피싱 탐지기법 조사 및 분류: 비정상적 상호작용

박지훈[1], 최상훈[2], 박기웅[3†]
[1]세종대학교 SysCore Lab. 석사과정
[2]세종대학교 SysCore Lab. 연구교수
[3]세종대학교 정보보호학과 교수
qkrwlgns1325@naver.com, csh0052@gmail.com, woongbak@sejong.ac.kr

A Survey and Classification of Phishing Detection Techniques: Anomaly Interactions

Ji-Hoon Park[1], Sang-Hoon Choi[2], Ki-Woong Park[3†]
[1,2]SysCore Lab., Sejong University
[3]Dept. of Computer and Information Security, Sejong University

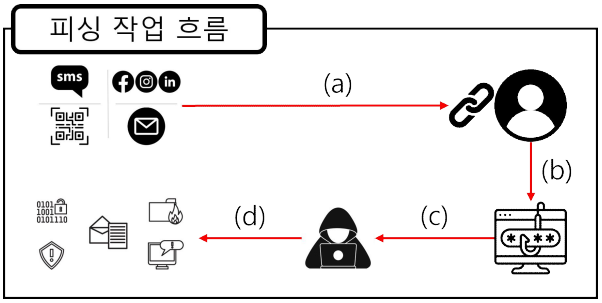
요 약

최근 피싱 키트의 고도화와 LLM의 등장, 그리고 다양한 통신 매체의 사용으로 인해 피싱 공격이 증가하고 있다. 이러한 피싱은 인간이 정상적인 상호작용인지 아닌지 구분하기 어렵게 하며 다양한 방식을 통해 탐지를 우회한다. 특히, 이와 같은 피싱으로 인해 개인 또는 단체 등 민감한 정보의 탈취가 발생하며, 정보 탈취 이후 추가적인 피해가 일어나므로 사용자의 정보가 탈취당하기 전에 피싱인지 아닌지 탐지를 수행해야 한다. 따라서 피싱으로부터 사용자를 보호하기 위해 다양한 매체들로부터 피싱 이메일, 메시지 등의 콘텐츠 수신과 피해자의 링크 클릭, 피해자 스스로 자격증명을 입력하는 과정에서 탐지하는 것이 중요하다. 본 논문에서는 피싱 방식별 위협 모델 소개와 탐지기법을 분류하고 각각의 피싱 탐지기법이 보유한 도전과제에 대해 소개한다.

1. 서론

피싱이란 개인정보를 의미하는 Private Data와 Fishing의 합성어로, 공격자가 신뢰할 수 있는 사람으로 가장하여 개인, 단체와 관련된 민감한 정보를 획득하기 위해 사용되는 공격이다. 이러한 피싱은 주로 인간의 감정 중 공포, 급박함, 호기심 등 동요를 일으키는 방식을 사용하며, 그 공격을 수행하는 매체 또한 다양하다. 이러한 피싱은 1990년대부터 시작하여 현재까지 끊임없이 이루어지고 있으며, 피싱의 절차 중 하나인 이메일을 사용한 피싱은 2024년을 기준으로 전년 대비 856% 증가하였다 [1]. 이와 같은 피싱의 증가 추세는 기술의 발전과 밀접한 연관성을 갖는다. 특히, GPT와 같은 LLM은 추가적인 학습 없이 글쓰기, 코드 생성이 가능하므로 이를 악용했을 경우 피싱 이메일, 피싱 웹 페이지 생성에 들어가는 노력을 대폭 줄일 수 있다.

이와 같은 피싱의 주된 목적은 데이터 탈취이다. 특히 자격증명 탈취를 통해 피해자의 자격증명으로 금융사기 및 신원사기를 일으키거나 금전적인 피해를 준다. 이외 개인 사진이나 영상과 같은 민감 데이터 탈취의 경우, 협박이나 지속적인 갈취로 이어진다. 즉 이러한 피해는 피싱으로 인해 발생하고, 개인을 대상으로 하거나 기업 또는 단체에 영향을 미치며 그 대상 범위가 매우 넓고 피해 규모 또한 다양하다. 이와 같은 피싱은 주로 소셜미디어의 메신저, 이메일, SMS, QR 등의 매체를 이용하여 불특정 다수를 대상으로 하거나 공개된 정보 수집을 통해 개인화될 수 있다. 또한 피싱의 절차로는 (a) 피해자가 다양한 매체로부터 링크를 수신, (b) 피해자의 링크 클릭으로 악성 웹 사이트 리디렉션, (c) 피해자가 악성 웹 사이트 내에서 본인의 자격증명을 입력하고 공격자에게 전송, (d) 공격자는 획득한 자격증명을 통해 추가 공격 수행으로 구분할 수 있으며, 이와 같은 피싱의 흐름은 그림 1과 같이 표현된다.



