自動化されたマルウェア動的解析システムで収集した大量APIコールログの分析

MWS 2013

藤野 朗稚[†]，**森 達哉**[†]

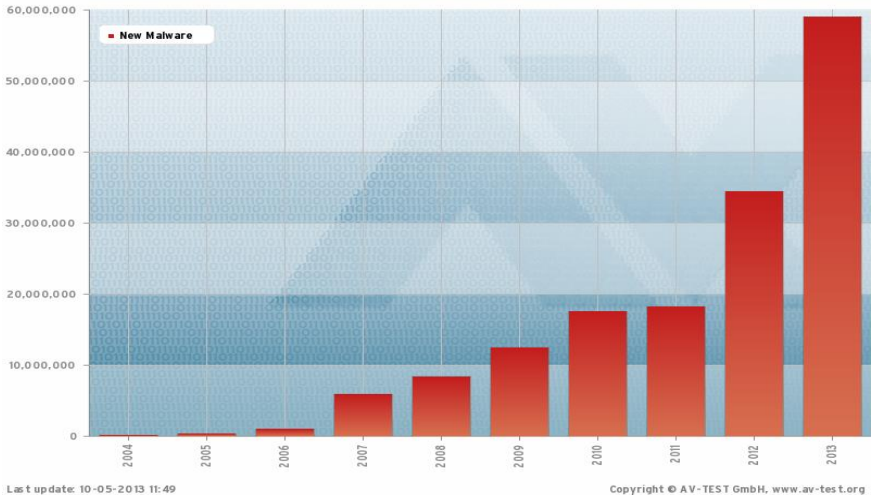
[†]早稲田大学 基幹理工学部 情報理工学科

目次

- 研究背景
- 提案手法
- 結果
- まとめ

研究背景

- マルウェアは日々驚くべき速さで増加している.
- 一日当たり20万個の新しいマルウェアが発見されている[1].



新たに発見されてきたマルウェア[2]

- このような状況の中，マルウェアに対応するためにはより効率のよい解析方法を確立する必要がある.

[1] Kaspersky Lab, ¥2012 by the numbers: Kaspersky Labnow detects 200,000 new malicious programs every day. "http://www.kaspersky.com/about/news/virus/2012/2012_by_the_numbers_Kaspersky_Lab_now_detects_200000_new_malicious_programs_every_day”.
[2] AV-TEST GmbH, The AV-TEST Institute registers over 200,000 new malicious programs every day. “http://www.av-test.org/en/statistics/malware/”

研究背景

マルウェアの解析方法には大きく分けて静的解析と動的解析の2種類がある.

静的解析:

マルウェアのソースコードを解析することでどのような挙動をするのかを詳細に調べるために行われる. 自動化が難しい.

長所: 動的解析よりも詳細に挙動がわかる.

短所: 解析に時間がかかる.

&

動的解析:

実際の感染者と類似した環境の中でマルウェアを動作させ, その結果を用いて解析する. 自動化がしやすい.

長所: 解析にかかる時間が少なくて済む.

短所: マルウェアに感染する可能性がある.

研究背景

- 大量に増え続けるマルウェアに対抗するために，動的解析を自動で行う技術が発達してきている．
  - Cuckoo sandboxというオープンソースのソフトウェアもそのような動的解析を自動化したシステムのひとつである．

