

# FbHash: A New Similarity Hashing Scheme for Digital Forensics

Timotej Knez  
63..

Sebastian Mežnar  
27192031

Jasmina Pegan  
63170423

## POVZETEK

nek povzetek

## Kategorija in opis področja

E.3 [Data encryption]

## Splošni izrazi

Hashing

## Ključne besede

Data fingerprinting, Similarity digests, Fuzzy hashing, TF-IDF, Cosine-similarity

## 1. UVOD

Živimo v obdobju shranjevanja ogromnih količin podatkov. Pri forenzičnih preiskavah se pogosto zgodi, da je pridobljenih datotek preveč za ročno pregledovanje. Digitalni forenziki se tako soočijo s problemom avtomatizacije preiskave datotek. Možna rešitev so algoritmi, kot so **ssdeep**, **sdhash** in **FbHash**, ki poskusijo filtrirati vnaprej znane "slabe" oziroma "dobre" datoteke. Ti algoritmi (angl. *Approximate Matching algorithms*) ugotavljajo delež ujemanja datotek s pomočjo (nekriptografskih) zgoščevalnih funkcij. Algoritma **ssdeep** in **sdhash** lahko preslepi aktivni napadalec, ki pametno napravi majhne spremembe na določenih mestih datoteke. Učinkovitega napada na algoritem **fbhash** ne poznamo.[4]

V 2. poglavju predstavimo predhodnike algoritma **FbHash**. V 3. poglavju podrobneje predstavimo algoritem **FbHash** in našo implementacijo. V 4. poglavju opišemo izvedene eksperimente in v 5. poglavju opišemo rezultate. V 6. poglavju povzamemo narejeno delo in rezultate.

## 2. SORODNA DELA

Prvi algoritem, namenjen iskanju približnih ujemanj, je bil objavljen leta 2002 pod imenom **dcfldd**. Ta algoritem je razvil N. Harbour kot izboljšano verzijo ukaza **dd**[6]. Izboljšana

različica tega algoritma je **ssdeep**. Pomembnejša predhodnika algoritma **FbHash** sta tudi **MRSH-v2** in **mvHash-B**. Obstaja še **bbhash**, ki pa je časovno potraten in ga ne bomo podrobneje opisali.

### 2.1 ssdeep

Algoritem **ssdeep** je implementacija kontekstno sprožene kosovno zgoščevalne funkcije (angl. *Context Triggered Piecewise Hash*, CTPH), ki jo je predstavil J. Kornblum septembra 2006 v članku [7]. Algoritem temelji na detektorju neželene elektronske pošte **spamsum**, ki lahko zazna sporočila, ki so podobna znanim neželenim sporočilom.

CTPH uporablja zgoščevanje po kosih (angl. *piecewise hashing*), kar pomeni, da se zgoščena vrednost izračuna na posameznih kosih fiksne dolžine. Za razliko od **dcfldd**, algoritem CTPH uporabi poljubno zgoščevalno funkcijo.

Drugi princip, ki ga uporablja CTPH, je zgoščevalna funkcija z drsečim oknom (angl. *rolling hash*), ki preslika zadnjih  $k$  zlogov (bajtov) v psevdonaključno vrednost. Vsakega naslednika je tako možno hitro izračunati iz predhodno izračunane vrednosti. Pri tem je uporabljena zgoščevalna funkcija **FNV**.

Postopek CTPH se začne z izračunom zgoščenih vrednosti z drsečim oknom. Ob določeni sprožilni zgoščeni vrednosti (angl. *trigger value*) se vzporedno s tem sproži še algoritem zgoščevanja po kosih. Ob ponovni pojavitvi sprožilne vrednosti se dotlej zbrane vrednosti druge zgoščevalne funkcije zapišejo v končni prstni odtis. Tako se ob lokalni spremembi v datoteki sprememba pozna le lokalno tudi v prstnem odtisu.

Sledi primerjava prstnih odtisov datotek, ki temelji na uteženi Levensteinovi razdalji (angl. *edit distance*), ki je nato še skalirana in obrnjena, da predstavlja 0 povsem različna prstna odtisa.

Algoritem **ssdeep**, ki je implementacija CTPH, se izkaže pri primerjavi podobnih besedilnih datotek in dokumentov [7]. Po drugi strani pa lahko aktivni napadalec popravi "slabe" datoteke na tak način, da se izognejo črni listi [4]. Ker je prstni odtis fiksne dolžine, je algoritem primeren le za relativno majhne datoteke podobnih velikosti.