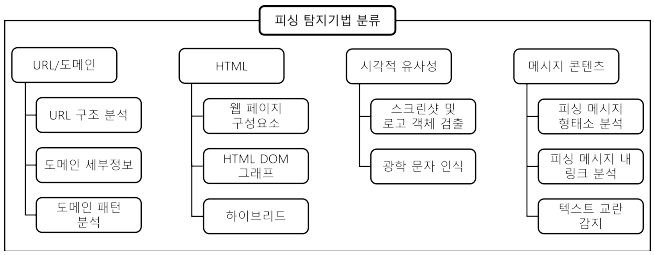
(그림 1) 피싱의 작업 흐름도

† 교신저자: 박기웅 (세종대학교 정보보호학과 교수)

이러한 피싱 피해 증가와 피싱 기술의 고도화로 피싱을 탐지하기 위한 다양한 연구가 수행되었다. 기존 탐지기법으로는 URL/도메인 기반, HTML 기반, 웹 페이지의 시각적인 유사성 기반이 존재하며, 이메일 내용을 포함한 메시지 콘텐츠 기반이 있다.

본 논문의 구성은 2장에서 기존 연구의 분석을 통해 탐지기법을 분류하고, 3장에서 기존 탐지기법이 보유한 주요 도전과제에 관하여 기술한다. 4장에서는 결론과 향후 연구 방향을 제시한다.

2. 피싱 탐지기법 분류



(그림 2) 피싱 탐지기법의 분류도

본 연구팀은 최근 피싱 피해 증가 추세에 따른 탐지기법의 동향을 파악하기 위해 기존 피싱 탐지 연구를 조사하였으며, 각 탐지기법과 요소를 그림 2와 같이 분류하였다.

2.1 URL/도메인 기반 탐지

과거의 URL 기반 피싱 탐지는 블랙리스트에 의존했으나, 현재 피싱이 도메인 생성 알고리즘(DGA)과 같은 기법을 사용하여 리스트에 없는 피싱 URL을 생성함에 따라 기존의 탐지를 더욱 어렵게 한다. 이러한 문제를 개선하기 위해, URL의 속성 및 구조의 특징으로 이루어진 데이터 세트를 모델에 학습시켜 피싱 사이트 여부를 예측한다.

이와 관련하여 URL/도메인 기반의 실시간 피싱 탐지를 위해 경량 모델을 사용한 RNT-J가 연구되었다 [2]. 해당 연구에서는 정상 웹 사이트와 피싱 웹 사이트 간 데이터 불균형 문제를 해결하기 위해 샘플링을 수행하였으며, 두 개의 모델을 결합한 특징 추출 및 주성분 분석을 통해 피싱 패턴 분류를 수행하였다. 해당 연구에서 제안한 모델은 98%의 정확도를 달성하여 피싱을 탐지하는 데 효과적임을 보인다. 이러한 데이터 세트와 모델을 사용함과 동시에 도메인 세부 정보 수집을 적용한 연구 [3], 단어 임베딩을 활용한 URL/도메인 기반의 탐지기법 또한 연구되었다 [4].

2.2 HTML 기반 탐지

최근 피싱 웹 페이지 제작에 LLM과 피싱 키트가 이용됨에 따라 피싱을 시도하기 위한 노력이 대폭 감소하였다. 특히 LLM을 사용한 웹 페이지 스크립트 생성이 가능하며, 이는 지식이 없는 사람도 웹 페이지 생성을 통한 피싱이 가능함을 시사한다 [5]. 또한, 피싱을 위한 백엔드를 제공하는 LLM과 피싱 키트는 지속적인 업데이트 및 서비스로 성능이 향상됨에 따라 이전보다 더 쉬운 피싱 공격이 가능함을 의미한다.

