LOGA-DM: UMA ABORDAGEM DE ANÁLISE DINÂMICA DE LOG COM BASE EM MINERAÇÃO DE DADOS

ROGER DA SILVA MACHADO[1]; RICARDO BORGES ALMEIDA[1]; ADENAUER CORRÊA YAMIN[1]; ANA MARILZA PERNAS[1]

[1]*Universidade Federal de Pelotas, Centro de Desenvolvimento Tecnológico – CDTec, {rdsmachado, rbalmeida, adenauer, marilza}@inf.ufpel.edu.br*

1. INTRODUÇÃO

O paradigma de computação ubíqua foi apresentado por Weiser (1991), que o definiu como um modo de se prover computação de uma forma transparente ao usuário, estando o modelo computacional plenamente integrado as demandas do usuário. Devido à característica de mobilidade e a decorrente troca de infraestrutura de acesso, o que resulta em um aumento do uso de diferentes redes de computadores e do número de aplicativos em execução nessas infraestruturas, a segurança de redes torna-se cada vez mais importante, potencializando a preocupação relacionada com a segurança desses sistemas, bem como da informação que trafega por eles.

Uma das tarefas relevantes para segurança da informação é a análise de log, que é uma técnica utilizada para compreender melhor o funcionamento do sistema, visando detectar tentativas de ataques, ações realizadas por um invasor, entre outras (HOEPERS; STEDING-JESSEN, 2003). Os arquivos de log contém uma ordem cronológica dos eventos gerados por determinadas aplicações, e as informações contidas num registro de log variam de acordo com a aplicação que gerou o evento.

Com o fato de os registros de log possuírem diferentes formatos e informações distintas em cada log, a tarefa de análise de log passa a não ser uma tarefa trivial, já que se torna necessário o conhecimento por parte do analista de rede sobre as informações contidas nos registros e seus respectivos formatos. Além disso, os arquivos de log tendem a possuir um tamanho significativo, já que são gerados inúmeros registros de suas atividades, o que torna a análise manual destes registros impraticável para analistas humanos. Entretanto, existem várias técnicas destinadas a reduzir a quantidade de registros a ser analisada, e assim facilitar a análise de log. Dentre as técnicas destinadas a tal tarefa, este trabalho explora as técnicas de mineração de dados.

Algoritmos de mineração de dados aplicados em registros de log automatizam a tarefa de filtragem, reduzindo a carga de trabalho do analista (GRÉGIO, 2008). Abordagens utilizando mineração de dados estão sendo pesquisadas e aplicadas à análise de log (CAMPOS; LIMA, 2012; CORREIA, 2004; GRÉGIO, 2008).

Este trabalho propõe uma abordagem que auxilie a análise de registros de log de aplicações e do tráfego da rede dos servidores utilizados no projeto AMPLUS (*Automatic Monitoring and Programmable Logging Ubiquitous System*). Com a utilização da abordagem melhora-se o entendimento a respeito dos sistemas computacionais por meio do monitoramento dos registros de log das aplicações que executam nos sistemas computacionais presentes no projeto. Além disso, com o monitoramento do tráfego da rede e com a aplicação de uma técnica de mineração de dados nos registros do tráfego de rede coletados, é possível detectar tentativas

de ataques contra os sistemas, dessa forma melhorando a segurança dos servidores do projeto AMPLUS.

2. METODOLOGIA

Sistemas Ubíquos possuem uma necessidade elevada de conectividade e, com isso, a segurança destes sistemas se torna cada vez mais importante. De forma a melhorar a segurança destes sistemas, este trabalho desenvolveu uma abordagem para facilitar a análise de log.

Existem tarefas básicas que devem ser providas de forma a prover a análise de log (CLEMENTE, 2012). Com isso, nesta abordagem se propôs tratar as seguintes funcionalidades:

- análise léxica, que é o processo de analisar as linhas de mensagens de log com objetivo de produzir uma saída formatada em um padrão mais adequado para futuro processamento;
- análise sobre eventos de log, onde é aplicada a técnica de árvores de decisão, com o intuito de classificar os registros coletados;
- transmissão, que consiste em transmitir os registros de log para um servidor central;
- armazenamento, que é o processo que compreende em armazenar os registros de log para futuras consultas.

Neste trabalho, optou-se por diferenciar os registros de log de aplicações dos registros de tráfego de rede, os quais seriam registros de log das camadas de rede e transporte. Com isso, o monitoramento do tráfego da rede é realizado de maneira diferente dos registros de log de aplicações, e este monitoramento possui como finalidade identificar possíveis tentativas de ataques aos servidores. A identificação das tentativas de ataque é realizada através da utilização da técnica de classificação por árvores de decisão.

A Figura 1 apresenta a arquitetura do analisador de log desenvolvido, apresentando os módulos implementados. Pode-se observar na Figura 1 que a arquitetura possui dois fluxos de dados, sendo o primeiro do coletor de log e o segundo do coletor do tráfego da rede, mostrando os tratamentos realizados de maneira diferente para os registros de log de aplicações e os registros de tráfego da rede.

Figura 1: Arquitetura proposta

Os registros de log de aplicações, após coletados, são repassados para o módulo de análise léxica/sintática que é responsável por realizar os processos de normalização e contextualização dos dados presentes nos registros através de expressões que foram desenvolvidas para os respectivos formatos dos registros de log. Os registros do tráfego da rede, depois de capturados, são repassados para o módulo de classificação, o qual é responsável por classificar as conexões no momento de sua captura, utilizando a técnica de árvores de decisão.