### 2.2 sdhash

Algoritem **sdhash** je opisal V. Roussev januarja 2010 v članku [8]. Glavna prednost tega algoritma pred predhodnimi je, da izbere statistično manj verjetne dele datotek kot izhodišče za računanje prstnega odtisa.

Postopek se začne z iskanjem statistično najmanj verjetnih delov datoteke. Izračuna se entropija skupin po  $k$  zlogov datoteke. Nato se izračuna rank vsake skupine glede na  $n$  sosednjih skupin. Izbrane so skupine, ki imajo rank večji ali enak postavljeni meji.

Sledi filtriranje skupin  $k$  zlogov, ki niso bistvene, povzročajo pa lažno pozitivne rezultate. Ocenili so, da je dobro zavreči skupine z oceno entropije pod 100 ali nad 990, ker so takšne skupine pogoste v datotekah tipa JPEG.

Nato se generira prstni odtis datoteke kot zaporedje Bloomovih filtrov, ki so verjetnostne strukture, uporabljene za prostorsko učinkovito predstavitev množic. Algoritem **sdhash** preveri za vsako izbrano skupino  $k$  zlogov, ali je že v množici, predstavljeni z Bloomovimi filtri. Če skupine ni v množici, jo algoritem doda.

Nazadnje algoritem primerja prstne odtise datotek, torej zaporedje Bloomovih filtrov. Za vsak filter, ki predstavlja prvo datoteko, se izračuna maksimalna ocena podobnosti s filtri, ki predstavljajo drugo datoteko. Rezultat je povprečje tako pridobljenih ocen podobnosti.

Algoritem **sdhash** doseže boljša priklic in preciznost kot **ssdeep** [4]. A tudi ta algoritem ima več pomanjkljivosti: nekatere datoteke ne more primerjati, primerjava datoteke same s seboj lahko vrne oceno med 50 in 100 ter prvih 15 zlogov sploh ne vpliva na končni prstni odtis. Poleg naštetega aktivni napadalec lahko spremeni "slabe" datoteke na tak način, da se izognejo črni listi oziroma "dobre" datoteke tako, da se obdržijo na beli listi [?].

## 2.3 MRSH-v2

Oktobra 2012 sta F. Breitinger in H. Baier predstavila algoritem **MRSH-v2** [3], ki se opira na predhodno razvit algoritem **MRSH** (angl. *multi-resolution similarity hashing*), ta pa temelji na algoritmu **ssdeep**.

Algoritem **MRSH** ima določene sprožilne točke  $-1 \bmod b$ , kjer  $b$  pomeni povprečno velikost bloka. Namesto zgoščevalne funkcije z drsečim oknom uporabi polinomsko zgoščevalno funkcijo **djb2**, kot primitiv pa **MD5**. Namesto konkatencije zgoščenih vrednosti **MRSH** kot prstni odtis uporabi seznam Bloomovih filtrov.

Algoritem **MRSH-v2** ponovno uporabi zgoščevalno funkcijo z drsečim oknom, kot **ssdeep**, namesto **FNV** pa uporabi funkcijo zgoščevanja **MD5**. Za večjo hitrost in v izogib napadu z dodajanjem sprožilnih točk je dodana tudi spodnja meja za velikost skupin zlogov  $\frac{b}{4}$ .

Algoritem **MRSH-v2** je po [3] časovno učinkovitejši od predhodnih algoritmov. Vključuje način za odkrivanje fragmentov in način za odkrivanje podobnih datotek. Po analizi leta 2014 [1], ki primerja **ssdeep**, **sdhash** in **MRSH-v2**, se v povprečju najboljše obnese **sdhash**, **ssdeep** in **sdhash** izkazujejo dobro preciznost, vsi trije algoritmi pa imajo relativno

slab priklic.

## 2.4 mvHash-B

Marca 2013 so F. Breitinger in sod. predstavili algoritem **mvHash-B** [2]. Ideja algoritma je, da majhne lokalne spremembe ne spremenijo končnega rezultata.

V prvem koraku se izvede večinsko glasovanje po bitih z nastavljivo mejo  $t$ . Vsakih  $k$  zlogov se tako preslika v ničle, če je število enic v zaporedju bitov manjše od  $t$ , sicer pa v enice.

Nato se zaporedje bitov zapiše na bolj kompakten način – enake zaporedne bite nadomestimo z dolžino takega niza. Če se zaporedje bitov ne začne z nekaj ničlami, bo prvo število v seznamu 0. Dobimo kodirano zaporedje dolžin nizov.

