機械学習を用いた偽サイト検知

Spoofed Website Detection using Machine Learning

栗原直樹・ネットワーク分科会・情報セキュリティ大学院大学

In recent years, damage by fake site has been rapidly increasing. It is difficult to simply detect fake sites as fake sites are cerebrating as if they are genuine. Police related organizations detect fake sites manually by present manually, so it is a problem to take considerable time to detect fake sites.In this research, in order to solve this problem, by developing a fake site detection method using machine learning, it is possible to automate the detection process that is now dependent on human hands to some extent and shorten the time taken for fake site detection aim. In this paper, we propose a fake site detection method using machine learning, and conducted three verification experiments as an evaluation of the proposed method. In the final verification experiment, we successfully identified 300 test data (150 fake sites / 150 regular sites) with a probability of 88.3%.

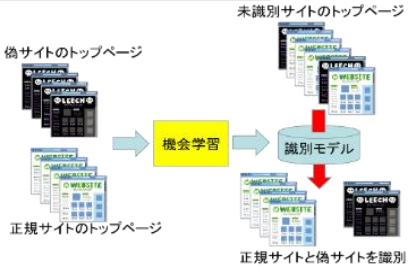
1.背景・課題

近年，偽サイトによる被害が急増している．偽サイトとは，クレジットカード番号等の個人情報や金銭を騙し取ることを目的とした，実際に存在しない会社・組織を騙ったサイトや，大手サイト・官公庁サイト等を模写した偽物のサイトの総称である．最近では，ふるさと納税サイトを騙ったサイトによる被害が急増している．偽サイトは，あたかも本物であるかの様に儀網しているため，偽サイトを単純に検知することは難しい．警察関連機関は，現状偽サイトを人手で検知しているため，偽サイトの検知までに相当の時間がかかる事が課題となっている．

2.研究目的

機械学習を用いて偽サイトの検知を効率化することである．すなわち，現在人手に頼っている検知プロセスを人工知能の技法の応用により半自動化することで，偽サイト検知にかかる時間の短縮を目指す．

具体的には，偽サイトと正規サイトのトップページに含まれる特徴を特定し，その特徴を機械学習させ，識別モデルを作成する．



3.提案手法・検証実験

提案手法

①偽サイト・正規サイトのHTMLデータ(トップページ)を収集

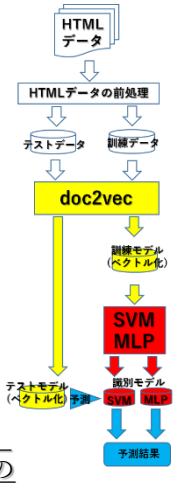
②HTMLデータの前処理(整形)
※識別に有用な特徴を抽出、ノイズ削除

③前処理を施したHTMLデータを機械学習させるために、doc2vecを用いてベクトル化(数値化)する
※自然言語はそのままでは学習できない

④機械学習で「教師あり学習」を行い識別モデルを作成する
※機械学習は、サポートベクターマシン(SVM)と多層パーセプトロン(MLP)
※偽サイトは「0」、正規サイトは「1」としてラベル付けを行う

⑤識別モデルを10分割交差検証で評価する

⑥訓練に使用していないデータを識別モデルに当てはめて、実際に識別ができるかテストを行う
※テストデータも同様にdoc2vecを用いてベクトル化したものを使用する
※偽サイトと判断した場合は「0」、正規サイトは「1」と出力する



提案手法を検証しつつ，主に②HTMLデータの前処理工程を変更し，偽サイト，正規サイトを識別できるHTMLデータ内の特徴を特定するために，3回の検証実験を実施．

	訓練データ	テストデータ	合計
偽サイト	4000件	150件	4150件
正規サイト	4000件	150件	4150件
合計	8000件	300件	8300件

検証実験に使用したデータ

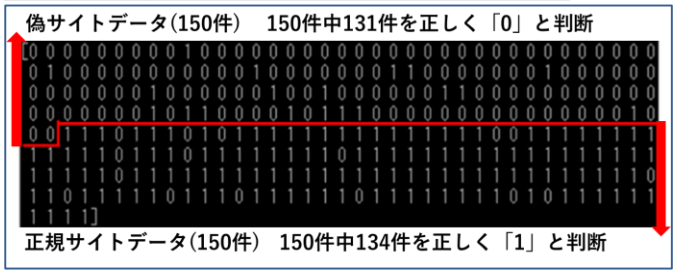
偽サイトは，警察庁から提供されたデータを使用

正規サイトは，日本の上場企業，大学，短期大学のサイトトップページを収集

4.結果

試行錯誤した結果，トップページのHTMLデータに記載された「コメントアウトされた文章」が識別に有効な特徴であることが判明した．

HTMLデータの前処理において，コメントアウトされた文章を抽出し，SVMを用いて機械学習させることによって作成した識別モデルを用いてテストしたところ，テストデータ300件(偽サイト150件/正規サイト150件)を88.3%の確率で正しく識別することに成功した．



テストデータ300件の予測結果

本提案手法の②において，コメントアウトされた文章を抽出し，④にあっては，SVMを選択(MLPは，ランダム性が強く精度が安定しない)することによって，偽サイトと正規サイトを高精度で識別するモデルを作成することができる．

しかし，全ての偽サイトを検知することができないため，偽サイトにも関わらず，正規サイトと判断される場合がある．そのため，現在行われている調査方法を継続しつつ，本提案手法の結果を参考にして，調査すべきサイトの優先順位を決める目的で使用することが望ましい．

5.課題

識別精度の向上
- コメントアウト以外の特徴を特定
- 訓練データの収集

実装に向けた準備
- 提案手法の簡略化
- ハイパーパラメータ選択の自動化