Securing the Web with R

Saeed Abu-Nimeh [1,*]

1. Websense Security Labs, San Diego, CA
*Contact author: sabu-nimeh @ websense.com

Keywords: Classifiers, Web security

R is widely referenced in several disciplines, such as statistics, economics, and biostatistics. However, it is rarely referenced in the network and web security world. When it comes to processing huge amounts of Web pages daily, to discover new attacks and infections, automating such processes is deemed vital.

Thanks to the abundance of classification packages in R, which helps in categorizing Web content in automated and easy ways. We frequently train and test classifiers including classification and regression trees (CART), naive bayesian classifiers (NB), support vector machines (SVM), and random forests (RF) to categorize Web pages based on their content. In addition, we use these classifiers to discover compromised Web pages that are injected with malicious content.

In this study, we go through the journey of crawling and mining Web pages, categorizing Web content, and discovering malicious and compromised Web pages. We show how we utilize various R packages including **e1071** [1], **rpart** [5], **randomForest** [3], **ROCR** [4], and **RMySQL** [2] in our experiments.

References

[1] Evgenia Dimitriadou, Kurt Hornik, Friedrich Leisch, David Meyer, , and Andreas Weingessel, *e1071: Misc functions of the department of statistics (e1071), tu wien*, 2009, R package version 1.5-22.

[2] David A. James and Saikat DebRoy, *Rmysql: R interface to the mysql database*, 2009, R package version 0.7-4.

[3] Andy Liaw and Matthew Wiener, *Classification and regression by randomforest*, R News **2** (2002), no. 3, 18–22.

[4] Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer, *Rocr: Visualizing the performance of scoring classifiers.*, 2009, R package version 1.0-4.

[5] Terry M Therneau and Beth Atkinson. R port by Brian Ripley., *rpart: Recursive partitioning*, 2008, R package version 3.1-42.