Kodirano zaporedje se razdeli v prekrivajoče se skupine fiksne dolžine. Te skupine so dodane v Bloomov filter. Nazadnje se primerja Bloomove filtre s pomočjo Hammingove razdalje, ocene pa se povpreči in odšteje od 100.

Algoritem **mvHash-B** naj bi bil hiter skoraj kot **SHA-1**, torej blizu zgornje meje učinkovitosti [2]. Vendar tudi ta algoritem ni varen pred aktivnim napadalcem, saj je možno popraviti "slabo" datoteko tako, da se izogne črni listi [5].

## 3. OPIS ALGORITMA

Avtorji v članku [4] predstavijo algoritem **FbHash-B**, ki je namenjen nestisnjenim (angl. *uncompressed*) in **FbHash-S**, ki je namenjen stisnjenim (angl. *compressed*) datotekam. Iskanje zelenih datotek poteka na sledeč način. Najprej datoteke, ki jih želimo preveriti zgoštimo s pomočjo ustreznega algoritma glede na njihov tip in tako dobimo bazo podatkov. Nato zgoštimo še ciljne datoteke in jih primerjamo z datotekami iz baze podatkov. Tako lahko datoteke, ki imajo s ciljnim dovolj podobnosti zavržemo (angl. *whitelist*) oziroma dodamo med iskane (angl. *blacklist*).

### 3.1 FbHash-B

**FbHash-B** je bolj primeren datotekam, ki niso stisnjene. V opisu algoritma bomo za boljšo preglednost uporabili sledečo notacijo:

- Del datoteke: niz  $k$  zaporednih bajtov.
- $ch_i^D$ : del datoteke  $D$ , ki se začne na  $i$ -tem bajtu.
- Frekvenca dela oziroma  $chf_{ch_i}^D$ : število pojavitev dela datoteke  $ch_i$  v datoteki  $D$ .
- Dokumentna frekvenca dela oziroma  $df_{ch}$ : število dokumentov, ki vsebujejo del datoteke  $ch$ .
- $N$ : število datotek v bazi podatkov.
- $RollingHash(ch_i)$ : vrednost rekurzivne zgoščevalne funkcije imenovane rolling hash na delu datoteke  $ch_i$
- $chw_{ch_i}^D$ : utež dela datoteke  $ch_i$  v datoteki  $D$
- $docw_{ch_i}$ : dokumentna utež dela datoteke  $ch_i$
- $W_{ch_i}^D$ : ocena dela datoteke  $ch_i$  v datoteki  $D$

Algoritem je predstavljen v treh korakih. Prvi zavzema računanje frekvence delov datotek, drugi računanje dokumentnih uteži delov datotek in tretji računanje zgostitvenih uteži.

### 3.1.1 Računanje frekvence delov datotek

V tem koraku se izračunajo frekvence delov datotek in njihove uteži. Vsaka datoteka vsebuje  $N_D - k$  delov datotek, kjer je  $N_D$  dolžina datoteke  $D$  v bajtih in  $k$  parameter. Tako ima  $i$ -ti del datoteke  $D$  obliko  $B_i^D B_{i+1}^D \cdots B_{i+k-1}^D = ch_i^D$ , kjer je  $B_j^D$   $j$ -ti bajt v datoteki  $D$ .

Najprej izračunamo  $RollingHash(ch_0^D)$  s formulo:

$$RollingHash(ch_0^D) = B_0^D \cdot a^{k-1} + \cdots + B_{k-1}^D \cdot a^0 \pmod{n},$$

kjer so  $a$ ,  $n$  in  $k$  parametri. Zaradi rekurzivne strukture funkcije  $RollingHash$  lahko zgostitve preostalih delov izračunamo na sledeč način:

$$RollingHash(ch_{i+1}^D) = a \cdot RollingHash(ch_i^D) - B_i^D \cdot a^k - B_{i+k}^D \pmod{n}.$$

Parametre  $a$ ,  $k$  in  $n$  izberemo na sledeč način. Za  $a$  izberemo neko konstanto med 2 in 255. Če za zalogo vrednosti funkcije  $RollingHash$  vzamemo 64-bitna števila, lahko  $k$  izračunamo na sledeč način:

$$B_i^D \cdot a^{k-1} \leq 2^{64} - 1$$

. Ker je maksimalna vrednost bajta in parametra  $a$  enaka 255 dobimo neenačbo  $255^k \leq 2^{64} - 1$  iz česar sledi, da je  $k$  manjši ali enak 7. Tako za  $k$  izberemo 7. Če želimo zagotoviti, da ne pride do trkov med različnimi deli datotek, moramo za  $n$  izbrati praštevilo, ki je večje od  $2^56 = 255 \cdot 255^{k-1}$ .