自動動的解析システム使用時のイメージ図

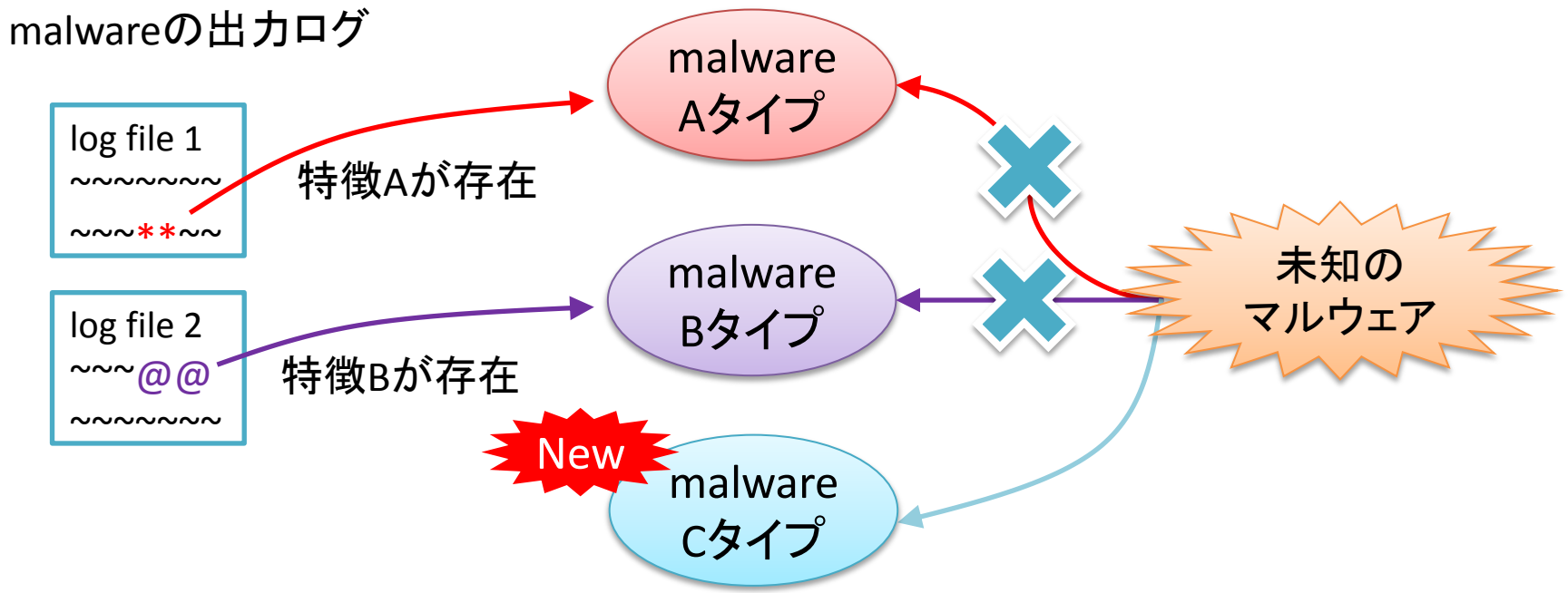
研究背景

- 大量のマルウェアに対して自動動的解析システムを使用した場合，ログファイルの量もマルウェアの数に比例して多くなってしまう.

解析するマルウェアの数が増えた場合

研究背景

- 膨大な数になったログを効率よく解析するためには，マルウェアの対策につながるような特徴を発見する必要がある.



マルウェアの特徴を発見できた場合

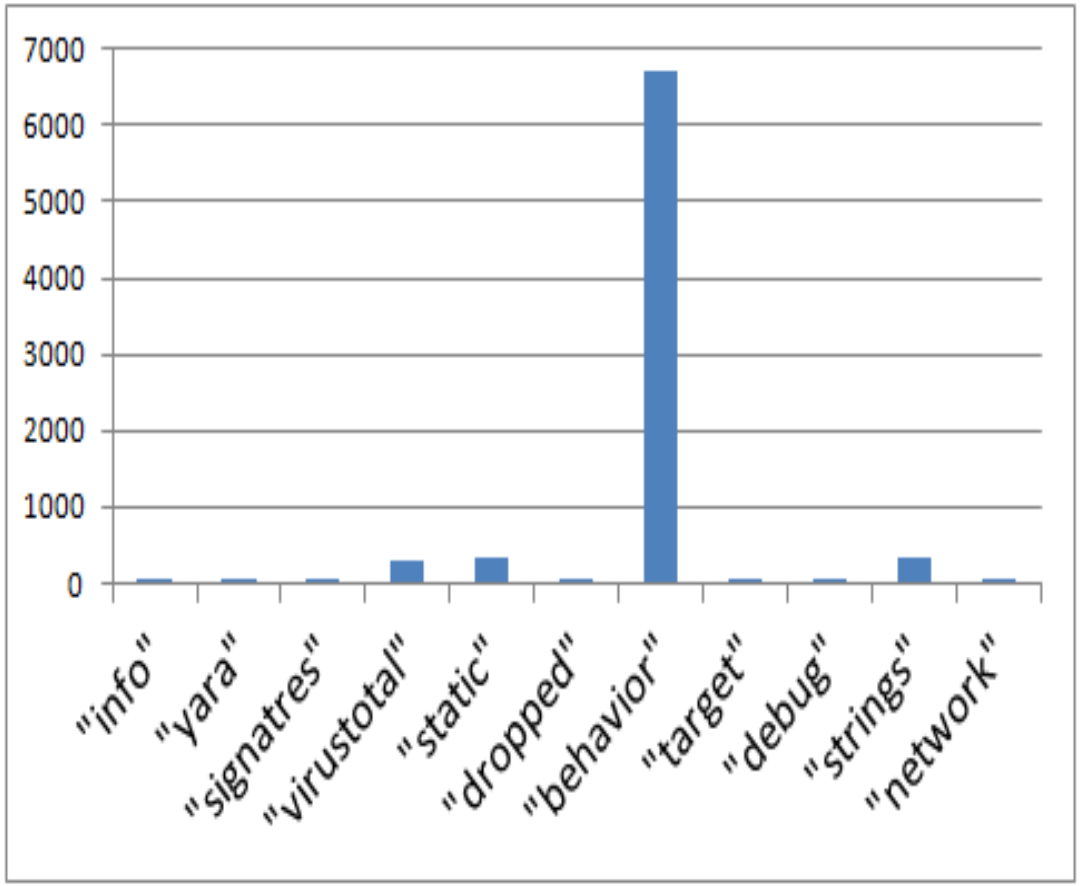
使用したデータについて

- FFRI-Dataset-2013
  - FFRI社より提供されたデータ.
  - 2012年9月から2013年3月に収取されたもの.
  - 検体数は，2644検体.
  - PE形式のものをCuckoo sandboxを用いて解析したログ.
  - ファイル形式は，json形式.

cuckoo sandboxのログのデータ構造と分布

e.g. ログのデータ構造

```
{
    “info”: {},
    “yara”: [],
    “signatures”: [],
    “virustotal”: {},
    “static”: {},
    “dropped”: [],
    “behavior”: {},
    “target”: {}
    “debug”: {},
    “strings”: [],
    “network”: {}
}
```



