EXTRACTION OF WEB USER-SESSIONS BY USING DENSITY BASED CLUSTERING TECHNIQUE: SECURITY ISSUES

Rajesh.Y [1], Vidya Sagar .V [2], Asha Varma.S [3]

1 Asst.Professor, CSE Department, ALIET, Vijayawada, India
raja_cse2006@yahoo.co.in
2 Asst.Professor, CSE Department, ALIET, Vijayawada, India
vidyasagar0058@gmail.com
3 Asst.Professor, CSE Department, ALIET, Vijayawada, India
ashavarmak@gmail.com

Abstract

In today's internet world mainly focuses on the definition and identification of "Web user-sessions". The classification of a user-session is not trivial, because several users has connected to web for performing various tasks. Conventional approaches rely on threshold based mechanisms. However, the mechanisms used previously are very sensitive to the value chosen for the threshold, which may be not easy to set correctly. Here, we applying Density based clustering technique, to define a methodology for identifying genuine Web user-sessions. Density-based approaches have the advantage of extracting clusters from a highly noisy environment, We define a density based clustering approach for grouping user-session of several TCP connections generated by the same source host, and we apply it to artificial traffic traces. We mainly proposed a security chain mechanism and also considering the security issues regarding sessions.

Index Terms: *User-session, Cluster, Clustering techniques, Density-based Clustering, security.*

---------- *** ----------

1. INTRODUCTION

The study of telecommunication networks has been often based on traffic measurements, which are used to create traffic models and obtain performance estimates. While a lot of attention has been traditionally devoted to traffic characterization at the packet and transport layers , few are the studies on traffic properties at the session/user layer This is due to the difficulty in defining the "session" concept itself , which depends on the considered application. Applications such as *telnet* or *ssh* typically generate a single TCP connection per single user-session, whereas application layer protocols such as HTTP, IMAP/SMTP and X11 usually generate multiple TCP connections per user-session. Also, the generally accepted conjecture that such sessions follow a Poisson arrival process might have reduced the interest in the user-session process analysis. User-session identification and characterization play an important role both in Internet traffic modeling and in the proper dimensioning of network resources. Besides increasing the knowledge of network traffic and user behavior, they yield workload models which may be exploited for both performance evaluation and dimensioning of network elements. Synthetic workload generators may be defined to assess network performance, e.g., benchmarking of server farms, firewalls, proxies or NATs, as in similarly, user-session characterization allows researchers to build realistic scenarios when assessing the performance of a complex network via simulation. Furthermore, network dimensioning problems are usually based on simple assumptions to permit analytical formulations and solutions. The validation of these assumptions can only be obtained by checking the model against traffic measurements. Finally, the knowledge of user-session behavior is important for service providers, for example to dimension access links and router capacity. Data mining is emerging as one of the key features of many homeland security initiatives. Often used as a means for detecting fraud, assessing risk, and product retailing, data mining involves the use of data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. In the context of homeland security, data mining is often viewed as a potential means to identify terrorist activities, such as money transfers and communications, and to identify and track individual terrorists themselves, such as through travel and immigration records. While data mining represents a significant advance in the type of analytical tools currently

available, there are limitations to its capability. One limitation is that although data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns. These types of determinations must be made by the user. A second limitation is that while data mining can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship. To be successful, data mining still requires skilled technical and analytical specialists who can structure the analysis and interpret the output that is created. Data mining is becoming increasingly common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. In the public sector, data mining applications initially were used as a means to detect fraud and waste, but have grown to also be used for purposes such as measuring and improving program performance. However, some of the homeland security data mining applications represent a significant expansion in the quantity and scope of data to be analyzed. Two efforts that have attracted a higher level of congressional interest include the Terrorism Information Awareness (TIA) project (now-discontinued) and the Computer-Assisted Passenger Prescreening System II (CAPPS II) project (now canceled and replaced by Secure Flight). As with other aspects of data mining, while technological capabilities are important, there are other implementation and oversight issues that can influence the success of a project's outcome. One issue is data quality, which refers to the accuracy and completeness of the data being analyzed. A second issue is the interoperability of the data mining software and databases being used by different agencies. A third issue is mission creep, or the use of data for purposes other than for which the data were originally collected. A fourth issue is privacy. Questions that may be considered include the degree to which government agencies should use and mix commercial data with government data, whether data sources are being used for purposes other than those for which they were originally designed, and possible application of the Privacy Act to these initiatives. It is anticipated that congressional oversight of data mining projects will grow as data mining efforts continue to evolve.

