Uso de inteligência artificial na construção de ferramentas de análise forense

Lecênio Hélio Trein Júnior[1]
André Peres[2]
Vanessa Lindemann[3]
Analúcia Schiaffino de Morales[3]
Fabiana Lorenzi[4]

RESUMO

Este artigo descreve o desenvolvimento de um mecanismo para investigações de incidentes de segurança. O mecanismo oferece este serviço através da disponibilização de um banco de dados de incidentes de segurança conhecidos, contendo características e ações a serem tomadas. Em situações de incidente de segurança, o responsável pela segurança pode utilizar o sistema, informando um conjunto de informações sobre o incidente e o sistema automaticamente apresentará o caso mais similar em conjunto com as ações a serem tomadas. Para a busca na base de dados, são utilizadas técnicas de raciocínio baseado em casos.

Palavras-chave: *análise forense; inteligência artificial; raciocínio baseado em casos, segurança.*

[1] Acadêmico do Curso de Ciência da Computação/ULBRA - Bolsista PROICT/ULBRA
[2] Professor do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS)
[3] Professora do Curso de Ciência da Computação/ULBRA
[4] Professora - Orientadora do Curso de Ciência da Computação/ULBRA (fabilorenzi@gmail.com)

ABSTRACT

This paper describes the development of a mechanism for computer security forensic investigations. The mechanism provides this service through the provision of known security incidents database, containing characteristics and actions to be taken in each situation. In situations of security incident, the security officer can use the system, setting a set of data regarding the incident and the system will automatically present the most similar case in conjunction with the actions to be taken. For the database search, case-based reasoning techniques are used.

Key words: *forensic analys, artificial intelligence, case-based reasoning, security.*

INTRODUÇÃO

Sempre que recursos computacionais são utilizados para armazenamento ou envio de informações sensíveis, corre-se riscos de segurança. Este fato torna-se evidente com o acompanhamento do crescimento de ataques aos sistemas computacionais. São exemplos de crimes que ocorrem no mundo digital o roubo de informações bancárias em conjunto com senhas, roubo de propriedade intelectual, falsificação de documentos, destruição de informações e paralização de serviços.

Torna-se necessário então que os profissionais de TI (Tecnologias da Informação) tenham consciência dos riscos de segurança existentes nos ambientes computacionais que estão sob sua reponsabilidade e sigam um processo consistente para a manutenção de um estado seguro. Nota-se, no entanto, que o preparo de muitas empresas e seus profissionais não são suficientes por questões diversas, tais quais: custos, tempo hábil para qualificação dos profissionais e planejamento.

Uma forma de contornar este problema é o desenvolvimento de mecanismos capazes de facilitar a identificação da ocorrência e as possíveis soluções para incidentes de segurança. Estes mecanismos devem auxiliar o profissional reponsável pela administração da segurança no rápido diagnóstico dos incidentes e nos procedimentos a serem adotados para contornar uma situação de vulnerabilidade.

Diante deste contexto, o objetivo do presente trabalho é o desenvolvimento de um sistema capaz de auxiliar o administrador de segurança computacional a identificar, prevenir e possivelmente contornar incidentes de segurança. Este sistema, denominado AFIA (Análise Forense com Inteligência Artificial), utiliza técnicas de Raciocínio Baseado em Casos (RBC) e conta com um banco de casos de incidentes e suas respectivas soluções. Os casos que compõem a base de conhecimento do sistema foram baseados em situações reais analisadas por especialistas em investigações forenses computacionais.

Raciocínio Baseado em Casos

O Raciocínio Baseado em Casos (RBC) é uma abordagem que prevê a solução de novos problemas através da adaptação de soluções que foram usadas para resolver problemas passados (KOLODNER, 1993).

Esta abordagem consiste em criar uma base de casos a partir de problemas que foram resolvidos no passado. Casos semelhantes normalmente possuem soluções semelhantes, portanto um sistema RBC indicará as possíveis soluções para o problema

comparando a semelhança das características do problema em análise com as características dos casos da base e indicando os casos semelhantes.

