Abstract

Web usage mining is the area of web mining which deals with extraction of useful knowledge from web log information produced by web servers. One of the most important tasks of Web Usage Mining (WUM) is web user clustering which forms groups of users exhibiting similar interests or similar browsing patterns. This paper presents results of clustering techniques for Web log data using K-means and Bisecting K-means algorithm. Clusters are formed with respect to similar IP address and packet combinations. The clustering framework is further used as an approach for intrusion detection from the log files. The system is trained first by labeling the classes and then tested to check for any intrusions. Recommendation output is generated which help in classifying the whether the input IP's are 'safe' or 'infected'. Comparison of both algorithms is done and performance is evaluated with respect to time and accuracy. From the experimental results, it is found that Bisecting K-means overcomes the major drawbacks of basic K-means algorithm.

References

- Oren Etzioni "The world wide Web: Quagmire or gold mine"; Communications
- Bamshad Mobasher, Chapter: 12, "Web Usage Mining in Data Collection and Pre-Processing", ACM SIGKDD 2007 Pages 450-483.
- Natheer Khasawneh and Hien-Chung Chan, "Active User-Based and Ontology-Based Weblog data preprocessing for Web Usage Mining", IEEE International Conference on Web Intelligence, 2006.
- M. Steinbach, G. Karypis, V. Kumar, "A comparison of document clustering techniques", In KDD Workshop on Text Mining, 2000

**Index Terms**

Computer Science, Web Services

**Keywords**

Web mining, Clustering, Bisecting K-means, Intrusion detection