LISTA DE FIGURAS

Figura 1: Exemplo de instância comportamental 39
Figura 2: Macro-estados x avaliação da confiança 39
Figura 3: Exemplo de SComp 40
Figura 4: Exemplo de SIntra 41
Figura 5: Exemplo de SInter 41
Figura 6: Procedimento de homogeneização 43
Figura 7: Pseudocódigo da função $\tau(x,y)$ 44
Figura 8: Pseudocódigo para a função DistânciaLevenshtein 46
Figura 9: Exemplo de cálculo de similaridade de seqüências 47
Figura 10: Exemplo de matriz de similaridade de seqüências 47
Figura 11: Pseudocódigo para a função DistânciaLevenshteinAdaptada 48
Figura 12: Gráficos Levenshtein x Levenshtein adaptado 49
Figura 13: Etapas de requisição de recursos Web 63
Figura 14: Primeiro diagrama de seqüência do usuário 64
Figura 15: Diagrama de seqüência do Agente 1 67
Figura 16: Diagrama de seqüência do Agente 2 67
Figura 17: Segundo diagrama de seqüência do usuário 78
Figura 18: Histograma de valores FN(%), FP(%), VN(%) e VP(%) – Abordagem 1. 78
Figura 19: Histograma de valores FN(%), FP(%), VN(%) e VP(%) – Abordagem 2 79
Figura 20: Histograma de valores FN(%), FP(%), VN(%) e VP(%) – Abordagem 3 79

LISTA DE TABELAS

Tabela 1: Dilema dos prisioneiros 29
Tabela 2: Exemplo de histórico comportamental do usuário 24 53
Tabela 3: Abordagem1 x Abordagem 2 x Abordagem 3 74
Tabela 4: Ocorrências de FN+FP e VN+VP 79

LISTA DE ABREVIATURAS E SIGLAS

AC	*Autoridade Certificadora*
AD	*Autoridade de Datação*
AR	*Autoridade de Registro*
DP	*Diretório Público*
FP	*Falsos-positivos*
FN	*Falsos-negativos*
HTTP	*Hypertext Transfer Protocol*
HTML	*Hypertext Markup Language*
IBL	*Instance Based-Learning*
ICP	*Infra-estrutura de Chave Pública*
IP	*Internet Protocol*
ITU-T	*International Telecommunication Union – Telecommunication Standardization*
KNN	*K-Nearest Neighbor*
NAT	*Network Address Translation*
PHP	*Hypertext Preprocessor*
PIN	*Personal Identification Number*
SComp	*Similaridade-comparativa*
SIntra	*Intra-similaridade*
SInter	*Inter-similaridade*
VP	*Verdadeiros-positivos*
VN	*Verdadeiros-negativos*
W3C	*World Wide Web Consortium*

SUMÁRIO

1.INTRODUÇÃO 12

1.1 Motivação 12

1.2 Caracterização do Problema 14

1.3 Objetivos da Dissertação 16

1.4 Procedimentos Adotados 17

1.5 Estrutura da Dissertação 18

2.CONCEITOS BÁSICOS 19

2.1 Autenticação 19

2.1.1 Autenticação via Credenciais Digitais 20

2.1.2 Autenticação via *Token* 22

2.1.3 Autenticação via Biometria 23

2.2 Confiança 24

2.2.1 Teoria da Confiança 24

2.2.2 Principais Formas de Confiança 26

2.3 Reconhecimento de Padrões de Comportamento 31

2.3.1 Tipos de Comportamento 32

2.3.2 Técnicas para o Reconhecimento 36

3.PROPOSTA PARA A AVALIAÇÃO COMPORTAMENTAL 38

3.1 Mecanismo de Segurança Baseado em Confiança 38

3.2 Cálculo da Confiança Comportamental 40

3.3 Abordagens para Quantificar a Confiança 42

3.3.1 Abordagem 1: Heurística de Carmo *et al.* 42

3.3.2 Abordagem 2: Distância de Levenshtein 45

3.3.3 Abordagem 3: Distância de Levenshtein Adaptada 46

3.4 Avaliação Experimental 49

3.4.1 *Web Site* Experimental 50

3.4.2 Aquisição da Assinatura Comportamental 52

3.4.3 Aplicação da Abordagem 1 53

3.4.4 Aplicação da Abordagem 2 54

3.4.5 Aplicação da Abordagem 3 57

3.4.6 Cálculo do Limiar de Confiança 60

4.PROJETO E IMPLEMENTAÇÃO DO WEB SITE 62
4.1 Descrição do Ambiente 62
4.2 Integração com o Servidor 62
4.3 Mecanismo de Captura dos Comportamentos 64
4.3.1 Agentes de Software 65
4.3.2 Coleta e Armazenamento dos Dados de Navegação 66
4.4 Construção das Páginas 68
4.5 Restrições do *Web Site* 68

5.ANÁLISE E APRESENTAÇÃO DOS RESULTADOS 69
5.1 Descrição dos Testes 69
5.2 Apresentação dos Resultados 70
5.2.1 Comparação das Abordagens 70
5.2.2 Limiar de Confiança 75
5.3 Discussão 76
5.3.1 Diretrizes 80

6.CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS 84
6.1 Resumo do Trabalho 84
6.2 Contribuições da Dissertação 86
6.3 Limitações e Dificuldades Encontradas 86
6.4 Trabalhos Futuros 87

REFERÊNCIAS BIBLIOGRÁFICAS 89

APÊNDICES

Apêndice A - Planos de Teste 95
Apêndice B - Telas do 1º web site 99
Apêndice C - Telas do 2º web site 101
Apêndice D - *Trust* por usuário 103
Apêndice E - *Trust* gerados 105
Apêndice F - Limiar por usuário 121

CAPÍTULO 1

INTRODUÇÃO

Este capítulo apresenta uma síntese da pesquisa documentada nesta dissertação abordando as motivações, a caracterização do problema, os objetivos da dissertação, os procedimentos adotados e a organização do texto.

1.1 Motivação

Grande parte do desenvolvimento de aplicações para a Internet hoje gira em torno da *Web,* que certamente é o serviço mais popular. No entanto, à medida que a *Web* cresce, também crescem os problemas de segurança associados a ela.

Segundo o relatório de ameaça de segurança na Internet da Symantec (2007), fraudes *on-line* e roubo de informações confidenciais, dominam o cenário atual. No primeiro semestre de 2007, 61% de todas as vulnerabilidades relatadas foram encontradas em aplicativos *Web* e os códigos projetados para roubar informações confidenciais representam 65% dos 50 principais códigos maliciosos reportados. De acordo com Forristal (2002) é nas aplicações que tipicamente uma empresa armazena seus dados sensíveis, como informações de clientes, incluindo nomes, senhas e informações de cartão de crédito, portanto, as aplicações *Web* são áreas de óbvio interesse para um ataque maligno.

Para terem acesso a essas aplicações, os usuários necessitam passar por um processo de autenticação. Esse processo geralmente é realizado através de credenciais na forma de "nome do usuário" e "senha". Mas conforme exposto por Schneier (2004), as senhas são facilmente roubadas por fraudadores que assumem a identidade de um usuário autorizado.

O processo de autenticação pode ser visualizado como sendo baseado em confiança (*Trust*), ou seja, uma vez autenticado, o usuário é considerado confiável para a aplicação. No entanto, em um cenário onde os dados de autenticação não são totalmente confiáveis, faz-se necessária a investigação de mecanismos complementares à autenticação. Com isso, é possível confirmar ainda mais a veracidade das informações inicialmente apresentadas, aumentando assim a confiança na identificação do usuário.

Atualmente, as propostas para incorporar mecanismos baseados em confiança, tanto em aplicações quanto em serviços *Web*, têm aumentado sensivelmente. A abordagem utilizada neste trabalho é fundamentada na proposta apresentada por Carmo *et al.* (2007), onde a confiança, definida informalmente como uma medida de quão certo o provedor de uma aplicação está a respeito da identidade de um usuário, é proveniente de uma avaliação comportamental.

Esse trabalho propõe que a avaliação do comportamento de um usuário seja contínua para permitir que a confiança e as restrições de acesso sejam deduzidas automaticamente em aplicações *Web*. A conduta contemporânea do usuário é analisada continuamente, considerando atividades pregressas, a fim de promover uma possível evolução do nível de confiança, dispensando assim a solicitação de credenciais em momentos pré-determinados.

O uso isolado da avaliação comportamental também pode ser útil para identificar condutas anômalas, tais como: um usuário autenticado no sistema que faz um uso legítimo para o abuso de recursos, uso esporádico por um colega de trabalho "que pede emprestada" uma estação de trabalho, ou ainda, um ataque automatizado lançado por um usuário relativamente ingênuo através de uma seqüência típica de ataque. Com o uso combinado de um mecanismo de troca de credenciais, a primeira condição naturalmente perde sua eficácia, pois é muito provável que um usuário autenticado também tenha sucesso numa troca de credenciais.

Considerando essa abordagem, a presente pesquisa propõe, na vertente de aplicações *Web*, estabelecer novos conhecimentos úteis para a avaliação comportamental e conseqüente estabelecimento da confiança.

1.2 Caracterização do Problema

Produzir conhecimentos para a avaliação comportamental de um usuário na *Web*, de forma a caracterizar e diferenciar o seu comportamento, não é uma tarefa fácil. A avaliação depende de fatores como, a determinação de um ambiente *Web* apropriado, a coleta e armazenamento de informações do comportamento do usuário, e a atribuição de uma medida de confiança a esse comportamento.

Trabalhos anteriores, como os que serão apresentados na Seção 3.4, relataram as experiências realizadas por meio da simulação dos comportamentos ou ainda pela adaptação de *logs* para permitir a avaliação comportamental. A dificuldade em simular comportamentos está em vislumbrar os possíveis cenários para criação das amostras. Em contrapartida, a decisão por adaptar *logs*, reputando informações registradas pode levar a conclusões errôneas. Acredita-se que a avaliação realizada em uma base de dados específica pode conduzir a conclusões mais propensas a acertos.

O problema inicia-se com a análise do contexto. Na análise procura-se definir uma coleção de páginas *Web*, onde os comportamentos possam ocorrer. Para isso, é preciso conhecer as informações que serão utilizadas e a ordem em que se encontram, e definir as informações que estão faltando, as irrelevantes e as que induzem os usuários a erros.

Definido o contexto, o problema que se segue é o de coleta e armazenamento de dados do comportamento de um usuário na *Web*. Normalmente, os arquivos de *log* gerados

automaticamente e armazenados em servidores *Web* são as principais fontes de informações sobre o comportamento dos 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 para identificação, que cada endereço IP é um usuário. Contudo, vários usuários podem utilizar um mesmo endereço IP, como por exemplo, quando o provedor fornece um IP dinâmico ou utiliza NAT (Network Address Translation).

Embora exista uma grande diferença em se identificar um usuário e analisar como ele se comporta em determinado contexto, o problema mencionado dificulta a reunião de um conjunto de instâncias comportamentais, relacionadas a um mesmo indivíduo. Essas instâncias vão dimensionar um ambiente propício a determinação da confiança, como será exposto na Seção seguinte.

A medida de confiança, segundo Carmo *et al.* (2007), é determinada considerando três fatores: 1) o quanto um comportamento se aproxima dos demais previamente capturados, 2) o quanto os comportamentos pertencentes à assinatura de um usuário, são semelhantes e, 3) o quanto a assinatura de um usuário é diferente das demais assinaturas de outros usuários.

Sendo assim, faz-se necessária a investigação de métodos que possam quantificar esses fatores a fim de estabelecer a confiança.

Dessa forma, no contexto desta dissertação, pretende-se examinar a seguinte questão: como gerar subsídios para a avaliação comportamental de um usuário que navega em um *web site?*

1.3 Objetivos da Dissertação

O objetivo geral deste trabalho é investigar abordagens para quantificar a confiança e obter uma medida para um determinado usuário. Na busca por uma melhor representação do valor de confiança os desempenhos das abordagens devem ser avaliados. Uma base de dados específica, gerada a partir da criação de um *web site* deve armazenar os comportamentos a serem mensurados.

O comportamento refere-se às informações de uso de um ambiente *Web*, tal como a maneira de navegar de um indivíduo, dada pela trilha de navegação, ou seja, pelo caminho percorrido por ele durante sua interação com o *web site*.

A trilha é interpretada como uma instância gerada para cada macro-estado (conjunto de páginas da aplicação *Web*) que uma vez identificada fará parte de uma assinatura (conjunto de instâncias associadas a um mesmo indivíduo). Através da assinatura, pode-se trabalhar com a confiança de forma analítica.

Conseqüentemente, o objetivo específico é propor diretrizes para direcionar a construção de *web sites* que utilizam a avaliação comportamental baseada em confiança como auxílio na identificação de um usuário. As diretrizes provêm de uma análise empírica, realizada por meio do *web site* construído como ambiente de teste. Assume-se que através da observação sistemática, ou alteração de algum elemento, pode-se obter as evidências para a formulação das diretrizes.

É importante observar que a confiança é utilizada de maneira unidirecional, sendo assim, não há o conceito de confiança do usuário na aplicação e sim apenas a confiança da aplicação no usuário.

1.4 Procedimentos Adotados

Os procedimentos que norteiam a elaboração desta pesquisa são descritos a seguir.

A primeira etapa compreendeu a definição do contexto do experimento que é o facilitador para a avaliação comportamental. Foi definido que os comportamentos ocorreriam no contexto de uma loja virtual, com um único macro-estado, para que os usuários fossem diferenciados pelo modo em que navegam.

A etapa seguinte foi a de projeto e implementação do *web site* da loja. Nesta etapa também foi implementado o mecanismo de coleta de dados através de agentes de software. Uma vantagem desse mecanismo é que o usuário não participa diretamente do processo, ou seja, ele não fornece diretamente os dados que serão usados para compor o seu histórico de navegação.

Assim, dois testes *on-line* foram realizados em períodos diferentes por um mesmo grupo de pessoas que receberam as instruções de acesso à loja. Com isto, os dados de navegação foram gerados e armazenados em um banco de dados. Esses testes serviram principalmente para coletar as amostras de uso do ambiente que caracterizam cada usuário, possibilitar a análise, tanto da técnica utilizada para coletar os comportamentos quanto da aplicação dos conceitos necessários à avaliação comportamental no desenvolvimento de *web sites*.

Coletadas as informações, técnicas para reconhecer padrões de comportamento foram investigadas como possibilidade de quantificar os três fatores de confiança. Estas foram comparadas como a finalidade de se buscar uma melhor representação. Ainda nesta etapa foi imprescindível a implementação de algoritmos baseados nas técnicas, a fim de estabelecer o valor de confiança para cada usuário.

A etapa final compreendeu a determinação de um limiar mínimo de confiança para cada usuário e a análise e interpretação dos resultados. Esta etapa incluiu ainda a definição e a reunião de um conjunto de diretrizes levantadas durante o desenvolvimento do *web site* experimental.

1.5 Estrutura da Dissertação

Este trabalho está organizado em seis capítulos, sendo o primeiro esta introdução.

O Capítulo 2 faz uma revisão dos aspectos teóricos necessários para a compreensão deste trabalho, expõe alguns trabalhos publicados na literatura que serviram de base e inspiração para o presente desenvolvimento e são descritos considerando as características importantes de cada um.

O Capítulo 3 apresenta, sucintamente, os conceitos relacionados às abordagens investigadas para quantificar os fatores de confiança e também a proposta deste trabalho para avaliar um comportamento experimentalmente.

O Capítulo 4 trata de detalhes do projeto e implementação do *web site* para teste e descreve o mecanismo de captura do comportamento contextual dos usuários.

O Capítulo 5 discute os testes e os resultados obtidos por meio da aplicação das abordagens investigadas para o cálculo da confiança.

O Capítulo 6 descreve as contribuições, as limitações e dificuldades encontradas e sugere novos trabalhos para continuação desta pesquisa.

CAPÍTULO 2

CONCEITOS BÁSICOS

São apresentados neste capítulo, alguns conceitos básicos sobre autenticação, confiança e reconhecimento de padrões de comportamento. É apresentada, também, uma revisão sistemática do estado da arte.

2.1 Autenticação

Segundo Lopez *et al.* (2004), a autenticação é vista como um serviço que prova que a identidade de um objeto ou sujeito é de fato a que ele diz ter. A autenticação de usuários em um sistema de computador pode ser realizada através de diferentes métodos, que apresentam características distintas. Rubin (1996), considera três tipos mais comuns, são eles:

- conhecimento (o que as pessoas sabem) - a autenticação baseada em conhecimento considera que somente uma pessoa tem acesso exclusivo a uma determinada informação. Um exemplo muito popular deste método é a utilização de segredos, como as senhas;

- propriedade (o que as pessoas têm) - esse método se baseia na utilização de elementos físicos que permanecem em poder de uma pessoa e que servem para autenticar sua identidade. Os dispositivos mais conhecidos são os cartões com tarja magnética e com código de barras, os *smartcard*s (conhecidos como cartões inteligentes) e os *tokens;*

- características (o que as pessoas são) - a autenticação baseada neste método utiliza as características pessoais de um indivíduo para identificá-lo. Uma forma de identificação

é através da biometria que agrega equipamentos e análise das características, tais como, formato da íris, impressão digital e geometria da mão ou face;

Para dificultar as fraudes no momento da autenticação, um ou mais métodos podem ser utilizados em conjunto. É comum a combinação de autenticação por propriedade com autenticação baseada em senhas, onde o portador de um cartão magnético pode precisar digitar uma senha para executar uma transação bancária, além de informar algum dado pessoal, como a data do seu nascimento.

2.1.1 Autenticação via Credenciais Digitais

As credenciais digitais de autenticação são as evidências apresentadas por um requerente para confirmar a sua identidade e/ou o direito de acesso a determinada informação. Existem vários tipos de credenciais, entre eles: certificados e nome e senha.

Nome e Senha

O mecanismo de autenticação mais popular, usado nos sistemas de computação, é a autenticação através de nome e senha. No entanto, sabe-se que as senhas podem ser descobertas, roubadas ou esquecidas. Soluções alternativas foram propostas para tentar tornar a autenticação mais segura. Segue abaixo uma breve descrição de algumas delas:

- senhas descartáveis (One-Time Passwords) - uma senha descartável é aquela que só é usada uma vez para estabelecer a autenticação. A cada processo de autenticação uma nova senha é solicitada. As senhas descartáveis podem ser classificadas como sincronizadas no tempo ou desafio/resposta. No método de autenticação baseado em tempo sincronizado, uma nova senha é calculada a cada *n* segundos. Em sistemas baseados em desafio/resposta, o sistema envia um desafio para o usuário que só pode

ser respondido caso ele detenha as informações de autenticação corretas. *Smartcards* e *tokens* são exemplos de dispositivos que utilizam essa solução (DEVEGILI e PARENTE, 2003; FIORESE, 2000).