Além da base de casos que representa situações anteriores, os sistemas RBC requerem mecanismos para tratar esse conhecimento. Para Aamodt e Plaza (1994), estes mecanismos podem ser representados por um ciclo de funcionamento dividido em quatro etapas principais, conforme ilustra a Figura 1: recuperação do caso ou casos mais similares; reutilização da informação e conhecimento presente no caso para solucionar o problema corrente; revisão da solução proposta, se necessário; retenção de parte da experiência obtida nos processos de modo a ser útil na solução de problemas futuros.

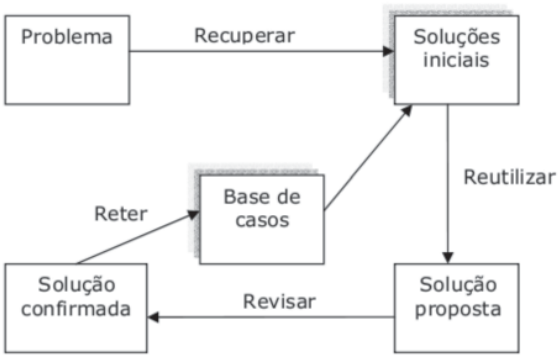

Figura 1. Ciclo de funcionamento de um sistema RBC

Cada etapa é constituída de um conjunto de tarefas que podem ser implementadas por meio de diferentes técnicas. As particularidades do domínio da aplicação definirão as etapas, modelagem e a linguagem a serem empregadas no desenvolvimento do sistema RBC.

Kolodner (1993) destaca as vantagens e desvantagens na utilização de um sistema RBC. As principais vantagens são:

- *permitir soluções de problemas de forma mais rápida e em domínios não conhecidos completamente;*
- *fornecer um meio de avaliar soluções mesmo não existindo um método algorítmico definido;*
- *realizar a interpretação de conceitos abrangentes e mal definidos;*
- *servir de alerta às pessoas, em função do conhecimento de experiências adquiridas, para que não repitam erros cometidos no passado e focar a atenção nas características principais do problema.*

Dentre as desvantagens, a autora destaca:

- *possibilidade de aplicação de casos não validados em situações novas, ocasionando resultados ineficientes, soluções erradas e mal avaliadas;*
- *existência de uma forte tendência dos casos influenciarem as pessoas na resolução de um novo problema e induzi-las ao erro;*
- *possibilidade de algumas pessoas, com pouca experiência, não utilizarem o conjunto de casos mais apropriado no processo de raciocínio, prejudicando a solução do problema.*

Através destas vantagens e desvantagens apresentadas pode-se concluir que a base de casos além de ser ampla, deve ser bem representada. Para Aamodt e Plaza (1994), o problema da representação está em decidir o que será armazenado no caso. Para isso, deve existir uma estrutura apropriada de descrição dos casos e uma forma de organização e indexação na memória. Kolodner (1993) apresenta diferentes critérios que devem ser considerados ao definir a representação de um caso, tais como: a

funcionalidade assegura uma representação em que somente os dados úteis ao sistema serão armazenados; a facilidade de aquisição de conhecimento assegura que somente informações com facilidade de aquisição de conhecimento sejam representadas no caso.

DESENVOLVIMENTO DO SISTEMA "AFIA"

Esta seção apresenta o desenvolvimento do sistema AFIA (Análise Forense com Inteligência Artificial) que tem como objetivos: criar uma base de casos de análise forense normalizada; sugerir ao usuário uma lista de características a serem analisadas durante uma investigação; e, por fim, através da técnica de RBC, encontrar rapidamente as possíveis soluções para corrigir e/ou previnir um ataque ocorrido baseando-se nos casos previamente cadastrados no sistema.

