Recuperação e Remanência de Dados

Allan da Silva Pinto, 115149

Série de Seminários

Disciplina de Análise Forense de Documentos Digitais

Prof. Dr. Anderson Rocha
anderson.rocha@ic.unicamp.br
http://www.ic.unicamp.br/~rocha

Organização

Organização

- Introdução e Motivação
- Remanência de dados
- Recuperação de dados remanescentes
- File Carving
- Trabalhos Correlatos
- Identificação e recuperação de arquivos JPEG com fragmentos perdidos [Sencar et al. 2009]
- Conclusões
- Referências

Introdução e Motivação

Introdução e Motivação

- Computação Forense Digital, segundo Edward Delp da Universidade de Purdue (Purdue Unversity);
  - "É o conjunto de técnicas científicas para a **coleção**, preservação, validação, identificação, análise, interpretação, documentos e apresentação de evidências derivadas de **meios digitais** com a finalidade de facilitar e/ou permitir a reconstrução de eventos, usualmente de natureza criminal";

Introdução e Motivação

- Evidências digitais podem ser encontradas em diversos equipamentos eletrônicos;
- Numa cena de um crime o perito pode encontrar:

© U.S. Department of Justice

Introdução e Motivação

- A coleta de evidências em meio digital necessita de técnicas que viabiliza a recuperação de evidências digitais;
- Construção física do dispositivo;
- Se o dispositivo estiver parcialmente ou totalmente danificado?
- Se as evidências foram apagados pelo criminoso?

Introdução e Motivação

- Caso Wellington Menezes
  - Para a Polícia Civil do RJ e para a Polícia Federal, o computador pessoal é uma peça-chave na investigação;
  - Provavelmente, o assassino não queria deixar rastros sobre seus hábitos na web e o modo como pesquisou e planejou o atentado;

© Info Online

Introdução e Motivação

- Injúrias no disco não destrói todos os dados, mas apenas dificulta o acessos a esses dados;
- Discos rígidos danificados podem terem seus dados recuperados em 80% dos casos [Mônica Campi 2011];
- E se o disco for formatado antes da destruição?

Remanência de Dados

- Oque é?
  - Refere-se aos dados residuais restantes no dispositivo de armazenamento após uma operação de remoção ou substituição de uma informação.
- Porque esse fenômeno ocorre?
  - Operações sobre os arquivos executadas pelo sistema operacional;
  - Característica intrínsecas do dispositivo de armazenamento;

Recuperação de Dados Remanescentes

- Recuperação pode ser feita por meio de Software
  - Funcionamento do dispositivo estiver em perfeitas condições;
  - Técnicas de File Carving;
- Recuperação pode ser feita por meio de Hardware
  - Dispositivo parcialmente danificado;
  - Microscópio de Força Magnética;
  - Dispositivos eletrônicos;

File Carving

- É uma técnica pela qual os arquivos são recuperados sem o uso da tabela de arquivos ou de qualquer outro tipo de metadado existente no dispositivo;
- Sistema de arquivo mantêm uma lista de blocos usados para armazenar cada arquivo;
- No processo de remoção dos arquivos os sistema operacional apenas o elimina da tabela de arquivos;

File Carving

- Habilidade de recuperar arquivos sem as informações da tabela de arquivo é um desafio;
- Fragmentação dos arquivos;
- Formatos sofisticados de arquivos;

Trabalhos Correlatos

- Recuperação de arquivos contíguos e fragmentado com validação rápida de objeto [Garfinkel 2007];
- Reconstrução automático de arquivos de imagens fragmentados usando algoritmo guloso [Pal et al. 2006];

Trabalhos Correlatos

- Recuperação de arquivos contíguos e fragmentado com validação rápida de objeto [Garfinkel 2007];
- Reconstrução automático de arquivos de imagens fragmentados usando algoritmo guloso [Pal et al. 2006];

