3E-6

ファイル類似度評価システムに関する考察

高田　慎也　　松村　隆宏　　元田　敏浩
NTT セキュアプラットフォーム研究所
takada.shinya@lab.ntt.co.jp

1.はじめに

類似するファイルを高速かつ高精度に見つけ出すことに対するニーズは高く、こうした分野で使用されるファイル類似度の評価方法としては、例えば、ファイルのエントロピー値を比較することで類似度を測定する方法の研究が盛んに行われている[1][2][3][4]。Mccreightらは、測定法をさらに発展させ、ファイルサイズで重みを付けた Weighted Entropy を使って、類似度を評価することを提案している[1]。しかしファイル全体のエントロピー値を使ったファイルの類似度評価は、無関係の2 つのファイルが偶然大きな類似度をとる事象が多々発生しうるという問題があった[1][3]。これに対して区分エントロピー値をファイルの区分ごとに計算し、得られるファイル区分エントロピースペクトルを比較することで、より詳細な類似度を判定する方式を提案してきた[5]。今回特に実行形式ファイルへの適用評価をすることで提案方式の有効性を検証する。

2. エントロピー値の計算方法

エントロピー値は閉域系における順序性の程度の指標値である。情報理論としてのエントロピー値は、電子データを２５６通りで表現されるバイトの集合とみなす。そして、そのバイト集合に偏りがある場合は、電子データが規則性のある状態（エントロピー値＝0）、反対に偏りが存在しない場合はランダムな状態（エントロピー値＝8）と見なす。そして、計算されたデータの"ランダムさ"は、「エントロピー値」という絶対値として表現される。エントロピー値の計算方式は、

$$E = -\sum_{i=0}^{255} P_i \log_2(P_i) \qquad [式\ 1]$$

で定義される。

3. エントロピー値を用いた類似度評価方式

Weighted Entropy を用いたファイルの類似度評価式は、Mccreight らの特許[1]には参考として

$$類似度\ 1 = \log(E_1 - E_2)\log(S_1 - S_2) \qquad [式\ 2]$$

で与えられている。しかしながら、この式は一例であって有意な値をとらない。例えば、E_1-E_2 が負の値を取る場合、対数計算が行えない点や、E_1-E_2 の値が 1 以下の場合、類似度が負の値となってしまう点等で実用には向かない。このため、類似度評価式として

$$類似度\ 2 = \frac{\sum_{i=1}^{n}|E1_i - E2_i|}{n} \qquad [式\ 3]$$

を本考察において上記式 3 を提案する。

この式では、比較対象の 2 つのファイルを始点から固定長でそれぞれ区分に分割し、各区分での区分エントロピー値をそれぞれ（$E1_i$, $E2_i$）求め、この例では値の差を取り、これをファイルの最後まで繰り返した後、差の平均を計算することで、さらに類似性 (D) を評価する。換言すれば個々のファイルの区分エントロピースペクトルを測定し、差の平均を求める。差の平均が 0 の時 2 つのファイルは一致し、差の増大とともに 2 つのファイルの類似度は低くなり、最大値は 8 となる。

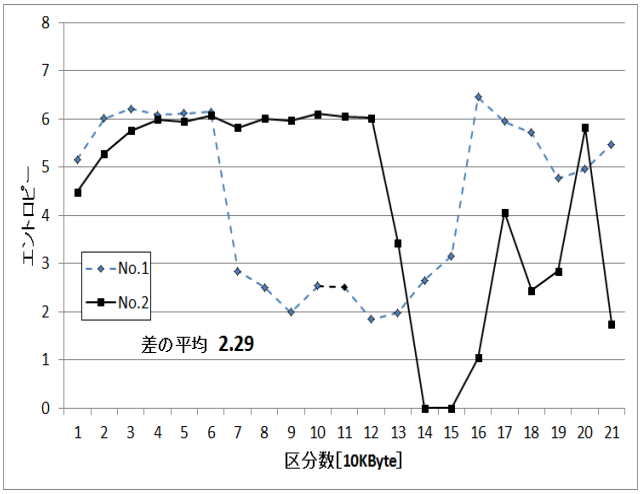
4.類似度評価方式の実行形式ファイルへの適用例 1

表 1 は類似度評価例 1 の対象とした 2 つのファイルのエントロピー値とファイルサイズを表している。

表 1.類似度評価例 1 の対象となる 2 つファイル

No.	File Name	Entropy	WEntropy	File Size
1	icwconn1.exe	5.445616	66.62589	205824
2	r200_001.exe	5.446029	66.60548	204864

図 1 .提案方式で評価した区分エントロピースペクトル



2 つのファイルは全く無関係にもかかわらずエントロピー値とファイルサイズがほぼ一致している。このため、これらを Mccreight らの方式（式 2）で評価した場合には、高い類似度を示すことが推測される。一方、図 1 は提案方式（式 3）でこれらのファイルの区分エントロピースペクトルを評価したものである。図のようにスペクトルは大きくことなり、差の平均も大きいことから、異なるファイルと識別できることが分かった。

5.類似度評価方式の実行形式ファイルへの適用例 2

図 2、図 3 は以下での提案方式の検証を目的としたモデルケースとして実行形式ファイルのコードの編集量の異なる 2 つのケースについて、提案方式（式 3）を用い区分エントロピースペクトルを評価した結果である。