- perguntas randômicas (Random Queries) - este método de autenticação é baseado em desafio/resposta. Primeiramente, é feito um cadastro do usuário, no qual ele responde a um questionário com perguntas variadas, como o número da identidade, CPF ou data de nascimento. Ao entrar com sua identificação, o sistema escolhe uma pergunta do questionário de forma aleatória e desafia o usuário. Se sua resposta coincidir com a previamente armazenada no questionário, os direitos de acesso são dados (FIORESE, 2000).

Certificados Digitais

Um certificado digital é um documento eletrônico assinado digitalmente por um componente de uma Infra-estrutura de Chave Pública (ICP) com a finalidade de associar uma entidade à sua chave pública. A entidade certificada pode ser uma pessoa (física ou jurídica), um dispositivo de hardware, um servidor *Web* ou mesmo um software A tecnologia utilizada é baseada na criptografia assimétrica, onde a chave pública é armazenada na identidade digital (ROCHA, 2001).

A recomendação X.509, criada pela *International Telecommunication Union - Telecommunication Standardization* (ITU-T), fornece uma base para definir formatos de dados e procedimentos para a distribuição de chaves públicas através dos certificados (HUNT, 2001).

Por meio dos componentes de uma ICP é possível gerenciar todo o ciclo de vida de um certificado que envolve: requisição, validação da requisição, emissão, aceitação pelo requerente,

utilização, suspensão, revogação, expiração e renovação. Os principais componentes e suas responsabilidades são descritos por Ardigo (2004) da seguinte forma:

- autoridade certificadora (AC) - responsável pela emissão, disponibilização, suspensão e revogação de certificados digitais;

- autoridade de registro (AR) - responsável pela conferência das informações prestadas pelo requerente;

- diretório público (DP) - responsável por disponibilizar em local público todos os certificados emitidos por uma AC, bem como listas de certificados revogados;

- autoridade de datação (AD) - responsável por conferir âncoras temporais a documentos eletrônicos.

2.1.2 Autenticação via *Token*

O *token* é um dispositivo que fornece autenticação híbrida, utilizando tanto autenticação por propriedade (o próprio dispositivo), como por conhecimento (senhas descartáveis).

Na abordagem mais conhecida, um algoritmo que executa tanto no *token* quanto no servidor, gera números idênticos que mudam no decorrer do tempo. Quando se deseja entrar no sistema, o usuário deve informar seu PIN (Personal Identification Number), seguido por um número mostrado no momento pelo *token*. Ao receber o PIN, o servidor localiza a chave do usuário e calcula qual deveria ser a senha de acesso naquele momento, comparando-a com a que o usuário enviou. Caso sejam iguais, o acesso é liberado (FIORESE, 2000; O'GORMAN, 2003).

2.1.3 Autenticação via Biometria

A biometria é o ramo da ciência que estuda a mensuração dos seres vivos (COSTA, 2006). O termo tem sido associado a medidas biométricas das pessoas como forma de identificá-las unicamente. Uma medida biométrica é definida pela *International Association for Biometrics* (2007) como: *"A medida de característica física ou de traço do comportamento pessoal com o objetivo de identificar ou verificar a identidade reivindicada"*.

A característica física é uma propriedade fisiológica relativamente imutável como o padrão da íris, as impressões digitais, a forma do contorno da mão e face. Um sistema, baseado nessa característica, pode ser classificado como de identificação ou verificação. Um sistema de identificação compara a informação biométrica apresentada por um indivíduo com toda a informação armazenada de todos os indivíduos, ou determinado conjunto deles, e declara se existe uma correspondência com algum deles ou não. O sistema de verificação compara a informação biométrica apresentada por um indivíduo com a informação armazenada em uma base de dados correspondente àquele indivíduo. Nesse caso, o sistema decide se existe uma correspondência entre a informação de entrada e a armazenada na base de dados (JAIN *et al.,* 2004).

Já a característica comportamental é um reflexo das atitudes psicológicas do indivíduo, se preocupa com a extração de características mais sutis, como o tipo de escrita do indivíduo, a maneira como assina seu nome, a forma como pronuncia determinadas palavras e o ritmo com que digita no teclado. O processo de analisar o ritmo com que um usuário digita em um terminal através do monitoramento das entradas do teclado, é apresentado nos trabalhos de Costa (2006) e Monrose e Rubin (2000). Segundo Jain *et al.* (2004), não se espera que esse comportamento biométrico seja único para cada indivíduo, mas que ele ofereça informações discriminatórias o suficiente para permitir a verificação da identidade.

2.2 Confiança

A maioria das decisões tomadas diariamente leva em consideração, algum fator de confiança e isso é verificável na interação com outra pessoa. Mesmo com os desenvolvimentos das tecnologias de rede, a confiança ainda existe, quando não é entre as pessoas é entre os aparelhos eletrônicos que são usados para interagir um com o outro.

Diversos autores têm trabalhado o conceito de confiança a partir de diferentes pontos de vista constituindo uma base teórica para as técnicas que tentam expressar a confiança do mundo real em vários formatos digitais (GRANDISON, 2003).

2.2.1 Teoria da Confiança

De acordo com Marsh (1994), a maioria dos trabalhos sobre confiança tem vindo de três áreas diferentes: sociologia, psicologia e filosofia. Marsh, também descreve quatro pesquisadores como os maiores colaboradores: Morton Deutsch, Niklas Luhmann, Bernard Barber e Diogo Gambetta.

O trabalho de Deutsch, segundo Marsh (1994), é baseado na psicologia e a confiança é definida da seguinte forma:

- um indivíduo confronta-se com um caminho ambíguo, um caminho que pode levar a um evento percebido como sendo benéfico (Va+) ou a um evento percebido como sendo maléfico (Va-);
- ele percebe que a ocorrência de (Va+) ou de (Va-) são possibilidades de comportamento de qualquer pessoa;
- ele percebe que a força de (Va-) pode ser maior que a força de (Va+);

- se ele escolhe tomar o caminho com tais propriedades, então ele fez uma escolha confiante; se ele escolhe não tomar o caminho, ele fez uma escolha desconfiada

Esta definição implica que a confiança é dependente dos indivíduos, sobretudo, da sua percepção e da análise dos benefícios do cenário dado. Escolher um caminho com tais propriedades, sempre envolve incertezas e ao mesmo tempo indica algum otimismo. É também interessante observar que a confiança nessa definição está ligada aos prejuízos percebidos (análise de custo/benefício).

Luhmann (2000), define a confiança como uma solução para problemas específicos de risco. Nessa definição, a confiança que uma entidade *A* tem em outra entidade *B* pode ser usada pela entidade *A* para mitigar o risco envolvido em um negócio com a entidade *B*. As idéias de Niklas Luhmann são baseadas na sociologia e semelhantes à de Deutsch quando descreve o cenário onde a confiança é usada. Toda vez que ocorrer uma situação de tomada de uma decisão, é preciso fazer algumas suposições, levando em conta a situação e o ambiente em particular e então fazer alguma escolha confiante.

Ainda segundo Marsh (1994), o trabalho de Bernard Barber, também baseado na sociologia faz uma ligação entre a confiança, a sociedade e as culturas, não como um valor do custo benefício percebido por um indivíduo, como as idéias de Deutsch, e sim a confiança com expectativa sobre o futuro. Barber apresenta três expectativas que envolvem alguns dos significados fundamentais da confiança. São elas:

- a expectativa de persistência e realização (cumprimento) das ordens (pedidos) naturais e sócio-morais;

- a expectativa da realização tecnicamente competente da regra (papel) daquele com quem se interage nos relacionamentos sociais e sistemas;

- a expectativa de que os parceiros na interação realizarão suas obrigações e responsabilidades (deveres) colocando os interesses de outros antes dos seus próprios.

Gambetta (2000), apresenta uma definição diferente de confiança, ele define a confiança em termos matemáticos como a seguir:

> *"Confiança (ou simetricamente desconfiança) é um nível particular da probabilidade subjetiva com a qual um agente avalia que outro agente ou grupos de agentes realizarão uma ação em particular. Quando nós dizemos que confiamos em alguém ou que alguém é confiável, nós implicitamente queremos dizer que a probabilidade de que ele realizará uma ação que é benéfica para nós é alta o suficiente para nós nos considerarmos comprometidos em alguma forma de cooperação com essa pessoa. Correspondentemente quando nós dizemos que alguém é desconfiável, isso implica que essa probabilidade é baixa o suficiente para abstermos de nos comprometer."* (GAMBETTA, 2000).

Nesta definição a confiança é modelada matematicamente e, portanto, torna-se mais concreta do que abstrata, comparada a outras definições apresentadas anteriormente. A confiança tem uma medida de *0* a *1,* onde *0* representa a desconfiança completa e *1* representa a confiança total. A confiança cega é um exemplo de confiança completa, onde um agente tem total confiança em outro, não importa qual.

Gambetta reconhece o fato de que a confiança é relevante somente quando há uma possibilidade de desconfiança, traição, saída ou deserção. Isso pode ser expandido dizendo que quando alguém é confiável (mas não completamente, pois do contrário a probabilidade seria *1*) há um risco de que a ação que ele realiza possa não ser benéfica.

2.2.2 Principais Formas de Confiança

Embora as seções anteriores resumam quatro diferentes definições de confiança, elas têm sido ligadas a perspectivas diferentes. Esta seção realça alguns dos conceitos de confiança e seu relacionamento.

Confiança e Cooperação

Gambetta (2000), relaciona a confiança com a cooperação no sentido que se a confiança for somente unilateral, então a cooperação não pode ter sucesso. Similarmente se houver desconfiança completa entre os agentes envolvidos, então não pode haver nenhuma cooperação entre eles. Um nível mais alto de confiança geralmente leva a uma probabilidade mais alta de cooperação.

O conceito de confiança ligado a cooperação também foi abordado por Jones e Marsh (1997). Esse trabalho toma como base os indivíduos e o contexto onde as interações entre eles ocorrem. Os seguintes conceitos são definidos:

- indivíduos - entidades identificadas numa relação de confiança;
- confiança inicial - toda confiança é derivada de uma confiança inicial. Essa confiança, genericamente, pode ser derivada de experiências em interações anteriores e pode mudar com o resultado de novas experiências;
- conhecimento - representação do conhecimento de outros indivíduos como indicativo de confiança que pode ser utilizado pelos demais indivíduos numa relação de confiança;
- confiança geral - dados dois indivíduos que se conhecem, ambos terão uma noção do quanto um pode confiar no outro, independente do contexto;
- confiança contextual - quando indivíduos se encontram em determinado contexto, há uma confiança contextual. Ou seja, para uma determinada situação um indivíduo pode ser mais confiável do que em outra situação ou contexto;

- confiança no tempo - o fato de que um indivíduo *A* tenha confiado em *B* no passado, não garante que venha a confiar em *B* no futuro. As suas interações e informações relevantes podem sempre levá-lo, a reavaliar sua confiança em *B*;

- importância - a interação num dado contexto pode possuir certo nível de importância. Por exemplo, em uma aplicação bancária pela Internet, o efeito de consulta ao saldo é menos crítico que o causado por uma transferência de valores;

- utilidade - indivíduos esperam algum retorno pela cooperação.

Confiança e Recomendação

Em qualquer sociedade de tamanho razoável é impossível todos se conhecerem e confiarem uns nos outros. Em uma situação onde não se sabe, se é possível confiar em alguém ou não, tende-se a perguntar a uma pessoa, que se conhece e confia, se a pessoa sob consideração é confiável ou não. Baseada na recomendação da terceira pessoa pode-se decidir se confia ou não nessa pessoa sob consideração. Normalmente, considera-se o quanto se pode primeiro confiar nessa terceira pessoa e o quanto esta terceira pessoa confia na pessoa sob consideração. Se a terceira pessoa não o conhece então ele pode pedir uma recomendação a outra pessoa que o conhece e assim por diante. Geralmente, quanto mais longa a corrente de recomendação se torna, mais difícil fica de confiar na pessoa sob consideração. A pessoa pode ser confiável só que o valor da confiança diminui.

O conceito de recomendação tem sido empregado em vários trabalhos (REATEGUI e CAZELLA, 2005; MOTTA, 1999; ARMADA, 2005). Sua aplicação abrange ambientes de comércio eletrônico onde as recomendações podem ser geradas para auxiliar os usuários a encontrar o conteúdo apropriado ao seu perfil mais rapidamente, e ainda, sistemas de

recomendação voltados para equipes de trabalho onde o processo de recomendação é baseado na avaliação do grupo.

Em outra vertente, Brainard *et al.* (2006) apresentaram o conceito de recomendação como um quarto método para autenticação ("alguém que as pessoas conhecem"). A rede social de um usuário é usada para prover um procedimento de autenticação de "emergência".

Confiança e Teoria de Jogo

A teoria de jogo é uma teoria matemática de barganha. A teoria de jogo mais popular é o dilema dos prisioneiros, que explica o significado da confiança e como um terceiro indivíduo confiável pode ajudar a ter um melhor resultado para os dois primeiros indivíduos (ROSS, 1997).

Esta situação pode ser explicada resumidamente com a ajuda de um exemplo. Considere que dois ladrões Bob e Al foram pegos pela polícia. Cada um tem que escolher se confessa ou não e se envolve o outro ou não. O cenário é o seguinte: 1) se Bob e Al não confessarem, então ambos serão presos por um ano, por porte de arma; 2) se Bob confessar e ao mesmo tempo Al confessar, então ambos ficarão presos por 5 anos cada; 3) entretanto, se Bob confessa, mas Al não confessa, então Bob fica livre e Al pega 10 anos. Similarmente o reverso acontece. Esse cenário do dilema dos prisioneiros pode ser representado como mostrado na seguinte Tabela 1:

Tabela 1: Dilema dos prisioneiros

		Al	
		Confessa	Não confessa
Bob	Confessa	5, 5	0, 10
	Não confessa	10, 0	1, 1

Se Bob e Al forem mantidos em celas diferentes e eles não têm qualquer meio para se comunicarem um com o outro, então eles tentarão escolher uma estratégia que minimizará o tempo em que eles têm que ficar na prisão.

Confiar na outra pessoa nesse cenário é difícil. Nesse caso, Bob e Al poderiam usar um terceiro indivíduo e negociarem através de Mike, que é confiável por ambos, assim eles podem tomar a estratégia que consegue os melhores resultados para os dois, e saírem da cadeia com um ano cada.

Confiança e Reputação

Segundo Dasgupta (2000), a confiança baseia-se na reputação e a reputação é adquirida através do comportamento com o passar do tempo em circunstâncias bem compreendidas.

A reputação é definida por Artz e Gil (2007), como uma avaliação apoiada na história das interações, observações ou experiência pessoal direta ou recomendada de uma entidade.

Em segurança de redes, o conceito de reputação é comumente associado aos protocolos de redes *peer-to-peer*. Os protocolos de reputação, como são conhecidos, geralmente baseiam-se em suas experiências anteriores ou de nós vizinhos para calcular um valor de confiança com relação a um determinado nó possuidor do recurso ou mesmo ao próprio recurso (PELISSARI, 2005).

De outra forma, o conceito também é utilizado pelos sistemas de reputação que auxiliam os usuários das comunidades virtuais a criarem relações de confiabilidade entre si (CRUZ e MOTTA, 2006).

Confiança e Segurança

O significado de confiança no ambiente de rede de computação pode ser ilustrado observando-se um cenário simples da vida real.

Quando se precisa de algum produto, é possível ir a um supermercado e adquirí-lo. Pagando pelo item, leva-se um recibo. Se houver algo de errado com a mercadoria, é possível devolvê-la ou trocá-la apresentando o mesmo recibo. O recibo prova que o item foi comprado naquela loja em

particular e que foi paga uma soma em dinheiro por ele. Se aquela loja não tem o produto procurado, então o funcionário do atendimento pode recomendar outra loja que ele confie. Se existe uma confiança no funcionário, então se pode ir até a outra loja e comprar o item. Ao invés de ir à loja pessoalmente, pode-se fazer a mesma compra via Internet. Mas como se pode confiar se está comprando na loja verdadeira, que se está pagando para a pessoa certa, que o item chegará após pagar por ele, que não há ninguém monitorando os detalhes do cartão de crédito ou dos dados credenciais?

Esses são os tópicos de segurança que um ambiente de rede tem que resolver antes que se deposite a confiança no sistema de compras pela Internet. E, são muito comuns em qualquer ambiente de rede onde se espera que se forneça certo nível de confiança em termos de segurança.

2.3 Reconhecimento de Padrões de Comportamento

O reconhecimento de padrões engloba uma vasta literatura e uma coleção de problemas, cujas soluções empregam técnicas provenientes de várias áreas.

Em Schalkoff (1992), o reconhecimento de padrões é definido como a ciência que realiza a descrição ou classificação (reconhecimento) de medidas. Segundo Gonzalez e Thomason (1978), um padrão é uma descrição quantitativa ou estrutural de um objeto ou alguma outra entidade de interesse. Um padrão pode ser, por exemplo, uma imagem de impressão digital, uma palavra escrita a mão, uma face humana ou um sinal de voz.

Ainda segundo Schalkoff (1992), as abordagens existentes para o reconhecimento de padrões são: sintática (ou estrutural), estatística e neural. Em Bishop (1995), a abordagem neural é considerada como uma extensão das técnicas utilizadas na abordagem estatística.

No reconhecimento de padrões estatísticos, as características são representadas na forma de *n*-uplas ou vetores, sendo utilizadas regras de decisão, teoria de probabilidades, funções discriminantes e outros procedimentos estocásticos. Na abordagem neural, as características também estão na forma de *n*-uplas ou vetores, no entanto as redes neurais possuem propriedades peculiares, tais como a capacidade de generalização, abstração e aprendizagem a partir de exemplos (COSTA, 2006).

No reconhecimento de padrões sintáticos, as características são da forma de sentenças de uma linguagem reconhecida por uma gramática de estrutura de frases. É também conhecido como reconhecimento estrutural de padrões, onde as características são estruturas de dados simbólicos, como cadeias de caracteres, para representação de padrões individuais. Porém, em vez de usar uma gramática, as classes de padrões são representadas por meio de protótipos (OGIELA e TADEUSIEWICZ, 2003 *apud* PIO *et al.*, 2006).

2.3.1 Tipos de Comportamento

A análise do comportamento leva em consideração dois tipos de comportamentos, o físico e o contextual. Os sistemas de segurança que se fundamentam na análise do comportamento físico buscam adquirir experiência de alguma característica pessoal do usuário como, por exemplo, a forma de teclar ou a forma de utilização do mouse. Sistemas de segurança que envolvem 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*.