Objetivo

- Levantamento de dados estatísticos sobre fragmentação de arquivos em diferentes dispositivos com diferentes sistema de arquivo;
- Recuperar arquivos bi-fragmentados que possuam cabeçalho *(Header)* e rodapé *(Footer);*
- Arquivos que possam ser validados e codificados
  - Validar: Tarefa de verificar se um arquivo obedece a estrutura do tipo do arquivo;
  - Codificar: Transformação da informação nos blocos associados aos arquivos em um formato original que descreve o seu conteúdo;

Fragmentação

- Foi analisado sistemas de arquivos ativos em 449 imagens de discos, sendo que 324 destas imagens continham mais do que 5 arquivos;
- A ferramentoa Sleuth Kit foi utilizado nesta análise;
  - Recuperação de 892GB de dados;
- 6% dos arquivos recuperados estavam fragmentados;
- Arquivos de interesse forense apresenta altas taxas de fragmentação (e.g. Jpeg, avi, doc, html, pst);

Fragmentação total dos discos

Tabela 1 – Distribuição de fragmentação de arquivos em discos com mais de 5 arquivos

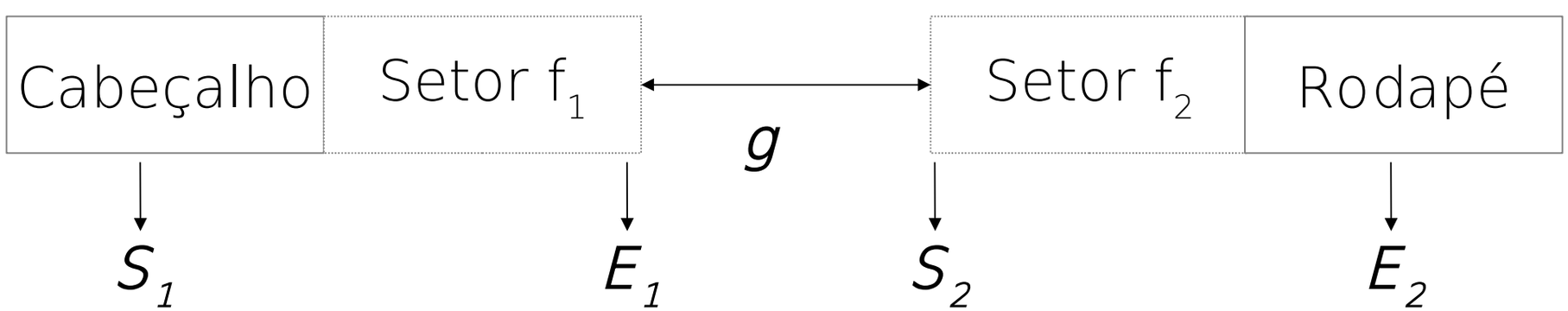
Fração de arquivos no discos que estão fragmentados	Total de discos	Total de arquivos referenciados
f = 0%	145	17.267
0% < f < 1%	42	459.229
1% < f < 10%	107	1.115.390
10% < f < 100%	30	412.297
	324	2.004.183

Fragmentação por arquivos

Tabela 2 – Fragmentação por arquivos encontrados nos discos			
Extensão	Numero de arquivos (2 fragmentos)	Numero de arquivos (3 fragmentos)	Numero de arquivos (> 3 fragmentos)
AVI	17	6	185
BMP	367	129	1630
EXE	2352	827	1100
GIF	2990	795	27581
HTML	4085	929	10330
JPEG	2999	57	13973
MPEG	4	3	22
PNG	175	93	300

Bifragment Gap Carving (BGC)

- Busca Exaustiva de todas as combinações de blocos entre o cabeçalho e o rodapé até que a validação/decodificação tenha sucesso.
- Seja f_1 o primeiro fragmento que se estende do setor S_1 ao setor E_1 e seja f_2 o segundo fragmento que se estende do setor S_2 ao setor E_2
- Seja g a distância entre os dois fragmentos, isto é, $g=S_2-(E_1+1)$.
- Iniciando com $g=1$, tente todas as valores para g até que $g=E_2-S_1$.
- Para todo g, tente todas os valores consistentes de E_1 e S_2.