ログの各項目の行数

ログの具体例

e.g. virustotalの構造の一部

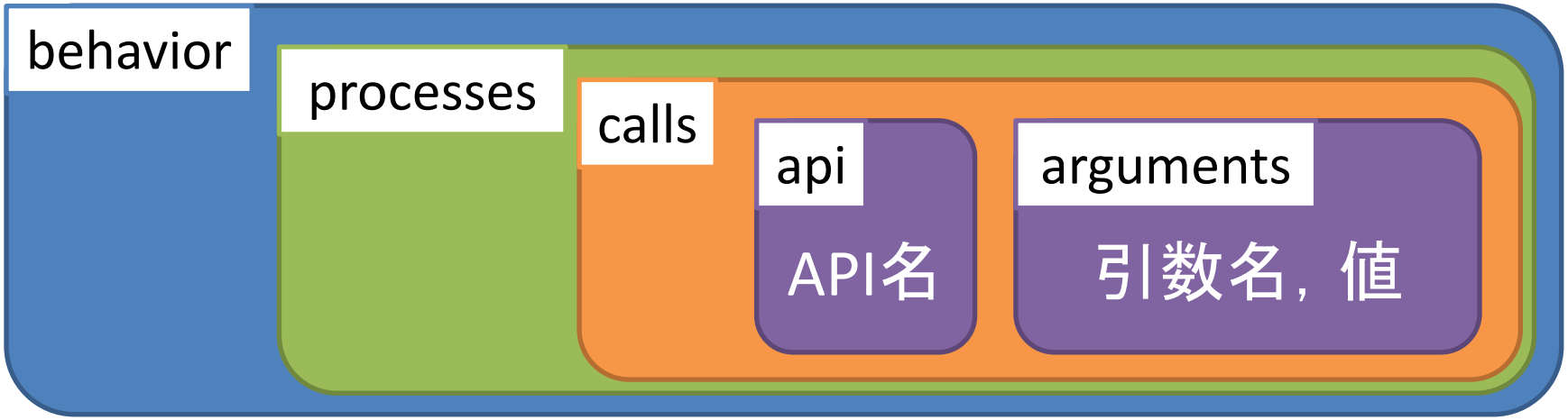
```
“scans”: {
    “AV NAME": {
        "detected": true or false,
        "version": “version",
        "result": “virus name",
        "update": “update“
    },
},
```

e.g. behaviorの構造の一部

```
"processes": [
  {
    "calls": [
      {
        "api": “API name",
        "arguments": [
          {
            "name": “name 1",
            "value": “value 1“
          },
          {
            "name": “name 2",
            "value":“value 2“
          }
        ]
      },
  },
]
```

研究背景

- マルウェアの特徴を発見するために，今回着目したのは出力ログの中のAPIコールに関するログである．
  - 下の図に示したbehavior以下の部分をAPIコールログと呼ぶことにする．