Comportamento Físico

No trabalho apresentado por Monrose e Rubin (1997), é exposta uma técnica para melhorar a autorização de acesso a recursos e a dados de um sistema via utilização do teclado. Este trabalho analisou a maneira de como o usuário tecla no teclado. As entradas foram monitoradas 100 vezes por segundo, com o objetivo de identificar um usuário baseado no seu ritmo habitual. Quando uma pessoa tecla, a pressão aplicada na teclada pode ser usada para construir uma assinatura única para aquele indivíduo.

O foco principal está no exame de um conjunto de características armazenadas sobre algum exemplo de população e seleção da que melhor representa o perfil desconhecido a ser identificado. Esse trabalho concentrou-se no problema de reconhecimento de perfis, acreditando que o sucesso no reconhecimento é diretamente transferido para a autenticação. Os perfis coletados foram representados na forma de vetores de características. Similaridades e diferenças foram calculadas usando distância euclideana normalizada e medida de probabilidade máxima.

Mais recentemente, outro trabalho envolvendo comportamento físico, foi proposto por Pusara e Brodley (2004). Os autores apresentaram uma técnica para re-autenticação, baseada na coleção de dados capturada via dispositivo do mouse. Este trabalho investigou também qual dos movimentos do mouse provê um modelo mais preciso para identificação do usuário.

Inicialmente foi definido um modelo de comportamento normal. Este modelo requer uma fase de treinamento inicial, durante o qual os dados da utilização do mouse pelos usuários são coletados, os parâmetros do modelo são selecionados e um modelo final é então produzido. Este modelo é então usado para monitorar a confiança continuamente. Para modelar o comportamento do movimento do mouse de um usuário foi necessária a captura do movimento do cursor e também dos eventos do mouse (*clicks* duplos ou simples do botão da esquerda, direita ou do meio e posição instantânea).

Depois de iniciar um modelo normal de comportamento é aplicado o modelo que detecta se o comportamento corrente é normal ou anômalo. Se o comportamento corrente do usuário desvia significantemente do modelo normal de comportamento, o sistema reconhece esse comportamento como anômalo e então pede ao usuário para autenticar novamente ou reporta a anomalia para o administrador. Para classificar o comportamento é empregada uma árvore de decisão, onde cada nó interno denota um teste em uma característica.

Comportamento Contextual

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.

Os autores desenvolveram um protótipo de sistema de detecção de intrusão por anomalia que emprega um *framework* IBL (Instance-Based Learning) para classificar comportamentos de usuários como normais e anormais.

Para adaptar a tarefa de detecção por anomalia para IBL, foi preciso definir um vetor de características e o conceito do mais próximo, ou seja, a similaridade de dois vetores.

Os dados de entrada do sistema são linhas de comando Unix e 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 desenvolvida a ferramenta *SimOffice,* que utiliza conceitos de confiança para controle de acesso em *web services*. A idéia principal foi a criação de 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.

Em um outro trabalho, 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.

Nesse caso, a confiança é uma medida probabilística quantificada com valores entre [*0,1*], onde *0* apresenta desconfiança completa e *1* a confiança total.

No domínio da mineração de dados, Onoda (2006) propõe o estudo do comportamento de navegação dos usuários, integrando *Web Usage Mining* e análise de comportamento. Enquanto que a análise do comportamento é realizada de forma individual, onde as informações coletadas ao longo do tempo servem para comparar um indivíduo consigo mesmo, em *Web Usage Mining* o estudo é voltado para grupos de indivíduos.

A partir da metodologia definida para analisar o comportamento de navegação, várias hipóteses sobre o comportamento dos usuários em um *web site* são levantadas. A conscientização sobre essas hipóteses permite o desenvolvimento de ações mais convenientes ao *web site*.

2.3.2 Técnicas para o Reconhecimento

Segundo Jain *et al.* (2000), o conhecimento de mecanismos que realizam o reconhecimento de padrões permite construir máquinas que podem observar o ambiente, aprender e distinguir padrões de interesse, verificar e justificar decisões sobre as classes dos padrões.

Aprendizado de Máquina

Aprendizado de máquina é um campo da Inteligência Artificial dedicado ao desenvolvimento de algoritmos e técnicas que permitam um sistema aprender com sua experiência. Isto é, um sistema pode obter informações que permitam que realizações futuras de uma tarefa tenham melhor desempenho (ARAUJO, 2004). Existem dois tipos básicos de aprendizado: indutivo e o dedutivo.

O aprendizado indutivo implica na obtenção de conclusões gerais a partir de fatos particulares, partindo do específico para o geral. O aprendizado dedutivo, ao contrário, parte do geral para o específico.

O problema do aprendizado indutivo é que as hipóteses geradas nem sempre preservam a verdade seja porque o número de observações pode ter sido insuficiente ou os dados relevantes para chegar à conclusão foram mal escolhidos. Mas, ainda assim, o raciocínio indutivo é fundamental e mostra-se útil principalmente para fazer generalizações sobre características específicas bem como encontrar agrupamentos a partir de características semelhantes (BARANAUSKAS e MONARD, 2000 *apud* ARAGÃO, 2004).

Aprendizado Baseado em Instâncias

Aprendizado Baseado em Instâncias (Instance-Based Learning - IBL) é uma técnica de aprendizado indutivo que utiliza dados armazenados, ao invés de um conjunto de regras induzidas, para classificar novos exemplos (ARAUJO, 2004).

Para classificar um novo exemplo, os algoritmos IBL utilizam uma função de distância ou similaridade para determinar o quão próximo um novo exemplo encontra-se de um exemplo já armazenado, e utiliza este(s) exemplo(s) mais próximo(s) para determinar a classe do novo exemplo.

Uma descrição do conceito inclui um conjunto de instâncias armazenadas e, possivelmente, alguma informação com relação aos seus desempenhos anteriores, durante o processo de classificação. Esse conjunto de instâncias pode mudar após cada instância de treinamento ter sido processada (AHA *et al.*, 1991).

Aprendizado Baseado em Árvore de Decisão

A árvore de decisão é também uma técnica, de aprendizado indutivo, muito utilizada. Uma árvore de decisão tem como entrada um objeto ou situação descritos por um conjunto de atributos, discretos ou contínuos, e retorna uma decisão (valor de saída previsto, de acordo com a entrada) (RUSSELL e NORVING, 2004).

Uma árvore de decisão alcança sua decisão executando uma seqüência de testes em algum atributo. Cada nó interno na árvore corresponde a um teste do valor de uma das propriedades, e as ramificações a partir do nó são identificadas com os valores possíveis do teste. Cada nó folha da árvore especifica o valor a ser retornado se aquela folha for alcançada

Este capítulo apresentou alguns conceitos básicos sobre autenticação, confiança e reconhecimento de padrões. Apresentou também trabalhos publicados na literatura científica em diversas áreas, que foram descritos considerando as características importantes de cada um, e serviram de base e inspiração para o desenvolvimento da proposta que será apresentada no capítulo seguinte.

CAPÍTULO 3

PROPOSTA PARA A AVALIAÇÃO COMPORTAMENTAL

É realizada, neste capítulo, a exposição dos principais conceitos necessários ao entendimento sobre o cálculo da confiança e as abordagens investigadas como forma de representar quantitativamente os seus fatores. Antes, faz-se necessário expor o contexto no qual a avaliação comportamental está sendo pesquisada. Em seguida, é apresentada a proposta deste trabalho para avaliar um comportamento experimentalmente.

3.1 Mecanismo de Segurança Baseado em Confiança

A avaliação comportamental é pesquisada no contexto da proposta de construção do mecanismo de autorização, baseado em confiança, apresentado por Carmo *et al.* (2007).

Nesse mecanismo, uma assinatura comportamental é representada pelo conjunto de instâncias comportamentais de um usuário, sendo gerada para cada macro-estado. Um macro-estado, por sua vez, é representado pelo conjunto de páginas de uma aplicação *Web*.

A Figura 1 mostra um exemplo de instância comportamental gerada para um único macro-estado. Cada estado do macro-estado {a, b, c, d, e, f, g, h, i, j, k, l, n} representa uma página da aplicação e a instância comportamental {a, c, f, g, k} é formada pelas páginas visitadas por um usuário durante sua interação com o ambiente.

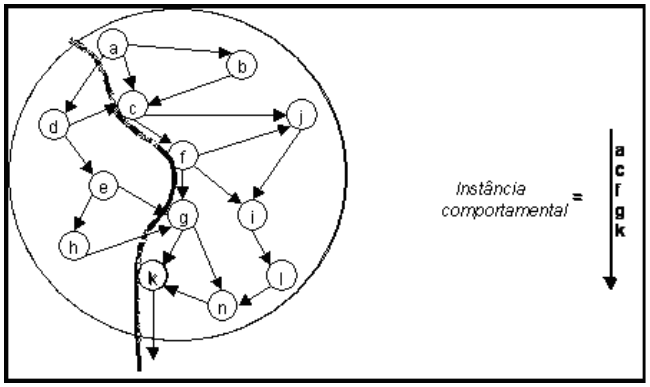

Figura 1: Exemplo de instância comportamental

A mudança de um macro-estado para outro (implicando em mudança de política de acesso) é feita pelo resultado de uma avaliação comportamental, ou, em caso de insucesso, por uma negociação explícita, via troca de credenciais como ilustrado na Figura 2.

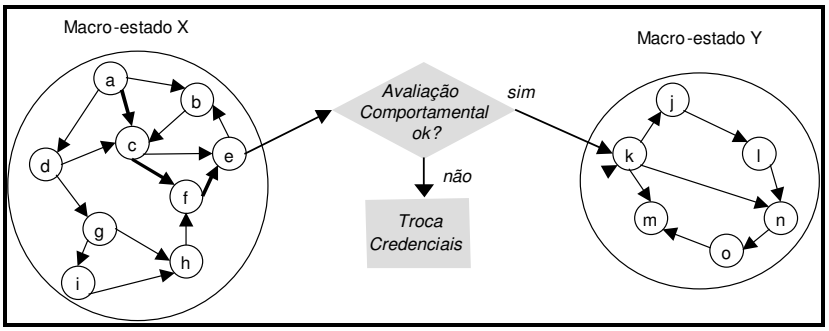

Figura 2: Macro-estados x avaliação da confiança

A avaliação por credenciais só é executada no caso de uma avaliação negativa do comportamento que tem como parâmetro a confiança. Assim, é necessário que um patamar de aceitação mínimo seja estabelecido para que a partir desse valor cada usuário possa ou não ser considerado confiável pela aplicação.

O conceito de confiança pode evoluir a partir do histórico de utilização e do comportamento corrente do usuário. Inicialmente, a confiança é a mesma para todos os usuários e evolui quando o usuário tem comportamento similar ao comportamento registrado no histórico anterior.

3.2 Cálculo da Confiança Comportamental

Para o cálculo da confiança, adotou-se o modelo apresentado por Carmo *et al.* (2007), que considera fundamentalmente a existência de três fatores interdependentes: a similaridade comparativa, a intra-similaridade e a inter-similaridade.

Similaridade Comparativa

Similaridade comparativa (SComp) representa a similaridade entre a instância atual coletada e o conjunto de instâncias que compõem a assinatura. Essencialmente, o seu valor espelha o quanto esse comportamento se aproxima dos demais previamente capturados. A Figura 3 ilustra um exemplo de *SComp*. A instância mais atual é comparada as demais instâncias da assinatura de um usuário que realiza uma navegação por uma aplicação.

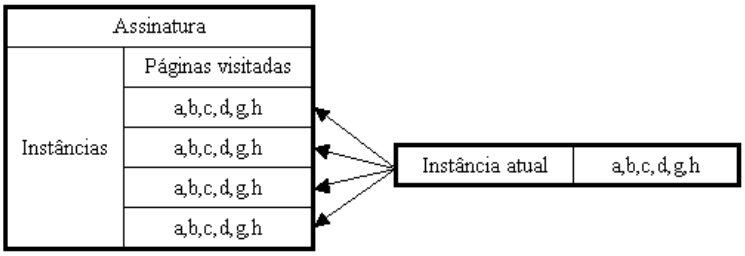

Figura 3: Exemplo de SComp

Intra-similaridade

Intra-similaridade (SIntra) está relacionada à qualidade da assinatura do usuário, sendo completamente independente da amostra atual de instância comportamental. Este fator representa se um usuário possui ou não um comportamento bem formado. Um comportamento bem formado é quando as instâncias pertencentes à assinatura são repetidas, ou levemente diferentes. O contrário quando todas as instâncias da assinatura são bem diferentes entre si, o comportamento é considerado mal formado. Naturalmente, uma assinatura mal formada dificulta o processo de validação do comportamento do usuário. A Figura 4 ilustra um exemplo de *SIntra*.

As instâncias da assinatura de um usuário que possui um comportamento bem formado são comparadas duas a duas.

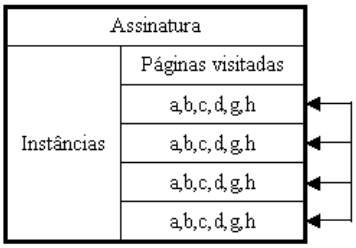

Figura 4: Exemplo de SIntra

Inter-similaridade

Inter-similaridade (SInter): traduz a qualidade da assinatura de um usuário em relação ao conjunto de assinaturas (de diferentes usuários), associado a um mesmo macro-estado. Uma dada instância comportamental pode ser altamente similar a uma assinatura bem-formada, mas mesmo assim, não expirar confiança devido a uma possível similaridade dela com as outras assinaturas existentes (dos diversos usuários do mesmo cenário). A similaridade das assinaturas dificulta o processo de diferenciação dos usuários. A Figura 5 ilustra um exemplo de *SInter*. As instâncias da assinatura de um usuário são comparadas as instâncias das assinaturas de outros usuários.

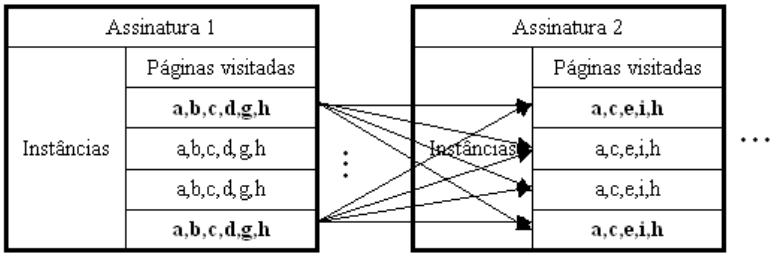

Figura 5: Exemplo de SInter

Os termos são utilizados junto à noção de "medida" para identificar o quanto os pares, instâncias ou assinaturas, são similares entre si. Esse modelo considera ainda que a confiança (*Trust*) pode ser expressa a partir do produto desses três fatores, conforme a Equação 1 abaixo:

$$Trust = SComp * SIntra * SInter \quad (1)$$

3.3 Abordagens para Quantificar a Confiança

Nesta seção são apresentados os conceitos teóricos necessários à compreensão das três abordagens aplicadas aos três fatores do *Trust* (Equação 1) para mensurar a confiança comportamental.

A primeira, é uma heurística proposta por Carmo *et al.* (2007), já conhecida para quantificar a confiança. A segunda, a Distância de Levenshtein, é originaria da Teoria da Informação e é utilizada atualmente para o reconhecimento de padrões de assinaturas ou ainda no domínio da detecção de intrusão (SCHIMKE *et al.*, 2004; UNTERLEITNER, 2006). A terceira abordagem, proposta do presente trabalho, realiza uma adaptação da Distância de Levenshtein. O algoritmo tradicional para medida de distância foi complementado para contemplar uma propriedade importante observada na heurística.

Outras abordagens, tais como Cadeias de Markov e Distância de Frobenius também foram investigadas como possibilidade de serem úteis para quantificar a confiança e foram descritas no trabalho de Costa e Carmo (2007). Os resultados dos experimentos realizados indicaram a Distância de Levenshtein como a mais promissora para calcular os três fatores do *Trust* e por este motivo foi selecionada para ser investigada profundamente.

3.3.1 Abordagem 1: Heurística de Carmo *et al.*

Na abordagem heurística é utilizado um cálculo de similaridade *S,* definido em função de duas instâncias comportamentais. Para o cálculo de *S*, o procedimento pontua pares de elementos idênticos e usa um cálculo acumulativo para dar um peso maior aos pares idênticos encadeados, ou seja, as subseqüências iguais. Para melhor compreensão, considere como exemplo as seguintes

instâncias: A={a,b,c,d,g,h} e B={a,b,c,f,g,s}. Os pares de elementos idênticos são: {a,b,c} e {g}, e os pares de elementos idênticos e encadeados apenas {a,b,c}.

No caso das instâncias de tamanhos diferentes, um procedimento de homogeneização é aplicado antes do cálculo de similaridade. O procedimento "estica" a menor instância até atingir o comprimento da maior como ilustrado na Figura 6.

Assim $\{z_0, z_1, \ldots, z_{m-1}\}$ é o conjunto de subseqüências iguais entre *x* e *y,* e $\{w_0, w_1, \ldots, w_{n-1}\}$ é o conjunto de subseqüências de tamanhos diferentes. Considerando, por exemplo, o par de instâncias X={a,h,d,e,e,g} e Y={a,b,c,d,e,f,g}, a aplicação da função retorna: 1) o conjunto Z={1,2,1} com 3 subseqüências iguais, duas com um elemento cada {a} e {g} e outra de 2 elementos {d,e} e 2) o conjunto W={2,1} significando uma subseqüência diferente de 2 elementos {b,c} e uma subseqüência diferente {f}.