Aquisição de Conhecimento

Uma das fontes de informação utilizadas para compor os casos da base de conhecimento do sistema AFIA foi o software Autopsy (SLEUTHKIT, 2010), através do qual foram extraídas informações de discos rígidos contendo sistemas invadidos. Para extrair as informações de um crime digital a partir de um disco rígido foram utilizados casos do The Forensic Challenge (HONEYNET, 2010). Durante o processo de análise dos casos do HoneyNet.org, foram descobertas técnicas de atacantes que mascaram algumas informações e dificultam a descoberta da invasão. Mesmo assim muitas características de ataques foram reveladas, formando a primeira visão sobre o que é e como ocorrem os ataques.

Na busca de novas referências, especialmente novos casos para compor a base de conhecimento do sistema desenvolvido, consultou-se um conjunto de especialistas em segurança. As principais conclusões extraídas a partir destas interações são representadas pelos tópicos descritos a seguir.

- *Todos seguem praticamente as mesmas etapas durante uma análise forense: coleta de evidências (preservar); aquisição da evidência (duplicação forense); metodologia de investigação (o que será alvo de análise); classificação dos dados coletados; filtragem dos dados que são efetivamente uma prova do fato investigado; redação do laudo.*

- *Não foi identificada uma técnica padrão para procura e análise de evidências, sendo que a grande variedade de software e hardware exige que o analista tenha um conhecimento muito variado e adapte-se a inúmeras situações.*

- *Os tipos de informações pesquisados podem ser classificados como dados voláteis que são data, hora, conteúdo da memória, configuração da rede, processos em execução, arquivos abertos, sessão de login (pesquisa realizada com o sistema atacado ainda em operação); dados não-voláteis, como registros temporários e de eventos, arquivos de configuração e outros arquivos.*

- *A maioria concorda que uma análise forense não deve restringir-se apenas à análise de discos rígidos, os resultados normalmente podem ser mascarados e são insuficientes para atingir uma conclusão precisa sobre o incidente. Nenhuma fonte de informação deve ser ignorada, se possível até informações dos usuários devem ser levadas em consideração.*

- *Durante o processo são utilizados vários softwares para análise das informações.*

- *Todos desconhecem a existência de uma base de casos formalizada. O que encontra-se na Internet são descrições de casos e desafios de análise forense que não exigem nenhuma formalização na resposta.*

Conclui-se que a aquisição de conhecimento é uma tarefa complexa pela falta de padronização das informações e a vasta variedade de técnicas para análise das evidências.

Representação dos Casos

Diante das fontes de informação apresentadas anteriormente, conclui-se que os casos da base de conhecimento não devem ser caracterizados apenas por um conjunto de atributos padrão. Deve ser construído um conjunto flexível, permitindo a inclusão de novos atributos conforme a necessidade. Através desta flexibilidade é possível analisar diversos tipos de informações, tanto de discos rígidos quanto de tráfego de rede e aplicar diversos algoritmos para calcular a semelhança entre casos.

Para permitir esta flexibilidade, o atributo de um caso terá um valor numérico para representar quantidades e um valor textual que será uma descrição, conforme os exemplos a seguir.

- *Atributo: "Serviços e/ou Portas Invadidas"*
  - *Valor Numérico: 3*
  - *Valor Descritivo: "http, ftp, TCP 465"*
- *Atributo: "Arquivos executáveis modificados"*
  - *Valor Numérico: 1*
  - *Valor Descritivo: "ssh (trocado por versão com back port)"*

Esta flexibilidade de atributos e valores será implementada na ferramenta seguindo o modelo UML da Figura 2.

A lista de atributos que irá caracterizar cada caso será criada a partir de um conjunto de questões padrão. Por isso a classe Atributo é derivada da classe Pergunta dentro do diagrama UML e cada atributo será vinculado a uma pergunta no banco de dados através de uma chave estrangeira.