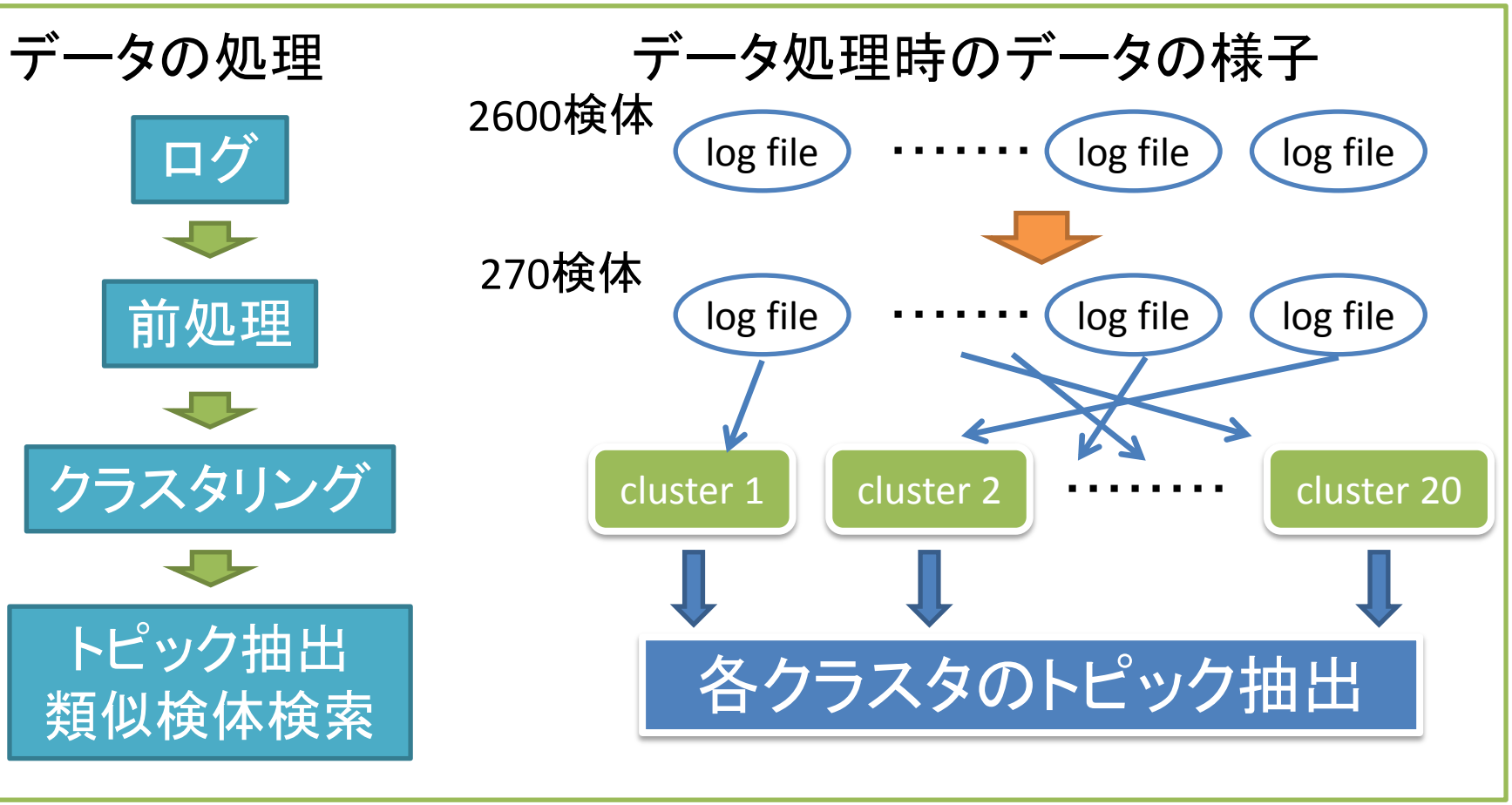
APIコールログのデータ構造の一部

目次

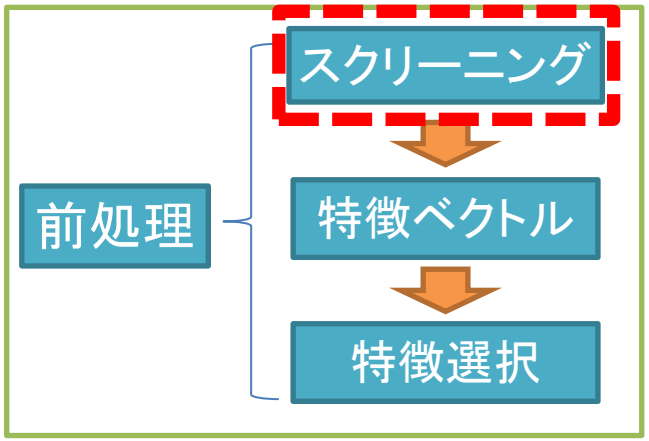
- 研究背景
- 提案手法
- 結果
- まとめ

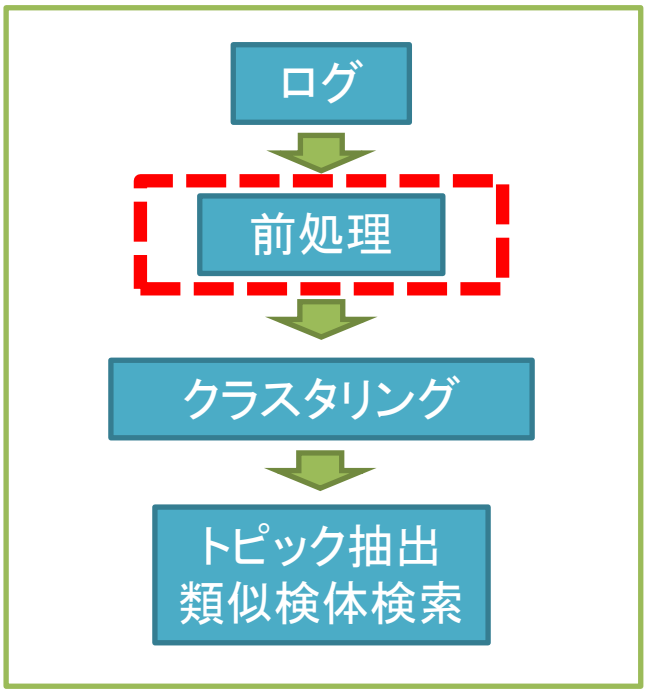
提案手法(概要)

- APIコールログを用いて特徴ベクトルを作成し，クラスタリングを用いて，データの分析をした.
  - クラスタリングは，k-meansとNMFを使用した.



提案手法(前処理)





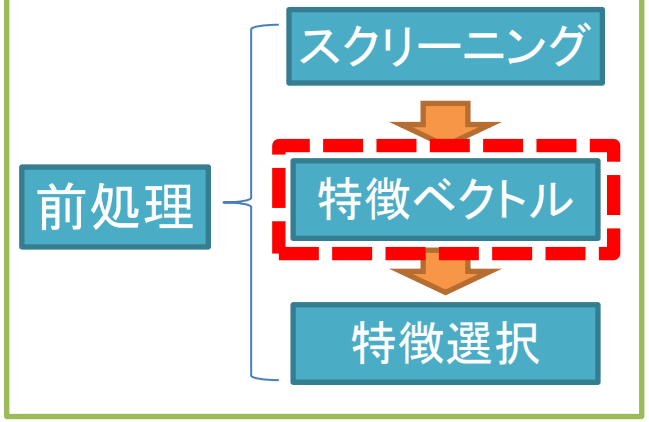
- スクリーニング

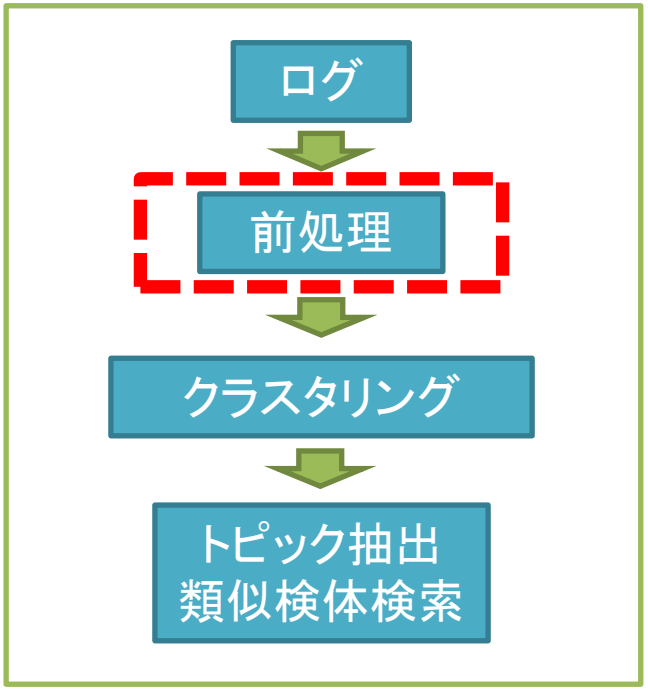
virustotal内のKasperskyの検査結果を使用しラベル付けをし，同一の種類の検体が10体以上存在する種類を選択した.

