利用者名：コンピュータサイエンス学部　准教授　宇田 隆哉

Scopus

Title: Fast Preprocessing by Suffix Arrays for Managing Byte n-grams to Detect Malware Subspecies by Machine Learning （機械学習による亜種マルウェア検出のためのバイト n-gram 処理に接尾辞配列を用いた高速処理）

Authors: Kouhei Kita and Ryuya Uda
（喜多 航平（東京工科大学　卒業生）、宇田 隆哉（東京工科大学　准教授））

Journal: Journal of Information Processing, Vol.32, No.2, pp.232-246, 2024
掲載年月：2024 年 2 月

研究概要：この論文では、亜種マルウェアのサンプルからバイト単位の n-gram を高速に抽出する手法を提案しました。この n-gram を並べ替える際に、接尾辞配列のアルゴリズムを利用しています。

研究背景：コンピュータの世界でウイルスと呼ばれているものがあります。この名前は、人間も感染して病気になるウイルスに由来しています。病気のウイルスは細菌ではなく、生きていません。生きている生物に入り込み、その生物の活動の際に自分自身もコピーさせることで増殖します。コンピュータのウイルスは、それ自体が実行可能なファイルではなく、実行されるプログラムの一部に入り込み、そのプログラムの実行時に自分自身をコピーさせて増殖するので、このように呼ばれるようになりました。現在では、実行可能なファイルそのものであってもウイルスと呼ぶ人もいますが、そのようなものは悪意のあるソフトウェア（Malicious Software）ということで、総称してマルウェア（Malware）という呼び方が定着してきました。マルウェアを見つける方法にはさまざまなものがあり、現在も研究が行われています。中でも、バイナリをバイト単位で扱い、それを n-gram とすることでマルウェアを検出するという研究が約 20 年前から行われてきました。そして、現在も人工知能に使われる技術として一般的になってきました、機械学習と組み合わせることで、高い精度でマルウェアを検出できるという研究結果がいくつも発表されています。それでは、マルウェアをバイト単位の n-gram で検出する方法は、広く流通しているマルウェア検出ソフトウェア（アンチウイルスソフトウェアという名前が一般的かもしれません）で採用されているのかといいますと、そうでもないようです。これについては、機械学習に使う数多くのサンプルから n-gram を抽出するのにとても時間が掛かることがボトルネックになっているといっている研究者がいます。実際、この処理をしてみると多くの時間が掛かりました。しかし、そうであれば、この処理さえ高速に行えれば、一般的な n-gram をマルウェア検出ソフトウェアに適用できます。

研究成果：この論文では、機械学習に使う数多くのサンプルからバイト単位の n-gram を高速に抽出する手法を提案しています。最も簡単な方法を論文の図 2 に示しましたので抜粋します。このように、比較対象の n-gram がそのファイルにあるかないか、順番に n-gram を比較していけばよいですが、これにはとても時間が掛かります。それではどうすればよいでしょうか。基本的な考え方は、n-gram を高速に抽出できるように、元のデータを事前に並べ替えておくということです。並べ替えたデータとの比較方法を論文の図 4 に示しましたので抜粋します。このように、順番に並べ替えられたデータと比較を行えば、どの位置にあるかがすぐにわかりますので、速く比較できます。ただし、並べ替えを行うには多くの時間が掛かりますので、そう単純な話ではありません。並べ替えられたデータから n-gram を高速に抽出できたとしても、並べ替えるのにより多くの時間が掛かってしまっては意味がありません。我々は接尾辞配列を使うことで、この問題を解決しました。

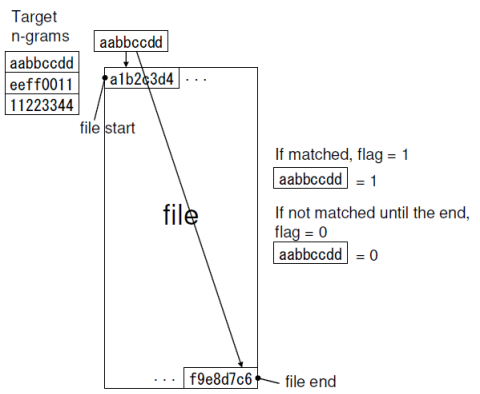

Fig. 2　Easiest n-gram Comparing

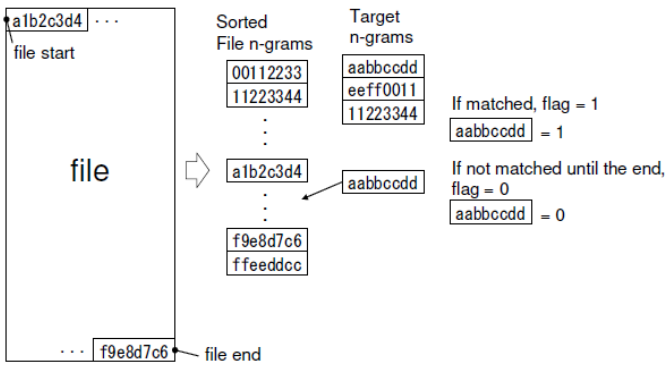

Fig. 4　Easy n-gram Comparing

社会的・学術的なポイント：与えられたデータを単純に接尾辞配列に変換しようとしますと、とても長い時間が必要です。しかし、どのようにすれば最も早く接尾辞配列が作れるかといったような課題には、長い年月を掛けて世界中の研究者が取り組んでくれています。我々はその成果を使うことで、データを高速に接尾辞配列に変換できました。こうなってしまえば、その接尾辞配列から高速に n-gram が抽出できます。どれくらい処理が早くなるかは、ファイルサイズとファイル数によりますので一概には言えませんが、10,000 バイトのファイル 100 個の場合には、従来手法で 7515.5 ミリ秒掛かっていたものが 338.7 ミリ秒になりましたので、22.19 倍速くなっています。詳しい話は簡単に説明できませんのでここでは省きますが、我々の手法では複数のサイズの n-gram を同時に抽出できる一方、亜種マルウェアにしか適用できないという条件があります。4 種類の亜種マルウェア（検体数はそれぞれ 967 個、13,232 個、1,950 個、936 個）に対して、良性ソフトウェア（検体数 1,500 個）を用意して実験したところ、見逃したマルウェアはゼロでした。一方で誤ってマルウェアと判断されてしまった良性ソフトウェアは 1 検体だけありました。

用語解説：

バイナリ：正確な説明が難しいのですが、2 進数で表現されたデータのことです。「0」と「1」が並んでいますのでこのように表現します。これに対して、「テキスト」というのが「abc」のような文字の並びで表現された形式です。

亜種マルウェア：悪いことをする内容については同様のことするのですが、検体ごとに動作が少しずつ異なったり、動作は完全に同じであってもファイルの中身が少し異なっていたりするマルウェアのことを指します。同じ種類のマルウェアであるのに、複数の検体があるということになります。つまり、バイナリの並び方を見ただけでは、どの種類のマルウェアか特定できないようになっています。