이와 같은 피싱 위협을 탐지하기 위해 웹 페이지를 구성하는 HTML을 기반으로 한 탐지 방법이 연구되었다. 정상 HTML과 피싱 HTML 데이터 세트를 사용하여 웹 페이지 내 텍스트, 링크, CSS 및 JS, 페이지 구조와 같은 구성 요소들을 특징으로 추출하였으며, 신경망 및 자연어처리를 결합하여 피싱을 탐지한다. 본 연구의 실험 결과로 97.18%의 정확도를 기록하여 피싱 활동을 식별하는 데 효과적임을 보인다 [6]. 이와 같은 HTML 기반의 탐지기법은 URL 탐지기법과 결합할 수 있으며, HTML과 URL 콘텐츠의 텍스트 데이터 세트와 신경망 및 단어 임베딩을 통해 구현된 WebPhish는 실험 결과로 98.1%의 정확도를 달성하였다 [7].

2.3 시각적 유사성 기반 탐지

시각적인 유사성을 이용한 피싱은 웹 페이지 디자인이나 로고와 같은 시각적인 콘텐츠를 이용하여 실제 웹 사이트와 유사한 피싱 페이지에 사용자의 자격증명 입력을 유도한다. 이와 같은 시각적 유사성을 이용한 피싱은 페이지 내에 경고 문구 등을 삽입하여 불안감을 조성하거나, 웹 브라우저 취약점을 악용하는 방식을 통해 공격의 성공률을 높인다.

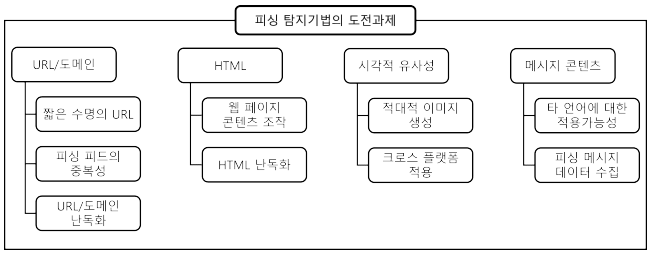
이러한 시각적 유사성을 이용한 피싱을 탐지하기 위해 웹 페이지 스크린샷과 로고의 특징이 사용된다. 관련 연구에서는 스크린샷을 입력으로 받아 객체 검출을 통해 다양한 크기의 로고를 식별하고, 처음 입력된 스크린샷과 식별된 로고로부터 특징 추출 및 결합을 통해 정상 웹 사이트와 피싱 웹 사이트를 식별하였다 [8]. 또한, 기존 안티피싱 솔루션의 느린 속도와 높은 컴퓨팅 리소스 요구량을 개선하기 위해 머신러닝과 광학 문자 인식(OCR)을 사용한 연구에서는 웹 페이지에 포함된 주요 색상 특징과 브랜드명을 기반으로 빠른 피싱 탐지 성능을 보여 실시간 탐지에 대한 가능성을 보인다 [9].

2.4 메시지 콘텐츠 기반 탐지

사이버 위협의 시작이 피싱 이메일과 관련되어 있다는 점과 전 세계적으로 피싱 공격이 증가하는 실태를 통해 피싱이 현재까지도 인기 있는 공격방식 중 하나임을 알 수 있다. 특히 피싱은 메신저, 이메일, SMS 등의 다양한 매체로부터 유입되고, 공개된 정보와 LLM을 사용하여 1센트가 안 되는 비용으로 개인화된 피싱 공격이 가능하므로, 피싱 탐지의 중요성이 대두된다 [10].

SMS를 사용한 피싱인 스미싱(Smishing)에서 메시지 내 텍스트를 기반으로 피싱을 탐지하는 연구가 진행되었으며, 특히 실제 스미싱 피해자의 메시지와 일반 메시지 데이터 세트를 통해 적대적 메시지를 생성하고 분류기를 사용하여 피싱 메시지를 탐지하는 방법이 제안되었다 [11]. 해당 논문에서는 스미싱 메시지의 98%가 이전에 보낸 메시지의 변형이라는 통계에 기인하였으며, 특정 메시지의 단어와 한글 문자, 메시지 구조에 기호, 문자, 공백 등을 포함한 적대적 메시지 생성기와 딥러닝 기반의 스미싱 분류기를 설계하여 피싱 메시지를 분류한다. 해당 연구의 분류기는 경량 모델의 크기를 가지며, 정확도 99%를 기록하여 피싱 탐지에 효과적임을 보인다. 이 외에도 메시지로부터 피싱을 탐지하기 위해 형태소 분석으로 동사 및 명사를 추출하여 탐지하는 연구가 진행되었다 [12].