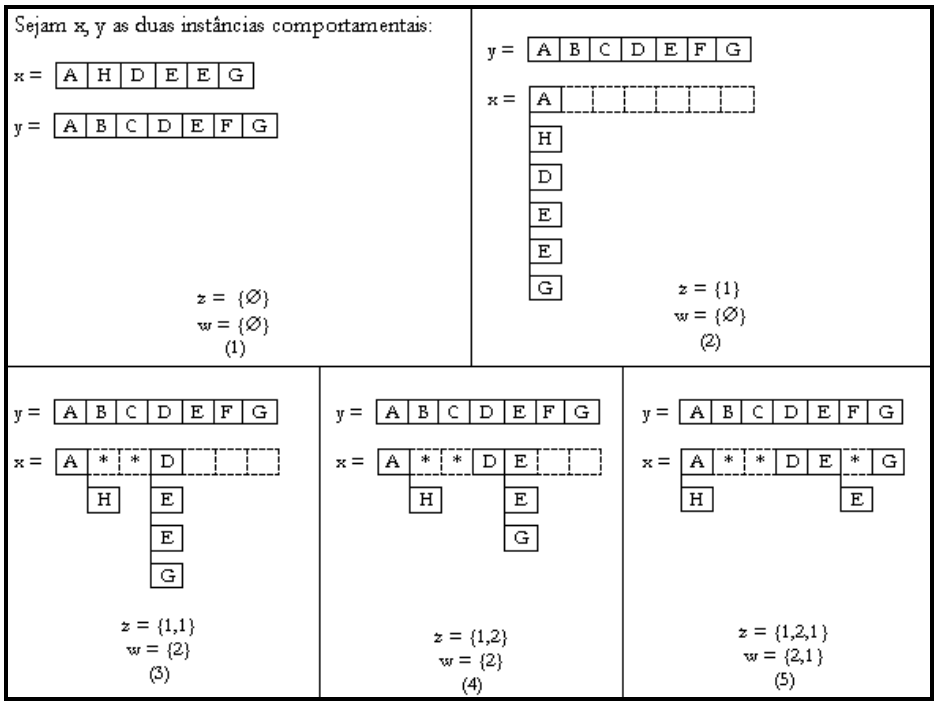

Figura 6: Procedimento de homogeneização

O processo de homogeneização é então expresso pela Equação 2:

$$\tau(X,Y) = \left(\{z_0, z_1, \ldots, z_{m-1}\}, \{w_0, w_1, \ldots, w_{n-1}\}\right) \quad (2)$$

O pseudocódigo da Figura 7, desenvolvido neste trabalho, descreve com detalhes o comportamento da função τ.

```
Função τ (X,Y)
Entrada
     x:      um vetor de caracteres de comprimento n;
     Y:      um vetor de caracteres de comprimento m | m >= n;
Saída
     z, w:   vetores de inteiros;
Variáveis
      aux, pt, pz, pw, px, py :  inteiro;
      achou:                     booleano;
Início
  pz, pw, px, py ← 1;
  se comprimento x ≠ comprimento y
    aux ← 0;
    para px de 1 até comprimento x faça
      se x(px) = y(py)
        se w(pw) > 0
              incrementa pw;
        incrementa z na posição pz
      senão
        se z(pz) > 0
              incrementa pz;
        aux ← py;
        achou ← FALSO;
        para pt de py até (comprimento y – comprimento x - px) faça
            se y(pt) = x(px)
               py <- pt;
               se (w(pw) > 0) e (pt = comprimento y –  comprimento x - px) e (pt ≠ comprimento y)
                  incrementa pw;
               senão
                  w(pw) ← w(pw) + py - aux;
                  incrementa pw;
               achou ← VERDADEIRO;
               incrementa z na posição pz;
               sai do laço de repetição
            se achou ← VERDADEIRO
               w(pw) ← py - aux;
            senão
               incrementa w na posição pw;
       incrementa py;
      senão
          para px de 1 até comprimento x faça
            se x(px) = y(px)
               se w(pw) > 0
                  incrementa pw;
               incrementa z na posição pz;
            senão
               se z(pz) > 0
                  incrementa pz;
             incrementa w na posição pw;
      Retornar z, w
Fim.
```

Figura 7: Pseudocódigo da função τ(x,y)

3.3.2 Abordagem 2: Distância de Levenshtein

A questão de comparar padrões reduz-se a comparar representações de seqüências com vários símbolos, e isto é conseguido usando três operações: inserção, remoção e substituição, que são aplicadas a cada símbolo para transformar uma seqüência na outra. Assim, atribui-se geralmente uma distância para as operações elementares no símbolo, e a distância do padrão é calculada em função destas distâncias.

Para utilizar esse conceito, deve-se escolher entre uma variedade de fórmulas de medidas de distância. No entanto, cada medida de distância foi desenvolvida para ser aplicada em diferentes tipos de dados.

A Distância de Levenshtein, também conhecida como Distância de Edição, é a medida de semelhança ou diferença entre duas cadeias de caracteres. O nome advém do cientista russo Vladimir Levenshtein, que considerou esta distância já em 1965. 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 de Levenshtein. Uma característica interessante deste método é a flexibilidade na adaptação do algoritmo a outras funções.

A Figura 8 abaixo exibe o pseudocódigo, desenvolvido neste trabalho, para a função *DistânciaLevenshtein* que usa duas cadeias, *Str1* de comprimento *tamStr1* e *Str2* de comprimento *tamStr2*, e calcula a distância entre elas.

```
DistânciaLevenshtein (Str1[1..tamStr1], Str2[1..tamStr2])
Entrada
   Str1, Str2:                    cadeias de caracteres;
Saída
   D[tamStr1, tamStr2]:           distância de Levenshtein entre duas cadeias;
Variáveis:
   D[0 .. tamStr1, 0 .. tamStr2]: matriz com tamStr1+1 linhas e tamStr2+1 colunas;
   tamStr1, tamStr2, i, j custo:  inteiro;
Início
   Para i de 0 até tamStr1 faça
         D[i,0] ← i;
   Para j de 0 até tamStr2
         D[0,j] ← j;
   Para i de 1 até tamStr1 faça
         Para j de 1 até tamStr2 faça
               Se Str1[i] = Str2[j] então custo ← 0;
                                    senão custo ← 1;
                D[i,j] ← min(D[i -1, j    ] +1,          //remoção
                             D[i    , j -1] +1,          //inserção
                             D[i -1, j -1] + custo);     //substituição
   Retornar D[tamStr1, tamStr2];
Fim.
```

Figura 8: Pseudocódigo para a função DistânciaLevenshtein

3.3.3 Abordagem 3: Distância de Levenshtein Adaptada

O algoritmo para cálculo da similaridade usado pela Distância de Levenshtein pode ser adaptado para atribuir, assim como na Abordagem 1, uma pontuação a subseqüências encadeadas.

Adotou-se como base o procedimento proposto por Lane e Brodley (1999), para o cálculo de similaridade entre duas seqüências. Este procedimento pontua pares de elementos idênticos e usa um cálculo acumulativo para dar um peso maior aos pares idênticos encadeados (subseqüências iguais). A similaridade entre duas seqüências de tamanho *n*, $A = (a_0, a_1, \ldots, a_{n-1})$ e $B = (b_0, b_1, \ldots, b_{n-1})$ é definida pelo par de funções:

$$\omega(a_i, b_i) = \begin{cases} 0 & \text{se } a_i \neq b_i \vee i < 0 \\ 1 + \omega(a_{i-1}, b_{i-1}) & \text{se } a_i = b_i \end{cases} \quad (6)$$

$$\Omega(A,B) = \sum_{i=0}^{n-1} \omega(a_i, b_i) \quad (7)$$

Duas seqüências são comparadas elemento por elemento. A área da curva, representada na Figura 9, exibe o peso contribuído para cada correspondência de similaridade.

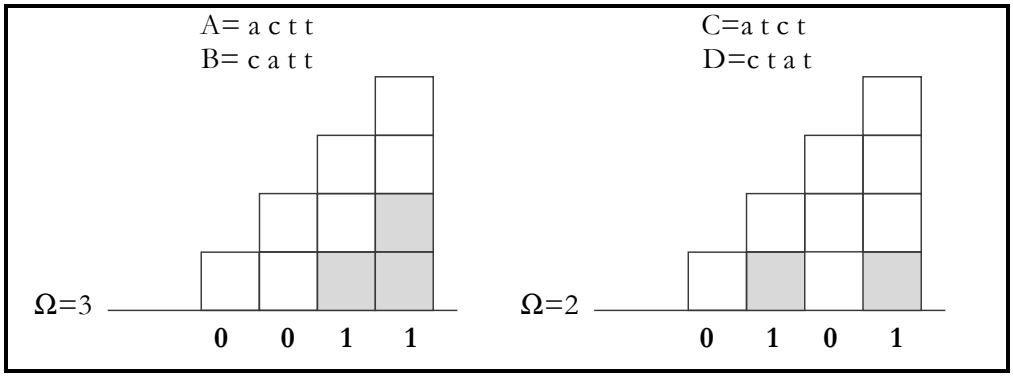

Figura 9: Exemplo de cálculo de similaridade de seqüências

Um procedimento de *backtrack* na matriz gerada pelo algoritmo de distância, conforme ilustrado na matriz da Figura 10, reproduz a seqüência de operações de edição realizadas para transformar a seqüência *A* na seqüência *B*. Pode-se obter o percurso de edição usando ponteiros.

Nos valores {0, 1, 2, 2, 2} pode-se observar que ocorreu uma operação de *0* para *1,* e outra de *1* para *2*.

		A	C	T	T
	0	1	2	3	4
C	1	**1**	1	2	3
A	2	1	**2**	2	3
T	3	2	2	**2**	2
T	4	3	3	2	**2**

Figura 10: Exemplo de matriz de similaridade de seqüências

Para obter as posições de correspondências de similaridade é feito o complemento da diferença entre os valores, desta forma o resultado é {0, 0, 1, 1}. O pseudocódigo da Figura 11, desenvolvido neste trabalho, mostra a formação da matriz e os novos procedimentos adotados.

```
DistânciaLevenshteinAdaptada (Str1[1..tamStr1], Str2[1..tamStr2])
Entrada
   Str1, Str2:                    cadeias de caracteres;
Saída
   Ω:                             valor da similaridade entre duas cadeias;
Variáveis:
   D[0 .. tamStr1, 0 .. tamStr2]: matriz com tamStr1+1 linhas e tamStr2+1 colunas;
   tamStr1, tamStr2, tamCaminho,
   k, i, j custo:                 inteiro;
   Caminho[1..tamCaminho]:        vetor dinâmico com os custos das operações de edição
   Posição[k]:                    vetor com as diferenças entre os custos das operações
                                  de edição
Início
   Para i de 0 até tamStr1 faça
         D[i,0] ← i;
   Para j de 0 até tamStr2
         D[0,j] ← j;
   Para i de 1 até tamStr1 faça
         Para j de 1 até tamStr2 faça
               Se Str1[i] = Str2[j] então custo ← 0;
                                    senão custo ← 1;
               D[i,j] ← min(D[i -1, j    ] +1,          //remoção
                            D[i    , j -1] +1,          //inserção
                            D[i -1, j -1] + custo);     //substituição

   Caminho ← backtrack (D);
   Para k de 1 até (tamCaminho –1) faça
         Posição[k] ← 1 - (Caminho[k+1] – Caminho[k]);

   Retornar Ω (Posição);
Fim.
```

Figura 11: Pseudocódigo para a função DistânciaLevenshteinAdaptada

A adaptação exige uma nova forma de interpretação dos resultados a serem gerados a partir do algoritmo. Para melhor compreensão desta observação, os gráficos da Figura 12 ilustram essa diferença. Considere a similaridade entre duas seqüências idênticas qualquer, o eixo *x* representa cada elemento de uma das seqüências e *y* as pontuações atribuídas, por cada uma das abordagens, a cada elemento das seqüências. Ao contrário da abordagem tradicional, onde os valores das correspondências entre duas seqüências têm característica linear, a abordagem adaptada tem característica exponencial como representado pela reta e curva, respectivamente.

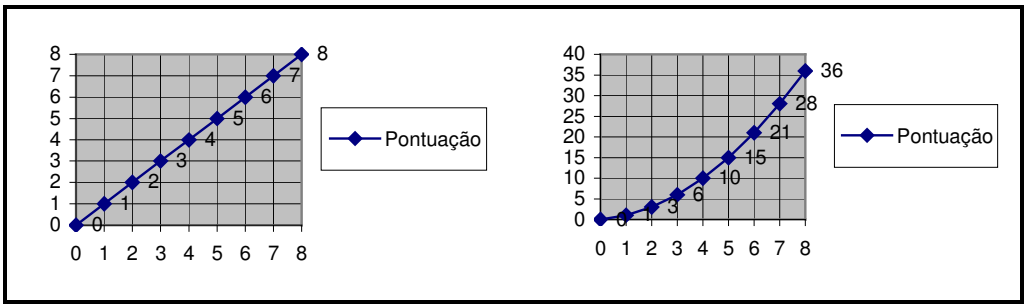

Figura 12: Gráficos Levenshtein x Levenshtein adaptado

3.4 Avaliação Experimental

A avaliação comportamental depende de fatores como: as informações do comportamento do usuário e a atribuição de uma medida de confiança a esse comportamento.

Em Carmo e Reis (2006), a avaliação foi realizada em uma base de instâncias comportamentais simuladas. Primeiramente foi definido um macro-estado totalmente interconectado. Em seguida, as assinaturas foram geradas considerando dois diferentes cenários: sem interseção e com interseção de instâncias. O objetivo foi gerar um conjunto de amostras representativas de possíveis comportamentos reais. Os valores de *Trust* foram então calculados considerando as amostras de cada cenário e como resultado verificou-se que um comportamento bem formado, assim como, sua dissimilaridade com outras assinaturas, são características fundamentais para o sucesso da avaliação.

Em um segundo trabalho, Carmo *et al.* (2007) avaliou um comportamento por meio das informações disponíveis no *log* de um servidor. Essa abordagem requereu a seleção e adaptação do *log*. O primeiro *web site* investigado (Biblioteca virtual da UFRJ) não se mostrou apropriado por não mostrar o IP real das máquinas que fizeram os pedidos de HTTP. O segundo, (Escola de Arquitetura e Urbanismo da UFRJ), embora bem estruturado para a adaptação macro-estatal, possuía pouca diversidade de trilhas de navegação para gerar assinaturas diferentes. O último *web*

site (Sistemas de Bibliotecas e Informações da UFRJ) não mostrou nenhum dos problemas acima mencionados e por isso foi escolhido para ser adaptado para a avaliação. A adaptação incluiu a definição de conjecturas, para identificar e delimitar uma instância comportamental.

A maioria das propostas semelhantes faz uso de análises empíricas, realizadas por meio de exemplos concretos de aplicações como ambientes de teste (PUSARA e BRODLEY, 2004; LANE, 2000). O grande problema desse tipo de abordagem é a seleção de um ambiente extremamente impróprio, gerando uma possível falsa avaliação negativa, ou, ao contrário, extremamente apropriado, o que também levaria a uma conclusão errônea.

Esta seção apresenta a proposta deste trabalho para o problema de avaliar o comportamento de um usuário em uma base de dados específica.

3.4.1 *Web Site* Experimental

Para fins da avaliação comportamental foi desenvolvido uma aplicação de um *web site* experimental para possibilitar a captura dos comportamentos e permitir a extração de requisitos para a implementação de *web sites*.

Segundo Nielsen (2000) pode-se controlar por onde o usuário vai navegar, contudo na *Web*, os usuários controlam fundamentalmente a navegação pelas páginas. É possível forçar os usuários a seguirem por caminhos definidos e evitar que estabeleçam *links* com determinadas páginas, mas os *web sites* que usam esta estratégia são considerados rígidos. Já os que são projetados visando a liberdade de movimento são vistos como uma melhor opção. Assim, o ambiente foi estruturado inicialmente no formato de um grafo fortemente conexo, permitindo que a partir de qualquer página se chegasse a qualquer outra.

Os comportamentos ocorrem em um único macro-estado para acesso e considerando que o interesse do usuário em visitar determinada página é um indicativo de seu comportamento, existe ainda a necessidade de se definir onde começa e termina uma instância comportamental.

Ainda segundo Nielsen (2000), 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 a página de conclusão da tarefa.

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 para a captura dos comportamentos de navegação dos usuários.

Neste *web site*, onde diferentes usuários podem navegar, a coleta dos dados é feita através de agentes de software que capturam somente as informações potencialmente úteis para a avaliação, tais como: IP, data e hora de acesso e a trilha de navegação.

O monitoramento da abertura das páginas, pelos agentes, possibilita a identificação das páginas mesmo 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 clique a clique (BRAINERD *et al.*, 2001; HU e ZHONG, 2005).

As informações da implementação do *web site* experimental são detalhadas no Capítulo 4.

3.4.2 Aquisição da Assinatura Comportamental

O problema da coleta de assinatura 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.

Utilizou-se um modelo adaptado de Aprendizado Baseado em Instâncias. Neste, um conceito é representado implicitamente por um conjunto de instâncias que exemplificam o conceito (dicionário de instâncias). No esquema típico uma instância desconhecida é classificada por meio de um algoritmo KNN (K-Nearest Neighbor). Nesse caso, necessariamente as representações das instâncias devem ser numéricas e de mesmo tamanho.

Na abordagem adaptada a classificação KNN foi substituída pela Distância de Levenshtein que permite fazer o reconhecimento de padrões simbólicos com comprimentos diferentes. Também, cada instância comportamental é diretamente classificada de acordo com o usuário gerador.

Esta abordagem requer o armazenamento de um conjunto completo de instâncias comportamentais que, em uma aplicação de um *web site,* são interpretadas como seqüências de páginas visitadas pelo usuário durante a sua interação com o ambiente (trilhas de navegação).

A assinatura por sua vez, pode ser interpretada como um conjunto de trilhas de navegação associada a um mesmo indivíduo. A assinatura, bem como os dados temporais e de localização, constituem o histórico de comportamento individual.

A Tabela 2 abaixo exibe o histórico comportamental de um usuário gerado a partir de sua interação com o *web site* experimental. Os dados registrados tratam-se dos códigos das tarefas realizadas, o código de registro do usuário, o endereço IP do computador, a data e hora dos acessos ao ambiente e a trilha de navegação realizada.

Tabela 2: Exemplo de histórico comportamental do usuário 24

Codigo	CodUsuario	IP	Data	Hora	Trilha
90	24	200.150.38.198	2006-12-04	10:07:00	Principal#RockNacional#Detalhes#ConcluirVenda#
129	24	200.150.38.198	2006-12-06	10:01:00	Principal#PopNacional#RockNacional#Detalhes#Conclu...
121	24	200.165.79.37	2006-12-05	18:45:00	Principal#RockNacional#Detalhes#ConcluirVenda#
156	24	200.150.38.198	2006-12-07	13:18:00	Principal#RockNacional#Detalhes#ConcluirVenda#
173	24	201.58.109.144	2006-12-08	15:27:00	Principal#RockInternacional#Detalhes#ConcluirVenda...
256	24	200.150.34.232	2007-02-05	19:26:00	Principal#RockNacional#Detalhes#ConcluirVenda#
257	24	200.150.34.232	2007-02-05	19:27:00	Principal#RockNacional#Detalhes#ConcluirVenda#
258	24	200.150.34.232	2007-02-05	19:27:00	Principal#RockInternacional#Detalhes#ConcluirVenda...
255	24	200.150.34.232	2007-02-05	19:26:00	Principal#RockNacional#Detalhes#ConcluirVenda#
254	24	200.150.34.232	2007-02-05	19:25:00	Principal#Lançamentos#Detalhes#ConcluirVenda#

3.4.3 Aplicação da Abordagem 1

Com a similaridade *S,* os três fatores independentes considerados no *Trust* são calculados. Para aplicação da Abordagem 1 aos fatores de confiança, os seguintes conceitos são necessários:

- ic_{atual}^{M} =denota a instância comportamental atual;
- ic_{i}^{M} = denota a *i*-ésima instância comportamental da assinatura;
- ac^{M} =denota a assinatura comportamental do macro-estado *M*;
- U^{M} =denota o conjunto total de assinaturas associadas a *M*.