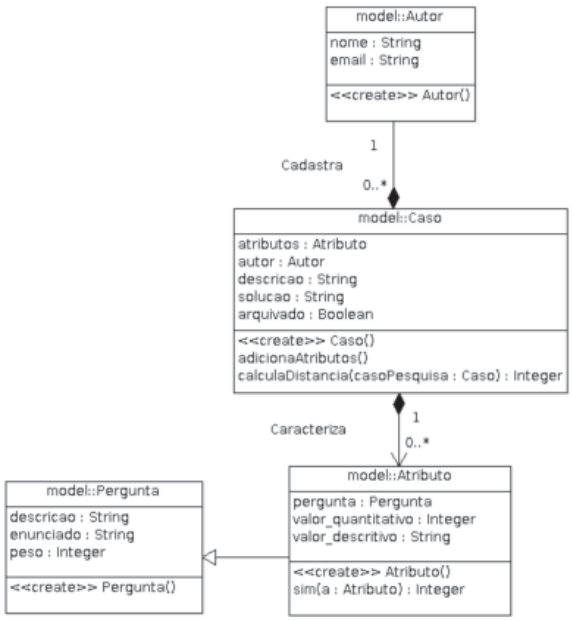

Figura 2. Modelo UML para representação dos casos

Recuperação dos Casos

A primeira proposta de algoritmo será o "vizinho-mais-próximo", calculado mesmo com uma quantidade de atributos variáveis. Somente um longo período de testes, com uma base de casos mais extensa, poderá definir se este realmente será o melhor algoritmo.

O cálculo da distância entre os casos da base e o novo problema baseia-se no princípio euclidiano de medidas conforme a equação a seguir.

$$D_{?q,c?} = \sqrt{\sum_{f=1}^{n} w_f * ?\text{sim}\ ?q_f, c_f?^2}$$

Onde:
q: novo problema;
c: caso da base;
f: atributo;
w: peso;
sim: função de similaridade local.

Conforme mencionado anteriormente, cada atributo possui dois valores: um numérico e outro textual. Neste primeiro momento, os valores numéricos não serão subtraídos como forma de quantificar a diferença entre eles, será comparada a igualdade como no caso dos valores texto, conforme os valores apresentados na Tabela 1.

Tabela 1. Similaridade de Atributos

	vdc != vdq	vdc == vdq
vqc != vqq	2	1
vqc == vqq	1	0

Onde:
vq: valor quantitativo;
vd: valor descritivo;
q: novo problema;
c: caso da base.

O objetivo da Tabela 1 é tornar relevante todos os valores de um atributo, pois qualquer atributo pode causar a seguinte dúvida: "O que é mais relevante? O número ou a descrição?". Por exemplo, pode ser considerado que: "É mais relevante o número de arquivos modificados ou uma lista dos nomes destes?" ou "É mais relevante o nome do usuário com alto número de senhas digitadas erradas ou a quantidade de tentativas?".

Implementação do Sistema

O sistema foi desenvolvido utilizando a linguagem *Ruby*, o *framework Rails* e o banco de dados *Sqlite*, visto que a parte de raciocínio foi realizada dentro do programa. A interface do sistema AFIA é *web* para permitir a distribuição da informação em uma intranet e talvez no futuro disponibilizar esta ferramenta para que alguma comunidade de especialistas melhorem sua base de casos.

A Figura 3 ilustra a página inicial do sistema AFIA, apresentado seus principais módulos.

Figura 3. Página inicial do sistema AFIA

Módulo Especialista

Somente o administrador do sistema terá acesso ao Módulo Especialista, que possui dois elementos fundamentais da ferramenta: Questionário Padrão e Base de Casos. O administrador deve ser um especialista que atuará como um avaliador dos atributos e dos casos, impedindo ambiguidades e duplicidades de significados.

