自然言語処理と畳み込みニューラルネットワークによる フィッシングサイト検知手法の改良

Proposal for Improvement Idea for the Phishing Site Detection Method Using Natural Language Processing and Convolutional Neural Network

1941027　戸村 悠綺
Yuki TOMURA
指導教員　秋葉 知昭

In recent years, the number of victims of phishing scams has increased with the spread of the Internet. There is a limit to the human response to all phishing scams, which are becoming highly developed. Therefore, Kasahara[1] and Kobayashi[2] proposed a method to automatically detect phishing sites by Image recognition technology and Natural language processing. However, the method by Image recognition yielded effective results against phishing sites, while the method by Natural language processing did not yield effective results against phishing sites.

The aim of study is made detection methods by Natural language processing, and it proposes an effective tool against for phishing sites. In this study, HTML source data collected from each site is preprocessed before use. This HTML source data is converted into a vector using TF-IDF and Doc2Vec, and analyzed by a Convolutional Neural Network to determine if it is a phishing site. The results showed that the proposed method better than previous studies. In particular, the method using TF-IDF and Convolutional Neural Networks was found to be effective against existing phishing sites.

1. 緒　言

近年では, 正規のサービスを騙った文面から偽のサイトに誘導し, 個人情報を搾取するフィッシング詐欺の被害件数の増加が問題となっている. 最近では, SMS を利用したフィッシング詐欺(スミッシング)が急増しており, 世相を反映した手口も存在する. フィッシング詐欺の対策として, ウイルス対策ソフトの導入や多要素認証といった多くの対策を行っているが, 人間が対応を行う以上, 常に後手に回ってしまい早期の被害を減らすことが困難となっている.

このような背景から, 笠原[1]は画像認識と自然言語処理を用いたフィッシングサイト検知手法を提案した. この手法はリアルタイムでフィッシングサイトを検知できることから早期の被害を防ぐことが期待できる. 結果は, 画像認識では高い精度が得られたが, 自然言語処理では実用に耐えうる精度を得られなかった. この結果をもとに, 小林[2]は自然言語処理による検知手法の改良を行ったが, 笠原[1]と同様, 実用的な精度を得ることができなかった. 本研究では, 自然言語処理によるフィッシングサイト検知手法の更なる改良を行うことで, 実用的な検知手法にすることを目的とする.

2. 検知手法の提案

2.1 検知手法の概要

本研究では, 畳み込みニューラルネットワーク(以下 CNN)の導入と HTML ソース の数値化手法の改良を行うことで, 精度を向上させる. フィッシングサイトか否かを判別する際は, ディープラーニングツールである Neural Network Console(以下 NNC)を使用する.

2.2 収集データ

解析を行うためには, フィッシングサイトと正規サイトの双方のデータを収集する必要がある. 本研究では, 先行研究[1][2]と同様にフィッシングサイトの HTML ソースを 200 件, 正規サイトの HTML ソースを 200 件使用する. また, 学習を行うにあたって HTML ソース全文を使用すると処理時間が膨大になり, 学習が困難になってしまうため HTML ソース内の URL のみを抽出したデータを使用した.

2.3 HTML ソースのベクトル化

本研究では, 2 つのベクトル化手法を用いた. 1 つ目は, TF-IDF と呼ばれる統計的な手法を用いた単語の重要度を図る手法である. 文書内での単語の出現頻度(TF)とある単語が含まれる文書の割合の逆数(IDF)を掛け合わせて求めることができる[3].

$$tfidf(t_i, d_j) = tf(t_i, d_j) \cdot idf(t_i) \quad (1)$$

$$tf(t_i, d_j) = \frac{\text{文書}d_j\text{内の単語}t_i\text{の出現回数}}{\text{文書}d_j\text{のすべての単語の出現回数の和}} \quad (2)$$

$$idf(t_i) = \log\left(\frac{\text{総文書数}}{\text{単語}t_i\text{が出現する文書数}+1}\right) \quad (3)$$

2 つ目は, Doc2Vec と呼ばれるニューラルネットワークを用いた手法である. 単語の分散表現が得られる Word2Vec を応用したモデルであり, 文書を低次元の実数値ベクトルで表現する手法である. Doc2Vec には, ２種類のモデルが内包されているが, 本研究では, 一般的に高い精度が得られるとされる PV-DM を使用する. 図 1 は PV-DM モデルの構造を示している. 入力値として欠如している文書と文書 ID を与え, その欠如

千葉工業大学　社会システム科学部　経営情報科学科
2022 年度卒業研究概要

した部分にどのような単語が出現するのかを学習していくことで文書の分散表現を獲得する.

