | 0,5560 | 0,4444 | 0,4744 | 0,117217788 |
| 28 | 2,5000E-17 | 0,1250 | 0,2778 | 0,4591 | 0,015942248 |
| 33 | 5,0000E-02 | 0,4889 | 0,4028 | 0,5378 | 0,105908373 |
| 34 | 1,0000E-12 | 0,5333 | 0,5944 | 0,4713 | 0,149399046 |
| 35 | 2,2200E-02 | 0,5238 | 0,5035 | 0,5986 | 0,157870753 |
| 36 | 1,0000E-08 | 0,6000 | 0,5852 | 0,4387 | 0,154036344 |
| 37 | 2,2220E-01 | 0,2222 | 0,0833 | 0,4149 | 0,007679492 |
| 40 | 5,5560E-01 | 0,5556 | 0,5972 | 0,4492 | 0,149046501 |
| 41 | 4,0000E-01 | 0,7037 | 0,5370 | 0,4035 | 0,152477364 |
| 42 | 1,1110E-09 | 0,6111 | 0,6157 | 0,4970 | 0,186998372 |
| 43 | 2,5000E-01 | 0,5000 | 0,3333 | 0,4294 | 0,07155951 |
| 45 | 1,0000E-05 | 0,2000 | 0,4667 | 0,4355 | 0,04064957 |
| 46 | 3,4711E-38 | 0,2137 | 0,4497 | 0,8665 | 0,083271421 |
| 47 | 1,0000E-08 | 0,1250 | 0,0000 | 0,4592 | 0 |
| 48 | 1,0000E-37 | 0,0123 | 0,4074 | 0,4419 | 0,00221437 |
| 50 | 1,0000E-08 | 0,1667 | 0,3333 | 0,4483 | 0,024908045 |
| 51 | 1,0000E-08 | 0,5714 | 0,4921 | 0,4355 | 0,122456477 |

De acordo com os testes realizados, algumas observações podem ser feitas:

- verificou-se que o endereço IP de fato não identifica o usuário com precisão, uma vez que, nas tarefas realizadas diariamente pelos participantes, vários IP's foram registrados para um mesmo usuário.
- devido a falta de permissão para a utilização do *Cookie* algumas trilhas não foram coletadas;
- problemas de condução foram verificados no ambiente estruturado inicialmente na forma de um grafo fortemente conexo;
- a alteração da estrutura do *site* por uma estrutura mais condutora, não comprometeu a liberdade de navegação dos participantes, uma vez que o contingente de páginas alteradas foram aquelas de acesso comum a todos os participantes e por isso não interessam à avaliação;
- o monitoramento dos eventos das páginas permitiu coletar os comportamentos mesmo em situações onde o navegador era utilizado ou quando várias páginas eram mantidas em aberto.

Comparando-se os resultados obtidos pela aplicação de Markov e Levenshtein à *SComp* constatou-se que a medida de similaridade dada por Markov considera a ordem de aparecimento das páginas na trilha e a similaridade máxima entre elas. Ao contrário, a Distância de Levenshtein considera a parcialidade de similaridade entre trilhas. Por este motivo optou-se pela abordagem de Levenshtein também para *SComp.*

Os valores apresentados pela aplicação de Markov são resultados da estratégia de se associar um valor mínimo (*0.0001*) de probabilidade a uma transição inexistente no histórico. Desta forma, procurou-se evitar o problema de se obter um resultado totalmente nulo, visto que este é dado pelo produto das probabilidades de transição.

Através dos resultados de *SInter,* observou-se uma outra interessante característica de diferenciação comportamental: além da distinção pelo interesse por determinados produtos específicos, existe também a diferenciação pelo tamanho da trilhas de navegação. Participantes que realizaram trilhas de navegação muito extensas foram os que obtiveram um valor mais alto de *SInter*.

## 6. Conclusões e Trabalhos Futuros

Este artigo descreveu uma proposta de construção de um *web site* experimental como gerador de subsídios para a avaliação comportamental e a investigação de técnicas de reconhecimentos de padrões para quantificar os fatores de confiança.

Destaca-se a proposta de utilização de agentes de software para formação dos históricos comportamentais e a resolução de diversos problemas conhecidos como: a identificação das trilhas, utilização do navegador e abertura de várias páginas. A linguagem de programação usada no desenvolvimento dos agentes permite ainda que o mecanismo seja reutilizado em outros ambientes. No entanto a utilização de *Cookies* é dependente do consentimento do usuário e é o fator limitante ao desempenho do mecanismo.

Outra importante contribuição deste trabalho é o procedimento proposto para quantificar os fatores de confiança. As abordagens Cadeias de Markov, Distância de Leveshtein e Distância de Frobenius foram investigadas e aplicadas ao cálculo de confiança.

