ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

Review Paper on Classification of Web Access Data Streams with Skewed Distribution

Sukhchain Singh, Dr. Rahul Malhotra

PG Student, Dept. of ECE, Guru Teg Bahadar Khalsa Institute of Engineering & Technology, Chapia Wali Malout, India

Director, Dept. of ECE, Guru Teg Bahadar Khalsa Institute of Engineering & Technology, Chapia Wali Malout, India

ABSTRACT: In recent years, there have been some interesting studies on predictive modeling in data streams. However, most such studies assume relatively balanced and stable data streams but cannot handle well rather skewed (e.g., few positives but lots of negatives) and stochastic distributions, which are typical in many data stream applications. One of fundamental problem in the task of mining streaming data is the concept drift over time. Such data streams may also exhibit high and varying degree of class imbalance, which can further complicate the task. In scenarios like these, class imbalance is particularly difficulty to overcome and has not been as thoroughly studied. Most of studies on classification of data streams assume relatively balanced and stable data streams but cannot handle well rather skewed streams which are typical in many data stream applications. Class imbalance in such skewed data stream can be seen in many real world applications. In such scenarios learning from skewed data streams results in classifier biased towards the majority class which results in classifier biased towards the majority class which results in misclassification of minority class examples, since in these scenarios minority class examples are too less than the majority class. The losses associated with misclassification of minority classes can be higher in some applications. In this project we present our preliminary work to deal with classification of the data streams with skewed distribution of classes.

I. INTRODUCTION

Many real applications, such as network traffic monitoring, credit card fraud detection, and web click stream, generate continuously arriving data, known as data streams. Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. It is used to understand customer behavior, evaluate the effectiveness of a particular Web site, and help quantify the success of a marketing campaign. It also allows looking for patterns in data through content mining, structure mining, and usage mining. Content mining is used to examine data collected by search engines and web spiders. Structure mining is used to examine data related to the structure of a particular Web site and Web Usage Mining is applied to many real world problems to discover interesting user navigation patterns for improvement of web site design by making additional topic or recommendations observing user or customer behavior.

Web usage mining has several applications and is used in the following areas:

1. It offers users the ability to analyze massive volume of click stream or click flow data,
integrate the data seamlessly, with translation and demographic data from offline sources.

2 Personalization for a user can be achieved by keeping track of previously accessed pages. These pages can be used to identify the typical browsing behavior of a user and subsequently to predict desired pages.

3 By determining access behavior of users, needed links can be identified to improve the overall performance of future accesses.

4 Web usage patterns are used to gather business intelligence to improve customer attraction, customer retention, sales, marketing, and advertisements cross sales.

5 Web usage mining is used in e-Learning, e-Business, e-Commerce, e-Newspapers, e-Government and Digital Libraries. The information gathered through Web mining is evaluated by using traditional data mining parameters such as clustering and classification, association, and examination of sequential patterns.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

II. WEB LOGS

A Web log file records activity information when a Web user submits a request to a Web Server. The main source of raw data is the web access log which we shall refer to as log file. As log files are originally meant for debugging purposes.
A log file can be located in three different places: i) Web Servers, ii) Web proxy Servers, and iii) Client browsers.

III. LITERATURE SURVEY

[1]A Survey on Web Usage Mining with Clicking Pattern in Grid Computing Environment (Volume 3, Issue 11, November 2013)
Web usage mining is a current and drastic research area in web usage mining focused on learning about web users and their interaction about web sites. The aim of web usage mining is to find user's access moves frequently and quickly from the massive web log data such as through frequent access paths, frequent access page groups and user clusters. Through web usage mining the whole registration information left by user access can be mined with the user access mode which provides foundation for decision making for big organizations. Web mining has become very crucial in those areas which are based upon web. So for managing massive web log data we need solid distributed environments like grids. Calculating user's browsing terms is a necessary operation of usage mining which generates exact usage data. In this paper we introduce a survey and analysis of latest web usage mining tricks and tactics in the form of algorithms with distributive grid environments for saving and using heavy data very easily from server in exact given time duration.
[2]Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining(Volume 2, Issue 9, September 2011)
Web usage mining refers to the automatic discovery and analysis of patterns in click stream and associated data collected or generated as a result of user interactions with web resources on one or more web sites. It consists of three phases which are data Preprocessing, pattern discovery and pattern analysis. In the pattern discovery phase, frequent pattern discovery algorithms applied on raw data. In the pattern analysis phase interesting knowledge is extracted from frequent patterns and these results are used for website modification. In this paper we are using the FP-growth algorithm for obtaining frequent access patterns from the web log data and providing valuable information about the user's interest.
[3] **Web Access Prediction Model using Clustering and Artificial Neural Network(Vol. 3 Issue 9, September-2014)**
The number of web users is increasing rapidly day by day. With the increasing number of web access requests and explosive growth of data sources available on the web, the network traffic also increases rapidly, which results in poor user latency and difficult for user to access the web. Thus web caching and pre-fetching of web pages play an important role in this scenario. However, without having a sophisticated mechanism regarding which pages are most likely to be visited, caching is meaningless. The idea is to improving the accuracy in web access prediction. This paper proposed a next web page access prediction model. The proposed model uses Feed Forward Artificial Neural Network and the concept of web session clustering. This model can also be used in other web oriented applications. The Neural Network is trained and tested with a large data set and better prediction accuracy is achieved.

