International Journal of Information and Computation Technology.
ISSN 0974-2239 Volume 3, Number 7 (2013), pp. 617-624
© International Research Publications House
http://www. irphouse.com /ijict.htm

Web Usage Mining: Identification of Trends Followed by the user through Neural Network

Priyanka Verma[1] **and Nishtha Keswani**[2]

[1]*Computer Science Department, The IIS University*
Jaipur, Rajasthan, India.
[2]*Computer Science Department, Central University*
Jaipur, Rajasthan, India.

Abstract

The main purpose of website is to provide user with information which is required by them. For proper website management it is important that we study individual trends followed by the users and then customize website according to customers. One such technique which can be used to mine data stored in web logs is web mining. Web mining has three main thrust areas Web Usage mining, Web Content mining, Web Structure mining. Web usage mining is mining of usage data captured through various logs stored on server, client or proxy. For best results we have to ensure that we analyze data stored in all three logs. This usage data includes preferences, trends followed by the user. Artificial neural network is information processing system made up of large number of processing elements called neurons. Their main function is to process data given by user. In this paper we would be studying what is web usage mining and then use artificial neural network to detect patterns or trends followed by user to through k means algorithm.

Keywords: Web usage mining, usage patterns, neurons, neural network.

1. Introduction

World Wide Web is warehouse of information. It is used by the user to get required information requested through queries. Sometimes user might not be satisfied with response given.This might be as pages which are requested by the user have not been indexed since they are not indexed they are not returned in response to query submitted by the user. To increase user satisfaction for requests made on web we need a new technique that will enable user to get required information easily, efficiently and correctly, that easily mines the required information within fraction of seconds."This extraction of Information on Internet or World Wide Web is called Web Mining" [3].It is technique of mining data on World Wide Web. Web mining has three major thrust areas

- Web Usage mining
- Web Content mining
- Web Structure mining

Web usage mining is mining of web log records to discover patterns of web pages accessed by the user. "It has become very critical for effective Web site management, creating adaptive Web sites, business and support services, personalization, and network traffic flow analysis and so on"[4].Neural Network is interconnected network of processing elements called artificial neurons or nodes. These neurons function similar to biological neurons i.e. they take some input process it and return output. In this paper we would be identifying trends flowed by the users through neural networks. Data will be taken from the logs stored on server side or client side or proxy. For best results it is important that we will have to consider data from all the three logs.

2. Web Mining

"Data mining is the process of analyzing data from different angles and summarizing it into useful information. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases."[1]. Web mining is the application of data mining techniques to discover patterns or trends followed by the user from the Web [2] .It is required as only small portion of information on web is relevant and giving user what he wants is important. It's again necessary to reduce time loss experienced by users while browsing for the required information.

2.1 Why web mining

Web mining is required as information stored on worldwide web is growing rapidly and giving user what he wants is very important.

2.2 Thrust areas of web mining

There are three main thrust areas of web mining. Patterns followed by the users are evaluated by these three techniques of Web Mining and then these patterns are analyzed to get a user desired output. Desired output is then fed into the user understandable GUI [8].

1. **Web Content Mining**
   Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content which can be text or multimedia or both[2]
2. Web Structure mining
   Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site[2].
3. Web usage mining
   Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications [2].

3. Web Usage Mining

Web usage mining is mining of web logs to discover access patterns of the pages accessed by the user.Analyzing regularities in web log records can help us to identify potential customers for ecommerce,help in customization of web pages,improving server performanace.Web server saves all entries of pages accessed in web logs.It includes URL requested,IP address,timestamp.These log files can also be created at client and proxy.Web log databases provide rich information about web dynamics and that’s why it is important to develop a technique that will help us to mine web log databases.This technique is web usage mining.Data stored in logs can be used to find most frequently accessed web pages,frequently accessed time periods.This data will help us in finding most potential customers to be targeted for marketing.It can also be done to find trends of web access.Web sites imnprove themselves by learning from user access pateerns.Web log analysis can also help to build customized web services for individual users.

3.1 Web usage mining process

There are four phases to perform web usage mining [4]

- Preprocessing→ It is a process of preparing data so that it can be used for Pattern Discovery and analysis. It includes Cleaning of Server Log files accompanied by identification of users sessions and user habits.

It consists of

- Data field extraction
- Data Cleaning
- User identification
- Session identification
- Pattern Discovery→After the data is preprocessed, this data is utilized for discovering homogeneous patterns.[8]

- Pattern Analysis→ Once the patterns are discovered then these patterns is evaluated and analysis is performed on these patterns and result generated is given to neural network for further processing.

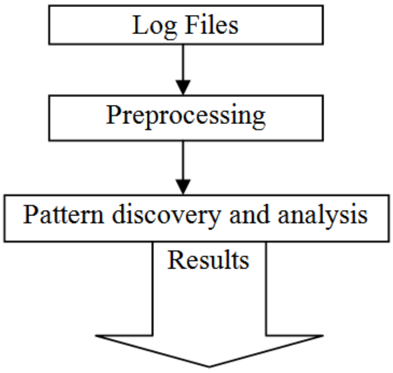

Figure 1: Web usage mining process.

3.2 Problems faced while performing web usage mining[9]

- Processing of logs that is cleaning of log files
- Cleaning of log files that is removing data that is not relavent
- Identification of user sessions
- Identification of user habits.

