Accurate classification of Web requests

Semester thesis proposal

David Gugelmann, gugelmann@tik.ee.ethz.ch, Networked Systems Group (NSG)

Background and motivation

Today's Web sites embed a large number of third-party services, such as advertisement and analytics services, content distribution networks, and services hosting Web templates and JavaScript libraries. This results in a large number of HTTP(S) requests to different services when browsing on the Web, and consequently in complex traffic patterns [1]. These complex traffic patterns make it difficult to automatically reconstruct user activities from Web traffic traces during a forensic investigation. One typically wants to distinguish between three kinds of HTTP(S) requests when reconstruction the Web activities of a client:

- Requests directly triggered by users during Web browsing, e.g., a request triggered by clicking on a link in the Web browser. These requests show which Web pages a user has visited.
- Requests that are not directly triggered by user clicks, but occur indirectly because of the user's Web browsing. Typically, these are the requests to first- and third-party services that load the objects embedded in a Web page.
- Requests that are not related to Web browsing. For example, these can be requests triggered by benign software to check for updates, but the requests can also result from malware activity.

There exist multiple approaches for classifying Web requests based on network traces. ReSurf [2] is an algorithm that prioritizes scalability over accuracy. It solely analyzes the Referer graph, timing information and object sizes, resulting in a fast processing of Web traces. We implemented ReSurf in our tool called Hviz [3] and tuned ReSurf's parameters, which resulted in an improved accuracy for the detection of user clicks for our dataset (recall: 81%, precision: 90%). The WebWitness system [4] presents additional heuristics to reconstruct dependencies and bridge gaps in the requests graph. ClickMiner [5] rather prioritizes accuracy over speed. It reconstructs user actions by replaying the recorded Web traffic in an automated browser. As a result, this approach is less scalable, but it achieves a false positive rate of only around 1% and can correctly reconstruct up to 90% of Web browsing activities.

Aim and tasks

The aim of this thesis is to analyze, combine, and complement the above presented heuristics in order to derive a highly scalable and accurate method for the classification of Web requests. The thesis consists of the following tasks:

- Background research on approaches for the classification of Web requests.
- Analyze the machine learning features used by existing approaches. Implement a selection of the existing features/methods.
- Derive additional features.
- Apply one or several machine learning approaches on the implemented features.
- Evaluate different combinations of existing/new features using Web browsing traces.
- Write a report and present the work.

More information

For more information on this thesis, please contact David Gugelmann (gugelmann@tik.ee.ethz.ch).

References

[1] M. Butkiewicz, H. V. Madhyastha, and V. Sekar, "Understanding website complexity: Measurements, metrics, and implications," in *Proc. IMC '11*, 2011, pp. 313–328. [Online]. Available: http://doi.acm.org/10.1145/2068816.2068846

[2] G. Xie, M. Iliofotou, T. Karagiannis, M. Faloutsos, and Y. Jin, "ReSurf: Reconstructing Web-Surfing Activity From Network Traffic," in *IFIP Networking Conference*, May 2013, pp. 1–9.

[3] D. Gugelmann, F. Gasser, B. Ager, and V. Lenders, "Hviz: Http(s) traffic aggregation and visualization for network forensics," *Digital Investigation*, vol. 12, Supplement 1, pp. 1–11, 2015, Proc. 2nd Annual DFRWS Europe (DFRWS 2015 Europe).

[4] T. Nelms, R. Perdisci, M. Antonakakis, and M. Ahamad, "Webwitness: Investigating, categorizing, and mitigating malware download paths," in *Proc. USENIX Security 15*, 2015.

[5] C. Neasbitt, R. Perdisci, K. Li, and T. Nelms, "ClickMiner: Towards Forensic Reconstruction of User-Browser Interactions from Network Traces," in *Proc. CCS '14*, 2014.