IV. PROBLEM FORMULATION AND OBJECTIVES

Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. It is used to understand customer behavior, evaluate the effectiveness of a particular Web site, and help quantify the success of a marketing campaign. It also allows looking for patterns in data through content mining, structure mining, and usage mining. Content mining is used to examine data collected by search engines and web spiders. Structure mining is used to examine data related to the structure of a particular Web site and Web Usage Mining is applied to many real world problems to discover interesting user navigation patterns for improvement of web site design by making additional topic or recommendations observing user or customer behavior.
For Example in banking web sites we want to analysis the basic web pattern of different option that how much time it requires to access . This can be done on the basis of age, profession and various other parameters. we present an

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

automated timeline tool for analyzing of the web servers for the time analysis. In the proposed system, we have developed tools that assist the server administrator and web administrator to improve their website by determining occurred link connections in the website. Firstly, we have obtained access log files, which are recorded in web server. The obtained log files were analyzed by proposed methodology. So, raw log files were pre-processed and the path analysis technique was used to investigate the web log files of URL information concerning access to electronic sources. The proposed methodology was applied to the user access log files in the web server. The results and findings of this experimental study can be used by the web administration and web designer in order to plan the upgrading and enhancement to the website. The work proposed here in this paper will focus on the analysis of events with visualizations for the web access log files events or simply web log file events. The proposed technique for the web server time line analysis will perform the following actions that provides support to the investigators. First, the experts will take the log files and by using the proposed tool for log files visualization and analysis. The proposed server timeline analysis and visualization tool presented in this paper supports many of the proposed solutions for automated forensic analysis, and it would be interesting to integrate some of these approaches with our work. It generates hypotheses before executing the process of reconstruction experiments and the problem of performing automated comparison of the results with the digital evidence.

V. OBJECTIVES

1. To understand the time for banking web site like HDFC bank.
2. Collect the data for various peoples.
3. Analysis for skewed ness for different classes of data

VI. TOOL USED

1. Hyper text markup language
2. Cascading style sheets (css)
3. Javascript
4. Php
5. Mysql
6. Wamp

VII. RESEARCH METHODOLOGY

This research methodology includes various steps

- in first step we will establish 25 question which can be surveyed for banking web site.
- in second step we will build a web site through which we will collect the data.
- Each page will be having one question and corresponding time of start of the attempt of the question and end of the question.
- Each question start time and end time will be noted on mysql server.
- Later on these mysql entries will be imported on to the excel sheet.
- Put the analysis for these data items

REFERENCES

[1]. K. Coar and D. Robinson. The WWW Common Gate-way Interface, Version 1.1. Internet Draft, June 1999.
[2]. J. Liberty and D. Hurwitz. Programming ASP.NET. O"REILLY, February 2002
[3]. Security Tracker. Vulnerability statistics April 2001-march 2002. http://www.securitytracker.com/learn/statistics.html, April 2002.
[4]. CERT/CC. "Code Red Worm" Exploiting Buffer Overflow In IIS Indexing Service DLL. Advisory CA-2001-19, July 2001.
[5]. M. Roesch. Snort - Lightweight Intrusion Detection for Networks.In Proceedings of the USENIX LISA "99 Conference, November 1999.
[6]. Carrier, B.D., Spafford, E.H.: Defining event recon-struction of digital crime scenes. J. Forensic Sci. 49 (2004)
[7]. Carrier, B.: An event-based digital forensic investiga-tion framework. In: Digital forensic research workshop (2004)
[8]. Chisum, W.J., Turvey, B.E.: Evidence dynamics: Lo-card"s exchange principle crime reconstruction. J. Be-hav. Profiling 1(1) (2000)
[9]. W. Vogels, "Eventually Consistent," ACM Queue, 4 Dec. 2008; http://queue.acm.org/detail.cfm?id=1466448.[10]. Stephenson, P.: Formal modeling of post-incident root cause analysis. Int. J. Digit. Evid. 2 (2003)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

[11]. Gladyshev, P., Patel, A.: Finite state machine approach to digital event reconstruction. Digit. Invest. 1 (2004)
[12]. Stallard, T.B.:Automated analysis for digital forensic science. Master˝s thesis, University of California, Davis (2002)
[13]. Stallard,T.,Levitt,K.N.:Automated analysis for digital forensic science: Semantic integrity checking. In: AC-SAC 160–169 (2003)
[14]. Abbott, J., Bell, J., Clark, A., Vel, O.D., Mohay, G.: Automated recognition of event scenarios for digital fo-rensics. In: SAC ˝06: Proceedings of the 2006 ACM symposium on applied computing pp. 293–300.ACMPress,NewYork (2006)