スクリーニングによって選択された種類

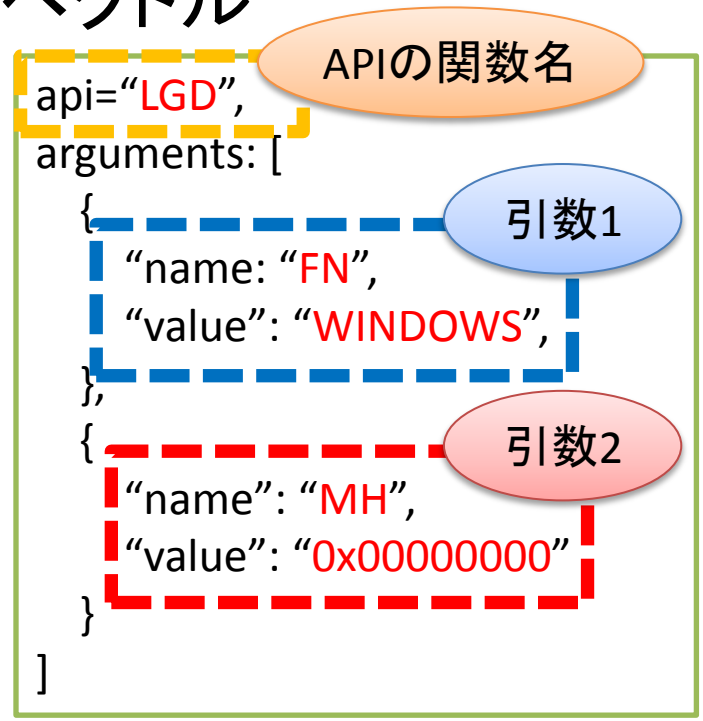
(1) B.Azbreg, (2) B.Buteraft, (3) B.Shiz, (4) B.Simda,
(5) T-Do.Agent, (6) T-Dr.Dorifel, (7) T-Dr.VB,
(8) T-PSW.Tepfer, (9) T-Spy.Zbot, (10) T.Agent,
(11) T.Diple, (12) T.Jorik.Vobfus, (13) T.Midhos,
(14) T.SelfDel, (15) T.Skillis, (16) T.VB,
(17) T.VBKrypt, (18) T.Vobfus, (19) T.Yakes,
(20) DangerousObject.Multi.Generic,
(21) W.AutoRun, (22) W.VBNA, (23) W.Vobfus, (24) W.WBNA,
(25) A.Gamevance, (26) A.iBryte, (27) A.iBryte.heur

提案手法(前処理)





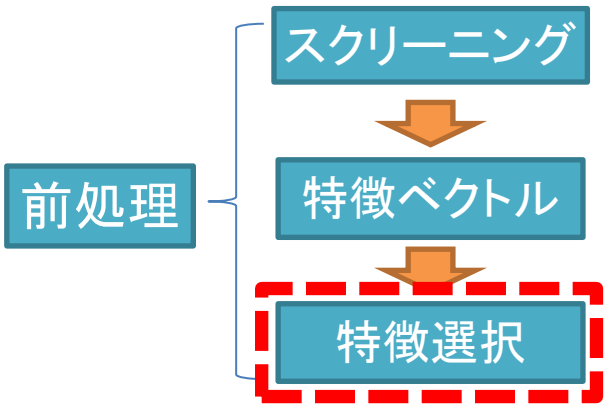
- 特徴ベクトル

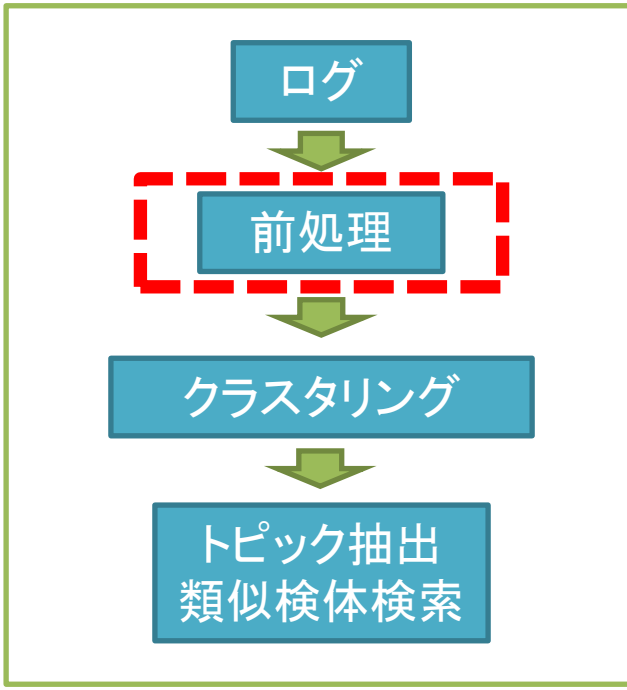


特徴ベクトル作成の具体例

Level 0	api	LGD
Level 1	api:n1:n2:: : :	LGD:"FN":"MH"
Level 2	api:n1:m(v1):n2:m(v2)::	LGD:"FN":"WINDOWS":"MH":MASK
Level 3	api:n1:v1:n2:v2:: : :	LGD:"FN":"WINDOWS":"MH":0x00000000

提案手法(前処理)