SComp representa a similaridade entre a instância atual coletada e o conjunto de instâncias que compõem a assinatura. Para o cálculo de *SComp* (Equação 8) aplica-se a função *S* entre a instância comportamental atual e cada uma das instâncias que compõem a assinatura, retendo o valor máximo obtido:

$$SComp^{M} = \max\left\{S(ic_{atual}^{M}, ic_{i}^{M}), \forall ic_{i}^{M} \in ac^{M}\right\} \quad (8)$$

SIntra representa o quanto as instâncias já pertencentes à assinatura são repetidas, ou levemente diferentes. Para *SIntra* (Equação 9) realiza-se a média entre todos os valores resultantes da aplicação da função *S* entre todas instâncias duas a duas de uma assinatura:

$$SIntra^{M} = \frac{\sum_{\forall ic_n^M \in ac^M} \sum_{\forall ic_m^M \in ac^M} S(ic_n^M, ic_m^M)}{C_{|ac^M|,2}} \quad (9)$$

SInter representa a dissimilaridade da assinatura de um usuário em relação as demais assinaturas de outros usuários. *SInter* é expressa pelo seguinte par de Equações 10 e 11:

$$\Phi(ac^M, ac_i{}^M) = \frac{\sum_{\forall ic_n^M \in ac^M} \max\{S(ic_n^M, ic_m^M), \forall ic_m^M \in ac_i{}^M\}}{|ac^M|} \quad (10)$$

$$SInter(ac^M) = 1 - \max\left\{\Phi(ac^M, ac_i^M), \forall ac_i^M \in U^M - \{ac^M\}\right\} \quad (11)$$

A Equação 10 realiza a comparação de uma instância da assinatura do usuário corrente e todas as instâncias da assinatura de outro usuário, retendo o valor máximo. Este procedimento é realizado para todas as instâncias da assinatura do usuário corrente. O número de valores obtidos é o número de instâncias da assinatura do usuário. Em seguida é calculada a média desses valores.

Os procedimentos acima citados são realizados para todas as assinaturas dos outros usuários e a Equação 11 retém o valor máximo das médias obtidas. Como *SInter* mede a dissimilaridade entre as assinaturas, este o valor deve ser subtraído por *1*.

3.4.4 Aplicação da Abordagem 2

O conceito de Distância de Levenshtein é aplicado aos três fatores do *Trust*. Para compreensão do processo, faz-se necessário a exposição das seguintes definições preliminares:

- u = usuário atual $u = 1,...,|H|$;
- i, j = índices para as instâncias onde $i, j = 1,\cdots,|ac^u| \ \forall \ ic_i^u \in ac^u$;
- ic_i^u =instância comportamental *i* do usuário atual;

- ac^u = assinatura comportamental do usuário atual ($ac^u \in H$);
- H = conjunto de assinaturas de todos os usuários;
- k = qualquer usuário onde $k = 1, \cdots, |H|$;
- *m, n, n_1, n_2, m_1, m_2* = índices que variam de acordo com as Equações 12, 13 e 14;
- $dist(ic_i^u, ic_j^k)$ = distância de Levenshtein entre a instância comportamental *i,* do usuário atual u e a instância comportamental *j,* do usuário k *;*
- $$dist_norm(ic_i^u, ic_j^k) = \begin{cases} \dfrac{dist(ic_i^u, ic_j^k)}{|ic_i^u|} & \text{se } |ic_i^u| \geq |ic_j^k| \\ \dfrac{dist(ic_i^u, ic_j^k)}{|ic_j^k|} & \text{se } |ic_i^u| < |ic_j^k| \end{cases}$$

  distância de Levenshtein normalizada entre a instância comportamental *i,* do usuário atual u e a instância comportamental *j*, do usuário k *;*

- $$\overline{dist_k^u} = \frac{\sum_{i=n_1}^{n_2} \sum_{j=m_1}^{m_2} dist_norm(ic_i^u, ic_j^k)}{m.n}$$

  distância média entre a assinatura comportamental do usuário atual u e a assinatura comportamental do usuário k .

O cálculo das distâncias entre as instâncias da assinatura do usuário e a instância mais atual (*SComp*), realiza basicamente os seguintes procedimentos:

- aplica-se o algoritmo para o cálculo da Distância de Levenshtein entre cada instância que compõe a assinatura do usuário e a instância atual;
- realiza-se a normalização pelo valor de comprimento da maior instância;

- calcula-se a média das distâncias: (soma das distâncias normalizadas)/(número de distâncias da assinatura).

A distância é um valor entre [*0,1*] onde *0* corresponde a similaridade máxima e *1* o contrário. Assim, quanto menor a distância, maior a similaridade entre as instâncias. Logo, *SComp* é expressa pela seguinte Equação 12:

$$SComp = 1 - \overline{dist_k^u} \qquad n_1 = 1 \qquad n_2 = \left|ac^u\right| \qquad m_1 = 1 \qquad m_2 = 1 \qquad k = u \tag{12}$$

Para o cálculo das distâncias, entre as instâncias da assinatura do usuário (*SIntra*), os seguintes procedimentos são realizados:

- aplica-se o algoritmo para o cálculo da Distância de Levenshtein entre todas as instâncias que compõem a assinatura do usuário;
- realiza-se a normalização pelo valor de comprimento da maior instância;
- calcula-se a média das distâncias: (soma das distâncias normalizadas)/(número de distâncias da assinatura).

Também neste caso, quanto menor a distância maior, a similaridade entre as instâncias. Logo, a Equação 13 é a seguinte:

$$SIntra = 1 - \overline{dist_k^u} \qquad n_1 = 1 \qquad n_2 = \left|ac^u\right| \qquad m_1 = i + 1 \qquad m_2 = \left|ac^u\right| \qquad k = u \tag{13}$$

Para o cálculo das distâncias entre a assinatura do usuário e as demais assinaturas (*SInter*), adotou-se os seguintes procedimentos:

- aplica-se o algoritmo para o cálculo da Distância de Levenshtein entre todas as instâncias que compõem a assinatura do usuário em questão e as instâncias das assinaturas dos demais usuários;
- realiza-se a normalização pelo valor de comprimento da maior instância entre duas assinaturas;
- calcula-se a média das distâncias: (soma das distâncias normalizadas)/(número de distâncias da assinatura);
- calcula-se a média das médias das distâncias: (soma das distâncias normalizadas)/(número de distâncias da assinatura -1).

Quanto menor a distância, menor a diferença entre as assinaturas. A Equação 14, a seguir, expressa *SInter*:

$$SInter = \frac{\sum_{k=1}^{|H|-1} dist_k^u}{|H|-1} \qquad \begin{matrix} n_1 = 1 & n_2 = \left|ac^u\right| \\ m_1 = 1 & m_2 = \left|ac^k\right| \\ k \neq u & \end{matrix} \tag{14}$$

3.4.5 Aplicação da Abordagem 3

Basicamente, a adaptação realiza duas modificações na forma tradicional de cálculo de similaridade: i) pontuação por Ω e, ii) normalização por $\Omega_{max..}$ Aplicadas aos três fatores do *Trust*, altera as seguintes definições apresentadas na Subseção 3.4.4:

- $dist(ic_i^u, ic_j^k)$ = valor de Ω entre a instância comportamental *i,* do usuário atual u e a instância comportamental *j,* do usuário k ;*m*

- $dist_norm(ic_i^u, ic_j^k) = \dfrac{dist(ic_i^u, ic_j^k)}{\Omega_{max}}$

valor de Ω normalizado entre a instância comportamental *i,* do usuário atual u e a instância comportamental *j*, do usuário k. Ω_{max} é o maior valor de Ω obtido considerando a similaridade total entre duas instâncias.

- $\overline{dist_k^u} = \dfrac{\sum_{i=n_1}^{n_2} \sum_{j=m_1}^{m_2} dist_norm(ic_i^u, ic_j^k)}{\Omega_{max}}$

Ω médio entre a assinatura comportamental do usuário atual u e a assinatura comportamental do usuário k.

Para o cálculo de *SComp* são realizados basicamente os seguintes procedimentos:

- aplica-se o algoritmo de cálculo da Distância de Levenshtein adaptado entre cada instância que compõe a assinatura do usuário e a instância atual;
- realiza-se a normalização ($\Omega_{max} = \dfrac{\Omega}{\Omega_{max}}$) dos valores obtidos anteriormente;
- calcula-se a média: $\dfrac{soma\ dos\ \Omega\ normalizados}{número\ de\ \Omega\ obtidos}$.

O Ω é um valor entre [*0,1*] onde *1* corresponde à similaridade máxima e *0* o contrário. Assim, quanto maior o valor de Ω , maior a similaridade entre as instâncias. Logo, *SComp* é expressa pela seguinte Equação 15:

$$SComp = \overline{dist_k^u} \qquad n_1 = 1 \qquad n_2 = \left|ac^u\right| \qquad m_1 = 1 \qquad m_2 = 1 \qquad k = u \tag{15}$$

Para o cálculo de *SIntra*, os seguintes procedimentos são realizados:

- aplica-se o algoritmo de Distância de Levenshtein adaptado entre todas as instâncias que compõem a assinatura do usuário;
- realiza-se a normalização ($\Omega_{\max} = \frac{\Omega}{\Omega_{\max}}$) dos valores obtidos anteriormente;
- calcula-se a média: $\frac{soma\ dos\ \Omega\ normalizados}{número\ de\ \Omega\ obtidos}$.

Também neste caso, quanto maior o valor de Ω, maior a similaridade entre as instâncias. Logo, a Equação 16 é a seguinte:

$$SIntra = \overline{dist_k^u} \qquad n_1 = 1 \qquad n_2 = \left|ac^u\right| \qquad m_1 = i+1 \qquad m_2 = \left|ac^u\right| \qquad k = u \tag{16}$$

Para o cálculo de *SInter*, adota-se os seguintes procedimentos:

- aplica-se o algoritmo para o cálculo da Distância de Levenshtein adaptado entre todas as instâncias que compõem a assinatura do usuário em questão e as instâncias das assinaturas dos demais usuários;
- realiza-se a normalização pelo valor de comprimento da maior instância entre duas assinaturas;
- calcula-se a média: $\frac{soma\ dos\ \Omega\ normalizados}{quantidade\ de\ \Omega}$

- calcula-se a média das médias dos **Ω**: $\frac{soma\ dos\ \Omega\ normalizados}{(quantidade\ de\ \Omega) - 1}$

Quanto menor o valor de **Ω**, mais dissimilares são as assinaturas. A Equação 17, a seguir, expressa *SInter*:

$$SInter = 1 - \frac{\sum_{k=1}^{|H|-1} dist_k^u}{|H|-1} \quad \begin{matrix} n_1 = 1 & n_2 = \left|ac^u\right| \\ m_1 = 1 & m_2 = \left|ac^k\right| \\ k \neq u & \end{matrix} \tag{17}$$

3.4.6 Cálculo do Limiar de Confiança

A quantificação da confiança torna possível o estabelecimento de um patamar de aceitação mínimo para cada usuário, assim procura-se determinar o valor a partir do qual um usuário pode ser considerado confiável por uma aplicação.

Na estratégia usada os valores para o limiar estão compreendidos entre [*0,1*] com variação de *0,001* (*0:0,001:1*). Para cada variação, é contabilizado o número de falsos-negativos (FN), falsos-positivos (FP), verdadeiros-positivos (VP) e verdadeiros-negativos (VN), para cada usuário em relação aos demais.

Acontecem FN quando um usuário legítimo é reconhecido como ilegítimo. Já os FP acontecem de forma inversa, o usuário é ilegítimo, mas é reconhecido como legítimo. VP acontecem quando um usuário legítimo é de fato reconhecido como legítimo. Já os FP acontecem de forma inversa.

Um número maior de ocorrências de FP é uma questão crítica, no entanto se o número de FN for muito alto, a utilização da avaliação comportamental perde seu objetivo, pois a não

identificação de um comportamento como sendo do usuário implica na realização da uma etapa "redundante" de troca de credenciais. Sendo assim, o problema está em minimizar FN + FP.

Para tanto, calcula-se os erros do Tipo I e do Tipo II, que são ferramentas estatísticas empregadas na realização de testes de significância. Um erro do Tipo I = $\left(\frac{FN}{FN+VP}\right)$ e acontece quando:

FN= é, digo que não é

VP= é, digo que é

Enquanto um erro do Tipo II = $\left(\frac{FP}{FP+VN}\right)$ e acontece quando:

FP = não é, digo que é

VN= não é, digo que não é

Em outras palavras, o erro do Tipo I mede a porcentagem de falsos negativos e o erro do Tipo II mede a porcentagem de falsos positivos. Pela própria definição, percebe-se que estas medidas se opõem, ou seja, quando o valor de uma aumenta, a outra diminui e vice-versa.

O melhor limiar é um valor para o qual a soma do erro do Tipo I e erro do Tipo II é mínima. Assim, a função que avalia o melhor limiar para cada usuário, é a seguinte Equação 18:

$$\min(\text{TipoI} + \text{TipoII}) \tag{18}$$

CAPÍTULO 4

PROJETO E IMPLEMENTAÇÃO DO WEB SITE

Neste capítulo são apresentadas as principais características do desenvolvimento do *web site* como experimento. Descreve a arquitetura, o mecanismo para coletar os comportamentos através de agentes de software e expõe os aspectos envolvidos no projeto de construção das páginas.

4.1 Descrição do Ambiente

O *web site* trata-se de uma loja virtual, cuja implementação foi baseada em padrões de software aberto e utilizou-se as seguintes tecnologias:

- Linguagem de programação PHP (Hypertext Preprocessor), a qual oferece meios para construir *web sites* movidos por uma base de dados;
- Linguagem *JavaScript* para permitir a coleta de dados através de agentes de software;
- Servidor *W*eb Apache, versão 1.3.23, distribuída pela *Apache Software Foundation*. Foi utilizado para registrar as chamadas HTTP (HyperText Transfer Protocol);
- Sistema de gerenciamento de banco de dados MySQL.

4.2 Integração com o Servidor

A arquitetura apresentada na Figura 13 representa a forma de interação entre os servidores, o repositório de informações, o usuário e os agentes utilizados. Com base nessa figura é possível identificar o caminho traçado quando é realizada uma requisição por um usuário.

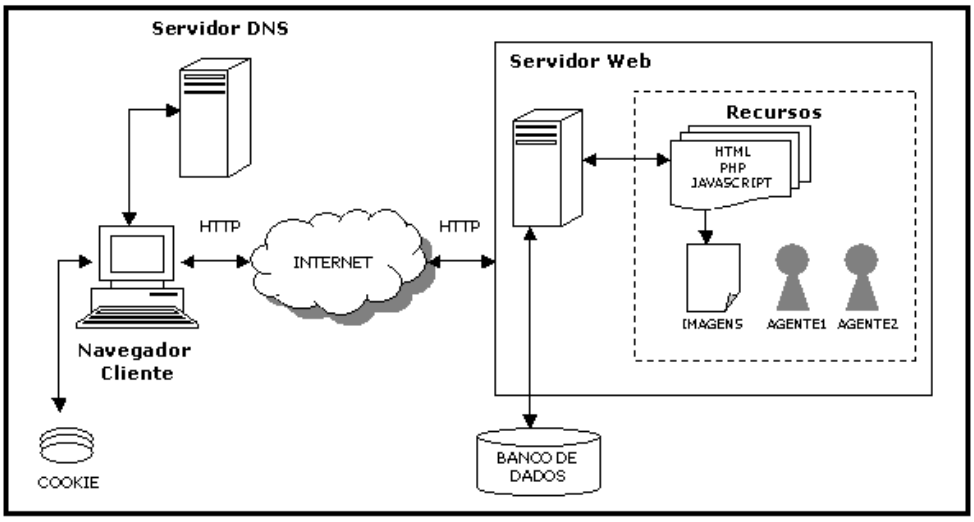

Figura 13: Etapas de requisição de recursos Web

Os componentes presentes na arquitetura, como o *cookie* de sessão e os agentes de *software*, são peças importantes para o tratamento de captura e armazenamento das informações comportamentais, como a trilha de navegação, dados temporais e de localização.

Esta arquitetura segue o modelo cliente-servidor, onde o cliente envia uma requisição HTTP ao servidor, que retorna as páginas HTML (HyperText Markup Language) contendo o código PHP e os agentes como resposta.

Em síntese, o usuário solicita acesso ao *web site* por meio do navegador. O servidor DNS envia o endereço do servidor *Web* para que o cliente possa solicitar os recursos desejados. Os agentes migram, para a máquina cliente, juntamente com alguma requisição. Eles possuem o conhecimento (código) e dados parciais para realização de suas tarefas em momento oportuno.

São de responsabilidade dos agentes, a execução do código responsável pela criação do *cookie* e a coleta e armazenamento das informações em repositório. O repositório constitui-se no banco de dados composto por um conjunto de tabelas, as quais possuirão as informações obtidas através do monitoramento de abertura das páginas.

4.3 Mecanismo de Captura dos Comportamentos

Os usuários geralmente despendem um tempo razoável fornecendo as informações requeridas por uma aplicação. Para minimizar esse problema, agentes de software podem ser usados para coletar as informações disponíveis de uso do ambiente, sem a participação direta do usuário. Desta forma, o usuário não participa do processo diretamente, fornecendo os dados que serão usados para compor seu histórico comportamental. O diagrama de seqüência da Figura 14 exibe a participação do usuário e dos agentes na formação do histórico.

Inicialmente o usuário realiza o seu cadastro na loja virtual, em seguida, se autentica com os dados informados no cadastro. Ao navegar pela loja um agente do tipo 1 vai coletando as páginas visitadas. Ao concluir a compra as informações capturadas são submetidas ao banco de dados por um agente do tipo 2.