Limitações

- Para arquivo com muitos fragmentos, o número de validações que precisam ser realizados grande;
- Não é escalável para distâncias muito grandes entre os dois fragmentos. (Neste experimento a maior distância entre dois fragmentos JPEG foi de 1272 setores);
- Arquivos que não possuem rodapé não são recuperados (e.g. BMPs);
- Nem sempre validação/decodificação bem sucedida implica que um arquivo foi reconstruída corretamente. Decodificadores vai dar um erro quando os dados nos blocos não estão em conformidade com as regras de decodificação;

Trabalhos Correlatos

- Recuperação de arquivos contíguos e fragmentado com validação rápida de objeto [Garfinkel 2007];
- Reconstrução automático de arquivos de imagens fragmentados usando algoritmo guloso [Pal et al. 2006];

Objetivo

- Reconstrução de imagens JPEG a partir de fragmentos recuperados;
- Assume que os fragmentos foram recuperados sem perda de dados e não há fragmentos corrompidos ou perdidos;

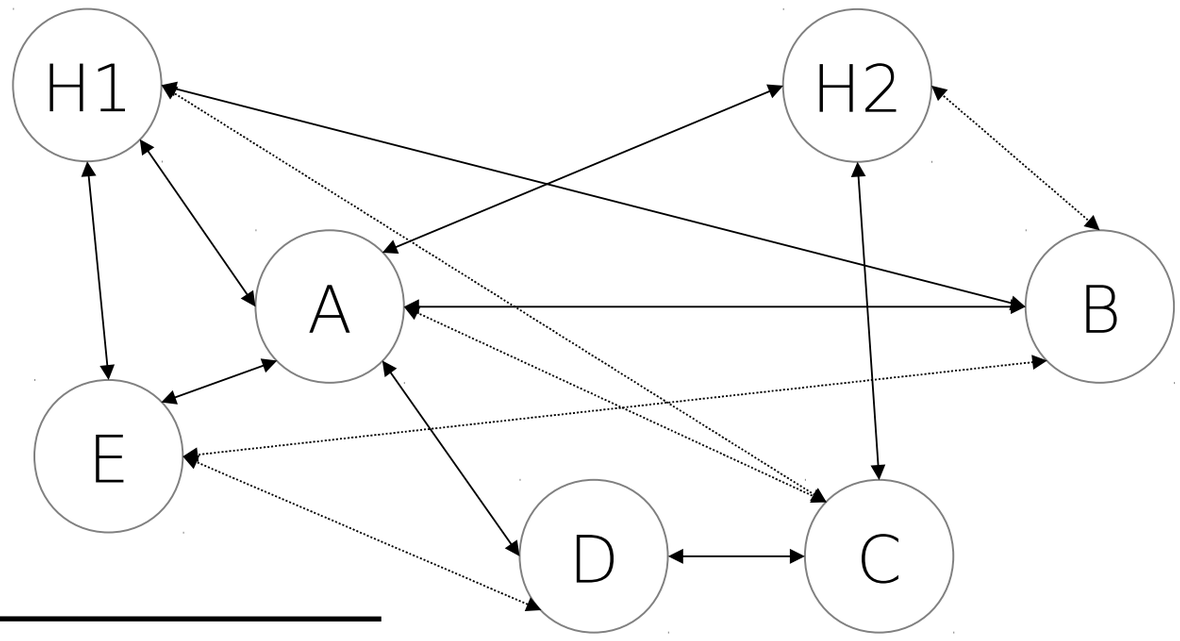
Definições

- Identificar pares de fragmentos que são adjacentes;
- Para quantificar a probabilidade da adjacência na imagem original é calculado pesos *C(i,j)* que representa a probabilidade do fragmento A_j ser subsequente a A_i.;
- Uma vez que esse pesos são calculados, a permutação dos fragmentos que leva a uma correta reconstrução da imagem pertence ao conjunto de todas as permutações possíveis;
- Problema de encontrar uma permutação que maximize

