This article describes an algorithm based on detecting all the IRL's pointed to by the underlying HTML of any web URL.

The key design features of this algorithm are detailed as follows. This particular anti-phishing approach has been designed to meet the following requirements.

High detection efficiency: To provide high detection efficiency, incorrect classification of benign sites as phishing (false-positive) should be minimal and correct classification of phishing sites (true-positive) should be high.

Real-time detection: The prediction of the phishing detection approach must be provided before exposing the user's personal information on the phishing website.

Target independent: Due to the features extracted from both URL and HTML the proposed approach can detect new phishing websites targeting any benign website (zero-day attack).

Third-party independent: The feature set defined in our work are lightweight and client-side adaptable, which do not rely on third-party services such as blacklist/whitelist, Domain Name System (DNS) records, WHOIS record (domain age), search engine indexing, network traffic measures, etc.

The paper demonstrates a good result based on application towards 60000+ URL's out of which 27000+ were Phishing URL's, with just 1.5% false positives.

Possible Shortcomings:

This algorithm seems to be highly resource intensive and uses Python code which could consume significant RAM/CPU and potentially slow down the performance of the system.

References

Aljofey, A., Jiang, Q., Rasool, A., Chen, H., Liu, W., Qu, Q., & Wang, Y. (2022). **An effective detection approach for phishing websites using URL and HTML features.** Scientific Reports, 12(1), 1–19. https://doi-org.nec.gmilcs.org/10.1038/s41598-022-10841-5