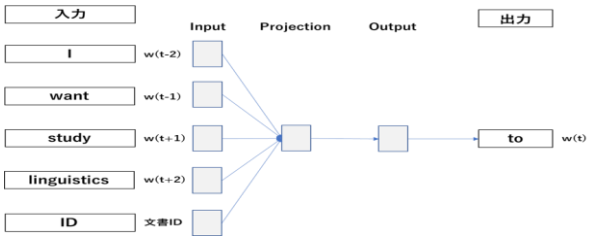

図 1　PV-DM モデル

2.4 自然言語処理アルゴリズム

NNC で構築した学習アルゴリズムのイメージを図2に示す. 本研究では, CNN を自然言語処理を施したデータに適用した. 前処理を行いベクトル化した HTML ソースを Input で入力し, Reshape プロパティで 3 次元配列に変換する. 変換されたベクトルから Convolution 層で特徴を抽出し, 活性化関数 PReLU で 0 以下の数値を補正する. 補正された数値を MaxPooling で特徴を残しつつ, 縮小する. この Input と Reshape プロパティを除いた一連の処理を複数回繰り返すことで, ベクトルの特徴を抽出していく. そこで出力された数値を全結合層 Affine で 1 次元の配列に変換する. 最後に, 活性化関数 Softmax で確率に変換し, CategoricalCrossEntropy で結果を出力する.

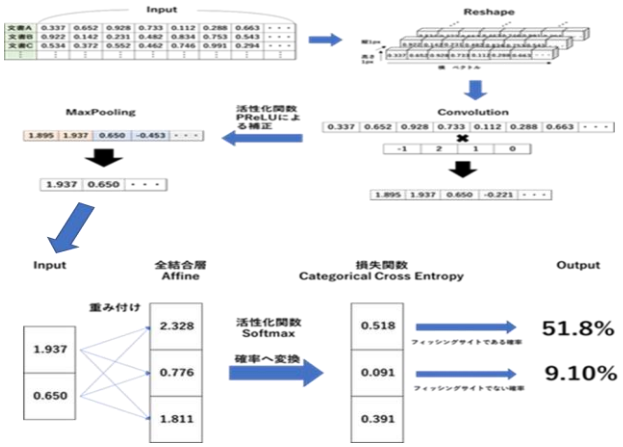

図 2 学習アルゴリズム

3. 結果及び考察

図 3, 4 は NNC で構築した学習アルゴリズムによる学習曲線を示している. 学習結果は, TF-IDF を用いた手法の正解率が 92.56%, 正規サイトに対する再現率(Recall)が 94.88%, フィッシングサイトに対する再現率が 90.61%とすべての項目で先行研究を大幅に上回る結果が得られた. Doc2Vec を用いた手法の正解率は 80.45%, 正規サイトに対する再現率が 80.95%, フィッシングサイトに対する再現率が 81.03%と先行研究で見られた再現率の極端な偏りこそ改善されたものの, 実用に耐えうるほどの精度を得ることはできなかった. 今後の改善点としては, Doc2Vec によるベクトル化の際のパラメータや畳み込み層の特徴マップ生成数といった各種パラメータの調整を行うことで, 正解率や再現率に変化が現れるのかを検証する必要があると考える.

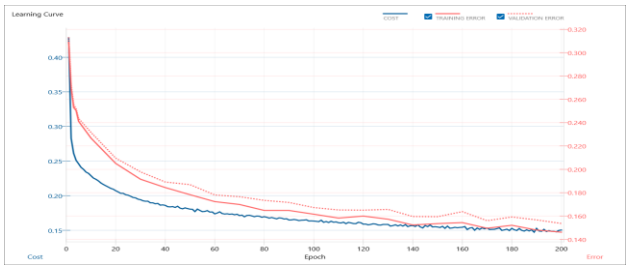

図 3　TF-IDF を用いた手法の学習曲線

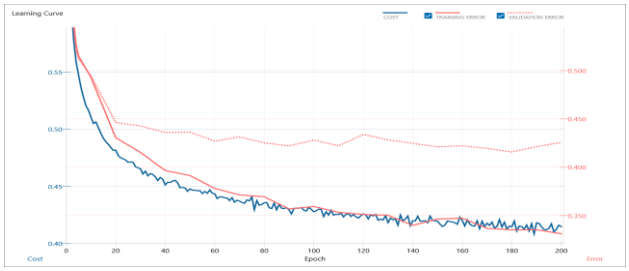

図 4　Doc2Vec を用いた手法の学習曲線

4. 結言

本報告では, TF-IDF と Doc2Vec を用いたフィッシングサイト検知手法を提案し, CNN による学習及び評価を行った. 評価結果としては, TF-IDF の正解率が 92.56%, Doc2Vec の正解率が 80.05%と先行研究の結果を上回る数値が得られた.

これらの結果から, TF-IDF と CNN による検知手法は現存するフィッシングサイトに対しては高度な対策効果が期待できる. しかし, 近年のフィッシング詐欺ではこのような高度なフィッシング対策システムにも対応した手口が見られてきている. そのため, 現状の結果に満足することなく, 今後も新たな自然言語処理手法を用いたフィッシングサイト検知手法あるいは新たなフィッシングサイト検知手法の開発を行っていくことで, フィッシング詐欺に対して先手を打ち, 早期の対応を行っていく必要があると考える.

文　献

[1] 笠原拓也 : パターン認識を用いたフィッシングサイトの検知手法の提案, 2018 年度千葉工業大学卒業研究(2018)

[2] 小林貴章:自然言語処理の深層学習によるフィッシングサイト検知手法の改良,2019 年度千葉工業大学卒業研究(2019)

[3] MIERUCA:【技術解説】単語の重要度を測る？TF-IDF と Okapi BM25 の計算方法とは, https://mieruca-ai.com/ai/tf-idf_okapi-bm25/(2018/5/8 時点)