$$T = \sum C\left(\pi\left(i\right),\pi\left(i+1\right)\right)$$

Definições

- Representação do problema por um grafo onde os nós representam os fragmentos e as arestas os pesos candidatos *C* entre fragmentos;
- No caso de haver k imagens fragmentadas é necessário encontrar k vértices disjuntos;



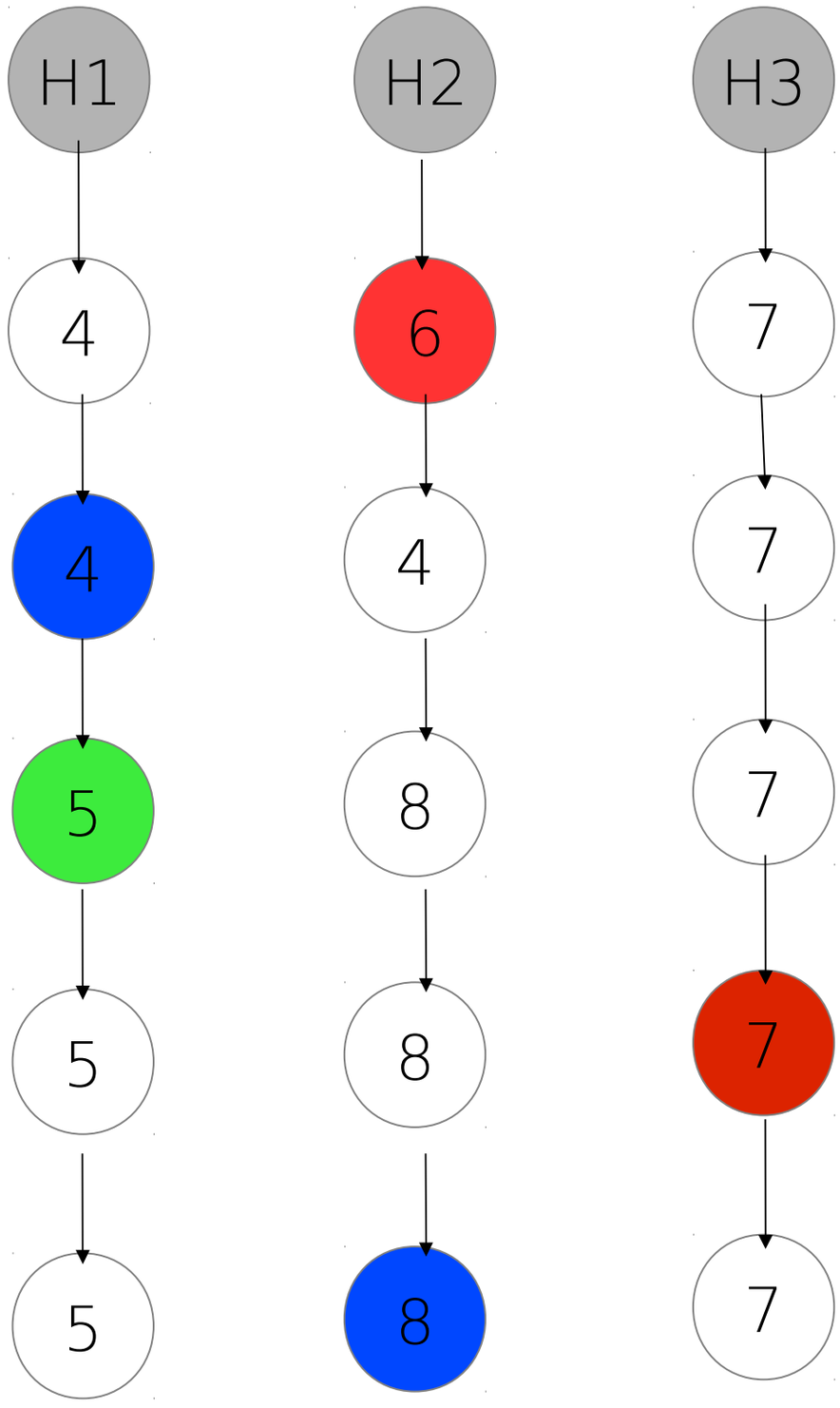
Grafo com sete fragmentos e dois caminhos disjuntos