3. 피싱 탐지기법의 주요 도전과제



(그림 3) 피싱 탐지기법이 보유한 도전과제

진화하는 피싱을 탐지하기 위한 최신 연구가 진행되고 있지만, 탐지 성능의 향상을 위해 아직 해결해야 하는 과제가 남아있다. 본 논문에서 분류한 피싱 탐지기법의 주요 도전과제는 그림 3과 같다.

3.1 URL/도메인 기반 탐지의 도전과제

피싱 URL/도메인 기반의 탐지 연구는 활발하게 이루어지고 있다. 하지만 기존 연구는 PhishTank, PhishStats, OpenPhish 등 다양한 소스로부터 얻은 데이터 세트를 사용하므로, 어떤 데이터 세트를 사용하는지에 따라 성능의 차이가 생겨 모델 간의 탐지 성능 비교가 어렵다 [13]. 또한, 피싱 페이지의 짧은 수명과 피싱 피드의 중복은 데이터 세트의 구축을 저해하기 때문에 이러한 데이터 세트 구축 역시 URL/도메인 기반의 탐지에서 주요 도전과제이다. 이 외에도 정상 페이지와 피싱 페이지의 데이터 불균형 역시 URL/도메인 기반의 탐지를 위해 해결해야 할 문제이다 [14].

3.2 HTML 기반 탐지의 도전과제

HTML 기반 탐지의 주요 한계점으로 HTML 콘텐츠의 최신 데이터 세트가 희소하다는 점을 지적하였다 [6]. 또한, HTML 콘텐츠의 복잡한 구조와 패턴을 분석 및 분류하기 위해 강력한 GPU가 필요하며, 이는 모바일과 같은 컴퓨팅 리소스가 제한된 환경에서 활용하기 어렵다는 것을 시사한다. 또한, 알려진 HTML 콘텐츠를 포함하는 피싱 탐지에는 뛰어나나, 웹 페이지 내 콘텐츠가 변조되는 경우 탐지에 어려움을 겪는다 [7].

3.3 시각적 유사성 기반 탐지의 도전과제

시각적 유사성 기반의 탐지기법은 정상 웹 사이트와 피싱 웹 사이트를 구별하기 위해 웹 사이트 로고와 같은 디자인 요소 및 텍스트에 의존하므로, 신뢰할 수 있는 웹 페이지의 디자인이 변경되는 경우 모델 재학습이 필요하다. 또한, 모바일 및 데스크톱 플랫폼에서는 경우 각각 별도의 웹 페이지를 제공하기 때문에 같은 데이터 세트를 사용할 수 없는 문제가 있다 [15]. 이 외에 LogoMorph와 같은 적대적 로고 생성을 통한 탐지 회피 기법이 연구되어 적대적 로고 및 스크린샷 생성에 대한 대비가 필요하다 [16].

3.4 메시지 콘텐츠 기반 탐지의 도전과제

메시지 콘텐츠 기반 탐지를 효과적으로 수행하기 위해 언어 처리 과정에서 형태소 분석 과정을 고려해야 한다. 일반적으로 텍스트는 여러 가지의 품사로 구성되어 있으므로, 모든 형태소를 사용한 분석이 어렵다 [12]. 또한, 피싱 메시지는 번역 또는 LLM을 통해 다양한 언어와 표현 방식으로 생성될 수 있으므로 다른 문화와 언어에도 적용할 수 있어야 한다. 이 외에도 피싱 메시지의 실제 데이터를 수집하는 것과 정상 및 피싱 메시지의 데이터 불균형 또한 탐지를 저해하는 요인으로 식별되었다.