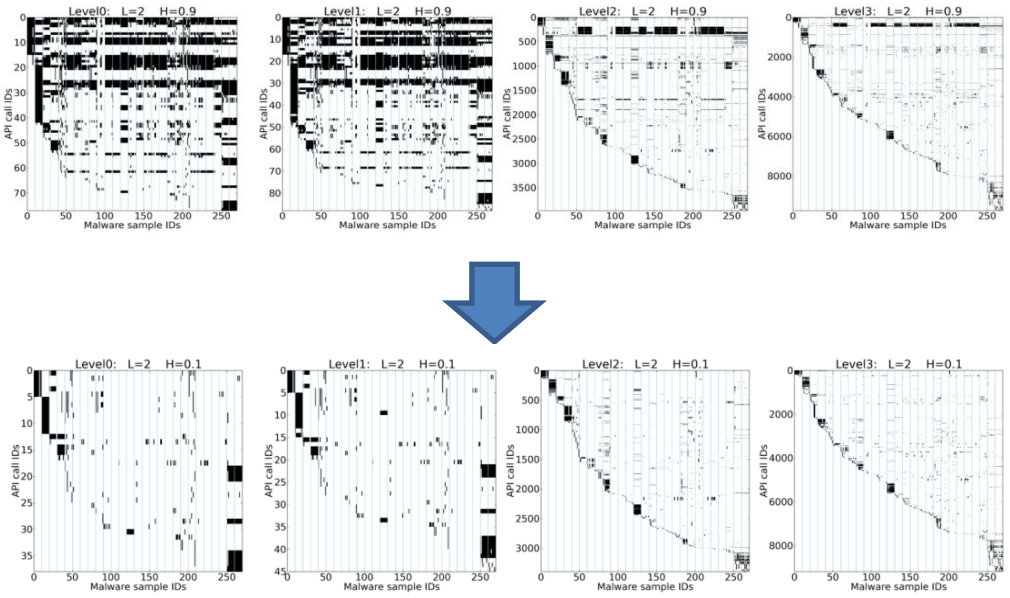




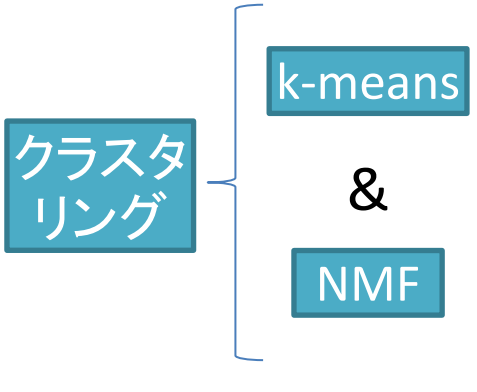
- 特徴選択

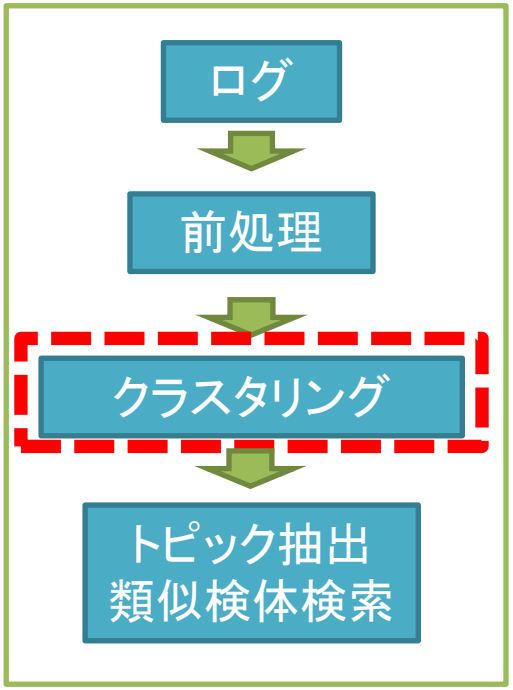
作成した特徴ベクトルの中から，出現頻度が高頻度のものと低頻度のものを除去する.

高頻度で出現する特徴ベクトルの除去



提案手法(クラスタリング)



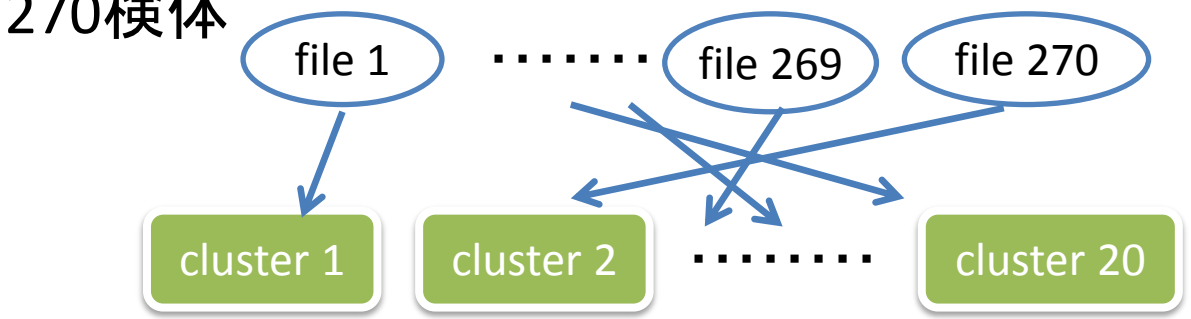


- クラスタリング

特徴ベクトルを用いることで，k-meansとNMFのそれぞれの手法でクラスタリングを行う.

トピック抽出はNMFの方でのみ可能.

クラスタの数は20とする.



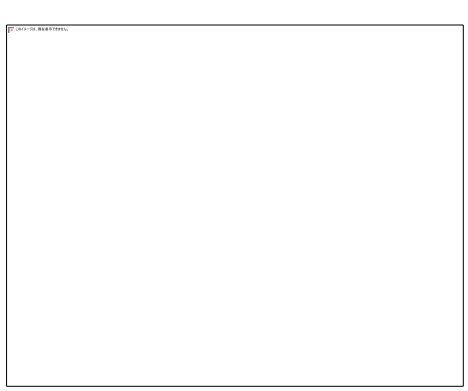
270検体のクラスタリングのイメージ図

NMF

- 非負値行列因子分解（NMF: Nonnegative Matrix Factorization）

0か正の値をもつ行列を解析する手法.