Questionário Padrão

O questionário padrão compreende um conjunto de perguntas que são transformadas nos atributos dos casos. A Figura 4 apresenta as perguntas que estão cadastradas no sistema. Cada pergunta, ou atributo, possui os seguintes campos:

- *descrição: nome do atributo, uma descrição resumida do que significa o atributo;*
- *enunciado: pergunta que irá esclarecer melhor o significado deste atributo e possíveis valores;*
- *valor: peso deste atributo, utilizado no algoritmo de procura pelos casos semelhantes.*

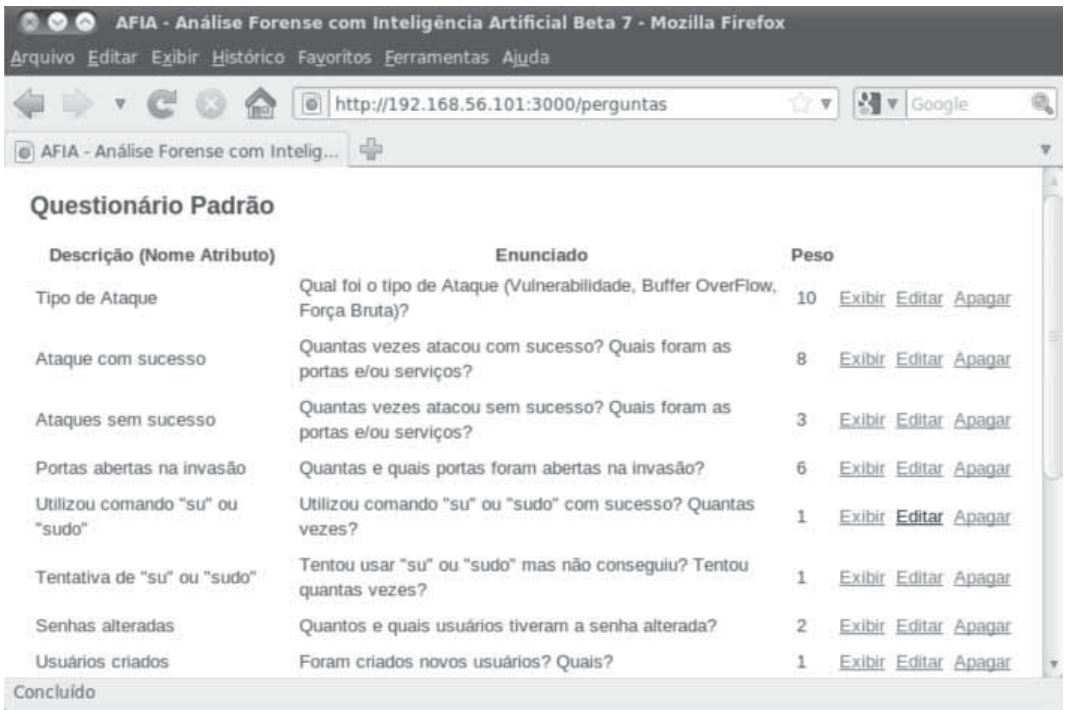

Figura 4. Questionário Padrão

O administrador deve estar consciente de que a mudança do valor do campo "valor" influencia bruscamente na eficiência da recuperação dos casos, mas recomenda-se que sempre que algum atributo for adicionado, ou excluído, ocorra uma reavaliação dos pesos de todas as perguntas.

Base de Casos

A Base de Casos compreende todos os casos cadastrados no sistema, ou seja, tanto os casos da base de conhecimento quanto os casos a serem investigados. A Figura 5 exemplifica esta lista de casos.