Parallel Unique Path (PUP)

- Temos um conjunto S com k cabeçalhos (B_{C1}, B_{C2}, ..., B_{CK}) referentes aos k arquivos. Inicialmente, esses blocos formam o conjunto de blocos correntes (B_{S1}, B_{S2}, ..., B_{SK});
- Os melhores casamentos entre os blocos correntes e os blocos a serem adicionados são adicionados numa lista T = (B_{T1}, B_{T2}, ..., B_{TK});
- Assumindo que o melhor casamento foi B_{ti}
  1. Adicione o bloco B_{ti} ao arquivo i-ésimo arquivo
  2. Substitua o bloco corrente no conjunto S para o i-ésimo arquivo
  3. Encontre um novo conjunto T de melhor casamento para o conjunto S
  4. Novamente, encontre o melhor casamento em T
  5. Repita 1 até todos os arquivos estarem completos.

Exemplo

- S=(H1,H2,H3),T=(4,6,7),R=(4,5,6,7,8)
- S=(H1,6,H3),T=(4,4,7),R=(4,5,7,8)
- S=(4,6,H3), T=(5,8,7),R=(5,7,8)
- S=(5,6,H3),T=(5,8,7),R=(7,8)
- S=(5,6,7),T = (5,8,7),R=(8)
- S=(5, 8, 7),T=(5,8,7),R=()



Limitações

- Os melhores fragmentos com o conjunto atual de fragmentos pode ser melhor para fragmentos que não foram processados até agora;
- Propagação de erros;
- Não trata o problema da recuperação. É assumido que a recuperação ocorreu com sucesso;

Identificação e recuperação de arquivos JPEG com fragmentos perdidos [Sencar et al. 2009]

Objetivo

- Recuperação de imagens JPEG usando informações do cabeçalho de um maneira mais eficiente
- Recuperação de imagens sem o cabeçalho disponível

Visão geral

- Atualmente poucas técnicas que permite recuperar arquivos fragmentados
- Eficientes quando a taxa de fragmentação é baixa e quando os fragmentos não estão muitos distantes
- Processos de recuperação não são tolerantes a falhas

Visão geral

- Técnicas existentes
  - Procura marcadores de início e fim de arquivo
  - Dados extraídos do primeiro bloco que contém o marcador de inicio da imagem é combinados com dados extraídos dos blocos consecutivos
- Bloco combinado é decodificado e caso a decodificação falhe tem-se o ponto de fragmentação
- Encontrar os outros fragmentos da imagem é computacionalmente caro
  - Casamento *(matching)* de padrões de bits para identificar sequências de blocos de dados que são mais prováveis conter outros pedaços do mesmo arquivo.