Strukturo datoteke lahko tako predstavimo kot razpršeno tabelo, kjer je ključ vrednost, ki jo dobimo s funkcijo  $RollingHash$  na določenem delu datoteke, vrednost pa število pojavitev pripadajočega dela datoteke (oziroma vrednosti, ki smo jo dobili s funkcijo  $RollingHash$ ).

Prvi korak zaključimo tako, da izračunamo uteži za dele besedila znotraj datoteke ( $chw_{ch_i}^D$ ) s pomočjo ene izmed funkcij iz 3.4.

### 3.1.2 Računanje dokumentnih uteži

Drugi korak je računanje dokumentnih uteži delov besedila. Te nakazujejo koliko informacije prinese posamezen del datoteke znotraj določene baze podatkov.

Najprej je potrebno izračunati dokumentno frekvenco za dele datotek, ki se v bazi podatkov pojavijo. To naredimo tako, da zgradimo razpršeno tabelo katere ključ je vrednost  $RollingHash(ch_i)$ , vrednost pa število pojavitev posameznega ključa. To s tabelo lahko sestavimo tako, da se sprehodimo čez vse ključe razpršenih tabel za posamezni datoteke in zgoščevalni tabeli prištejemo 1, če se ključ v datoteki pojavi.

Iz frekvence posameznega dela nato izračunamo njegovo utež ( $docw_{ch_i}$ ) s pomočjo ene izmed funkcij iz 3.4.

### 3.1.3 Računanje zgostitvenih ocen

V tretjem koraku izračunamo zgostitvene ocene za posamezne datoteke. S temi ocenami lahko kasneje datoteke med seboj primerjamo, kot je predstavljeno v 3.3.

V vsaki datoteki izračunamo oceno pomembnosti za posamezen del s formulo:

$$W_{ch_i}^D = docw_{ch_i} \cdot chw_{ch_i}^D.$$

Zgostitveno oceno za datoteko predstavimo kot razpršeno tabelo, kjer za ključ vzamemo vrednosti funkcije  $RollingHash$ , ki smo jih dobili v datoteki, vrednosti pa so ocene pomembnosti ( $W_{ch_i}^D$ ).

## 3.2 FbHash-S

Med testiranje različnih datotek, so avtorji članka ugotovili, da algoritem **FbHash-B** na stisnjenih datotekah (tipov docx, pptx, pdf, zip, ...) ne deluje najbolje. Tako datoteki, ki imata 90 % podobnost v datoteki tipa docx, nimata veliko podobnosti na nivoju bajtov. Zato so razvili tudi verzijo algoritma **FbHash-S**, ki dosegla dobre rezultate tudi na stisnjenih datotekah.

Algoritem **FbHash-S** deluje na sledeč način. Najprej datoteko razširimo (angl. *uncompress*) in uporabimo algoritem **FbHash-B** na vsaki izmed dobljenih datotek. Podobnost med datotekami je izračunana tako, da med sabo primerjamo slike in besedila. Končna podobnost je izračunana kot povprečje podobnosti med posameznimi datotekami.

Zaradi dodatnega dela je algoritem **FbHash-S** časovno bolj zahteven od algoritma **FbHash-B**.

## 3.3 Primerjanje dokumentov

Podobnost med datotekama izračunamo s pomočjo kosinusne razdalje [9]. Najprej poskrbimo, da ima vsaka izmed datotek izračunano tabelo ocen pomembnosti. Podobnost med datotekama tako izračunamo s formulo:

$$Similarity(D_1, D_2) = \frac{\sum_{i=0}^{n-1} W_{ch_i}^{D_1} \cdot W_{ch_i}^{D_2}}{\sqrt{\sum_{i=0}^{n-1} W_{ch_i}^{D_1}} \cdot \sqrt{\sum_{i=0}^{n-1} W_{ch_i}^{D_2}}} \cdot 100$$

Pri implementaciji lahko opazimo, da lahko za vsak dokument vnaprej izračunamo  $\sum_{i=0}^{n-1} W_{ch_i}^D$  in da je v števcu dovolj, da zmnožimo ocene pomembnosti, ki so v preseku ključev.

## 3.4 Uporabljene funkcije za uteževanje

Tukaj bomo predstavili različne funkcije, ki smo jih preiskusili za iskanje primernih uteži posameznih delov datoteke.