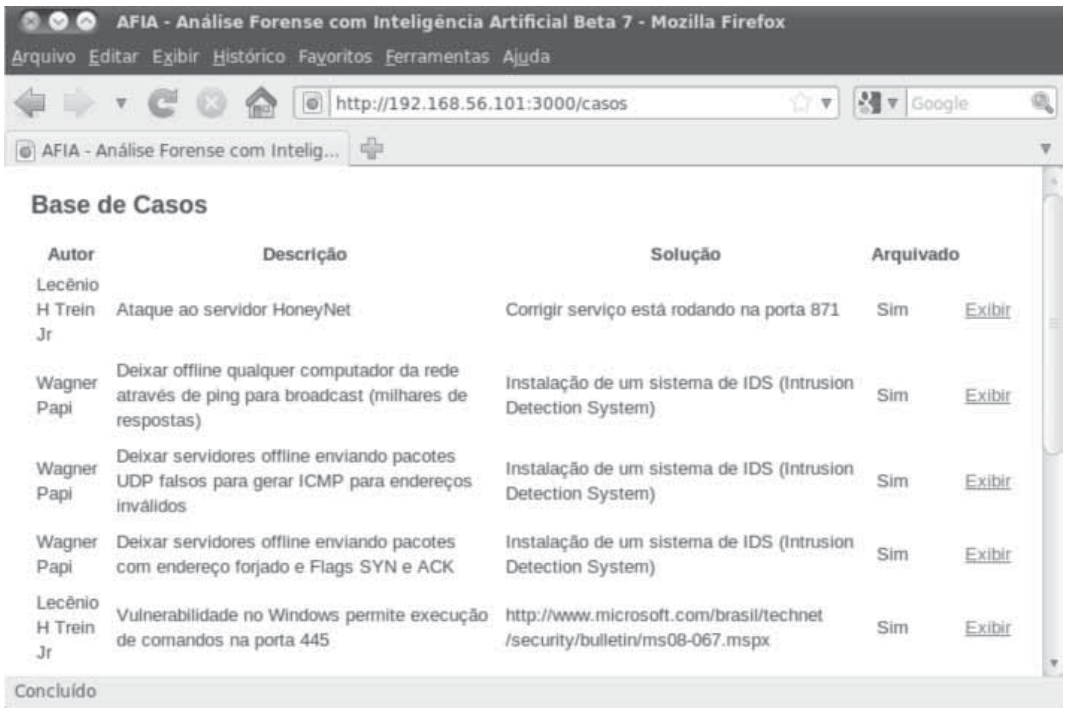

Figura 5. Base de Casos

Cada caso possui os seguintes campos:

- *autor: usuário que cadastrou o caso;*
- *descrição: uma descrição resumida do caso;*
- *solução: local onde se encontra a solução ou a descrição da mesma;*
- *arquivado: quando definido como "sim", define que este caso pertence a base de conhecimento, caso contrário é um caso a ser analisado, uma análise, que precisa de uma solução.*

Quando o administrador clica no *link* "Exibir", será redirecionado para a página ilustrada na Figura 6. Esta página exibe todas as informações do caso e permite acesso para a edição destas informações. Como observa-se na Figura 6, os atributos são exatamente os itens do questionário padrão, mas ao invés da pergunta, a descrição é apresentada.

