氏名：宮本　大輔

論文タイトル：A machine learning approach for detecting fraudulent websites

論文内容の要旨：

This dissertation presents machine learning based detection methods against phishing. Phishing is a fraudulent activity defined as the acquisition of personal information by tricking an individual into believing the attacker is a trustworthy entity. Phishing attackers lure people by using ``phishing email'', as if it were sent by a legitimate corporation. The attackers also attract the email recipients into a ``phishing site'', which is the replica of an existing web page, to fool a user into submitting personal, financial, and/or password data. Strategies against phishing tackle to protect users from fraud.

My research motivation is developing a method for detection of phishing sites to prevent users from browsing phishing sites. Currently, existing detection methods are far from suitable. URL filtering-based detection methods could not deal with new types of phishing attacks, i.e, spear phishing. Conversely, heuristics-based detection methods have a possibility to identify these sites. When users browse a site, the methods calculate the likelihood of being a phishing site for the site. The methods also classify the site as phishing if the likelihood is greater than the discrimination threshold. The problem in heuristics-based detection methods is that the detection accuracy is not high. Accordingly, users would become distrusting the system and would ignore the notification from detection systems.

In this dissertation, I employ machine learning algorithms to improve the detection accuracy, machine learning can facilitate the development of algorithms or techniques by enabling computer systems to learn. As my preliminary experiment, I investigate whether a machine learning algorithm is available or not. I construct a training dataset by analyzing 50 phishing sites reported on Phishtank.com and the same number of legitimate sites with 8 heuristics, namely, Age of Domain, Known Images, Suspicious URL, Suspicious Links, IP Address, Dots in URL, Forms, and TF-IDF-Final heuristics. I then let AdaBoost, one of the typical machine learning algorithms, study from the training dataset in a supervised-learning manner. I also construct a testing dataset composed of 50 phishing sites and the same number of legitimate sites, and classify them based on the model derived from the training dataset. By

comparing with the existing method, AdaBoost can provide higher detection accuracy in almost of all cases. In some cases, overfitting problems are observed. To avoid the overfitting problems, I attempt to increase the number of URLs in dataset by both implementing all heuristics and monitoring Phishtank.com periodically.

In my performance evaluation, I employ 9 machine learning techniques including AdaBoost, Bagging, Support Vector Machines, Classification and Regression Trees, Logistic Regression, Random Forests, Neural Networks, Naive Bayes, and Bayesian Additive Regression Trees. I let these machine learning techniques combine heuristics, and also let machine learning-based detection method(MLBDM)s distinguish phishing sites from others. I analyze our dataset, which is composed of 1,500 phishing sites and the same number of legitimate sites. These 1,500 URLs of phishing sites are reported Phishtank.com during November, 2007 -- February, 2008, and are verified as phishing sites by registered users of Phishtank.com. I then classify them using the machine learning-based detection methods, and measure the performance. In my performance evaluation, I decide f1 measure, error rate, and Area Under the ROC Curve (AUC) as performance metrics along with my requirements for detection methods. The highest f1 measure is 0.8771, the lowest error rate is 11.96%, and the highest AUC is 0.9543, all of which are observed in the case of AdaBoost.

Next, I check whether or not MLBDMs are available even if the dataset or the set of heuristics are different. I test another dataset which contains phishing sites reported in different time period. I also use another dataset which contains 1,277 URLs of phishing sites, 223 URLs which are not phishing sites but treated as phishing, and 1,500 URLs of legitimate sites. I also change a set of heuristics and observe the performance. All results show that almost of all MLBDMs outperform the traditional detection method.

I then discuss utilization methods for MLBDMs. First, I explore a way for deciding the discrimination threshold for each user. Within my preliminary algorithm, I confirm that changing threshold can customize the detection strategy. Next, I argue the another approach which aims to cover the weak points of users by existing heuristics with machine learning techniques. The key idea of this approach, named ``HumanBoost'', is employing users' past trust decision as a new heuristic. As my pilot study, I conduct subject within

test by calling 10 subjects. Subjects browse 14 emulated phishing sites and 6 legitimate sites, and check if the site seems to be a phishing site or not. By using such types of subjects' judgments as a new heuristic, I let AdaBoost to incorporate the heuristic into existing 8 heuristics. The results shows that the average error rate in the case of HumanBoost was 9.5%, whereas the average error rate of subjects was 19.0% and that in the case of AdaBoost was 20.0%.

Finally, I propose HTTP Response Sanitizing(HRS) which is a countermeasure against phishing. When a phishing prevention system focuses on reducing false negative errors, false positive errors would increase even if the system employs MLBDMs. My proposed HRS is designed to reduce the loss of convenience arisen from false positive errors. HRS removes all input forms from the sites. While users can browse the rest of content, the loss of convenience would be lower than the existing method which filters whole suspected web pages. I implement HRS-capable proxy servers and verify the function of removing by browsing 100 actual phishing sites. The performance overhead is 3.59 millisecond given content size is 10Kbytes and the content involves 13 HTML tags to be removed. I also compare HRS among existing countermeasures.

This dissertation demonstrates that machine learning algorithms are available for detecting phishing sites. Since MLBDMs contribute to improve the detection accuracy, users will believe that the notification from detection methods. Accordingly, users can easily avoid phishing sites.

氏　名	宮本　大輔

（論文審査結果の要旨）

本博士論文では，機械学習を用いた高精度なフィッシング(phishing)サイトの検知手法を提案している．本博士論文では，フィッシング攻撃を対策するという目的のもと，既存のフィッシングサイト検知手法の検知精度の低さを問題として捉える．また，フィッシングサイトか否かを決定するための判断材料が先行研究にて提案されているが，これらの判断材料の組み合わせ手法が検知精度の低さの原因であることを示している．この主張を検証するため，本博士論文では以下の判別実験を行っている：（１）実際に報告されたフィッシングサイトと，先行研究などに示される非フィッシングサイトを収集し，各判断指標により分析を行う，（２）検知方式には正確性及び各ユーザへの適応能力を要件とし，性能評価の指標としてF1値，エラー率，AUC値を用いる，（３）複数種類の機械学習手法を用い，その性能を比較評価する，（４）四交差検定を10回行い，その平均を用いることにより評価指標値の偏りを防ぐ．これらの判別実験の結果，機械学習を用いたフィッシングサイト検知は，既存の検知手法よりも高精度であることが確認されている．また，この傾向は異なる年月に取得したフィッシングサイトにおいても有効であるか否か，また異なる判断材料の組み合わせにおいても有効であるか否かが検証され，その結果，機械学習方式の有効性が示されている．なお，予備実験ではさらに機械学習方式の問題点が過学習である事，及び過学習が実際の検知率にもたらす影響について議論を行い，大量のフィッシングサイトを収集する手法についても議論している．さらに本博士論文では，機械学習に基づいた検知アルゴリズムを有効に活用する手法として，人間の判断結果を一つの判断材料として機械学習による検知に組み込むことにより，検知の精度を向上できるという提案が行われ，10人の被験者による初期実験においてその主張を裏付ける結果が得られたことを報告している．さらに，検知結果に基づいた対策手法に関して，ユーザの安全性を高めつつ，利便性にも考慮されたフィルタリング方式の設計及び実装を行い，このフィルタリング方式が有効であることを示している．以上により、本博士論文は研究内容について新規性並びに有効性があることが認められ、博士（工学）の学位を授与するにあたって十分な内容であると認められる。