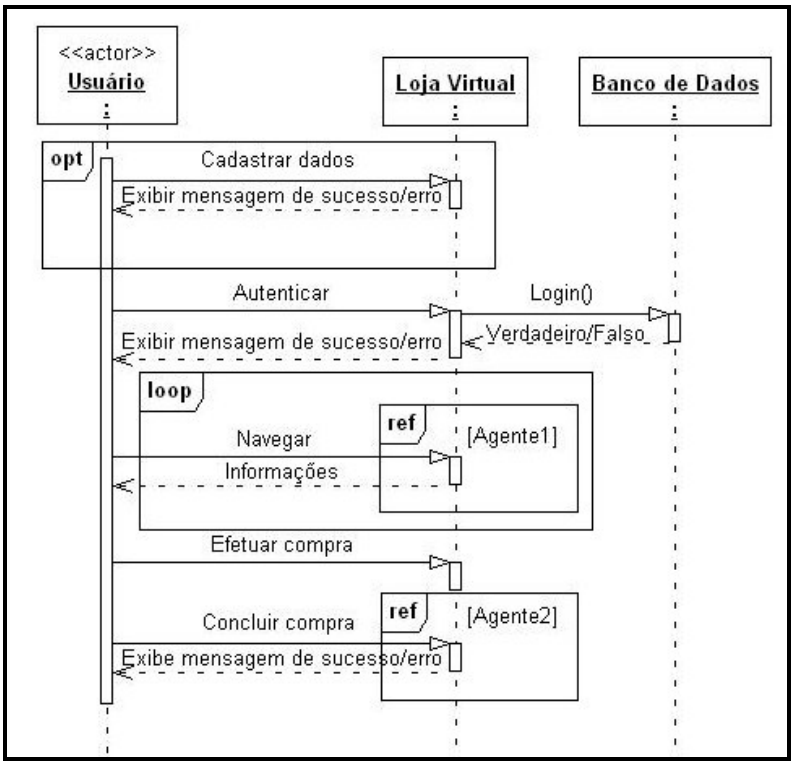

Figura 14: Primeiro diagrama de seqüência do usuário

4.3.1 Agentes de Software

O significado da palavra "agente" pode variar conforme o contexto em que se insira. Segundo alguns autores como Wooldridge e Jennings (1999) não existe um consenso no emprego do termo agente, por se aplicar a diversas áreas de investigação e desenvolvimento.

A definição utilizada neste trabalho é a de Smith *et al* (*apud* FRANKLIN e GRAESSER, 1996), que se refere a um agente como sendo uma entidade dedicada a um propósito específico que possui algumas características próprias que a diferencia de outros tipos de entidades tradicionais.

O conjunto de características associada ao agente de software é um fator importante para indicar o seu grau de complexidade. Quanto maior for o número de características associadas a um agente, maior será a variedade de comportamentos. Para o desenvolvimento dos agentes deste trabalho são destacadas algumas características, tais como: autonomia, mobilidade e reatividade (ETZIONI e WELD, 1995; CAZELLA, 1997).

A característica autonomia, define que um agente deve ser capaz de agir sem a intervenção direta do usuário ou de outro agente, sendo responsável pela decisão de qual ação executar.

A reatividade é a capacidade que um agente tem de perceber seu ambiente e responder a cada mudança ocorrida. Todas as informações referentes ao seu comportamento se encontram no ambiente onde está inserido. Os agentes reativos não apresentam memória de seus atos, não planejam ações futuras e não se comunicam com outros agentes, mas cada agente pode perceber a ação realizada por outro agente através da modificação do ambiente em que está inserido.

A característica de mobilidade é a capacidade que um agente tem de transportar-se de uma máquina especifica para outra máquina, e através de diferentes sistemas de arquiteturas, e plataformas.

A tecnologia de agentes vem sendo aplicada principalmente em tarefas que necessitavam da intervenção do usuário. Essa tecnologia abrange aplicações de gerência de sistemas de redes, interfaces, correio eletrônico entre outras.

4.3.2 Coleta e Armazenamento dos Dados de Navegação

Existem dois mecanismos de captura de informações através de agentes de software que colaboram para a realização de suas ações. O primeiro agente vai coletar as trilhas de navegação, direto do computador do usuário e armazená-las em um c*ookie* de sessão que é criado no momento que a página é carregada no cliente, caso ele não exista. Um *cookie* de sessão é armazenado somente para a sessão de navegação atual, sendo excluído do computador quando o navegador é fechado.

O primeiro tipo de agente trabalha com as páginas selecionadas para serem capturadas durante uma iteração. Este agente apresenta reações no contexto em que está inserido, iniciando uma série única de ações próprias para atingir sua meta.

A Figura 15 do diagrama ilustra as ações do primeiro agente, bem como a ordem de execução das mesmas ao longo do tempo. O agente cria um *cookie* caso não exista, em seguida, coleta a página visitada pelo usuário e armazena no *cookie.* Este procedimento é realizado a cada página visitada pelo usuário.

O segundo agente é enviado com a requisição para abertura da página selecionada como conclusão da tarefa. Tem a função de capturar os dados temporais e de localização, juntamente com as trilhas armazenadas no *cookie* para submetê-los ao banco de dados, em próxima requisição, já esperada para confirmação de conclusão da tarefa.

Este agente aproveita as requisições necessárias a navegação para execução de suas ações, dispensando requisições extras para submeter as informações capturadas ao banco de dados, ao final, destrói o *cookie*. A Figura 16 ilustra o diagrama de seqüência das suas ações.

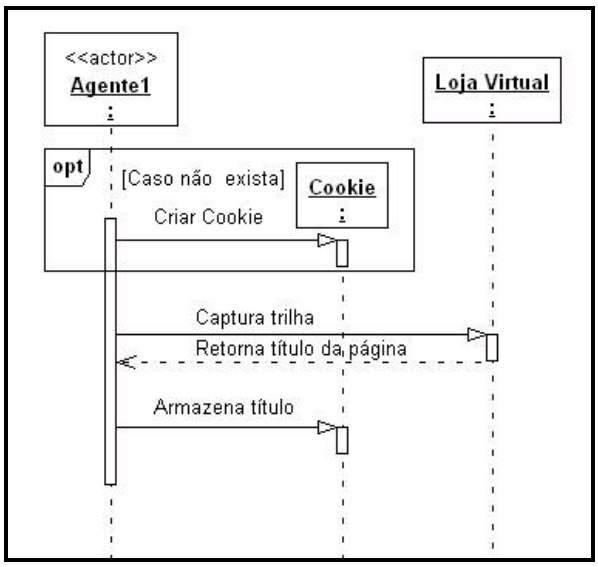

Figura 15: Diagrama de seqüência do Agente 1

As informações referentes ao comportamento de cada agente estão inseridas no ambiente e suas reações dependem da percepção do ambiente. Cada agente tem seu objetivo e não há memorização das suas ações.

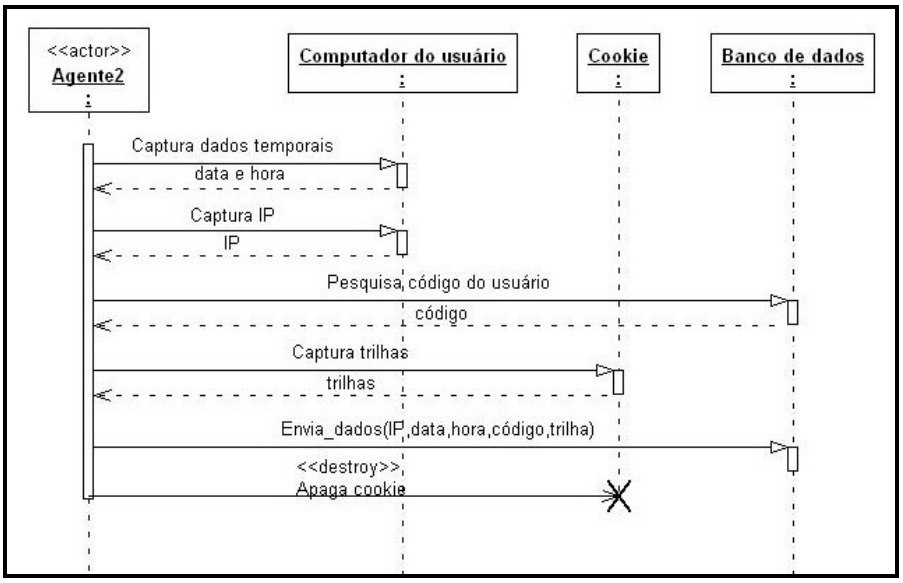

Figura 16: Diagrama de seqüência do Agente 2

4.4 Construção das Páginas

Uma página de um *web site* tanto define as estratégias para a realização de uma tarefa, como conduz, orienta, recepciona, alerta e responde ao usuário durante as interações.

O conjunto de critérios ergonômicos de usabilidade definidos por Bastien e Scapin (1993) foi observado durante o projeto de desenvolvimento para atribuir qualidades às páginas de forma que estas satisfizessem os usuários quando da sua utilização, pois, os usuários jamais chegariam perto das páginas destinadas a execução de uma tarefa, a menos que estas fossem implementadas de acordo com suas necessidades e que contivessem um esquema de navegação que os permitisse descobrir o que desejam.

O grau de usabilidade das páginas foi avaliado com o emprego da técnica de Análise da Tarefa x Usuário, onde são realizados o levantamento das atividades desempenhadas pelo usuário e a construção de um protótipo. O protótipo é validado após sucessivas apresentações e modificações, procurando adequá-lo o mais próximo das atividades do usuário (ARAGÃO, 2001).

4.5 Restrições do *Web Site*

A implementação do ambiente desconsidera o conceito de personalização onde um *web site* é personalizado a cada novo acesso, com as preferências de compra do usuário. Assim, optou-se por reconstruir o trajeto por onde o usuário de fato desejou navegar, diminuindo a possibilidade de uma navegação induzida. Um outro fator limitante é a utilização de *cookies* que depende de consentimento do usuário. Para que um comportamento seja coletado, o navegador do usuário deve estar habilitado para gravação dos *cookies*.

CAPÍTULO 5

ANÁLISE E APRESENTAÇÃO DOS RESULTADOS

Em concordância com a avaliação comportamental proposta no Capítulo 3, é feita uma análise comparativa das abordagens e, em seguida, os resultados são apresentados e discutidos caracterizando cada uma delas, ressaltando suas contribuições e limitações. Antes, o processo de composição da base para os experimentos é exposto. Certas diretrizes podem ser indicadas para servir como parâmetro e viabilizar a criação de *web sites* baseados em confiança. Como resultado dessa afirmação, foram obtidas oito diretrizes, também expostas neste capítulo.

5.1 Descrição dos Testes

Dois testes *on-line* com o *web site* foram realizados em períodos diferentes por um grupo de pessoas que receberam as instruções de acesso à loja virtual. O Apêndice A apresenta o planejamento para execução do teste, incluindo a abrangência, os procedimentos a serem adotados, recursos e cronograma das atividades do teste. Identifica os itens, as funcionalidades a serem testadas e as tarefas a serem realizadas.

Na primeira fase do teste foi solicitada, aos participantes, a tarefa de efetuar compras diárias de qualquer produto de sua preferência. O objetivo 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 foi coletar uma maior quantidade de dados para

a criação do histórico individual e verificar a eficácia das alterações realizadas. O Apêndice B e o Apêndice C exibem as imagens das páginas do *web site* implementado para os dois testes.

A definição da quantidade ideal de classes (assinaturas) a serem consideradas em uma amostragem, mesmo em problemas clássicos de reconhecimento de padrões, não segue uma regra muito precisa. Nesse contexto, costuma-se citar um conjunto pequeno ou grande. Dentro do escopo deste trabalho, escolheu-se 25 como o número total de assinaturas.

Assim, de 25 participantes obteve-se como retorno 21 históricos e, portanto, 21 assinaturas e 179 instâncias comportamentais, sendo que quatro históricos foram considerados inválidos por apresentarem uma quantidade de registros inferior a 40% das tarefas solicitadas.

5.2 Apresentação dos Resultados

Nesta seção são apresentados os resultados obtidos experimentalmente. Os experimentos foram realizados visando comparar as três abordagens. A primeira comparação é realizada para verificar a superioridade da Abordagem 2, e conseqüentemente da Abordagem 3, com respeito ao seu desempenho em diferentes situações para o cálculo de similaridade. A segunda comparação, mais geral, faz uma caracterização das principais propriedades que definem as diferenças e semelhanças de cada abordagem.

5.2.1 Comparação das Abordagens

Sejam as cadeias de caracteres a= CCC, b=FFF e c=CCC. Fazendo uma analogia, cada cadeia de caractere pode ser interpretada como uma instância, onde a instância mais atual {c} de um usuário é similar a apenas uma instância de sua assinatura $ac^u = \{a,b\}$.

Aplicando a Abordagem 2 ao cálculo de *SComp* (Equação 12) obteve-se o resultado de 50% de similaridade entre a cadeia *c* e as demais cadeias, como mostrado pelos cálculos descritos abaixo:

CCC→CCC

1) $\overline{dist_u^u} = \dfrac{\sum_{i=1}^{2} \sum_{j=1}^{1} dist_norm(ic_a^u, ic_c^u)}{2*1}$

$dist_norm(i_a^u, i_c^u) = \dfrac{dist(ic_a^u, ic_c^u)}{|ic_a^u|} = \dfrac{0}{3} = 0$

FFF → CCC

$\overline{dist_u^u} = \dfrac{\sum_{i=2}^{2} \sum_{j=1}^{1} dist_norm(ic_b^u, ic_c^u)}{2*1} =$

$dist_norm(i_b^u, i_c^u) = \dfrac{dist(ic_b^u, ic_c^u)}{|ic_b^u|} = \dfrac{3}{3} = 1$

2) $\overline{dist_u^u} = \dfrac{0+1}{2} = 0{,}5$

3) *SComp* = 1 – 0,5 = 0,5

As mesmas cadeias de caracteres são consideradas na aplicação da Abordagem 1 a *SComp* (Equação 8) com o resultado de 100% de similaridade entre as cadeias. Os cálculos realizados abaixo demonstram o processo de obtenção deste resultado.

1) τ(a,c) = ({1,1,1}, {∅}) τ(b,c) = ({∅},{1,1,1})

τ(a,a) = ({1,1,1},{∅}) τ(b,b) = ({1,1,1},{∅})

2) $\delta(\tau(a,c))$ = (2 *1-1) + (2 *1-1) + (2 *1-1) = 3
$\delta(\tau(a,a))$ = (2 *1-1) + (2 *1-1) + (2 *1-1) = 3

$\delta(\tau(b,c))$ = -((2 *1-1) + (2 *1-1) + (2 *1-1)) = -3
$\delta(\tau(b,b))$ = (2 *1-1) + (2 *1-1) + (2 *1-1) = 3

3) S'(a,c) = $\frac{\delta(\tau(a,c))}{\delta(\tau(a,a)} = \frac{3}{3} = 1$

S'(b,c) = $\frac{\delta(\tau(b,c))}{\delta(\tau(b,b)} = -\frac{3}{3} = -1$

4) S(a, c) = $\frac{S'(a,c)+1}{2} = \frac{1+1}{2} = 1$

S(b, c) = $\frac{S'(b,c)+1}{2} = \frac{-1+1}{2} = 0$

5) $SComp^M = \max\{1, 0\} = 1$

Quando ocorre uma situação como exemplificado acima, verifica-se que a utilização da média usada pela Abordagem 2, por representar com mais exatidão a similaridade, se mostra mais apropriada para simbolizar o valor de similaridade, do que o resultado apresentado pela Abordagem 1 que utiliza o valor máximo.

Considere agora as seguintes cadeias de caracteres a=ABCDEFG, b=ACEG e c= ABFG. Aplica-se a Abordagem 1 para o cálculo de similaridade entre as cadeias *a* e *b*, e *a* e *c*. Os seguintes resultados são obtidos:

ABCDEFG →A_C _ E_ G

1) $\tau(a,b) = (\{z_0, z_1, \ldots z_{m-1}\}, \{w_0, w_1, \ldots w_{n-1}\}) \mid n,m \in N$

τ(a,b) = ({1,1,1,1}, {1,1,1})

τ(a,a) = ({1,1,1,1,1,1,1},{∅})

2) $\delta(Z, W) = Soma(Z) - Soma(W) \mid Soma(c_0, c_1, \ldots, c_{m-1}) = \sum_{i=0}^{m-1} 2c_i - 1$

$\delta(\tau(a,b))$ = ((2 *1-1)+ (2 *1-1)+(2 *1-1)+(2 1-1)) – ((2 *1-1)+(2 *1-1)+(2 *1-1)) = **1**

$\delta(\tau(a,a))$ = (2 *1-1)+ (2 *1-1)+(2* 1-1)+(2* 1-1) + (2 *1-1)+(2 *1-1)+(2 *1-1) = **7**

3) S'(a,b) = $\frac{\delta(\tau(a,b))}{\delta(\tau(a,a)} = \frac{1}{7} = 0{,}1428$

4) S(a, b) = $\frac{S'(a,b)+1}{2} = \frac{0{,}1428+1}{2} = 0{,}5714$

ABCDEFG →AB_ _ _FG

1) τ(a,c) = ({2,2}, {3})
   τ(a,a) = ({1,1,1,1,1,1,1},{∅})

2) $\delta(\tau(a,c))$ = ((2 *2-1)+ (2 *2-1)) – (3 *2-1)) = **1**
   $\delta(\tau(a,a))$ = (2 *1-1)+ (2 *1-1)+(2* 1-1)+(2* 1-1) + (2 *1-1)+(2 *1-1)+(2 *1-1) = **7**

3) S'(a,c) = $\frac{\delta(\tau(a,c))}{\delta(\tau(a,a)} = \frac{1}{7} = 0{,}1428$

4) S(a, c) = $\frac{S'(a,c)+1}{2} = \frac{0{,}1428+1}{2} = 0{,}5714$

Na situação exemplificada observa-se que quando os valores de δ forem iguais, o valor da similaridade também será, e assim o valor de similaridade não é maior para a cadeia que possui subseqüências encadeadas.

A comparação entre as abordagens quanto ao seu desempenho, quando não existe uma abordagem *benchmark*, pressupõe o teste da hipótese nula de que todas as abordagens são igualmente adequadas para representar a confiança, contra a hipótese alternativa de que uma é melhor que a outra.

Assim, foram estabelecidas algumas propriedades para comparação, embora a escolha de uma abordagem como a melhor para representar o cálculo da confiança seja subjetiva, já que algumas dessas propriedades podem ser mais importantes que outras, dependendo da aplicação. A Tabela 3 descreve as principais propriedades de cada abordagem.