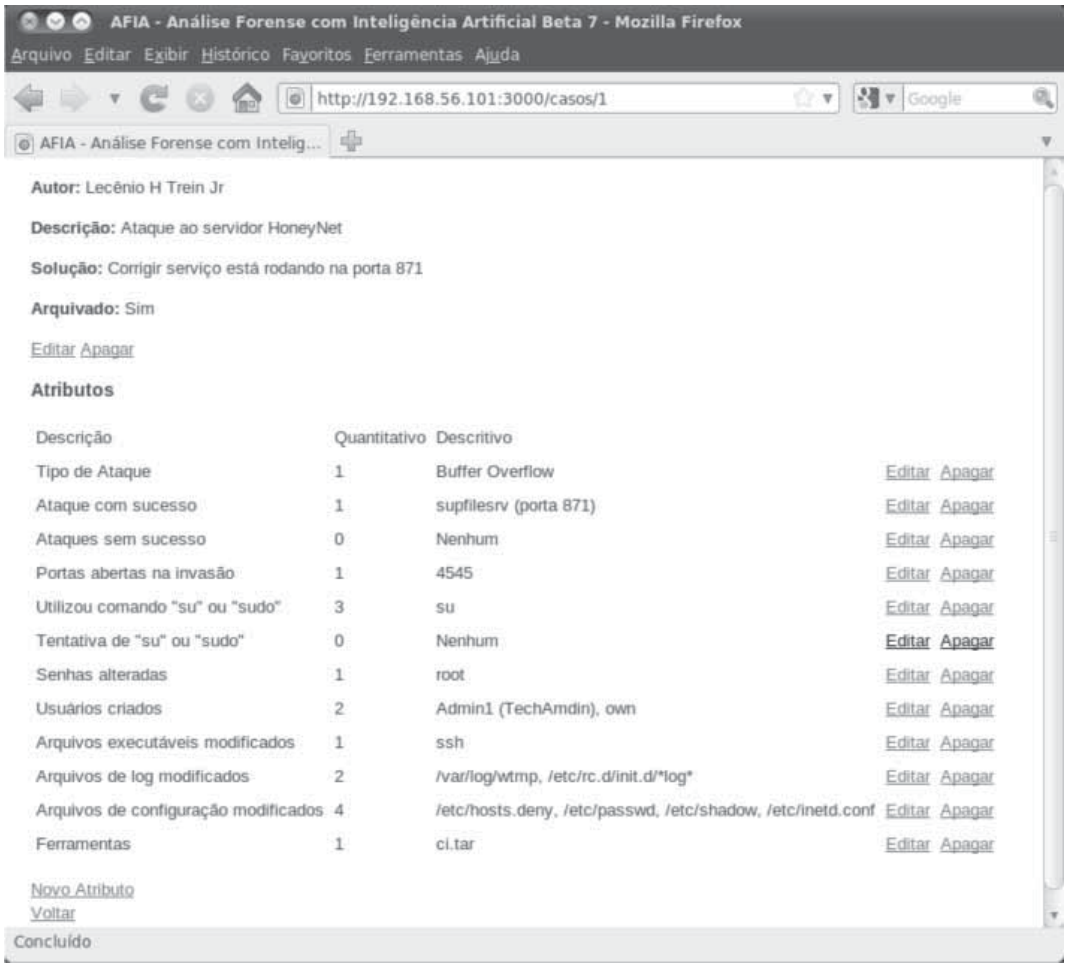

Figura 6. Exemplo Caso 1

Módulo Usuário

O Módulo Usuário é utilizado pelos analistas e é muito semelhante à base de casos, com a diferença de que não são disponibilizados os casos arquivados, ou seja, não acessa a base de conhecimento. Os casos disponíveis são os casos a serem solucionados, que nesta parte do sistema são chamados de "Investigações" conforme a Figura 7.

Figura 7. Investigações

Quando algum caso é considerado útil para a base de casos, os administradores também poderão acessá-lo através da página de "Base de Casos" do módulo "Especialista" e arquivar o caso, tornando-o mais uma parte da base de conhecimento e deixando de ser exibido neste módulo.

Ao clicar no *link* "Pesquisa", o sistema irá recuperar todos os casos da base e ordená-los pela distância entre os casos, ou seja a semelhança entre eles, conforme ilustra a Figura 8.

