BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2026
http://excel.fit.vutbr.cz

Rekurzivní streamová extrakce multimediálních dat z obrazů pevných disků

Michal Novák*

Abstrakt

Multimediální obsah představuje v digitální forenzní analýze klíčový zdroj důkazních informací, zejména při vyšetřování trestné činnosti související například s drogovou kriminalitou, obchodem se zbraněmi či mravnostními delikty. V praxi se forenzní analýza provádí nad obrazy pevných disků, které zachycují kompletní stav úložiště v daném okamžiku. Vyhledávání relevantních dat je však značně komplikováno tím, že multimediální soubory se často nacházejí uvnitř víceúrovňově vnořených struktur, jako jsou archivy, dokumenty nebo jiné kontejnery. Tato práce se zaměřuje na návrh a implementaci nástroje pro rekurzivní a streamovou extrakci multimediálních dat z obrazů pevných disků bez nutnosti jejich předchozího rozbalování. Součástí je také návrh syntetické datové sady a evaluace řešení z hlediska správnosti, výkonu a odolnosti vůči složitým či poškozeným strukturám.

*xnovak3g@stud.fit.vut.cz, *Fakulta informačních technologií, Vysoké učení technické v Brně*

1. Úvod

V digitální forenzní analýze jsou data zajišťována ve formě obrazů pevných disků, které představují bitově přesnou kopii obsahu fyzického úložiště. Tyto obrazy jsou vytvářeny pomocí specializovaných nástrojů a ukládány ve standardizovaných formátech, jako například *RAW* [1], *EWF (E01)* [2] nebo *AFF4* [3], s cílem zachovat integritu dat a umožnit jejich opakovanou analýzu bez zásahu do původního média.

Multimediální data uložená v těchto obrazech představují významný zdroj důkazních informací, například při vyšetřování drogové kriminality, obchodování se zbraněmi nebo mravnostních deliktů. Tato data se však často nenacházejí jako samostatné soubory, ale jsou ukryta ve víceúrovňově vnořených strukturách, jako jsou archivy, dokumenty nebo jiné kontejnerové formáty, například obrazový soubor uložený v dokumentu formátu *Word*, který je následně zabalen v archivu *ZIP*.

Získání těchto dat je proto technicky náročné. Klasický postup založený na postupném rozbalování jednotlivých vrstev vede k vysoké režii, duplicitnímu zpracování dat a výrazné závislosti na výkonu diskového úložiště. S rostoucí velikostí analyzovaných dat se tento problém dále prohlubuje.

Cílem této práce je analyzovat současné přístupy a navrhnout nástroj, který umožňuje efektivní extrakci multimediálních dat z obrazů pevných disků, včetně hluboce vnořených struktur Obrázek 2, a to bez nutnosti jejich předchozího ukládání na disk.

2. Existující řešení

Současné forenzní nástroje umožňují práci s obrazy disků a rekonstrukci souborových systémů bez nutnosti jejich připojení do operačního systému, čímž poskytují robustní základ pro analýzu dat. Mezi běžně používané nástroje patří například *Autopsy*. Tyto a podobné nástroje jsou primárně orientovány na interaktivní analýzu diskového obrazu a rekonstrukci souborového systému. Jejich podpora pro automatizovanou a efektivní extrakci multimediálního obsahu z hluboce vnořených struktur je však omezená.

Typickým přístupem k automatické extrakci je sekvenční zpracování souborů s využitím externích nástrojů. Jednotlivé objekty jsou extrahovány na disk a následně znovu zařazeny do fronty k zpracování. Tento postup vede k vysokým nárokům na I/O operace, zbytečné materializaci dat a omezené škálovatelnosti při práci s rozsáhlými datovými sadami.

Podobný princip využívají i nástroje vyvinuté v rámci projektu *FACIS*, konkrétně nástroje *Maxtor* [4]

a *Extractor* [5], které umožňují extrakci dat z obrazů disků a jejich další zpracování. Tento přístup je založen na opakované materializaci dat na datovém úložišti, což vede k rychlému nárůstu objemu dat. Zejména při zpracování víceúrovňově vnořených struktur může být pro analýzu potřeba několikanásobně větší diskový prostor než velikost původního vstupu. Nevýhodou tohoto přístupu je také absence předběžné informace o požadované kapacitě úložiště, což může při jednoproudovém zpracování vést k vyčerpání dostupného diskového prostoru v průběhu analýzy.

3. Architektura a implementace řešení

Navržené řešení je založeno na kombinaci rekurzivního a streamového zpracování dat Obrázek 3. Základní myšlenkou je analyzovat obsah souborů přímo během jejich čtení, bez nutnosti jejich úplného uložení na disk.