Tabela 3: Abordagem1 x Abordagem 2 x Abordagem 3

	Heurística de Carmo *et al.*	Distância de Levenshtein	Distância de Levenshtein Adaptada
Formato das entradas	Padrões simbólicos de tamanhos iguais	Padrões simbólicos de tamanhos iguais ou diferentes	Padrões simbólicos de tamanhos iguais ou diferentes
Representação da similaridade	Valor máximo de similaridade	Valor médio de similaridade	Valor médio de similaridade
Forma de pontuação	Pontua subseqüências iguais e encadeadas	Pontua subseqüências iguais	Pontua subseqüências iguais e encadeadas
Flexibilidade	Calcula a similaridade	Calcula a similaridade ou a dissimilaridade	Calcula a similaridade ou a dissimilaridade
Abrangência	Aplica-se aos 3 fatores de *Trust*	Aplica-se aos 3 fatores de *Trust*	Aplica-se aos 3 fatores de *Trust*
Complexidade do algoritmo	O(*mn*)	O(*mn)*	O(*mn)*

As três abordagens fazem o reconhecimento de padrões simbólicos, no entanto, as abordagens 2 e 3 permitem fazer o reconhecimento de padrões simbólicos com comprimentos diferentes. Na Abordagem 1 as representações das instâncias devem ter o mesmo tamanho, senão, é necessário um procedimento de homogeneização para igualar o tamanho das instâncias. Também, as três abordagens, se aplicam à quantificação dos três fatores de *Trust* tornando a métrica homogênea.

A Abordagem 2 difere das demais abordagens por não atribuir uma pontuação maior quando ocorrem subseqüências iguais, propriedade adaptada na Abordagem 3.

Na Abordagem 1, a função *S*, foi desenvolvida para calcular a similaridade entre duas instâncias e não há registros de que possa também calcular a dissimilaridade, como é o caso do algoritmo usado pelas abordagens 2 e 3. Também se diferencia das outras abordagens, pelo valor que utiliza para representar a similaridade, o valor máximo.

O três algoritmos usados para o cálculo de similaridade podem ser implementados com uma complexidade de tempo O(*nm*), como pode-se observar diretamente nos pseudocógicos das Figuras 7, 8 e 11. A operação, de maior custo, executada pelos algoritmos de distância, é a que constrói a matriz. Na abordagem heurística a operação de maior custo é executada pela função de

homogeneização τ. O algoritmo de programação dinâmica para calcular a Distância de Levenshtein não é conhecido por ter a menor complexidade de tempo, no entanto, permite com facilidade a adaptação de novas funções.

A função de pontuação Ω utilizada no algoritmo adaptado (Abordagem 3), pode ser implementada seqüencialmente utilizando uma estrutura auxiliar de uma pilha para que a complexidade não ultrapasse o valor de O(*nm*).

5.2.2 Limiar de Confiança

A quantificação da confiança permite que seja estabelecido o valor de confiança para cada usuário. O Apêndice D exibe os resultados de *Trust* calculados por meio das três abordagens, considerando a última trilha de navegação como a mais atual.

Devido a pouca quantidade de registros (trilhas) disponíveis, a estratégia foi gerar valores de *Trust* obtidos usando-se a cada vez uma das trilhas da assinatura como a mais atual. A quantidade de valores de *Trust,* por usuário, são tantos quantos forem a quantidade de trilhas de sua assinatura. O Apêndice E exibe a tabela com os valores gerados por meio desta estratégia.

Para compor a base de trilhas que representam os valores de *Trust* dos outros usuários a ser comparada com a do usuário corrente foi selecionada a trilha mais representativa de cada assinatura. A trilha mais representativa é a que tem o maior valor de similaridade entre as trilhas que compõem uma assinatura. Logo, para a Abordagem 2, a trilha mais representativa é a que tem $\max\{S(ic_i, ic_j) \forall ic \in ac\}$ e para as abordagens 2 e 3 é a que tem $\min\{dist(ic_i^u, ic_j^k)\}$ onde $k=u$.

A idéia básica é selecionar a trilha de navegação que melhor caracteriza o comportamento de cada usuário. Cada trilha é comparada com a assinatura de um usuário em estudo, simulando um comportamento ilegítimo. Para tanto, a trilha mais representativa é considerada como a trilha mais atual no cálculo de *SComp*. O Apêndice E exibe as tabelas com os valores gerados por meio desta estratégia. A quantidade de valores de *Trust* são tantos quantos forem o número de assinaturas.

Em seguida o limiar foi calculado para cada usuário através da metodologia descrita na Subseção 3.4.6. Desta forma, uma avaliação comportamental pode usar este valor estabelecido como referência. O Apêndice F exibe as tabelas com os valores do limiar para cada usuário.

$$\text{Avaliação Comportamental} = \begin{cases} \text{verdadeiro} & \text{se } \mathit{Trust} \geq \mathit{limiar} \\ \text{Falso} & \text{se } \mathit{Trust} < \mathit{limiar} \end{cases}$$

5.3 Discussão

De acordo com os testes realizados com o *web site*, algumas observações puderam 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 à 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. Para solucionar este problema, a estrutura foi alterada no segundo teste, conforme descrito detalhadamente no diagrama de

seqüência da Figura 17. Assim, o usuário pode, inicialmente, navegar pela loja e no momento em que realiza uma compra é solicitado o seu cadastramento e/ou a autenticação. Quando o usuário não realiza a autenticação, as informações coletadas não são associadas ao código de registro do usuário no banco de dados, situação evidenciada no primeiro teste com a estrutura do *web site* e evitada por esta alteração.

- a alteração da estrutura do *web 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;

- durante a realização dos testes, o ambiente se revelou eficiente para permitir aos usuários total liberdade de navegação. Este aspecto foi importante para possibilitar a geração de amostras que caracterizassem os comportamentos dos usuários.

- 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;

- verificou-se que quanto mais elementos de navegação estiverem disponíveis, uma maior variedade de comportamentos pode ser identificada.

Os resultados apresentados na Seção 5.2. apontam para uma superioridade da Abordagem 2 se comparada a Abordagem 1 principalmente em relação à representação do valor de confiança e as propriedades: flexibilidade e formato de entrada. Entretanto, esta abordagem tem desempenho inferior ao considerar a pontuação de subseqüências encadeadas, propriedade adaptada na Abordagem 3.

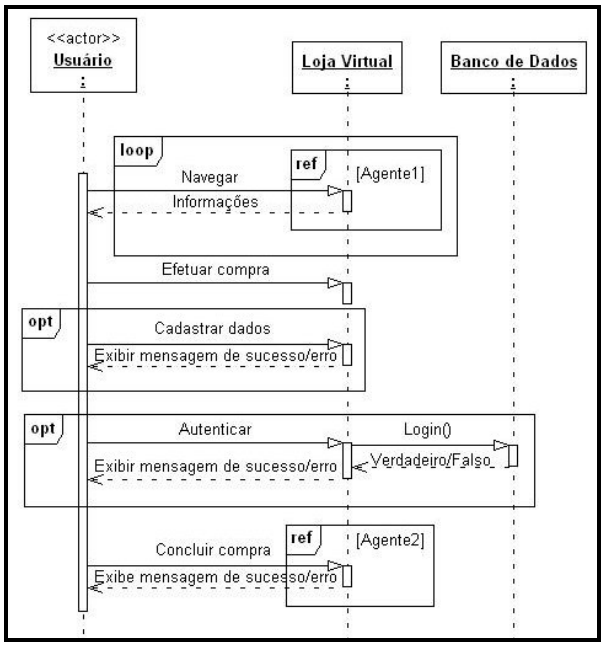

Figura 17: Segundo diagrama de seqüência do usuário

Considerando ainda, a natureza complexa dos procedimentos usados para o cálculo da similaridade, a Abordagem 2 pode ser considerada mais simples, uma vez que dispensa a utilização de uma função suplementar para igualar o tamanho das instâncias, como é o caso da função de homogeneização usada pela função τ na Abordagem 1.

As Figuras 18, 19 e 20 exibem os histogramas para cada uma das três abordagens. O eixo *y* corresponde ao número de FN, FP, VN e VP contabilizados a partir da determinação do limiar de confiança para cada usuário. O eixo *x* exibe o código de registro dos usuários no banco de dados.

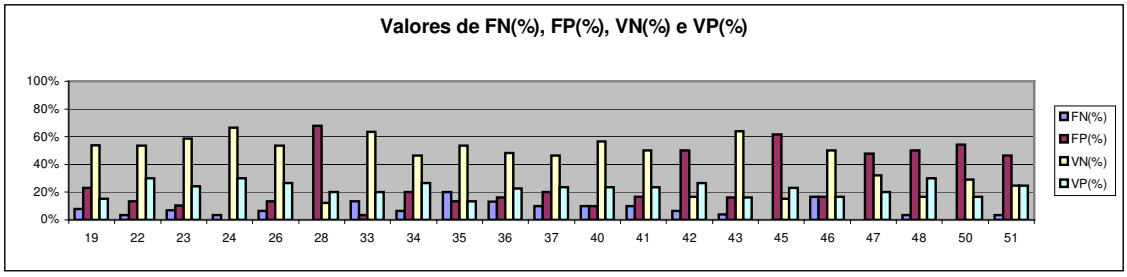

Figura 18: Histograma de valores FN(%), FP(%), VN(%) e VP(%) – Abordagem 1.

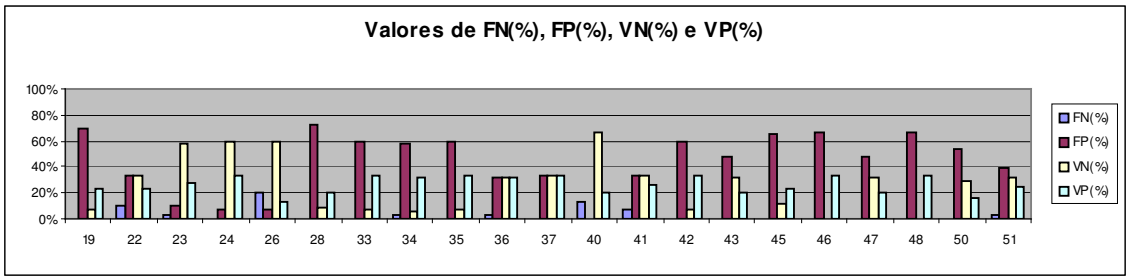

Figura 19: Histograma de valores FN(%), FP(%), VN(%) e VP(%) – Abordagem 2

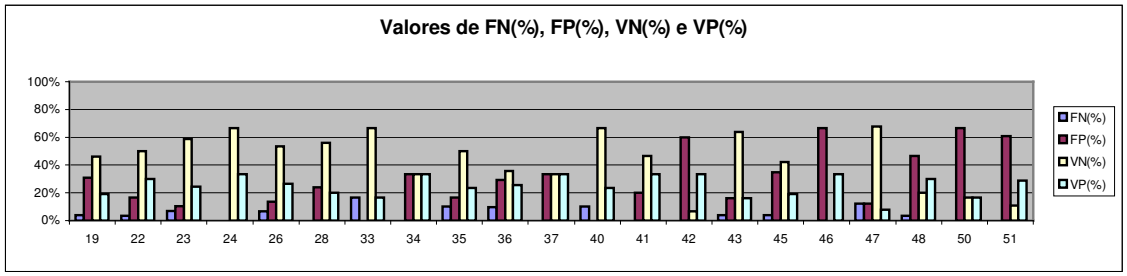

Figura 20: Histograma de valores FN(%), FP(%), VN(%) e VP(%) – Abordagem 3

Em média a Abordagem 3 gerou melhores resultados que as duas abordagens anteriores uma vez que minimizou as ocorrências de FN e FP (32%) e maximizou VN+VP(68%). A Tabela 4 exibe os resultados para a comparação.

Tabela 4: Ocorrências de FN+FP e VN+VP

	Abordagem 1	Abordagem 2	Abordagem 3
FN + FP	0,3407	0,4699	0,3243
VN + VP	0,6593	0,5301	0,6757

A similaridade das assinaturas dificulta o processo de diferenciação dos comportamentos dos usuários. De maneira geral, um problema de reconhecimento de padrões de comportamento bem definido e suficientemente delimitado apresentará pequenas variações intraclasse (SIntra) e grandes variações interclasses (SInter). Assim, se as assinaturas são muito similares é esperado que a quantidade de FP seja alta.

A estrutura informacional do *web site,* organizada por gêneros, favoreceu a diferenciação comportamental. No entanto, os resultados expostos pela avaliação revelaram a dificuldade em se

projetar ambientes que realizam essa tarefa eficientemente. Mesmo com a disposição de mais elementos de navegação ou a captura de mais características de cada usuário na tentativa de promover a diferenciação, a existência de usuários com as mesmas preferências, torna a ocorrência de FP inevitável.

As assinaturas onde também se observou a falta de regularidade no comportamento, (comportamento mal formado), tais como as assinaturas dos usuários 42, 46, 50 ou 51, as ocorrências de FP foram mais evidentes. Nestes casos, é impossível traçar um padrão que caracterize as assinaturas desses indivíduos, resultado ainda potencializado pela similaridade com as demais instâncias das outras assinaturas.

No entanto, a adaptação da Distância de Levenshtein possibilitou constatar que a pontuação de subseqüências encadeadas é uma propriedade importante e que contribui para uma melhora significativa dos resultados.

5.3.1 Diretrizes

Existem na literatura uma série de diretrizes consagradas, entre elas as apresentadas por Nielsen (1993) e W3C (World Wide Web Consortium) para a criação de interfaces. De acordo com Dias e Carvalho (*apud* Hix e Hartson, 1993), diretrizes são recomendações aplicáveis amplamente aos projetos de interação com o usuário, embora possam ser escritas de forma genérica, o seu claro entendimento pode servir como uma orientação para um projeto. São algumas vezes elaboradas e validadas empiricamente, apoiadas em opiniões baseadas na experiência de seus criadores.

Após análise empírica, realizada no ambiente de teste, foram reunidas oito diretrizes que não se referem especificamente a *web sites* de compra e acredita-se com adaptações, podem ser aplicadas a outros gêneros. Tais diretrizes são apresentadas e exemplificadas como a seguir:

Diretriz 1:	Definir um macro-estado.
Justificativa:	Um contingente de páginas deve ser definido para que os comportamentos possam ocorrer no mesmo espaço navegacional.

É no macro-estado que os comportamentos ocorrem, assim, é necessário definir um contingente de páginas para que passando pela avaliação comportamental ocorra a transição de um macro-estado a outro. Para o experimento foi definido um único macro-estado para todos os usuários.

Diretriz 2:	Delimitar a instância comportamental pelas tarefas que os usuários desejam realizar.
Justificativa:	A delimitação da instância possibilita que cada comportamento seja coletado individualmente a cada vez.

No experimento a delimitação foi baseada na tarefa de compra que se inicia com a página principal do ambiente e termina na página destinada para confirmar a compra.

Diretriz 3:	Definir as páginas que não refletem o comportamento que possa ser diferenciado.
Justificativa:	As páginas de acesso comum a todos os usuários não interessam a avaliação comportamental e podem não fazer parte do macro-estado.

Páginas como a de "Cadastro" ou "Login" que não refletem um comportamento que possa ser diferenciado não é utilizado pela avaliação comportamental. Uma vez que não fazem parte da definição do macro-estado podem ser alteradas de localidade ou mesmo excluídas sem comprometimento para a avaliação.

Diretriz 4:	Definir o nível de granularidade dos comportamentos.
Justificativa:	A granularidade deve ser definida para o conhecimento do nível de detalhamento do comportamento.

Como exemplo, pode-se citar o comportamento de ver detalhes dos cds. A página de "Detalhes" é a mesma para todos os usuários, no entanto, seu conteúdo é especifico para cada produto. Uma granularidade mais fina do comportamento de ver detalhes seria "DetalhesNomeCD", que nesse caso é desnecessária.

Diretriz 5:	Definir as páginas de navegação obrigatória.
Justificativa:	Importante para se ter a informação de que um usuário inicia e finaliza uma tarefa com êxito.

As páginas de acesso obrigatório são as páginas definidas como início e fim da instância comportamental. No experimento, a navegação que não alcança a página definida como obrigatória para a confirmação da compra não é coletada. É importante para que todos os comportamentos iniciem da mesma forma mesmo com o uso da opção de histórico do navegador.

Diretriz 6:	Definir estratégia de agrupamento dos comportamentos derivados de várias páginas mantidas em aberto.
Justificativa:	Um usuário pode manter algumas páginas abertas e optar por navegar por elas ao invés de utilizar os *links*.

A estratégia implementada faz utilização de um único repositório de dados (*cookie*) para colecionar os comportamentos originados de várias páginas mantidas em aberto.

Diretriz 7:	Definir estratégia de associar os comportamentos aos seus respectivos usuários.
Justificativa:	A associação do comportamento ao seu usuário gerador possibilita a formação das assinaturas.

Para a identificação dos comportamentos como sendo provenientes de um mesmo usuário utilizou-se uma combinação das técnicas, de autenticação e c*ookie*s. Após o usuário ser identificado, os comportamentos são associados ao seu código de registro no banco de dados.

Diretriz 8:	Definir estratégia de captura dos comportamentos derivados do navegador.
Justificativa:	Os comportamentos derivados do uso do navegador também são importantes para avaliação comportamental.

Os navegadores possuem opções de navegação através dos botões de "voltar" ou "seguir". A estratégia para coletar os comportamentos deve considerar essa possibilidade de uso. No *web site*, a estratégia utilizada foi a de monitorar o evento de abertura das páginas.

CAPÍTULO 6

CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

É apresentado, neste capítulo, o resumo da pesquisa descrita nesta dissertação, expondo as suas contribuições, problemas revelados ao longo do seu desenvolvimento e sugestões para prosseguimento do trabalho.

6.1 Resumo do Trabalho

A pesquisa desenvolvida é fundamentada na proposta apresentada por Carmo *et al.* (2007), onde a confiança, definida informalmente como uma medida de quão certo o provedor de uma aplicação está a respeito da identidade de um usuário, é proveniente de uma avaliação comportamental.

A fim de produzir conhecimentos úteis para a avaliação de um usuário que navega em uma aplicação *Web* foi construído um *web site* como gerador de subsídios. Este ambiente experimental permitiu a captura das informações de comportamentos a serem mensurados e a extração de requisitos para a implementação de *web sites*.

Na análise do contexto foi definido que os comportamentos ocorreriam em um único macro-estado formado pelo contingente de páginas de uma loja virtual. As informações registradas a partir das páginas foram coletadas e armazenadas constituindo um histórico comportamental. Foi necessário ainda delimitar uma instância, a granularidade das informações a serem utilizadas e definir uma estratégia para agrupar e associar os comportamentos a seus respectivos usuários.

A coleta das assinaturas comportamentais foi formulada como uma tarefa de aprendizagem para caracterizar o comportamento típico de um usuário. Para tanto, foi adaptado o modelo de

Aprendizado Baseado em Instâncias, onde o formato representacional dos dados de entrada são as seqüências de páginas visitadas pelo usuário durante a sua interação com o *web site*.