下図に示すように，行列Xを2つの行列に分解し，XとTVの距離を最小化する.



行列X

分解

APIコール	クラスタ 1	クラスタ 2
1	0	2.83
2	2.23	0
3	0	1.41

行列T

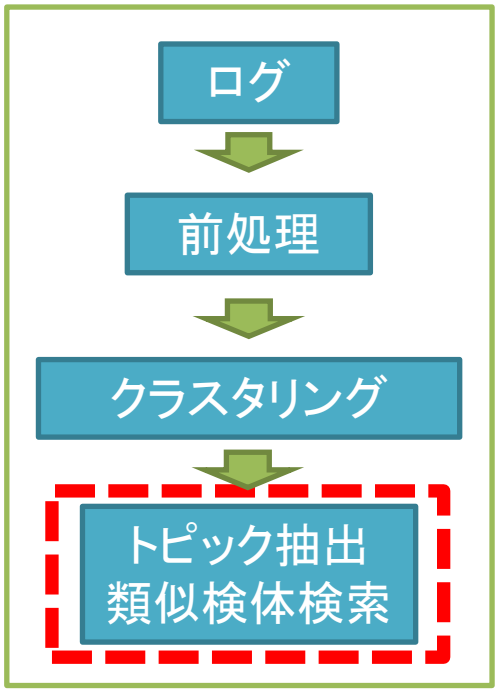
×

クラスタ	malware 1	malware 2	malware 3	malware 4
1	2.23	0	2.23	0
2	0	2.83	0	2.12

行列V

NMFのアルゴリズムイメージ図

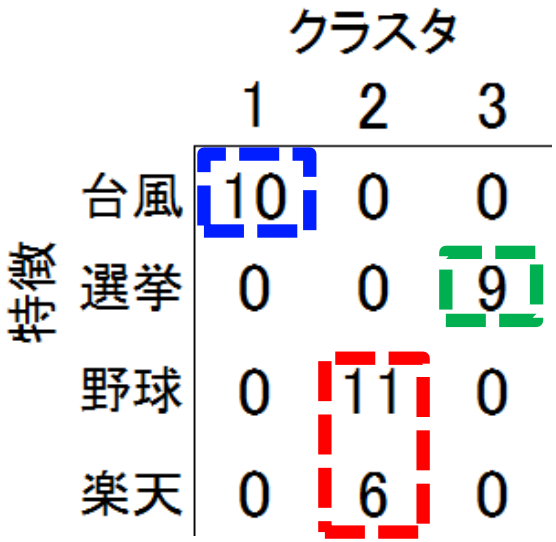
提案手法(トピック抽出)



- トピック抽出

NMFを用いて，同時生起しやすい単語をトピックとして抽出する.

トピックを構成する特徴を使うことで，類似の検体を検索することができると期待できる.



今年の野球では，楽天がリーグ優勝した.

楽天がリーグ優勝するのは，初めてのことだ.

類似検体検索のイメージ図

目次

- 研究背景
- 提案手法
- 結果
- まとめ

結果(クラスタリング)

- NMFのクラスタリング結果

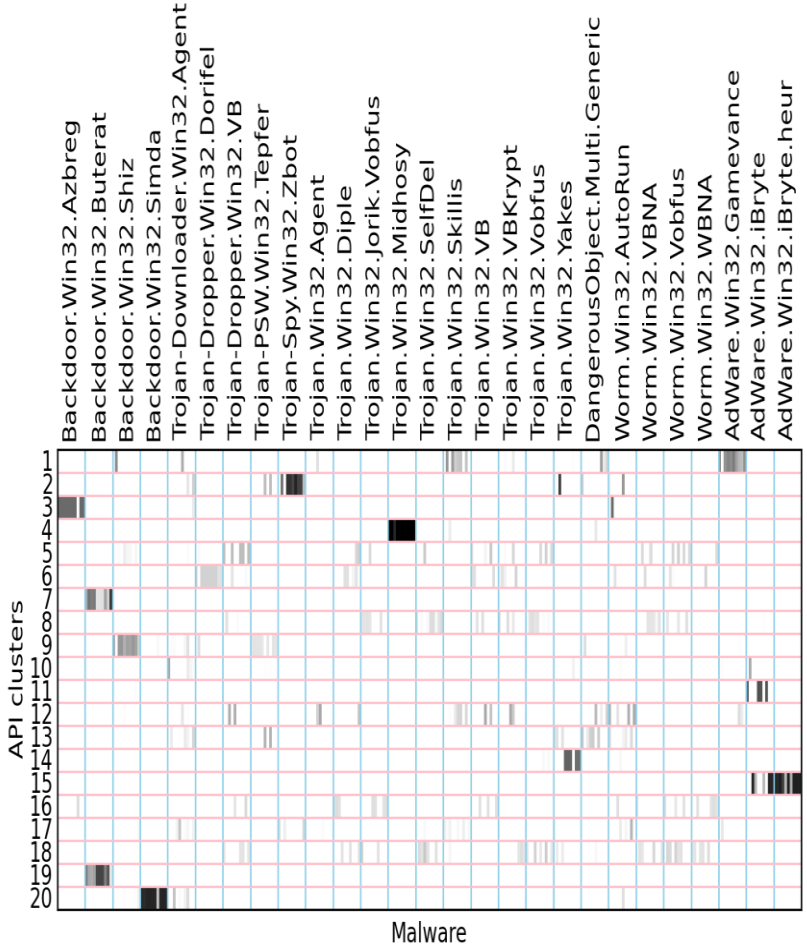
Backdoor 4種,

Adware.iBryte2種,

Trojanの一部

は綺麗にクラスタに分離された.

