URLBlackList Lite: Uma lista enxuta de catalogação baseada na URLBlackList

Nilson Mori Lazarin, Tielle da Silva Alexandre

CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca
Av. Roberto da Silveira, 1900 – 28.635-000 – Nova Friburgo – RJ – Brazil

Abstract. *The URLBlackList is one of the cataloging lists available for content controlling; however, it has domains unsolved and anomalies, which need to be repaired to optimize the content controlling process. Thus, the objective of this work is to present an automatized process of refinement of the URLBlackList. As solution, it is presented a web system capable of segmenting domains from the URLBlackList in two groups: the solved domains and the unsolved domains, besides it purges any anomaly found. After the refinement process, the system will allow the generation of a lean list composed only by solved domains. A comparative and qualitative analysis between the list generated by the URLBlackList Lite and the URLBlackList is also shown.*

Resumo. *As ferramentas de controle de conteúdo, tais como Proxy, são altamente dependentes de uma boa lista de catalogação de sites. Uma das listas disponíveis para download é a URLBlackList, entretanto, ela possui domínios cadastrados que não estão registrados (não resolvíveis via DNS), além de outras anomalias que afetam o custo computacional no processo de controle de conteúdo. Este trabalho apresenta uma ferramenta que através de um processo automatizado de refinamento, da URLBlackList, e segmentação dos domínios pertencentes à mesma em dois grupos: os domínios resolvíveis e os não resolvíveis, além de expurgar qualquer anomalia encontrada. Após o processo de refinamento, a ferramenta possibilita a geração de uma lista enxuta composta somente por domínios resolvíveis. Uma análise comparativa e qualitativa entre a lista gerada pela URLBlackList Lite e a lista da URLBlackList também será apresentada.*

1. Introdução

Um servidor *Proxy* atua como um intermediador entre os computadores de uma rede local e a Internet, analisando todas as requisições recebidas [MORIMOTO 2009]. O processo de filtragem de conteúdo por meio de um servidor *Proxy* ocorre através da comparação das requisições do cliente com uma lista de *Uniform Resource Locator* (URL) ou domínios. Para isso, regras de acesso são configuradas de forma a autorizar ou não o acesso à determinada página solicitada pelo usuário. Quando um servidor *Proxy* recebe uma requisição de uma página, inicia-se um processo de comparação entre a requisição e todas as linhas de domínio presentes na *URLBlackList* [FOROUZAN 2008].

O projeto *URLBlackList* (2015) é uma lista de catalogação composta por um conjunto de listas de URL, domínios e expressões subdivididas em, aproximadamente,

100 categorias. A lista de catalogação da *URLBlackList* foi escolhida para o estudo e a análise desse projeto por ser uma lista de catalogação extensa, popular entre os administradores de redes e por disponibilizar, frequentemente, uma versão gratuita. Entretanto, a lista da *URLBlackList* apresenta domínios catalogados que estão inativos e que consequentemente consomem processamentos desnecessários por parte do servidor *Proxy*.

Outro fato da *URLBlackList* é a sua origem norte-americana, muitas vezes não atendendo o contexto em redes brasileiras e provocando certas distorções de entendimento. Por exemplo, a categoria *Guns* faz referência às armas de pequeno porte e de livre comércio, enquanto que a categoria *Weapons* contempla artefatos de guerra. Como no Brasil qualquer porte de arma deve passar por um processo de legalização, estas categorias poderiam ser unidas adequando-as ao contexto brasileiro e assim, diminuindo a duplicidade de registros na lista.

Na *URLBlackList*, um domínio pode ser encontrado em mais de uma categoria, ocasionando uma redundância de verificação no servidor *Proxy.* Caso seja necessário estabelecer um bloqueio de acesso a conteúdos pornográficos, provavelmente, deverá ser incluso as categorias *Adult*, *Sexuality* e *Porn*, entretanto, vários domínios existem em concomitância nestas categorias (e.g. *sexwork.com*). Além disso, é possível encontrar endereços IP onde deveriam existir apenas domínios. Destaca-se ainda a existência de domínios com anomalias de formato, como por exemplo, *googlex..com* e *relato-sexo..com* (possuem dois pontos consecutivos). Esses domínios ocasionam falhas de resolução por ser tratarem de domínios inválidos.