1.
2. **2. Introduction to Web**
3.
4. **2.1 Web Applications**
5.
6. Broadly defined, a web-based software system consists of a set of web pages and components that interact to form a system that executes using web server(s), network, HTTP and a browser, and in which user input (navigation and data input) affectthe

state of the system. A web page can be either static, in which case the content is fixed, or dynamic, such that its content may depend on user input. Web applications may have a number of characteristics, including an integration of numerous technologies; modularization into reusable components that may be constructed by third parties; a well-defined, layered architecture; dynamically generated pages with dynamic content; and typically extend an application framework. Large web-based software systems can require thousands to millions of lines of code, contain many interactions among objects, and involve significant interaction with users. In addition, changing user profiles and frequent small maintenance changes complicate automated testing.

2. Definition of User Sessions

The grouping of IP packets into TCP connections is determined by the TCP protocol. However, the way in which TCP connections should be combined into user sessions is a matter of choice. Our definition of user session was based only on information contained in the TCP and IP packet headers. For each TCP connection, the host IP address initiating the connection was defined to be the user for that connection. The TCP connections corresponding to a given user were then grouped into user sessions as follows:

- A user session begins when a user who has been idle opens a TCP connection.
- A user session ends, and the next idle period begins, when the user has had no open connections for consecutive seconds

2.1 Algorithm

The key idea of the DBSCAN algorithm is that, for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points, that is, the density in the neighborhood has to exceed some predefined threshold. This algorithm needs three input parameters:

1) k is the neighbor list size;

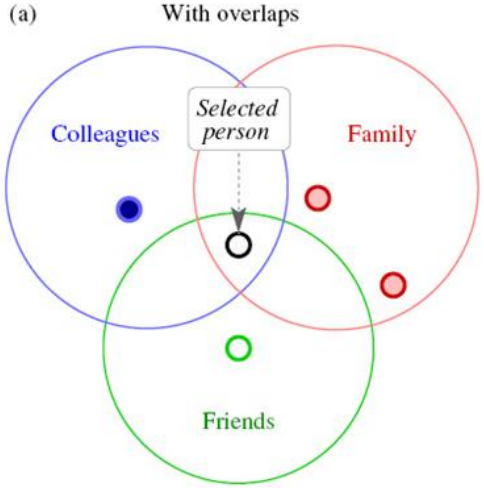
2) X is the radius that delimitate the neighborhood area of a point (Epsneighbourhood);

ISSN: 2250–3676

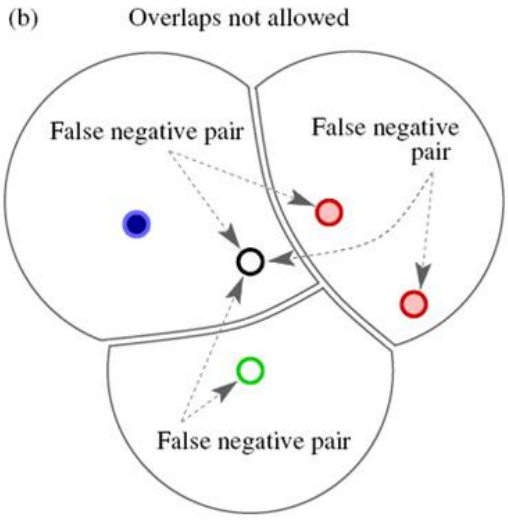
[IJESAT] INTERNATIONAL JOURNAL OF ENGINEERING SCIENCE & ADVANCED TECHNOLOGY Volume-2, Issue-4, 1156 – 1159

3) MinPts is -the minimum number of points that must exist in the X-neighbourhood.

The clustering process is based on the classification of the points in the dataset as core points, border points and noise points, and on the use of density relations between points directly density-reachable, density-reachable, and density-connected to form the clusters.