Worm系全般を含むVobfus, VB, VBNA, WBNAはファミリー毎のクラスタを持たなかった.



NMFによるクラスタリング

結果（トピック抽出）

クラスタ番号4のTrojan.Win32.Midhosの重要度が高い上位5個のトピック

- "RegOpenKeyExW":"Registry":"0x80000002":"SubKey“
  :"Software*n*Microsoft*n*COM3":"Handle":"0x000000f4"
- "LdrGetProcedureAddress":"ModuleHandle":
  "0x77cf0000":"FunctionName":"CallNextHookEx":"Ordinal":"0"
- "NtCreateSection":"SectionHandle":"0x00000070":"DesiredAccess":
  "0x00000004":"ObjectAttributes:":"FileHandle":"0x0000006c"
- "LdrLoadDll":"Flags":"522168":"FileName":"uxtheme.dll":
  "BaseAddress":"0x58730000"
- "LdrLoadDll":"Flags":"522348":"FileName":
  "C:*n*WINDOWS*n*system32*n*uxtheme.dll":"BaseAddress":"0x58730000"

上記のトピックをすべて持つ検体を
全検体から検索したところ，15体が該当し，
10体は今回使用した検体
5体は今回使用しなかった検体
であった．5検体の種別は右図に示す．

検体	Kaspersky	他のAVソフト
No.1	HEUR: Trojan. Win32.Generic	Midhos ファミリー
No.2		
No.3		
No.4		
No.5		Midfosファミリー

目次

- 研究背景
- 提案手法
- 結果
- まとめ

まとめ

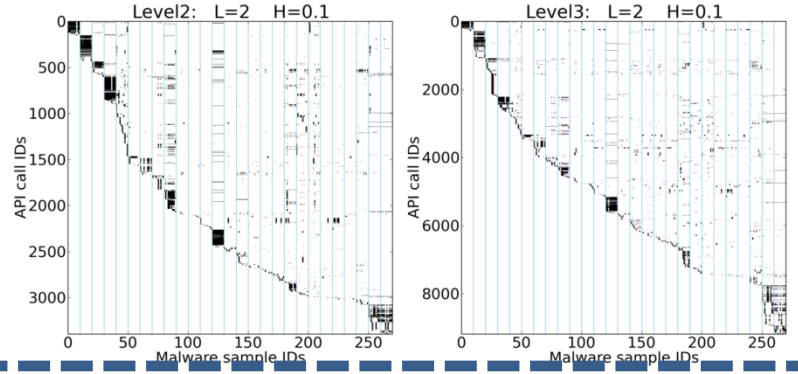
- APIコールを機械学習を用いて分析するための前処理方法を提示した.
- APIコールで分類可能なマルウェア種別と不可能なものが存在する.
- NMFを利用することで, トピックを自動抽出し, 類似する検体の検索に応用可能であることを示した.

今後の課題

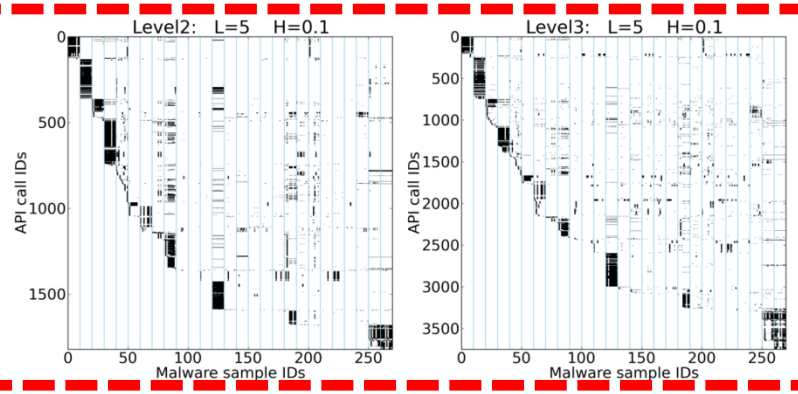
- 未使用のbehaviorの項目の活用.
- APIコールの時間的流れの使用.
- 未知のマルウェアの分類.
- クラスタの精度評価

補足資料

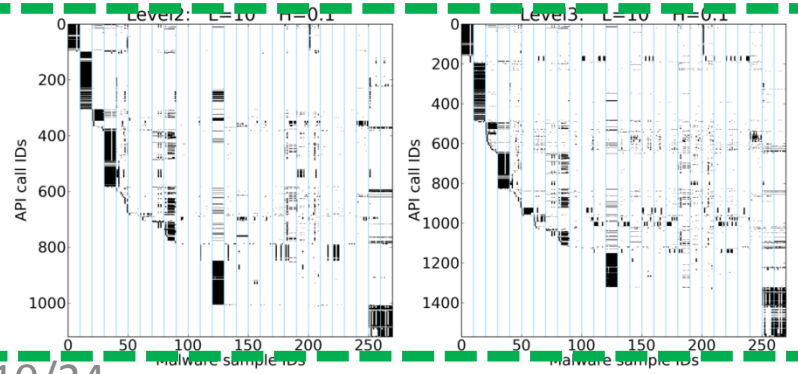
特徴選択時の図（低頻度）



2以下の数の検体で使用されたものを削除



5以下の数の検体で使用されたものを削除



10以下の数の検体で使用されたものを削除

k-means

- k-meansとは，クラスタリングの手法のひとつである.

e.g. 3つのクラスタに分類する場合.