Portanto, o objetivo desse trabalho é apresentar um processo capaz de refinar a lista da *URLBlackList* expurgando ou tratando as adversidades encontradas com intuito de fornecer uma lista de catalogação enxuta que proporcionará a otimização do processo de filtragem de conteúdo realizado por servidor *Proxy.* Para isso uma ferramenta, a *URLBlackList Lite,* foi desenvolvida com o intuito de implementar o processo de refinamento possibilitando gerar, após o refinamento, uma lista catalogada enxuta apenas com domínios resolvíveis. Além disso, a ferramenta desenvolvida possibilita uma reclassificação das categorias da *URLBlackList* a fim de proporcionar uma melhor adequação a diversos contextos.

Este trabalho está estruturado da seguinte forma: na seção 2 será apresentada a *URLBlackList Lite* e a sua metodologia; na seção 3, serão descritos os experimentos realizados com a *URLBlackList Lite* comparados com a *URLBlackList*; a conclusão é apresentada na seção 4.

2. Metodologia

Nesta seção serão apresentados os métodos e as etapas para solucionar os problemas identificados na *URLBlackList,* possibilitando: a eliminação de endereços IP existentes; a segmentação a *URLBlackList* em domínios resolvíveis e não resolvíveis; o tratamento dos domínios com formatos ilegais e as anomalias de resolução; a análise e o tratamento dos domínios redundantes; e a adequação da lista a vários contextos.

A metodologia utilizada no projeto *URLBlackList Lite* é dividida em: processo de refinamento; processo de análise de redundância; e processo de reclassificação das categorias. Os processos de refinamento e análise de redundância serão os responsáveis por refinar, tratar e gerar a *URLBlackList Lite*, enquanto que o processo de

reclassificação adequará a *URLBlackList* ao contexto brasileiro. A figura 1 ilustra a metodologia proposta evidenciando as etapas que um domínio percorre até o final do processo.

2.1. Processo de Refinamento

O processo de refinamento consiste na análise sequencial dos domínios pertencentes à *URLBlackList*, usando a ferramenta DIG do pacote BIND para a resolução de domínio, com o objetivo de segmentar a lista da *URLBlackList* em dois grupos de domínios distintos: os domínios resolvíveis e os não resolvíveis [BIND 2015]. Esse processo tem como *input* a *URLBlackList* descompactada e recebe para processamento um domínio por vez. O processo de refinamento é dividido em duas etapas: a identificação dos IP e a resolução de domínio.

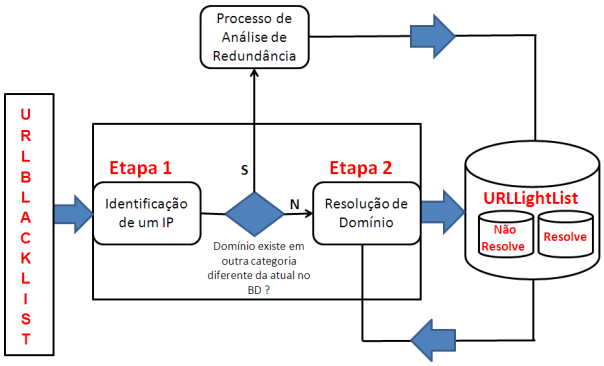

Figura 1. A metodologia da *URLBlackList Lite*.

A identificação de IP visa identificar e expurgar os endereços IP presentes nas listas de domínio da *URLBlackList*. Para isso, todos os registros da *URLBlackList* passam, inicialmente por esta etapa, para que um IP não seja resolvido e pertença a uma lista que deveria conter apenas domínios. Sendo assim, apenas domínios prosseguem para a próxima etapa, onde é identificado se um domínio já existe na base de dados. Se o domínio não existir no armazenamento, o mesmo é direcionado para a segunda etapa do processo de refinamento. Se for identificado que esse domínio já existe no armazenamento, ele é direcionado para o processo de análise de redundância.

A resolução de domínio pode receber como *input* os domínios oriundos de dois casos. No primeiro caso, recebe como *input* os domínios que passaram pela etapa de identificação de IP e não existem no armazenamento. Neste caso, estes domínios são submetidos à etapa de resolução de domínio e se houver uma resposta do comando de resolução (não ocorrer erros), então estes domínios serão direcionados para o armazenamento, sendo relacionados a um dos segmentos (resolvíveis ou não resolvíveis) de uma categoria. Se houver uma ocorrência de erro no processo resolução esses domínios serão expurgados. O resultado desse caso é uma lista enxuta, livre de anomalias e composta por dois segmentos, os domínios resolvíveis e os domínios não resolvíveis.

No segundo caso, o processo de resolução de domínios pode receber como *input* os domínios existentes no banco de dados da *URLBlackList Lite* (Figura 1). Neste caso, cada domínio armazenado passa pela etapa de resolução de domínio podendo mudar de um segmento resolvível para um não resolvível e vice e versa. É importante ressaltar que esse processamento apenas atualiza o *status* de resolução dos domínios a fim de se manter uma base de dados consistente.