Trabalhos em andamento incluem: técnicas para estabelecimento de um limiar de confiança mínimo para cada usuário e a definição de um conjunto de diretrizes que vão guiar a construção de *web sites* que desejam utilizar a avaliação comportamental como meio de aumentar a confiança na identificação do usuário.

## 7. Referências


Carmo, L.F.R.C., Oliveira, B.G and Braga, A.C.B. (2007).“Trust Evaluation for Web Applications based on behavioral Analyses”. In: 22th International Information Security Conference.. New approaches for security privacy and trust in complex enviroments (IFIP 07), Sandton, South Africa

Lane, T. and Brodley, C.(1999). “Temporal Sequence Learning and Data Reduction for Anomaly Detection”, ACM Transactions on Information and System Security, New York, v.2, p. 295–331

Platzer, C. (2004). “Trust- Based security in web services**”**, Masther’s Thesis, Information Institute, Tecnical University of Vienna, Austria

Véras, L.M.A e Ruggiero, W.V. (2005). “Autenticação Contínua de Usuários em Aplicações Seguras na Web”. In: V Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais (SBC 2005), Florianópolis. p.40-53.

Onoda, M. (2006). “Metodologia de mineração de dados para análise do comportamento de navegar num Web Site.”, Dissertação (Doutorado em ciências em engenharia civil), Universidade Federal do Rio de Janeiro, Rio de Janeiro.

El-ramly, M.and Stroulia, S. (2006). “Analysis of Web-usage behavior for focused Web sites: a case study”. Disponível em: http://www3.interscience.wiley.com/cgi-bin/abstract/107582383/ABSTRACT. Acesso em 4 nov. 2006

Deshpande, M. and Karypis, G. (2004). “Selective Markov models for predicting Web page accesses”. ACM Transactions on Internet Technology (TOIT), p.163-184.

Anderson, C. R., Domingos, P. and Weld, D. S. (2002). “Relational Markov models and their application to adaptive web navigation”.In: Proceedings of the Eighth ACM

SIGKDD international Conference on Knowledge Discovery and Data Mining Edmonton, Alberta, Canada (KDD 02). p.143-152.

Schimke, S., Vielhauer, C. and Dittmann, J. (2004).”Using adapted Levenshtein distance for on-line signature authentication”.In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR04), v.2, p.931-934.

Cheng, Q, Liu, K, Yang, J and Wang, H. (1992).“A robust algebraic method for human face recognition”, In: Proceedings 11th IAPR International Conference on Pattern Recognition, Conference B: Pattern Recognition Methodology and Systems, p.221-224.

Unterleitner, M.C. (2006).“Implementation of a Computer Immune System for Intrusion- and Virus Detection” Disponível em: http://www.iaik.tugraz.at/teaching/11_diplomarbeiten/archive/unterleitner/thesis14.pdf.

Acesso em 13 de nov. 2007.

Kolman, B. (1998). “Introdução a álgebra Linear com Aplicações”, 6 ed. Rio de Janeiro: Prentice Hall do Brasil, p.357.

Rabiner, L. R. (1989). **“**A tutorial on hidden Markov models and selected applications in speech recognition”, In: Proceedings of the IEEE vol. 77, p.257 – 286

Winston, W. L. (1994). "Operations Research - Applications and Algorithms", Duxbury Presss.

Haykin, S. (2001). “Redes Neurais: Princípios e Prática**”**. 2 ed. Porto Alegre: Bookman.

Navarro, G. 2001. “A guided tour to approximate string matching**”**.(2001) ACM Comput. Survey (CSUR), v.33, p. 31-88.

Golub, G. H.and Loan, C. F. V. (1996). “Matrix Computation”, 3 ed., Baltimore, Johns Hopkins University Press.

Nielsen J. (1999). “Do Interface Standards Stifle Design Creativity?”. Disponível em:

http://www.useit.com/alertbox/990822.html. Acesso em 17 de mar. 2006.

Nielsen, Jakob. (2000). “Projetando Websites”. Rio de Janeiro: Campus, p.416

Shahabi, C., Zarkesh, A. M., Abidi, J. and Shah, V. **(**1997). “Knowledge discovery from users Web-page naviagation**”**.In: Proceedings of 7th International Workshop on Research Issues in Data Engineering (RIDE 97), p. 20-29.

Brainerd, J.and Becker, B. (2001). **“**Case study: e-commerce clickstream visualization ” IEEE Symposium on Information Visualization (INFOVIS 01), p.153 – 156.

Hu, J. and Zhong, N. (2005). "Clickstream Log Acquisition with Web Farming," IEEE/WIC/ACM International Conference on Web Intelligence, p. 257-263.