Após os registros serem processados, eles são encaminhados para um servidor central onde são armazenados em um banco de dados, para possibilidade de posterior análise.

3. RESULTADOS E DISCUSSÃO

Com o objetivo de testar o módulo de classificação, optou-se por utilizar um conjunto de treinamento e um conjunto de teste disponível em (KDD, 1999), sendo este um dos principais conjuntos de dados utilizados para este tipo de trabalho. Nos testes realizados cada conexão pode ser classificada em uma das cinco conexões presentes no conjunto de treinamento, sendo elas: normal, DOS (*Denial of Service*), R2L (*Remote to Local*), U2R (*User to Root*), proibing.

Foram desenvolvidos dois classificadores utilizando a técnica de árvores de decisão, sendo que o primeiro trabalha com todos os atributos presentes no conjunto de treinamento e o segundo trabalha com um grupo reduzido de atributos.

Na Tabela 1 é apresentada uma comparação entre os resultados obtidos entre o classificador utilizando todos os 41 atributos presentes nos arquivos e o classificador que utiliza somente 5 atributos.

Tabela 1: Resultados obtidos pelos classificadores

Categoria	Classificador com todos os atributos	Classificador com atributos reduzidos
Normal	98,18%	98,68%
DOS	99,99%	99,93%
R2L	17,95%	53,85%
U2R	25,71%	15,38%
Proibing	99,20%	68,66%
Acertos Geral	98,07%	97,68%

De forma geral, os resultados foram semelhantes. As diferenças notadas são nas categorias Proibing e U2R, onde, na categoria Proibing, o classificador com atributos reduzidos teve um desempenho relativamente pior, o que se deve em grande parte à eliminação dos atributos calculados, sendo analisadas as demais conexões em uma janela de 2 segundos, já que esta categoria de ataque costuma gerar uma variedade de conexões em um intervalo pequeno de tempo. Enquanto que, na categoria U2R, o classificador com atributos reduzidos alcançou um desempenho superior em relação ao outro classificador. Acredita-se que esta melhora se deve ao fato da eliminação dos atributos, pois possivelmente alguns destes atributos estavam dificultando o aprendizado das classificações das conexões da categoria U2R.

Apesar do classificador com atributos reduzidos ter alcançado resultados um pouco inferiores em relação ao outro classificador, ele apresenta a vantagem de poder ser aplicado no momento da coleta das conexões, não sendo necessário outro tipo de análise para calcular valores de outros atributos. Com os resultados alcançados, o classificador demonstra ser de grande utilidade, pois pode ser

utilizado como um filtro para diminuir o número de registros que devem ser analisados por parte do administrador da rede. Além disso, pode ser utilizado para detectar ataques a rede, facilitando ao administrador a detecção deste tipo de tentativa de intrusão.

4. CONCLUSÕES

Com o intuito de automatizar a análise de log, este trabalho desenvolveu uma abordagem para coleta dos registros de log e aplicação de expressões de maneira a definir o formato do registro, visando realizar a normalização e a contextualização dos dados presentes no mesmo. Ainda, foi realizada a modelagem da arquitetura de um servidor central para armazenamento dos registros coletados em um banco de dados, onde poderão ser mantidos os registros para análises posteriores, mantendo um histórico do funcionamento do sistema.

Na descrição da arquitetura da abordagem desenvolvida foi caracterizada a opção de tratar de forma distinta os registros de log de aplicações dos registros da camada de rede e de transporte. Sendo assim, passou-se a ter dois fluxos de dados na arquitetura desenvolvida.

Um dos diferenciais deste trabalho é a possibilidade de classificar as conexões no momento de sua captura, ao contrário de outros trabalhos que se propõem a realizar a classificação somente de registros históricos. Nos testes realizados o desempenho do classificador referente à taxa de acertos foi satisfatório, demonstrando que o classificador utilizando a técnica de árvores de decisão pode ser utilizado para classificar as conexões capturadas, trazendo um novo mecanismo para facilitar a tomada de ações por parte do administrador dos sistemas.

5. REFERÊNCIAS BIBLIOGRÁFICAS

CAMPOS, L. M. L.; LIMA, A. S. Sistema para Detecção de Intrusão em Redes de Computadores com uso de Técnica de Mineração de Dados. **V Congresso Tecnológico Infobrasil, Fortaleza.** [S.l.], 2012.

CLEMENTE, R. G. **Uma arquitetura para processamento de eventos de log em tempo real**. 2008. Dissertação de Mestrado — Pontifícia Universidade Católica do Rio de Janeiro - PUC-RIO.

CORREIA, L. J. **Mineração de dados em arquivos de log gerados por servidores de páginas Web**. 2004. Monografia de Graduação Bacharelado em Ciência da Computação - Universidade Regional de Blumenau/FURB, Blumenau/SC.

GRÉGIO, A. R. A. **Aplicação de Técnicas de Data Mining para a Análise de logs de Tráfego TCP/IP**. 2008. Dissertação Mestrado do Curso de Pós- Graduação em Computação Aplicada — Instituto Nacional de Pesquisas Espaciais/ INPE, São José dos Campos/SP.

HOEPERS, C.; STEDING-JESSEN, K. **Análise e Interpretação de logs**. NIC BR Security Office (NBSO) Comitê Gestor da Internet no Brasil.

KDD Cup. Acessado em 01 out. 2013. Online. Disponível em: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.