To clusters a web log information, our DBSCAN implementation starts by identifying the k nearest neighbors of each point and identify the farthest k nearest neighbor (in terms of Euclidean distance)1. The average of all this distance is then calculated. After that, for each point of the dataset the algorithm identifies the directly density-reachable points (using the Eps threshold provided by the user) and classifies the points into core or border points. It then loop trough all points of the dataset and for the core points it starts to construct a new cluster with the support of the GetDRPoints() procedure that follows the definition of density reachable points.



In this phase the value used as Eps threshold is the average distance calculated previously. At the end, the composition of the clusters is verified in order to check if there exist clusters that can be merged together. This can append if two points of different clusters are at a distance less than Eps.

2.1.1 SNN algorithm

The SNN algorithm, as DBSCAN, is a density-based clustering algorithm. The main difference between this algorithm and DBSCAN is that it defines the similarity between points by looking at the number of nearest neighbours that two points share. Using this similarity measure in the SNN algorithm, the density is defined as the sum of the similarities of the nearest neighbours of a point. Points with high density become core points, while points with low density represent noise points. All remainder points that are strongly similar to a specific core points will represent a new clusters.

3. PROPOSED METHODLOGY

3.1 Security Chain Mechanism:

The mechanism which tends to prove that the user is authenticated and takes the valid cookies and provides the Authentication for sessions.

Step 1:

When Unknown user logins our mechanism checks for validity of the users request.

Step 2:

Then the user Request is authenticated and takes the valid cookies from client side and authenticated sessions will be generated.

Step 3: If the session expires moves to step 1.

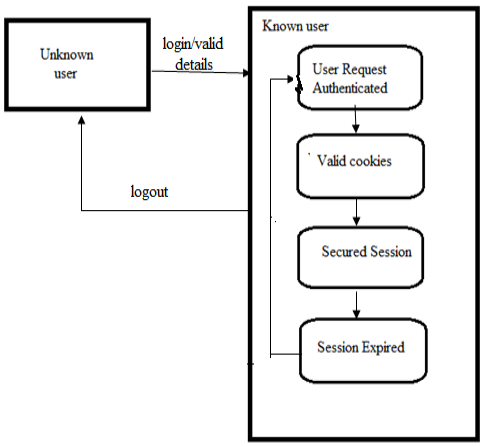


Fig: Security chain mechanism

3.2 Identify user-sessions:

The following three-step algorithm is run to identify user-sessions:

1) An initial clustering is obtained using a partitional algorithm;

2) A DBSCAN algorithm is used to aggregate the clusters and to obtain a good estimation of the final number of clusters and to indentify the genuine user-session

3) A density based algorithm (SNN) is used to find the high density points (core points) and low density points (Noise) in web traffic.

4. CONCLUSION

Clustering techniques were applied to a large set of real Internet traffic traces to identify genuine Web user-sessions. We define a density based clustering approach for grouping user-session of several TCP connections generated by the same source host, and also., The clustering algorithm proposed in this paper can be helpful in studying traffic properties at the user level, and could be easily extended to deal with other types of user-sessions. we are also considering security issues.

5. REFERENCES:

[1] **Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu**, *" A Density- Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise"*, The Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA, 1996

[2 **Levent Ertoz, Michael Steinback, Vipin Kumar**, "Finding Clusters of Different Sizes, Shapes, and Density in Noisy, High Dimensional Data", Second SIAM International Conference on Data Mining, San Francisco, CA, USA, 2003

[3] **V. Paxson,** "Empirically derived analytic models of wide-area TCP connections," *IEEE/ACMTrans. Netw.*, vol. 2, no. 4, pp. 316–336, Aug. 1994.

[4] **C. Nuzman, I. Saniee, W. Sweldens, and A. Weiss**, "A compound model for TCP connection arrivals, with applications to LAN and WAN," *Computer Networks, Special Issue on Long-Range Dependent Traffic*, vol. 40, no. 3, , Oct. 2002.

[5] **F. D. Smith, F. H. Campos, K. Jeffay, and D. Ott**, "What TCP/IP protocol headers can tell us about the web," *SIGMETRICS Perform. Eval. Rev.*, vol. 29.