doi:10.20396/revpibic.v0i0.id

Aprimoramento de técnicas de hash de similaridade para investigações forenses

André S. Kameyama*, Marco Aurelio A. Henriques.

Resumo
Atualmente as investigações forenses têm que lidar com grandes quantidades de dados devido ao avanço da tecnologia, tornando-se impraticável a análise manual de cada caso. Neste trabalho, nós mostramos como melhorar o desempenho e a precisão de uma das ferramentas de pareamento aproximado mais consolidadas na área, o sdhash. Nossos resultados mostram que a ferramenta modificada é capaz de identificar similaridades entre diferentes artefatos com maior facilidade e rapidez.

Palavras-chave:
Função Hash, Resumo Criptográfico, Hash de Similaridade.

Introdução

A ferramenta de pareamento aproximado sdhash[1] é utilizada para detecção de similaridade entre objetos digitais de forma eficiente, por meio do uso de representações compactadas, chamadas de hashes de similaridade ou resumo. Esta ferramenta extrai características (*features*) únicas dos objetos e as codifica em resumos, os quais são mais tarde comparados para se quantificar o quão similares dois objetos são.

Com o intuito de melhorar o processo de identificação de objetos similares, o presente trabalho implementou uma nova versão do sdhash contendo algumas modificações que visam aprimorar tanto sua eficiência, como sua precisão. Foram modificados o processo de seleção de *features* (a fim de melhorar a precisão do mesmo) e a forma de codificar *features* (substituição da atual função de hash, SHA-1, por outra mais eficiente, FNV[2]). A fim de validar as melhorias propostas, utilizamos uma base de dados gerada localmente para podermos controlar o tamanho e o grau de similaridade de seus objetos.

Resultados e Discussão

Durante a geração de resumos com o sdhash, são extraídas *features* de um dado objeto e armazenadas em um vetor temporário. Assim, uma janela de tamanho fixo, iniciada na primeira *feature*, percorre todo o vetor até alcançar a última, onde em cada iteração, as *features* pertencentes à janela competem entre si (de acordo com um critério preestabelecido) e a vencedora é selecionada. A janela se move uma posição no vetor e o processo se repete. Porém, as primeiras e as últimas *features* competem um número menor de vezes em relação às demais, causando uma perda de precisão no processo. A implementação de uma janela circular visa solucionar este problema, uma vez que todas as *features* irão participar o mesmo número de iterações. Nós constatamos que esta mudança ocasionou um aumentou de aproximadamente 2% no número total de *features* selecionadas ao passo que o processo de seleção como um todo se tornou mais justo.

Após a seleção das *features*, o sdhash calcula o hash de cada uma delas por meio da função de hash SHA-1. Contudo, esta função é computacionalmente custosa e o cálculo da mesma sobre as diversas *features* que são selecionadas torna o processo lento. Desta forma, substituímos a função de hash por outras mais eficientes, como o MD5 e FNV. Para verificar o impacto dessa troca, foram realizados vários experimentos em diferentes tipos de arquivo. Nos testes foram medidos o tempo de execução e a precisão da ferramenta. O indicador de precisão utilizado foi o desvio padrão da distribuição das *features* em um histograma, de forma a simular a colisão entre *features* dada pela troca da função de hash. Quanto mais próximo o desvio das novas funções está em relação ao desvio de referência (SHA-1), mais próximas estão as distribuições e menor o número de colisões pela troca de função.

Observamos que ambas as funções de hash trouxeram uma diminuição no tempo de execução, sendo que o FNV apresentou o melhor resultado (aproximadamente 62%). Em todos os casos o desvio padrão permaneceu o mesmo, portanto podemos adotar essa troca sem perda de qualidade. Em relação à sensibilidade a mudanças nos tipos de arquivos não constatamos variações significativas em nenhum caso.

Conclusão

Este trabalho buscou aprimorar uma das funções mais utilizadas em investigações forenses para a busca de objetos similares, o sdhash. Mostramos que esta ferramenta possui imprecisões na geração de seus resumos e, através da implementação de uma estrutura de janela circular na etapa de seleção de *features*, conseguimos solucionar o problema, tornando o processo mais justo e aumentando a sensibilidade do mesmo em relação a detecção de mudanças no começo e final de objetos. Também atuamos na melhoria do desempenho do sdhash, trocando a atual função de hash SHA-1 pela FNV, resultando em uma diminuição no tempo de execução sem perda de precisão.

Agradecimentos

Os autores gostariam de agradecer ao programa PIBIC do CNPq pelo suporte financeiro parcial.

---

[1] Roussev, V. (2010, January). Data fingerprinting with similarity digests. In IFIP International Conference on Digital Forensics(pp. 207-226). Springer, Berlin, Heidelberg.
[2] G. Fowler, L. Noll, P. Vo, Fowler/Noll/Vo (FNV) Hash, ONLINE http://isthe.com/chongo/tech/comp/fnv/ - acessado em 11/07/2018.