ケース 1 の図 2 は比較先実行形式ファイルのコード全体に対して比較元実行形式ファイルの 64%のコードが含まれる 2 つの実行形式ファイルを比較したものである。スペクトル形はほとんど一致し、差の平均も 0.021 と極めて小さい値となり異なるファイルとは識別できないことが分かった。

一方ケース 2 の図 3 は比較先実行形式ファイルのコード全体に対して比較元実行形式ファイルの 16%のコードが含まれる 2 つの実行形式ファイルを比較したものである。スペクトル形は一致せず、差の平均も 0.69 とケース 1 よりも大きな値となり異なるファイルと識別できることが分かった。

6.結果の評価

適用例 1 より、提案方式（式 3）による実行形式ファイルの類似度評価は、Mccreight らの方式（式 2）よりも誤検出が少ないことが分かった。大量のファイルについて類似度評価を行う場合には、どうしてもエントロピー値やファイルサイズが似通った値をとるケースが発生してしまう。このような場合にあっても提案方式（式 3）はより正確に類似度を評価することができた。

適用例 2 より、提案方式（式 3）では図 2 のマイナーバージョンアップ級のコードの変更量に対しては実行形式ファイルを極めて類似度が高いものと判定することが分かった。一方図 3 のメジャーバージョンアップ級のコードの変更量に対しては実行形式ファイルを類似度が低いものと判定することが分かった。

これらの性質を利用した応用例として、実行形式ファイルとそれを使用するファイルの認証認可への応用が期待できる。ファイル管理上好ましくない実行形式ファイルを実行することで、ファイルの内容が漏えいしたり、意図と異なる改変が加えられてしまったりすることを防ぐため、あらかじめファイルを実行できる実行形式ファイルを規定しておき、提案方式を適用することで実行形式ファイルの冗長性を持った認証をすることが可能となる。実用上メジャーバージョンの場合にのみ規定を更新し、マイナーバージョンアップでは規定を更新する手間が省けることが期待される。

図 2.ケース 1 の実行形式ファイルの区分エントロピースペクトル
マイナーバージョンアップ級

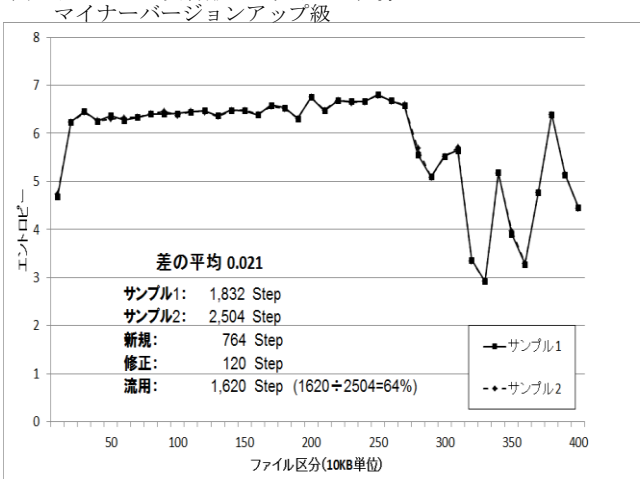
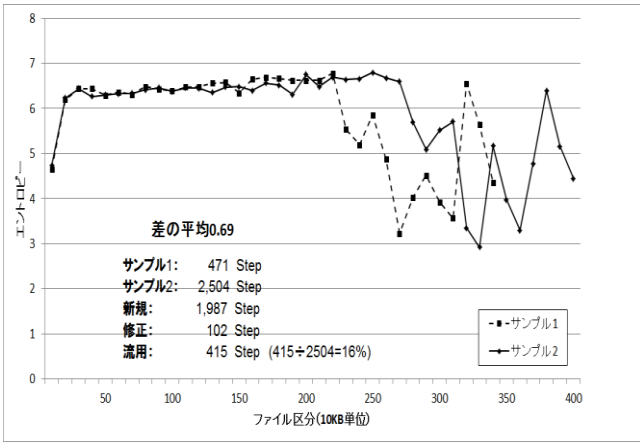


図 3 ケース 2 の実行形式ファイルの区分エントロピースペクトル
メジャーバージョンアップ級



7.今後の予定

今後、類似度判定の適用領域として、6 章で述べた、実行形式ファイルの勝手な差し替えによるスプーフィング対策の実現性を検証したい。また提案方式をツールとして実装し、いろいろな人に使ってもらうことで、提案方式の適用領域拡大に関する知見を収集したい。

8.参考文献

[1] Mccreight et al. “System and method for entropy-based near-match analysis.” 国際特許 WO2010/107659 A1
[2]Davis et al.Guidance Software “Utilizing Entropy to Identify Undetected Malware”
[3] 松本ら “エントロピーとフォレンジック” http://www.netagent-blog.jp/archives/51451285.html: 2010
[4]高田他”類似度を用いたファイル追跡に関する一手法の提案” CSS2012
[5]高田他” ファイルのエントロピー測定による類似度評価の新手法に関する提案” 第 60 回 CSEC 研究会