### 3.4.1 funkcije za računanje uteži dela besedila

Najprej si oglejmo funkcijo za računanje uteži dela besedila, ki je predstavljena v članku [4]:

$$W_{ch_i}^D(chf_{ch}^D) = 1 + \log_{10} \left( \frac{chf_{ch}^D}{n} \right),$$

pri čemer je  $n$  število vseh dokumentov. Relativna frekvenca dela besedila je število med 0 in 1, torej je njegov logaritem med  $-\infty$  in 0. Menimo, da je to napaka (ali pa so avtorji članka uporabili drugačno funkcijo normiranja kot mi, saj

je nikjer ne opišejo), zato smo uteži dela besedila ( $W_{chi}^D$ ) izračunali nekoliko drugače.

Funkcijo smo popravili, da je rezultat število na intervalu med 0 in 1. Izračunamo relativno frekvenco dela in za 1 povečan rezultat logaritmiramo:

$$W_{chi}^D(chf_{ch}^D) = \log_2 \left( 1 + \frac{chf_{chi}^D}{n} \right).$$

Ker je ulomek povečan za 1 število med 1 in 2, smo vzeli logaritem z bazo 2, da kot utež dobimo število med 0.5 in 1.

Predlagamo še drugo funkcijo za računanje uteži, ki je razmerje med frekvenco dela besedila in številom takih delov besedila  $n$ . Tako je rezultat na intervalu  $[0, 1]$ :

$$W_{chi}^D(chf_{ch}^D) = \frac{chf_{chi}^D}{n}.$$

### 3.4.2 funkcije za računanje dokumentne uteži dela besedila

Članek [4] definira dokumentno utež dela besedila tako:

$$docw_{chi}^D(df_{ch}) = \log_{10} \left( \frac{1000m}{df_{ch}} \right),$$

kjer je  $m$  je število vseh dokumentov. Notranji ulomek je število med 0 in  $1000m$ , torej je rezultat število med 1 in  $3 \log_{10} m$ . Mi smo dokumentno utež dela besedila izračunali kot relativno frekvenco dokumentov, ki vsebujejo posamezen kos  $ch_i$  in rezultat logaritmirali:

$$docw_{chi}^D = \log_{10} \left( \frac{m}{df_{ch}} \right).$$

Dobljeno število je očitno nenegativno in je med 1 in  $\log_{10} m$ .

(Če bi želeli normirano vrednost:

$$docw_{chi}^D = 1 - \log_{10} \left( \frac{df_{ch}}{m} \right).$$

)

## 4. DOSEŽKI ČLANKA

## 5. NAŠI EKSPERIMENTI

## 6. REZULTATI

## 7. ZAKLJUČEK

## 8. ZAHVALA

Mogoče zahvala avtorjem za narjeno delo al kej.

## 9. REFERENCES

- [1] Automated evaluation of approximate matching algorithms on real data. volume 11, pages S10 – S17, April 2014. Proceedings of the First Annual DFRWS Europe.
- [2] F. Breiting, K. P. Astebøl, H. Baier, and C. Busch. mvhash-b - a new approach for similarity preserving hashing. In *2013 Seventh International Conference on IT Security Incident Management and IT Forensics*, pages 33–44, March 2013.
- [3] F. Breiting and H. Baier. Similarity preserving hashing: Eligible properties and a new algorithm mrsh-v2. In *Digital Forensics and Cyber Crime*, pages 167–182, October 2012.
- [4] D. Chang, M. Ghosh, S. K. Sanadhya, M. Singh, and D. R. White. Fbhash: A new similarity hashing scheme for digital forensics. In *The Digital Forensic Research Conference*, volume 29, pages S113–S123. DFRWS, July 2019.
- [5] D. Chang, S. Sanadhya, and M. Singh. Security analysis of mvhash-b similarity hashing. *Journal of Digital Forensics, Security and Law*, 11(2):2, 2016.
- [6] N. Harbour. Dcfldd. defense computer forensics lab. *online*, 2002.
- [7] J. Kornblum. Identifying almost identical files using context triggered piecewise hashing. *Digital Investigation*, 3:91–97, September 2006. The Proceedings of the 6th Annual Digital Forensic Research Workshop (DFRWS '06).
- [8] V. Roussev. Data fingerprinting with similarity digests. *IFIP Advances in Information and Communication Technology*, 337:207–226, September 2010. Advances in Digital Forensics VI. DigitalForensics.
- [9] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. In *INFORMATION PROCESSING AND MANAGEMENT*, pages 513–523, 1988.