Architektura nástroje je navržena jako datová pipeline, kde jednotlivé fáze komunikují pomocí asynchronních kanálů. Tento přístup umožňuje paralelní zpracování a zároveň zajišťuje nízkou vazbu mezi komponentami.

Klíčovou komponentou je modul rekurzivní extrakce, který pro každý vstupní datový proud:

- určí skutečný typ souboru na základě jeho binární struktury,
- zvolí odpovídající specializovaný extraktor (např. pro obrázky, video, archivy nebo dokumenty),
- v případě kontejnerového formátu vytvoří nové datové proudy,
- tyto proudy opětovně zpracovává stejným způsobem.

Celý proces je navržen jako jednoprůchodový, kdy jsou data analyzována již během čtení. Současně probíhá výpočet kryptografických hashů a sběr metadat (např. velikost, typ nebo geolokační informace), což umožňuje jejich další využití při forenzní analýze. Navržená architektura rovněž podporuje obnovení zpracování po jeho přerušení, a to bez nutnosti opětovného zpracování již analyzovaných dat, což oproti dosavadním řešením představuje významné zlepšení.

Výsledkem je jednotný mechanismus, který dokáže transparentně zpracovávat jak běžné soubory, tak i hluboce vnořené struktury bez nutnosti jejich explicitního rozbalování.

4. Porovnání a vyhodnocení

Pro ověření navrženého přístupu byla vytvořena syntetická datová sada diskových obrazů, která simuluje různé scénáře vnoření multimediálních dat, včetně kombinací archivů, dokumentů a dalších kontejnerových struktur. Testování se zaměřilo na správnost extrakce, výkonnost zpracování a schopnost nástroje pracovat s hluboce vnořenými nebo nestandardními daty.

Naměřené výsledky ukazují, že navržený přístup umožňuje výrazně efektivnější zpracování ve srovnání s tradičními metodami založenými na ukládání mezivýsledků na disk Obrázek 4. Dosažené zrychlení se zobrazených příkladech pohybovalo v rozmezí 1,75× až 5,6× v závislosti na charakteru dat, přičemž konkrétní hodnoty jsou ovlivněny zejména typem zpracovávaných souborů a hloubkou jejich vnoření. Ke zlepšení dochází především díky snížení počtu I/O operací, lepšímu využití paralelismu a odstranění nutnosti spouštět externí procesy.

Zároveň byla potvrzena schopnost nástroje pracovat s komplexními a částečně poškozenými strukturami bez přerušení zpracování.

5. Přínosy práce

Implementované řešení přináší zrychlení a zefektivnění extrakce multimediálních dat oproti tradičním přístupům, především díky odstranění nutnosti ukládání mezivýsledků na disk. Tím se eliminuje problém „nafouknutí“ dat a výrazně se snižují nároky na diskový prostor. Současně dochází k rozšíření podpory diskových obrazů o formát *AFF4* a k efektivnímu využití paralelismu při zpracování. Implementace navíc umožňuje obnovení extrakce po přerušení, což v dosavadních řešeních nebylo možné.

Dalším přínosem je redukce počtu potřebných nástrojů, kdy navržené řešení sjednocuje funkcionalitu nástrojů Maxtor a Extractor do jednoho systému. Implementovaný přístup je prakticky použitelný při forenzní analýze reálných digitálních důkazů.

6. Závěr

Práce představuje přístup k extrakci multimediálních dat z obrazů pevných disků založený na kombinaci rekurzivního a streamového zpracování, který eliminuje potřebu ukládání mezivýsledků na disk.

Navržené řešení je implementováno jako modulární architektura podporující paralelní zpracování a bylo experimentálně ověřeno na syntetické datové sadě. Výsledky potvrzují jeho praktickou použitelnost při zpracování digitálních důkazů a potenciál pro další rozvoj v oblasti automatizované forenzní analýzy.

Literatura

[1] Forensics Wiki contributors. Raw image format. online. https://forensics.wiki/raw_image_format/.

[2] Joachim Metz. Ewf specification. online. https://github.com/libyal/libewf/blob/main/documentation/Expert%20Witness%20Compression%20Format%20(EWF).asciidoc.

[3] Mike Cohen and Bradley L Schatz. Aff4. online. https://github.com/aff4/.

[4] Ondřej Ryšavý, Jan Pluskal, Dušan Kolář, and Dominika Regéciová. Maxtor – multimedia artifact extractor, 2022. https://www.fit.vut.cz/research/result/c180411/.

[5] Dominika Regéciová, Dušan Kolář, and Jan Pluskal. Extractor, 2023. https://www.fit.vut.cz/research/result/c186760/.