4. Artificial Neural Network

“An Artificial Neural Network (ANN) is an information-processing structure. The key element of this structure is large number of highly interconnected processing elements called neurons working together to solve a specific problem. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Neural Network has few advantages

- Adaptive learning: An ability to learn how to do tasks based on the data given for initial training
- Self-Organization: An ANN can create its own organization or representation of the information it receives during learning time. This is called as SOM (self organization methods).
- Real Time Operation: ANN computations can be carried out in parallel.
- Fault Tolerance via Redundant Information Coding: Partial destruction of a network leads to the corresponding degradation of redundant information. ” [5]

4.1 Structure of data in web logs [6]

The log files are text files that can range in size from 1KB to 100MB, depending on the traffic at a given a website.The data will be taken for any particular website at given time

There are various fields in the log data which includes

- IP address:This is the IP address of the machine that contacted our site.
- Username etc:This is the user that requested that website
- Timestamp:It is the timestamp of the visit
- Access request:It is the request made
- Result status code: This is whether Url was sucessfully returned or not.A number is saved stating whether request was sucessfully answered or not.
- Bytes transferred: The number of bytes transferred after request was responeded to by the server.
- ReferrerURL:This is the page refered by the user
- User agent:It is the software that the user is using to access the websiteIt is actually browser used by the user.

4.2 Performing web usage mining

The web log data considered for evaluation is collected from any particular web server for specific period. Initially the log file consists of raw log entries with noisy entries like gif, jpeg etc which are not necessary for web log mining. So data cleaning is performed to remove the unnecessary log which will reduce the processing in determining the web usage pattern. After the data cleaning process is performed, then users are identified by using IPaddress and UserAgent fields. Unique users are identified after applying the algorithm and sessions whose paths are completed to form transactions are found out. Completed transactions are represented in a user transactions-urls matrix format. In this paper clustering of URLs is done by using k means to find out occurrence of each unique URL in each cluster. Clustering is a technique to search hidden patterns that exists in datasets. It is a process of grouping data objects into disjoint clusters so that the data in each cluster are similar, yet different to the other clusters. A popular clustering method that minimizes the clustering error is the k-means algorithm. It is attractive in practice, because it is simple and it is generally very fast. It partitions the input dataset into k clusters. Each cluster is represented by an adaptively changing centroid(also called cluster centre), starting from some initial values named seed-points. k-Means computes the distances between the inputs (also called input data points) and centroids, and assigns inputs to the nearest centroid

Steps for Clustering [10],[11]

1. After cleaning and identification of sessions we get multiple transactions.Each transaction consists of multiple URLS.
2. Number of input neurons to ANN is the most common pages of website
3. We calculate alpha which is threshold value it is actually similarity between any two transactions.
4. Select any one transaction as the centroid out of multiple transactions identified.

5. Choose one more transaction and compute the distance between centroid and this transaction ,if this distance is smaller then alpha then second transaction is another cluster.Again choose another transaction compute distance and like this continue forming clusters

Input: Dataset D of N log entries and threshold α

Output : Initial points

```
Algorithm
Select one transaction from dataset Fix C1 is the center of first cluster

K = 1, Ck= X1
For i = 2 to N

Cm :d(Xi , Cm) = MAXi<j<=k d(Xi,Cj)
If d(Xi , Cm) < α then k = k+1 Ck = Xi
Else i= i+1
End if
End for
```

The above algorithm automatically select s initial points based on given threshold α.

6. Next step is to reduce number of clusters formed for this initial points of each cluster are choosen as centroids,then similarity between centroids is calculated using city bliock measure then again threshold value is assigned and is compared with similarity value if similarity value is greater than threshold then both the clusters are the same and can be merged as one

5. Conclusion

In this paper, we studied the use of the neural network to classify the data stored in logs and access patterns or trends followed by user. Before studying patterns we have used k means clustering algorithm to cluster users.The data obtained after mining would ultimately help us in customizing our web sites according to the user.i.e. it will help in better web site management, personalization of web pages.

References

[1] http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[2] http://en.wikipedia.org/wiki/Web_mining

[3] Mrs.Bhanu Bhardwaj (2012), “Extracting Data Through Web mining”, International Journal of Engineering Research & Technology (IJERT),Vol. 1 Issue 3,

[4] Sonali Muddalwar Shashank Kawar (2012) ,“Applying artificial neural network in web usage mining”, Vol 1 Issue 4, International Journal of Computer Science and Management

[5] O. Etzioni(1996), “The world wide Web: Quagmire or gold mine.” Communications of the ACM, Vol. 39 No. 11, pp. 65 -68, Nov. 1996.

[6] http://www.web-datamining.net/usage/

[7] http://en.wikipedia.org/wiki/Web_mining

[8] Anshuman Sharma (2012), “Web usage mining using neural network” International Journal of Reviews in Computing

[9] Ketki Muzumdar, Ravi Mante, Prashant Chatur(2013),Neural Network Approach for Web Usage Mining, International Journal of Recent Technology and Engineering (IJRTE), Volume-2, Issue-2, May 2013

[10] V.Chitraa, Dr.Antony Selvadoss Thanamani(2012), An Enhanced Clustering Technique for Web Usage Mining, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 1 Issue 4, June – 2012.