Open Science Index, Computer and Information Engineering Vol:10, No:04, 2016 publications.waset.org/abstracts/39697.pdf

Building User Behavioral Models by Processing Web Logs and Clustering Mechanisms

Authors : Madhuka G. P. D. Udantha, Gihan V. Dias, Surangika Ranathunga

Abstract : Today Websites contain very interesting applications. But there are only few methodologies to analyze User navigations through the Websites and formulating if the Website is put to correct use. The web logs are only used if some major attack or malfunctioning occurs. Web Logs contain lot interesting dealings on users in the system. Analyzing web logs has become a challenge due to the huge log volume. Finding interesting patterns is not as easy as it is due to size, distribution and importance of minor details of each log. Web logs contain very important data of user and site which are not been put to good use. Retrieving interesting information from logs gives an idea of what the users need, group users according to their various needs and improve site to build an effective and efficient site. The model we built is able to detect attacks or malfunctioning of the system and anomaly detection. Logs will be more complex as volume of traffic and the size and complexity of web site grows. Unsupervised techniques are used in this solution which is fully automated. Expert knowledge is only used in validation. In our approach first clean and purify the logs to bring them to a common platform with a standard format and structure. After cleaning module web session builder is executed. It outputs two files, Web Sessions file and Indexed URLs file. The Indexed URLs file contains the list of URLs accessed and their indices. Web Sessions file lists down the indices of each web session. Then DBSCAN and EM Algorithms are used iteratively and recursively to get the best clustering results of the web sessions. Using homogeneity, completeness, V-measure, intra and inter cluster distance and silhouette coefficient as parameters these algorithms self-evaluate themselves to input better parametric values to run the algorithms. If a cluster is found to be too large then micro-clustering is used. Using Cluster Signature Module the clusters are annotated with a unique signature called finger-print. In this module each cluster is fed to Associative Rule Learning Module. If it outputs confidence and support as value 1 for an access sequence it would be a potential signature for the cluster. Then the access sequence occurrences are checked in other clusters. If it is found to be unique for the cluster considered then the cluster is annotated with the signature. These signatures are used in anomaly detection, prevent cyber attacks, real-time dashboards that visualize users, accessing web pages, predict actions of users and various other applications in Finance, University Websites, News and Media Websites etc.

Keywords : anomaly detection, clustering, pattern recognition, web sessions
Conference Title : ICINA 2016 : International Conference on Information, Networking and Automation
Conference Location : Istanbul, Türkiye
Conference Dates : April 19-20, 2016