2.2. Processo de Análise de Redundância

Após a primeira etapa do processo de refinamento, um mecanismo de identificação de domínios já existentes no armazenamento é realizado direcionando os domínios redundantes para o processo de análise de redundância. Dessa forma, esse processo tem por objetivo tratar as redundâncias de um domínio que estão associados a mais de uma categoria na *URLBlackList*, evitando assim, que um servidor *Proxy* verifique o mesmo domínio mais de uma vez.

O método de tratamento de redundância será embasado na atribuição de pesos para cada categoria existente na *URLBlackList*. O peso é um valor numérico associado à categoria e que representa o grau de perversidade dessa categoria. Entende-se por grau de perversidade: o quão perverso representa o conteúdo das páginas de uma categoria, em relação às políticas de acesso definidas pelo administrador de redes. As categorias que possuem conteúdos cujos acessos são inadmissíveis pelas políticas de acesso podem receber como peso um valor entre 7 e 10 (perversos). As categorias que possuem conteúdos que são considerados admissíveis pelas políticas de acesso mesmo contendo certas restrições podem receber como peso um valor entre 5 a 6 (moderado). Já as categorias que possuem como peso um valor entre 1 e 4 (baixo), o administrador de redes não possui preocupação em relação ao acesso do usuário a esse conteúdo.

Supondo que o administrador de redes deseje bloquear o acesso a conteúdos pornográficos atribuindo os valores 9, 10 e 10 como peso para as categorias *Sexuality*, *Porn* e *Adult*, respectivamente; e que a análise de redundância ocorra nessa mesma ordem. O domínio *sexinfo101.com* se encontra na lista de domínio destas três categorias. Sendo assim, a primeira vez que esse domínio passar pelo processo de refinamento será resolvido e direcionado para o armazenamento pertencendo ao segmento de domínios resolvíveis da categoria *Sexuality*. Na segunda vez que o domínio *sexinfo101.com* passar pelo processo de refinamento (associado à categoria *Porn*), este domínio não será submetido à etapa de resolução de domínio, entretanto, este será direcionado para o processo de análise de redundância.

Neste processo, a categoria que possuir o maior grau de perversidade (peso) prevalecerá sobre a categoria com o menor grau. Logo, o domínio *sexinfo101.com* será associado à categoria *Porn*, por possuir maior grau de perversidade. A terceira vez que o domínio *sexinfo101.com* passar pelo processo de refinamento (associado à categoria *Adult*), este também será direcionado para a análise da redundância. Neste caso, o domínio não mudará de categoria, pois o grau de perversidade das categorias é o mesmo. Ao final do processo, o domínio *sexinfo101.com* estará associado somente à categoria *Porn*, eliminando as redundâncias existentes.

Caso seja estabelecida uma regra para bloquear o acesso somente aos domínios pertencentes à categoria *Sexuality* ou *Adult,* o domínio *sexinfo101.com* não será bloqueado já que o mesmo foi desassociado dessas categorias. Neste caso, o grau de

perversidade da categoria deve ser modificado, de forma que essa categoria possua o maior peso. Para o exemplo acima, se for desejado bloquear somente a categoria *Adult*, o grau de perversidade da categoria *Porn* deverá ser alterado para um valor inferior a 10 como peso. Quando o processo de redundância for acionado, os domínios redundantes serão associados à categoria *Adult* já que esta possui o maior peso.

2.2. Processo de Reclassificação de Categorias

O processo de reclassificação de categorias recebe como *input* as categorias da *URLBlackList* e tem como objetivo possibilitar a correção de certas distorções de entendimento relacionadas a interpretação do significado de cada categoria. Isso ocorre porque as categorias da *URLBlackList* estão atreladas a linguagem e a cultura americana. Esse processo possibilitará que as categorias da *URLBlackList Lite* estejam adequadas a especificidade cultural brasileira.

Esse processo possibilita a escolha das categorias, que devem ser unificadas em uma única categoria devido ao contexto a que se referem. Sendo assim, esse processo receberá as categorias selecionadas pelo administrador de redes e o resultado é uma única categoria, nomeada de forma a atender a interpretação do significado e que possui o conjunto dos domínios pertencentes a cada categoria unificada. Por exemplo, para as categorias *Guns* e *Weapons* o resultado obtido é a criação de uma categoria Armas.

3. Experimentos e Resultados

Nesta seção, serão apresentados os resultados dos experimentos realizados com o objetivo de proporcionar uma análise comparativa entre as listas da *URLBlackList* e da *URLBlackList Lite*; comprovar a compatibilidade de execução da ferramenta nos sistemas operacionais *Windows* e *Linux*; e para demonstrar o índice de redução dos registros de domínios da *URLBlackList*.

