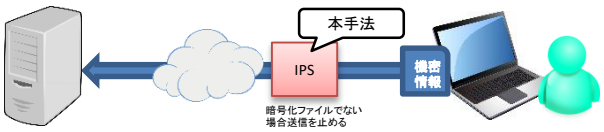
類似度による暗号化ファイルの検出
Detection of encrypted file by similarity

楠美 淳弥・マネジメント分科会・情報セキュリティ大学院大学

Abstract: By detecting that it is an encrypted file, it is thought that it is possible to prevent the risk of information leakage to sending an important file to the outside with an unencrypted plain text file. Therefore, in this research we propose a new method to detect encrypted files.
In this method, we measure the similarity by using the Jaro - Winkler distance the data of each when file is separated by a certain length, thereby detecting whether it is a plaintext file or an encrypted file.

【背景・目的】

誤送信による情報漏えいを対策として送信するファイルをファイルのデータ間の類似度に着目することで平文，暗号文検出を行う．暗号文であった場合，解読が難しい脆弱ではない暗号方式を採用しているかその暗号方式の特徴をもとに判別する手法を提案する．



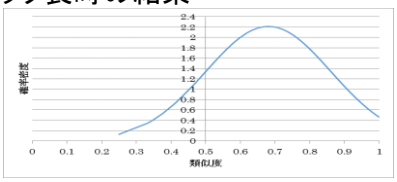
【類似度解析】

RC4で暗号化したファイルと安全性の高いAESで暗号化したファイルと平文ファイルを用意した．そのデータを1,2,4,8,16,32,64,128,256,512,1024の長さのブロックに20~100ブロック分割し，分割したブロック同士をJaro-Winkler距離という二つ文字列の違いがどの程度あるか測るアルゴリズムを用いて解析を行った．

解析の統計を取った時の平文，暗号文で大きく結果が異なったデータを下記に示す．

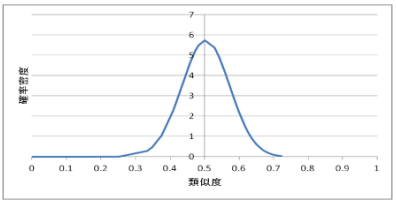
・平文ファイル8ブロック長時の結果

平均値　　:0.6800
不偏分散値:0.0325
標準偏差値:0.1803



・暗号化ファイル8ブロック時の結果

平均値　　:0.5000
不偏分散値:0.0048
標準偏差値:0.0695



【平文/暗号文識別】

機械学習(サポートベクターマシン)を用いて，類似度解析結果の統計情報を学習データにし，平文ファイルと暗号ファイルの識別が行えるか実験を行った．

ブロック長	検出精度(ブロック数)								
	100	90	80	70	60	50	40	30	20
1	100%	98.88%	96.88%	92.77	88.88%	78.88%	80%	86.67%	71.11%
2	100%	100%	100%	100%	100%	100%	96.11%	72.22%	63.89%
4	100%	100%	100%	100%	100%	100%	100%	99.44%	73.89%
8	100%	100%	100%	100%	100%	100%	100%	100%	99.33%
16	100%	100%	100%	100%	100%	100%	100%	100%	100%
32	100%	100%	100%	100%	100%	100%	100%	100%	100%
64	100%	100%	100%	100%	100%	100%	100%	100%	100%
128	100%	100%	100%	100%	100%	100%	100%	100%	100%
256	100%	100%	100%	100%	100%	100%	100%	100%	100%
512	100%	100%	100%	100%	100%	100%	100%	100%	100%
1024	100%	100%	100%	100%	100%	100%	100%	100%	100%

【暗号方式判定】

暗号化に利用した，OpeanSSLでファイルを暗号化した際の特徴や，RC4，AESの各暗号化方式の暗号化の特徴を調査するために暗号化のファイルサイズや暗号化にかかる処理時間を調査した．

ファイル形式	平文ファイルサイズ(byte)	AESファイルサイズ(byte)	暗号化所要時間(秒)	RC4ファイルサイズ(byte)	暗号化所要時間(秒)
Word(docx)	160026	160048	0.004	160042	0.004
Word(docx)	2594869	2594896	0.038	2594885	0.03
text	129	160	0.001	145	0.001
text	307	336	0.001	323	0.001
Execl(xlsx)	10712	10736	0.002	10728	0.001
Execl(xlsx)	11957	11984	0.002	11973	0.001
pdf	675863	675888	0.007	675879	0.006
pdf	23711968	23712000	0.341	23711984	0.289
pptx	472527	472544	0.009	472543	0.008
pptx	266371	266400	0.005	266387	0.004

【検出のプロセス】

(1)暗号文の類似度を調べ平文／暗号文を判別する．
　⇒平文を不許可
(2)暗号文についてサイズを調べ，16バイトの倍数かどうか調べる．
　⇒倍数でない場合は不許可(RC4と判定)
(3)もとの平文に比べ，暗号文のサイズが16バイト大きいかどうか調べる．
　⇒16バイト大きい場合は不許可(RC4と判定)