4. 결론 및 향후 연구

본 연구팀은 최근 피싱 기술의 고도화와 세계적으

로 피싱 위협이 증가하는 것을 식별하였고, 이에 따라 피싱을 탐지하기 위해 사용되는 다양한 탐지기법을 조사하였다. 조사를 통해 최근 피싱 탐지기법을 URL/도메인, HTML, 시각적 유사성, 메시지 콘텐츠 기반의 접근법으로 분류하였으며, 각 탐지기법이 가진 한계점과 도전과제를 식별하였다. 특히, 탐지를 위해 모델을 학습할 때, 새로운 데이터에 대한 재학습 필요성과 정상 데이터 및 피싱 데이터의 불균형 문제가 공통적으로 식별되었다. 이에 따라 알려지지 않은 데이터에 대한 탐지 성능이 주요 한계로 거론되었으며, 데스크탑과 모바일 플랫폼 간 적용 및 실시간 탐지 가능성과 같은 주요 도전과제가 남아있음을 확인하였다. 향후 피싱 공격 절차에서 발생하는 리디렉션의 탐지기법에 관해 연구하고자 한다.

Acknowledge

본 논문은 과학기술정보통신부의 재원으로 정보통신기획평가원(IITP)의 정보통신방송기술 국제공동연구(Project No. RS-2022-00165794, 50%), 국방ICT융합연구(Project No. 2022-11220701, 30%), 정보통신방송혁신인재양성사업(Project No. 2021-0-01816, 20%)의 지원을 받아 수행된 연구임.

참고문헌

[1] SOCRadar, Phishing in 2024, https://socradar.io/phishing-in-2024-4151-increase-since-chatgpt/

[2] Alsubaei, F. S., Almazroi, A. A., & Ayub, N. (2024). Enhancing phishing detection: A novel hybrid deep learning framework for cybercrime forensics. IEEE Access.

[3] Rangasamy, Gokul, et al. "Phishion: Phishing Detection Application." 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE). IEEE, 2024.

[4] Aravena, Lucas Torrealba, et al. "More than Words is What you Need-Detecting DGA and Phishing Domains with Dom2Vec Word Embeddings." 2024 8th Network Traffic Measurement and Analysis Conference (TMA). IEEE, 2024.

[5] Roy, S. S., Naragam, K. V., & Nilizadeh, S. (2023). Generating phishing attacks using chatgpt. arXiv preprint arXiv:2305.05133.

[6] Çolhak, Furkan, et al. "Phishing Website Detection through Multi-Model Analysis of HTML Content." arXiv preprint arXiv:2401.04820 (2024).

[7] Opara, C., Chen, Y., & Wei, B. (2024). Look before you leap: Detecting phishing web pages by exploiting raw URL and HTML characteristics. Expert Systems with Applications, 236, 121183.

[8] Wang, Mengli, et al. "Phishing webpage detection based on global and local visual similarity." Expert Systems with Applications 252 (2024): 124120.

[9] Pandey, P., & Mishra, N. (2023). Phish-Sight: a new approach for phishing detection using dominant colors on web pages and machine learning. International Journal of Information Security, 22(4), 881-891.

[10] Hazell, J. (2023). Spear phishing with large language models. arXiv preprint arXiv:2305.06972.

[11] Seo, Jae Woo, et al. "On-Device Smishing Classifier Resistant to Text Evasion Attack." IEEE Access (2024).

[12] Kim, Siyoon, et al. "Detection of Korean Phishing Messages Using Biased Discriminant Analysis under Extreme Class Imbalance Problem." Information 15.5 (2024): 265.

[13] Skula, I., & Kvet, M. (2024). A Framework for Preparing a Balanced and Comprehensive Phishing Dataset. IEEE Access.

[14] Vamsi, P., U. Muthaiah, and C. H. Roshan Vardhan. "Defending the Digital Frontier: URL-Based Phishing Detection Extension." International Conference on Computational Intelligence in Data Science. Cham: Springer Nature Switzerland, 2024.

[15] Jain, A. K., Debnath, N., & Jain, A. K. (2022). APuML: an efficient approach to detect mobile phishing webpages using machine learning. Wireless Personal Communications, 125(4), 3227-3248.

[16] Hao, Q., Diwan, N., Yuan, Y., Apruzzese, G., Conti, M., & Wang, G. (2024). It Doesn't Look Like Anything to Me: Using Diffusion Model to Subvert Visual Phishing Detectors. In 33rd USENIX Security Symposium (USENIX Security 24) (pp. 3027-3044).