A *URLBlackList Lite* é uma ferramenta *web* que visa disponibilizar as listas de catalogação enxuta e a própria ferramenta para *download* a fim de possibilitar que o administrador de redes configure a ferramenta de acordo com as políticas de acesso vigentes em uma determinada organização. A ferramenta possibilita a geração de uma lista de catalogação enxuta para cada operador de acordo com as reclassificações de categorias estabelecidas por este. A figura 2 ilustra as telas da ferramenta *URLBlackList Lite*; a primeira tela mostra o menu da ferramenta exibindo as funcionalidades disponíveis, como analisar um arquivo da *URLBlackList*, gerenciar o grau de perversidade das categorias, verificar a versão e realizar o *download* de uma lista da *URLBlackList*. A segunda tela mostra a interface para reclassificar as categorias da *URLBlackList.*

Figura 2. Telas da ferramenta *URLBlackList Lite*.

Os primeiros experimentos foram realizados no sistema operacional Linux devido ao fato do utilitário DIG ser nativo desse sistema operacional. Nessa experimentação inicial, a necessidade de que o sistema registrasse todo o seu funcionamento em um *log* foi evidenciada a fim de capturar qualquer comportamento desconhecido e o experimento foi executado via terminal do sistema operacional Linux. Em uma segunda fase de experimentação, a *URLBlackList Lite* teve o seu funcionamento verificado no ambiente Windows através da instalação do utilitário DIG neste sistema operacional [BIND 2015]. Depois que esses experimentos foram realizados, a *URLBlackList Lite* teve a compatibilidade de execução nos sistemas operacionais *Linux* e *Windows* devidamente comprovada.

A tabela 1 mostra os resultados obtidos com o processo de refinamento realizado pela ferramenta *URLBlackList Lite* tendo como *input* a versão da *URLBlackList* disponibilizada em 14 de maio de 2015. O processo analisou noventa e nove categorias, 2.901.374 domínios presentes nessa versão da *URLBlackList* e foi executado em, aproximadamente, cinco dias. Durante a execução do processo ocorreram algumas interrupções devido à falta de energia elétrica e conectividade, sendo assim, o processo foi reiniciado algumas vezes.

Recomenda-se que o processo de refinamento seja executado quando uma nova versão da lista da *URLBlackList* for disponibilizada. Como o tempo de execução do processo de refinamento é prolongado, uma nova versão da *URLBlackList* pode ser disponibilizada enquanto o processo da *URLBlackList Lite* ainda estiver executando uma versão anterior da *URLBlackList*. Quando julgar necessário, o administrador de redes poderá executar um processo de atualização do *status* de resolução dos domínios existentes no banco de dados.

Tabela 1. Resumo dos Resultados Obtidos

URLBlackList	2.901.374	100%
URLBlackList Lite Resolvível	1.202.452	41,44%
URLBlackList Lite Não Resolvível	756.096	26,06%
Domínios com formato ilegais/redundantes/IP	942.826	32,50
Índice redução (%)	58,56%	

O índice *URLBLackList* representa a quantidade de domínios encontrados na lista da *URLBLackList*; já os índices *URLBlackList Lite – Resolvível* e *URLBlackList*

Lite – Não Resolvível representam o quantitativo, respectivamente, de domínios resolvíveis e não resolvíveis identificados na lista da *URLBlackList*; o índice de domínios com formato ilegais/redundantes/anômalos/IP engloba a quantidade de domínios que foram expurgados devido as anomalias, IP e redundâncias encontradas na lista da *URLBlackList*; por fim, o índice de redução representa o percentual de registros retirados da *URLBlackList* revelando assim, o quanto a lista ou uma categoria ficou mais enxuta após o processo de refinamento da *URLBlackList Lite*.

Analisando a tabela 1, a ferramenta *URLBlackList Lite* identificou 756.096 domínios como não resolvíveis proporcionando uma redução de cerca de 26% em relação a lista da *URLBlackList*. Sendo assim, retirando os domínios não resolvíveis dos registros da *URLBlackList*, esta passou a ter 2.145.278 registros. O quarto índice, representa uma redução de 32,5 % nos registros da *URLBlackList* através do tratamento de redundância. Esse índice também abrange a quantidade de expurgos de domínios com formatos ilegais e com outras anomalias identificadas na etapa de resolução. Aplicando mais esse percentual de redução a *URLBlackList*, esta foi reduzida para 1.202.452 registros, totalizando um índice de redução de 58,56% proporcionado pelo processamento da *URLBlackList Lite*.