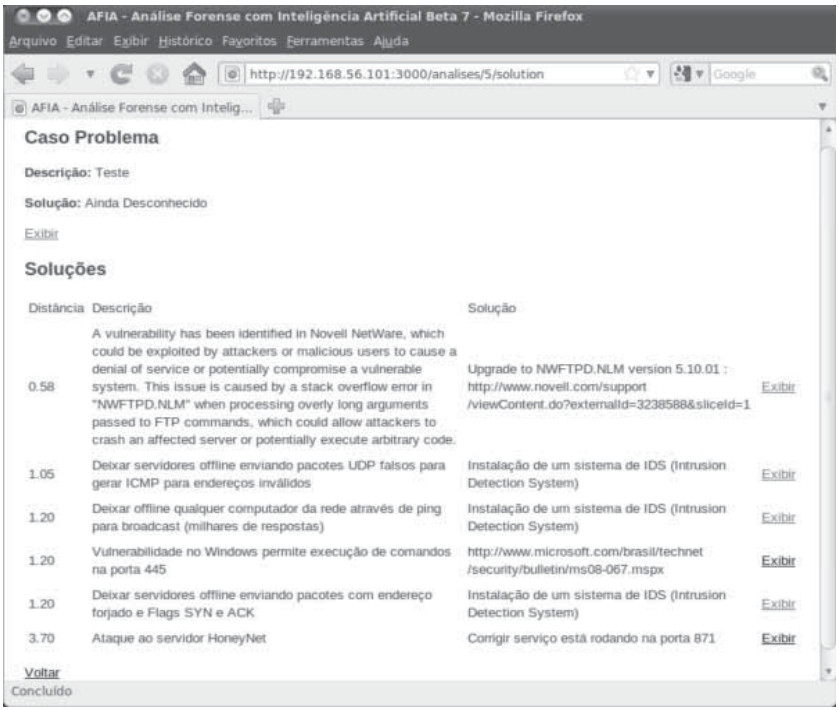

Figura 8. Soluções ordenadas por distância (semelhança)

CONCLUSÃO

Devido a uma grande variedade de ataques e constantes mudanças nas tecnologias da informação, o trabalho dos investigadores forenses requer uma larga base de conhecimento e experiência. Este fato motivou a criação de um sistema que armazene este conhecimento de forma estruturada e que consiga comparar um novo caso de ataque com casos já conhecidos. Isto permite uma análise das soluções já utilizadas em incidentes similares, possivelmente indicando mais de uma solução para o caso em análise.

Por outro lado, este fato foi também o grande desafio do presente trabalho. O domínio é amplo e composto por informações que não podem ser simplificadas apenas por "sim" ou "não", números ou palavras. O sistema AFIA permite o cadastro de casos com atributos dinâmicos, adaptando-se às mais diversas necessidades, com valores compostos por número e texto, permitindo combinações de valores de múltiplos significados.

Esta diversidade de atributos possíveis também cria a necessidade de disponibilizar ao analista uma lista de atributos a serem analisados, desenvolvidos e definidos no projeto como "Questionário Padrão".

Apesar da modelagem e dinâmica proposta no sistema, o principal desafio do projeto ainda persiste em formar sua base sólida de casos. A Internet confirma-se como a mais ampla e atualizada fonte de informações, especialmente pelas inúmeras comunidades sobre segurança e análise forense. Contudo, essa informação não está formalizada, as publicações fazem referência a técnicas e não apresentam descrições de casos reais.

A solução para a etapa de aquisição do conhecimento, que possibilitará a formação da base de casos, pode estar na abertura do projeto para alguma comunidade de especialistas interessados na formalização e divulgação de seu conhecimento através do sistema proposto.

Com base nos testes realizados durante o projeto propõe-se alguns itens a serem investigados no futuro, tais como disponibilizar o sistema para alguma comunidade de especialistas para formar uma boa base de casos; e implementar um mecanismo de filtragem para exibir apenas os casos mais próximos, visto que a base de casos aumenta durante a utilização da ferramenta.

REFERÊNCIAS

AAMODT, A.; PLAZA, E. Case-Based Reasoning: Foundational Issues, Methodical Variations and System Approaches. **AI Communications**, v. 7, n. 1, 1994.

HONEYNET. **The Forensic Challenge.** Disponível em: <http://old.honeynet.org/challenge/index.html> Acesso em: 19 ago. 2010.

KOLODNER, J. **Case-based reasoning**. San Mateo: Morgan Kaufman, 1993.

SLEUTHKIT. **Sleuth kit (TSK) & Autopsy**: Open Source Digital Investigation Tools. Disponível em: <http://www.sleuthkit.org> Acesso em: 10 abr. 2010.