2D-06

エントロピーと DP Matching を用いた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]。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]。また方式をさらに実行形式ファイルへ適用することで、類似実行形式ファイルの検索における提案方式の有効性を検証してきた[6]。今回は従来の方式に DP Matching を加えることで方式の更なる改善をめざす。

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、

$$E = -\sum_{i=0}^{255} P_i \log_2(P_i)$$

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

$$\text{類似度 } 1 = \log(E_1 - E_2)\log(S_1 - S_2)$$

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

$$\text{類似度 2(差分平均)} = \frac{\sum_{i=1}^{n} |E1_i - E2_i|}{n}$$

上記式を用いる。

この式では、比較対象の 2 つのファイルを区分に分割し、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、差分平均を計算することで類似性を評価する。差分平均が 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.評価方式の実行形式ファイルへの適用例

図 1 は類似度評価方式を Adobe Acrobat（AcroRd32.exe）の Ver.11.0.0 と Ver.11.0.0.01 に適用した結果の図である。2 つのスペクトルの差分平均は 0.11 となり類似度は高いと推測することができる。次に AcroRd32.exe について各 Major Version ごとにサンプルからその Minor Version を検索した結果を表 1 に示す。ここで検索は、「c:¥配下の全 exe ファイルでサイズがサンプルの 85%～115%でかつ差分平均（スレッショルド）が“0.30”以下を検索結果とする」との条件で行った。スレッショ

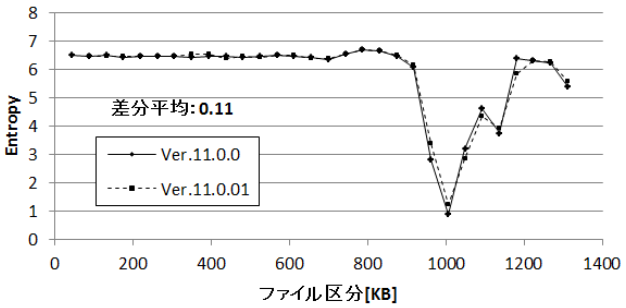

図 1. AcroRd32.exe の Ver.11.0.0 と Ver.11.0.01 の区分エントロピースペクトル

表 1. AcroRd32.exe の検索結果

サンプル	マイナーバージョンアップ数	差分平均評価対象ファイル数	適合率（Precision）	再現率（Recall）
Ver9.3.0	18	201	100%	50%
Ver10.1.0	12	88	100%	75%
Ver11.0.0	9	91	100%	22%

ルドを 0.30 としたのは、適合率を重視し、一番条件の厳しい Ver.11.0.0.0 の適合率が 100%となるものを選択した。結果は一様に再現率が低いものとなり、検索方式としては改善の余地があることが分かった。

5.類似度評価方式の更なる改善

図 2 は類似度に差が見られた Ver11.0.0 と Ver11.0.05 の区分エントロピースペクトルを表したものである。特

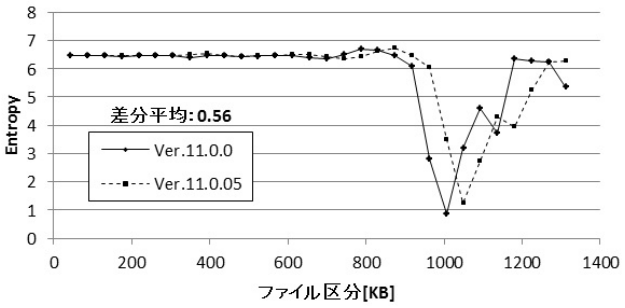

図 2.AcroRd32.exe の Ver.11.0.0 と Ver.11.0.05 の区分エントロピースペクトル

徴として、スペクトルがずれ、ピーク位置があっていないことが類似度評価の低下を招いていることが分かる。これを解決するために、いわゆる DP Matching を 2 つのスペクトル対に対して行うことを試みた。ここで DP Matching は以下の式で表現される。

$X = (x_1, x_2, \ldots, x_n)$、$Y = (y_1, y_2, \ldots, y_n)$について

動的計画法により以下を計算

$$D(X,Y) = f(n,m)$$

$$f(t,i) = \|x_t - y_i\| + \min \begin{cases} f(t,t-1)\ X\ Stutter \\ f(t-1,i)\ Y\ Stutter \\ f(t-1,i-1)\ noStutter \end{cases}$$

$$f(0,0) = 0, \qquad f(t,0) = f(0,i) = \infty$$

DP Matching を施したスペクトルを図 3 に示す。2 つのスペクトルは図 2 より一致度が高くなりその差分平均は 0.56 から 0.052 へと大きく改善された。また第 4 章と同条件での Minor Version の検索結果を表 2 に示す。

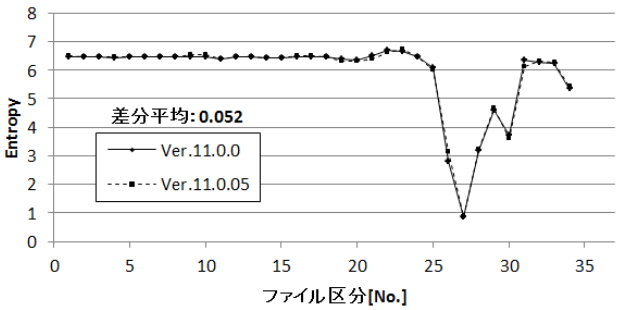

図 3.AcroRd32.exe の Ver.11.0.0 と Ver.11.0.05 の区分エントロピースペクトル(DP Matching あり)

表 2. AcroRd32.exe の検索結果(DP Matching あり)

サンプル	マイナーバージョンアップ数	差分平均評価対象ファイル数	適合率 (Precision)	再現率 (Recall)
Ver9.3.0	18	201	100%	100%
Ver10.1.0	12	88	100%	100%
Ver11.0.0	9	91	100%	100%

6.結果の評価

第 5 章の表 2 の結果から DP Matching により適合率、再現率がともに 100%に改善された。このように DP Matching を適用することで非類似ペアの非類似性はそのまま保ったまま、類似ペアの類似性を強調して検出することが可能であることが分かった。これにより適合率 100%を維持しつつ、再現率 100%を達成し類似ペアの取りこぼしを抑えられることが分かった。

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(もしくはできない)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(Minor Version Up の場合は検索条件の変更を必要としない冗長性)を持った認証をできるようになることが期待される。

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（主にブラックリストへ応用）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。さらには検証対象を Acrobat 以外にも広げ、スレッショルドの適正値を探りたい。

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
[6]高田他” ファイル類似度評価システムに関する考察” 第 76 回情報処理学会全国大会