A tabela 2 mostra um *ranking (top10)* com as dez categorias analisadas que obtiveram os maiores índices de redução pelo processo de refinamento da *URLBlackList Lite*. Analisando a tabela 2 constata-se que: a categoria *Verisign* foi reduzida em 100%, pois os dois domínios existentes nessa categoria são não resolvíveis; as categorias *Malware*, *Phishing*, *Homerepair* possuem uma quantidade maior de domínios não resolvíveis do que domínios resolvíveis; as categorias *Malware*, *Arjel*, *Hacking*, *Aggressive*, *Spyware* e *Cleaning* apresentaram uma quantidade expressiva de domínios expurgados (índice de domínios com formato ilegais/redundantes/anômalos/IP) e a categoria *Adult* revelou um alto índice de redundância com a categoria *Porn* o que foi observado no arquivo de *log* da ferramenta.

Tabela 2. Resultado Obtido por Categoria

Categoria	*URLBlackList*	*URLBlackList Lite* Resolvível	*URLBlackList Lite* Não Resolvível	Domínios com formato ilegais/redundantes/IP	Índice redução (%)
Verisign	2	0	2	0	100%
Malware	340.030	25.633	53.176	261.221	92%
Arjel	69	7	4	58	90%
Hacking	581	90	25	466	85%
Phishing	121.388	18.711	82.098	20.579	85%
Aggressive	433	105	19	309	76%
Spyware	193	49	45	99	75%
Adult	997.238	272.664	195.824	528.750	73%
Cleaning	178	55	14	109	69%
Homerepair	21	7	9	5	67%

4. Conclusão

Esse trabalho apresentou uma ferramenta capaz de realizar um processo de refinamento na lista da *URLBlackList*, segmentando os domínios presentes nessa lista em domínios resolvíveis e não resolvíveis, expurgando IP e domínios com erros de resolução e tratando as redundâncias de domínios quando este pertencer a mais de uma categoria na lista da *URLBlackList*. Após o processo de refinamento, a ferramenta possibilita a geração da *URLBlackList Lite*, que é uma lista enxuta composta apenas por domínios resolvíveis, a qual será utilizada por um servidor *Proxy* para realizar um serviço de controle de conteúdo das páginas *web* acessadas por um usuário de uma determinada rede local.

Os resultados obtidos com os experimentos realizados se mostraram promissores já que a lista da *URLBlackList* foi reduzida em 58,56%, ou seja, os resultados comprovam o fato de que se a lista da *URLBlackList Lite* for usada por um servidor *Proxy* ao realizar o serviço de controle de conteúdo proporcionará uma otimização dos recursos computacionais utilizados por este serviço. Portanto, a ferramenta *URLBlackList Lite* conseguiu refinar a lista da *URLBlackList* e atingir o objetivo de gerar uma lista de catalogação efetivamente mais enxuta, a lista da *URLBlackList Lite*.

Como trabalhos futuros, o processo de refinamento poderá ser implementado empregando-se *Threads* para que mais de uma categoria seja analisada ao mesmo tempo, aumentando assim, a performance da ferramenta e diminuindo o tempo gasto em uma análise. Ao processo de analisar banco de dados e arquivo, um botão de parada deverá ser disponibilizado para que o usuário possa interromper o processo de análise quando desejar. Poderá ser implementado uma funcionalidade de agendamento para que a ferramenta dispare uma análise de banco de dados, automaticamente e um aperfeiçoamento do tratamento de erro do *time out*. Além disso, um experimento poderá ser realizado em um servidor Proxy, como o Squid, utilizando a lista de catalogação da *URLBlackList Lite* e a lista da *URLBlackList* a fim de mensurar e comparar o desempenho proporcionado pela processo de refinamento.

5. Referências

BIND. BIND. Disponível em: < https://www.isc.org/downloads/bind/>.Acesso em: 12 de junho de 2015.

FOROUZAN, Behrouz A. Comunicação de Dados e Redes de Computadores – Porto Alegre: Bookman, 2008.

MORIMOTO, Carlos Eduardo. Servidores Linux, guia prático – Porto Alegre: Sul Editores, 2009.

URLBLACKLIST. UrlBlackList. Disponível em: <http://urlblacklist.com/?sec=search>. Acesso em: 12 de junho de 2015.

KRASNER, G. E., POPE, S. T. A description of the Model-View-Controller user interface paradigm in the Smalltalk-80 System. 1988. Disponível em: www.create.ucsb.edu/~stp/PostScript/mvc.pdf. Acesso em: 05 de julho de 2015.