Compressão JPEG

- Baseada na transformada discreta dos cossenos (DCT)
  - Dividir imagem em blocos disjuntos com 8x8 pixels.
  - Aplicar DCT de duas dimensões em cada bloco para obter os coeficientes (DC e AC).
  - Quantização dos coeficientes para descartar informações menos importantes, dando origem a tabela de quantização.
  - Coeficientes quantizados são reorganizado por meio de um escaneamento zigue-zague e estes passam por um codificador por entropia.
- Para imagens coloridas espaço de cor original é convertida para o espaço de cor YcbCr
  - Na compressão, estas componentes são processadas separadamente

Armazenamento de arquivos JPEG

- Do ponto de vista da recuperação de arquivos, o aspecto mais importante é o formato usado para encapsular os bytes da imagem
- Toda imagem JPEG é armazenada como uma série de blocos de imagem comprimida (*MCU – minimum coded unit*)
- Cada MCU consiste de uma quantidade de blocos de cada componente da cor
  - MCU = YcbCr
  - MCU = YYcbCr
  - MCU = YYYcbCr

Armazenamento de arquivos JPEG

- Para distinguir tabelas de codificações de diferentes imagens é usado as frequências de ocorrências de certos padrões de bits
- Premissa do método: frequência da ocorrência de um dado padrão de n-bits em m-bits sequenciais deves esperar que haja $m/2^n$ casamentos
- Se m-bits sequenciais não forem aleatórios então para certos padrões de n-bits é esperado um viés do número esperado de casamentos randômicos
- Para capturar esse grau de não-aleatoriedade, é usada características de frequência de padrões de bits que são construídas a partir de palavras da tabela de codificação.

Objetivo

- Recuperação de imagens JPEG usando informações do cabeçalho de um maneira mais eficiente
- Recuperação de imagens com fragmentos perdidos e sem o cabeçalho disponível

Recuperação de fragmentos corrompidos

- No padrão JPEG, marcas de reinício *(restart markers)* são fornecidos com o meio de detecção e recuperação de erros em bitstream.
- Existem 8 marcas únicas de reinício sendo cada uma representada por dois bytes (0xFFD0-0xFFD7).
- Estas marcas são inseridas periodicamente. (O número de MCUs entre essas marcas está no cabeçalho)
- Em arquivos JPEG, os coeficientes DC de todas as cores são codificadas com valores da diferença ao invés do valor absoluto. Quando a marca é alcançada,esta diferença é inicializada para 0 e o bitstream é sincronizado.
- No caso de um erro de bitstream pode-se computar o número de saltos de MCUs para alcançar a próxima marca de reinicio.

Recuperação de fragmentos sem cabeçalho

- Sem um cabeçalho válido, uma imagem JPEG não pode ser recuperada.
- Autores propõem uma construção de um pseudo cabeçalho para uma imagem.
- É assumido que as imagens armazenadas se relacionam de algum modo.
- Utiliza informações de outras imagens
  - Comprimento e altura da imagem;
  - Tabela de quantização;
  - Informações sobre MCUs;
  - Tabela de codificação (Huffman);

Conclusões

Conclusões

- Poucos métodos desenvolvidos na área;
- Problema da fragmentação de arquivos é ainda um desafio;
- Nem todos os tipos de arquivos podem ser recuperados;
- Arquivos com codificações sofisticadas são difíceis de serem recuperados;

Referências

Referências

1. [Mônica Campi 2011] **Campi, Mônica** (2011). *Polícia tenta recuperar HD de assassino do Rio acessado em 20 de Outubro de 2011 disponível em www.http://info.abril.com.br/noticias/tecnologia-pessoal/policia-do-rj-tera-como-recuperar-hd-queimado-08042011-17.shl*
2. [Garfinkel 2007] **Garfinkel, Simson L.** (2007*). Carving contiguous and fragmented files with fast object validation,* Digital Investigation, Elsevier, Volume 4, Se. 2007, Pages 2-12
3. [Pal et al. 2006] **Memon, N.; Pal, A.;** *Automated reassembly of file fragmented images using greedy algorithms,* Image Processing, IEEE Transactions on , vol.15, no.2, pp.385-393, Feb. 2006
4. [Sencar et al. 2009] **Husrev T. Sencar, Nasir Memon**, *Identification and recovery of JPEG files with missing fragments*, Digital Investigation, Elsevier, Volume 6, September 2009, Pages S88-S98

Obrigado!