Para implementar esse conceito foi necessária a identificação do padrão de comportamento contextual de navegação do usuário no *web site* (trilha de navegação). Para a identificação das trilhas, coletadas pelos agentes, e provenientes de um mesmo usuário utilizou-se uma combinação de autenticação por credenciais (nome e senha) e c*ookie*s.

Formadas as assinaturas a confiança pôde ser estabelecida. O cálculo da confiança considera fundamentalmente a existência de três fatores interdependentes: a similaridade comparativa, a intra-similaridade e a inter-similaridade.

Para quantificar os fatores de confiança, a abordagem heurística de Carmo *et al.*(2007) e a Distância de Levenshtein foram selecionadas para serem investigadas, para tanto, seus desempenhos foram examinados experimentalmente.

Com o intuito de se definir uma melhor representação para a confiança, foi proposta uma adaptação no algoritmo da Distância de Levenshtein. O benefício dessa nova abordagem é a capacidade de pontuação das subseqüências encadeadas.

A quantificação da confiança tornou possível o estabelecimento de um patamar de aceitação mínimo para cada usuário. Com a estratégia proposta para esse fim, o valor a partir do qual um usuário pode ser considerado confiável por uma aplicação foi determinado.

Segundo Nielsen (2000), na Internet não existe uma padronização, mas certas convenções podem ser utilizadas e respeitadas. Sendo assim, diretrizes podem ser sugeridas para a construção de *web sites* direcionados a avaliação comportamental. Como resultado desta afirmação, foram obtidas oito diretrizes derivadas de uma análise empírica, realizada no *web site* de teste.

6.2 Contribuições da Dissertação

Uma das contribuições desta pesquisa foi a elaboração de um conjunto de diretrizes para guiar a construção de *web sites* destinados a utilizar a avaliação comportamental como meio de aumentar a confiança na identificação do usuário. Acredita-se que as informações apresentadas podem subsidiar um projetista.

Tendo ciência de que as diretrizes aqui sugeridas representam apenas uma pequena amostra do potencial de recomendações para o objetivo proposto neste trabalho, novas sugestões devem ser incentivadas.

A apresentação dos resultados, a partir da proposta de avaliação experimental, contribuiu para verificar a viabilidade do uso das abordagens investigadas. Os resultados revelados foram considerados úteis para a determinação do sucesso da avaliação comportamental.

Outra importante contribuição é a apresentação de uma nova abordagem para quantificar a confiança. Os resultados obtidos, por meio da adaptação da Distância de Levenshtein e sua aplicação aos fatores de confiança, mostraram-se mais satisfatórios comparados às outras abordagens.

6.3 Limitações e Dificuldades Encontradas

No decorrer do desenvolvimento desta pesquisa algumas limitações e dificuldades foram encontradas e são aqui relatadas.

Embora tenha abrangido vinte e um participantes, amostra razoável de população, o objetivo da pesquisa era trabalhar com um número maior de participantes, o que não foi possível devido à dificuldade de se angariar pessoas disponíveis a colaborar em mais de um teste.

O número de participantes, contudo, não é um fator tão crítico quanto o número de registros por usuário. Assim, também se evidencia como limitante a quantidade de registros obtidos, que é um aspecto pertencente a este tipo de aplicação, e que por isso houve a necessidade de se definir uma estratégia para geração de novos valores.

Os resultados desta estratégia, todavia, permitiram a extensão da pesquisa, gerando novos conhecimentos a partir da tentativa de definição de um limiar de confiança para cada usuário.

Apesar dos cuidados metodológicos descritos, visando a imparcialidade na análise dos resultados, a pesquisa empírica baseia-se apenas na percepção do analisador, outro fator limitante.

Portanto, as conclusões apresentadas refletem somente as evidências observadas na pesquisa empírica, que se restringem à população pesquisada e estão sujeitas às imprecisões inerentes a um estudo deste tipo.

6.4 Trabalhos Futuros

Alguns aspectos relacionados foram identificados durante o desenvolvimento do trabalho, mas não foram tratados com a profundidade necessária. Esses aspectos apresentam-se como trabalhos futuros, sendo importantes para o prosseguimento do trabalho, tais como:

- prolongar os experimentos para *web sites* personalizados - considerando a tendência atual no desenvolvimento de ambientes que utilizam o conceito de recomendação seria importante avaliar qual o seu impacto na formação dos comportamentos, sobretudo para os valores de SIntra e SInter.

- estender a pesquisa a mais de um macro-estado - o foco desta pesquisa é na avaliação comportamental, portanto, um único macro-estado foi considerado suficiente

para permitir a coleta das informações necessárias de serem investigadas. Entretanto, no vislumbre do mecanismo como mostrado na Figura 2 da Seção 3.1 é necessário a definição de mais macro-estados.

- buscar definir formas de representar um comportamento originado por um usuário legítimo mas destinado a outro usuário - nem todos os comportamentos, principalmente, ao considerar ambientes de comércio eletrônico, são reflexos do comportamento do usuário em particular. Um indivíduo pode efetuar uma operação destinada a outro indivíduo, como por exemplo, presenteá-lo com algum produto. Assim, acredita-se que um comportamento coletado em condições como essa possa influenciar o resultado da confiança.

- buscar definir um limiar menos rigoroso podendo ser modelado segundo conjunto nebuloso - enquanto a relação entre um conjunto clássico só admite duas possibilidades "pertence " e "não pertence", os conjuntos nebulosos permitem quantificar de maneira mais granular a relação entre conjunto e elementos. É uma ferramenta que não utiliza a precisão a rigor, mas é tolerante a falhas e verdades parciais.

- investigar outras variáveis – a variável utilizada para o cálculo de confiança foi a trilha de navegação, no entanto, outras variáveis podem ser investigadas, como informação adicional e complementar ao estabelecimento da confiança;

- investigar a aplicação de modelos mentais para melhorar a diferenciação comportamental – projetistas geralmente se preocupam em antecipar os percursos e ações dos usuários, para ajudá-los a realizar as suas ações. Essa antecipação dos percursos e das ações dos usuários de interfaces *Web* refere-se ao entendimento das diversas estruturas cognitivas (ou modelos mentais).

REFERÊNCIAS BIBLIOGRÁFICAS

AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-Based Learning Algorithms**. Journal Mach. Learn**. v.6, n.1, p. 37-66 , jan. 1991.

ARAUJO, R.M. **Aprendizado de máquina em sistemas complexos multiagentes: estudo de caso em um ambiente sob racionalidade limitada.** 2004. Dissertação (Mestrado em Computação) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2004.

ARAGÃO A .L. **Utilização de aprendizado de máquina para a adaptação de estruturas em hipermídia adaptativa** 2004. Dissertação (Mestrado em Ciências da Computação e Matemática Computacional). Universidade de São Paulo, São Paulo, 2004.

ARDIGO, J. D. **Modelo de Infra-estrutura de Chaves Públicas como Organização Virtual para Processos de Avaliação Somativa à Distância**. 2004. Tese (Doutorado em Engenharia de Produção) - Universidade Federal de Santa Catarina, Florianópolis, 2004.

ARMADA, R.N. **Um ambiente de apoio a externalização e mapeamento do conhecimento em equipes distribuídas de trabalho**. Dissertação (Mestrado em Informática) - Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2005.

ARTZ, D.; GIL, Y. 2007. **A survey of trust in computer science and the Semantic Web**. Web Semantics: Science, Services and Agents on the World Wide Web, v. 5, jun.2007, p.58-71.

BASTIEN, J.M.C.; SCAPIN, D.L. **Critérios Ergonômicos para Avaliação de Interfaces Homem-Computador.** 1993 Disponível em: http://www.labiutil.inf.ufsc.br/CriteriosErgonomicos/Abertura.html Acesso em: 10 set. 2007.

BISHOP, C. M., **Neural Networks for Pattern Recognition**. Oxford University Press, 1995.

BRAINARD, J.; JUELS, A.; RIVEST, R. L.; SZYDLO, M; YUNG, M. 2006. **Fourth-factor authentication: somebody you know.** In: Proceedings of the 13th ACM Conference on Computer and Communications Security, 2006, Alexandria, Virginia, p. 168-178.

BRAINERD, J.; BECKER, B.; SOFTWARE, B. N. **Case study: e-commerce clickstream visualization**. Proceedings of the IEEE Symposium on Information Visualization, 2001, p.153-156.

CARMO, L.F.R.C; REIS JUNIOR, C.A. Avaliação da Confiança por Análise Comportamental para Autorização Evolutiva em Aplicações Seguras na Web. In: Simpósio Brasileiro de Redes de Computadores, 24, 2006, Curitiba. **Anais do XXIV Simpósio Brasileiro de Redes de Computadores**. SBRC, Curitiba: 2006.

CARMO, L.F.R.C.; OLIVEIRA, B.G; BRAGA, A.C.B. **Trust Evaluation for Web Applications based on behavioral Analyses**. In: 22th International Information Security Conference, 2007, Sandton, South Africa.

CAZELLA, S. C.. **Uma arquitetura para coordenar a interação de agentes na Internet**. 1997. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal do Rio Grande do Sul, Porto Alegre: 1997.

COSTA, C.R.N. **Autenticação Biométrica via teclado numérico baseada na dinâmica da digitação: Experimentos e Resultados.**. 2006. Dissertação (Mestrado em Engenharia Elétrica e Computação) - Universidade Estadual de Campinas, Campinas, 2006.

COSTA, D.; CARMO, L.F.R.C. Reconhecimento de padrões de comportamento individual baseado no histórico de navegação em um Web Site. In: Simpósio Brasileiro em Segurança da Informação e Sistemas de Computadores, 7 , 2007, Rio de Janeiro. **Anais do VII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais.** SBSeg, Rio de Janeiro:2007, p.18-29.

CRUZ, C.C.P; MOTTA, C.L.R. **Um Modelo de Sistema de Reputação para Comunidades**. In: Simpósio Brasileiro de Informática na Educação, 17, 2006, Brasília. **Anais do XVII Simpósio Brasileiro de Informática na Educação**. SBIE, Brasília: 2006. p.92-101.

DASGUPTA, P. **Trust as a Commodity.** 2000. Disponível em: <http://www.sociology.ox.ac.uk/papers/dasgupta49-72.pdf>Acesso em: 15 abr. 2007.

DEVEGILI, A. J.; PARENTE, R. V. Autenticação em HTTP baseada em desafio/resposta. In: Simpósio Brasileiro de Redes de Computadores, 21., 2003, Natal. **Anais do III Workshop em Segurança de Sistemas Computacionais**. WSeg, Natal: 2003.

ETZIONI, O.; WELD, D. S. **Intelligent Agents on the Internet: Fact, Fiction and Forecast**. IEEE Expert: Intelligent Systems and Their Applications, 1995, v.10, n.4, p. 44-49.

FIORESE, M. **Uma Proposta de Autenticação de Usuários para Ensino a Distância.** 2000. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2000.

FORRISTAL, J. **Site Seguro Aplicações Web.** Rio de Janeiro: Alta Books, 2002.

FRANKLIN, S.; GRAESSER, A.**Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents.** In Proceedings of the Workshop on intelligent Agents, Agent theories, Architectures, and Languages, Springer Verlag, 1996.

GAMBETTA, D. **Can We Trust Trust?** 2000. Disponível em: <http://www.sociology.ox.ac.uk/papers/ gambetta213-237.doc>Acesso em: 24 fev. 2007.

GONZALEZ, R.C.; THOMASON, M.G. **Syntactic Pattern Recognition**. Addison Wesley, 1978.

GRANDISON, T.W.A. **Trust Management for Internet Applications.** 2003. Ph.D. Thesis. University of London, 2003.

HU, J.; ZHONG, N. **Clickstream Log Acquisition with Web Farming**, IEEE/WIC/ACM International Conference on Web Intelligence, 2005, p.257-263.

HUNT, R. **PKI and Digital Certification Infrastructure**. In: Proceedings of the 9th IEEE International Conference on Networks, 2001, Washington, DC, p. 234-239.

International Association for Biometrics. **Glossary of Biometric Terms**. Disponível em:<http://www.afb.org.uk/docs/glossary.htm> Acesso em: 21 abr. 2007.

JAIN, A. K.; DUIN, R.P.W.; MAO, J. **Statistical pattern recognition: A review.** IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, v.22, n.1**,** p.4-37.

JAIN A. K; ROSS A.; PRABHAKAR S. **An Introduction to Biometric Recognition**. IEEE Transation.on Circuit and Systems for Video Technology, 2004, v.14, p. 4-20.

JONES, S.; MARSH, S. Human-Computer-Human Interaction: Trust.in CSCW. **ACM SIGCHI Bulletin**, v.29,p. 36-40, jul.1997.

LANE, T.; BRODLEY, C. **Temporal Sequence Learning and Data Reduction for Anomaly Detection.** ACM Transactions on Information and System Security, 1999, v.2, n.3, p. 295-331.

LANE, T. **Machine learning techniques for the computer security**. 2000. Ph.D. thesis, Purdue University, 2000.

LOPEZ, J.; OPPLIGER, R.; PERNUL, G. **Authentication and authorization infrastructures (AAIs): a comparative survey**, Computers & Security, v.23, out. 2004, p. 578-590.

LUHMANN, N. 2000. **Familiarity, Confidence, Trust: Problems and Alternatives.** Disponível em: <http://www.sociology.ox.ac.uk/papers/luhmann94-107.pdf>Acesso em: 21 jun. 2007.

MARSH, S.P. **Formalising Trust as a Computational Concept.** 1994. Ph.D. Thesis. University of Stirling, 1994. Disponível em: http://www.cs.stir.ac.uk/research/publications/techreps/pdf/TR133.pdf > Acesso em: 23 out. 2007.

MONROSE, F.; RUBIN, A. D. 2000. **Keystroke dynamics as a biometric for authentication**. Future Generation. Computer Systems, v.16, fev. 2000, p. 351-359.

MONROSE, F.; RUBIN, A. **Authentication via Keystroke Dynamics**. In: 4th ACM Conference on Computer and Communication Security, 1997, Zurich, Switzerland, p. 48-56.

MOTTA, C.L.R. **Um Ambiente de Recomendação e Filtragem Cooperativas para Apoio a Equipes de Trabalho**, Tese (Doutorado em Ciências em Engenharia de Sistemas e Computação) - Universidade Federal do Rio de Janeiro, Rio de Janeiro, 1999.

NAVARRO, G. 2001. **A guided tour to approximate string matching**. ACM Computing. Surveys, 2001, v.33, p. 31-88.

NIELSEN, J. **Projetando Websites**. Rio de Janeiro: Campus, 2000.

NIELSEN, J. **Usability Engineering**, Academic Press, 1993.

O'GORMAN, L. **Comparing passwords, tokens, and biometrics for user authentication,** Proceedings IEEE, 2003, v. 91, n. 12, p. 2021-2040.

ONODA, M. **Metodologia de dados para análise do comportamento de navegação num Web Site.** 2006, Tese (Doutorado em Ciências em Engenharia Civil) - Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2006.

PELISSARI, F. R.; RIGHI, R. R.; WESTPHALL, C. M. RBRP: Protocolo de Reputação Baseado em Papéis para Redes Peer-to-Peer. In: Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais, 5.,.2005, Florianópolis. **Anais do V Simpósio Brasileiro em**

Segurança da Informação e de Sistemas Computacionais . SBSeg, Florianópolis: 2003. p.257-270.

PIO, J. L. S; OLIVEIRA, C. J. S.; ARAÚJO, A. A.; VIEIRA, N. J **Reconhecimento Sintático e Estrutural de Padrões**. 2006 Disponível em: <http://bibliotecadigital.sbc.org.br/?module=Public&action=PublicationObject&subject=238&publicationobjectid=69> Acesso: 29 de jul. 2007.

PLATZER, C.**Trust-Based Security in Web Services**.2004. Ph.D Thesis.University of Vienna, Austria, 2004.

PUSARA, M.; BRODLEY, C.E. **User Re-Authentication via Mouse Movements**, In*:* CCS Workshop on Visualization and Data Mining for Computer Security, 2004, Washington, DC, p.1-8.

REATEGUI, E; CAZELLA S. C. **Sistemas de Recomendação.** 2005. **Disponível em: <**http://www.sbc.org.br/bibliotecadigital/?module=Public&action=SearchResult&author=781> Acesso em: 03 out. 2007.

ROCHA, J.L.F. **Proteção de Software por Certificação Digital.** 2001. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Santa Catarina, Florianópolis, 2001.

ROSS, D. **Game Theory, Stanford Encyclopaedia of Philosophy**,.1997. Disponível em <http://plato.stanford.edu/entries/game-theory/> Acesso em: 20 de jun. 2007.

RUBIN, A. D. Independent One-Time Passwords. **USENIX Journal of Computer Systems,** fev. 1996.

RUSSELL, S; NORVING, P. **Intêligencia Artificial.** 2 ed. Rio de Janeiro: Campus 2004.

SCHALKOFF, R.J., **Pattern Recognition**: **Statistical, Structural and Neural Approaches**, John. Wiley and Sons, 1992.

SHAHABI, C.; ZARKESH, A. M.; ABIDI, J; SHAH, V. **Knowledge discovery from users Web-page naviagation.** .In: Proceedings of 7th International Workshop on Research Issues in Data Engineering, 1997, Birmingham, England, p.20-29.

SCHIMKE, S.; VIELHAUER, C.; DITTMANN, J. 2004. **Using Adapted Levenshtein Distance for On-Line Signature Authentication.** In: Proceedings of the 17 th International Conference on Pattern Recognition, 2004, Washington, DC, p.931-934.

SCHNEIER, B. Customers, Passwords, and Web Sites. **IEEE Security and Privacy Magazine**, v. 2, n. 4, p. 88, jul./aug. 2004.

SYMANTEC Internet Security Threat Report. 2007. Disponível em: <http://eval.symantec.com/mktginfo/enterprise/white_papers/ent-whitepaper_internet_security_threat_report_xii_09_2007.en-us.pdf> Acesso em: 23 out. 2007.

UNTERLEITNER, M.C. **Implementation of a Computer Immune System for Intrusion and Virus Detection**. 2006. Ph.D.Thesis, Graz University of Technology, Austria, 2006. Disponível em:<http://www.iaik.tugraz.at/teaching/11_diplomarbeiten/archive/unterleitner/thesis14.pdf> Acesso: 29 de jul. 2007.

VÉRAS, L.M.A ; RUGGIERO, W.V. **Autenticação Contínua de Usuários em Aplicações Seguras na Web**, In: Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais, 5 , 2005, Florianópolis. **Anais do V Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais.** SBSeg, Florianópolis: 2005. p.40-53.

WOOLDRIDGE, M. J.; JENNINGS, N. R. **Software Engineering with Agents: Pitfalls and Pratfalls**. IEEE Internet Computing, 1999, v.3, p.20-27.