ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

"Clustering Algorithm Employ in Public Log Data": An Overview

Sachin Pardeshi[1], Parag Patil[2]

Assistant Professor, Dept. of Computer Engineering, R.C.Patel Institute of Technology, Shirpur, India[1]

Assistant Professor, Dept. of Computer Engineering, J.T.Mahajan College of Engineering, Faizpur, India[2]

Abstract: This paper aims to present technique to make private log information public and apply clustering algorithm on collected log file to extract knowledge from public and free log files.

Keyword: —cluster; free log; web mining

I. INTRODUCTION

Web Usage Mining is the application of data mining technique to discover information from the web log data in order to understand and better serve the needs of Web based applications. Web log gathers the identity or origin of web users along with their browsing behavior at a Web site [1,3]. But the problem is companies do not share their private log information with other people. If all the people get this log information, all people get opportunities to extract knowledge from that log data and accordingly they can make changes in their web site to better serve the people.

On internet some web site have heavy traffic and some similar service provider web site have scarcity in traffic, but if they get the information of another web site who have huge demand they gets opportunities to make changes accordingly to their web site. Many researcher wants a log file for their research work but they does not get these log files for their research work because log files of different web site are private they are not public.

This paper aims to present technique to make private log information public and apply improved Apriori algorithm on collected log file to extract knowledge from public and free log files.

This paper has been organized as follows: Section 2 present technique for free log, Section 3 discussed Apriori Algorithm, Section 4 Introduces association rule mining, Section 4 describing methodology, Section 5 provides illustrative example.

II. TECHNIQUES FOR FREE LOG

A. A Tool for User Behaviour Information

Public log information gives opportunity to the people to develop a web site by analyzing log information of other web sites[2]. It can be helpful for researchers if they get all different web site log information from this tool. It can be help full for analysis and to design a different API.

To test this concept, we have developed Google Chrome extension.

- *Google Chrome Extension:*

Google Chrome supports two ways of installing external extensions.

i. Using preferences .json file

ii. Using the windows registry .crx file

We have implemented a tool using .json file. Example is as shown below.

```
{
 "name": "WEB LOG GENERATOR",
 "description": "Generates User Web Activity Log.",
 "version": "1.1",
 "background_page": "background.html",
 "permissions":[
        "webRequest",
        "webRequestBlocking",
        "tabs",
        "http://*/*",
        "https://*/*"
 ],
 "browser_action": {
   "default_title": "WEB LOG GENERATOR",
   "default_icon": "off_16x16.png"
 }
}
```

Example 1 .json file

As shown in above example application logic implemented in html page "background.html" embedded with Java Script. Java Script is the technology used to process the information, manage the events, and communicate with the servers and services [2].

The most significant way to gather users' behavior information when the users' browsing the web pages. A tool implemented that is Google Chrome Extension.

The performance parameter used to at client side that is Fetch Response and Fetch Transmission time.

- *Fetch Transmission Time:* The time between the requests is sent and the first byte of the response is received.

- *Fetch Response Time:* The time between requests is sent and last byte of the response is received.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

Sometime user is reluctant to share his private behavior information. Also avoid government law problems. The user can decide way of sharing his behavior information. Our extension provides the play and pause functionality. If user is clicked on the pause button of the extension, Extension does not share users' behavior information with the public.

Google Chrome extension forwards the information like URL, IP address, time, request ID and type to the local database. From local database, it can be uploaded to public server.

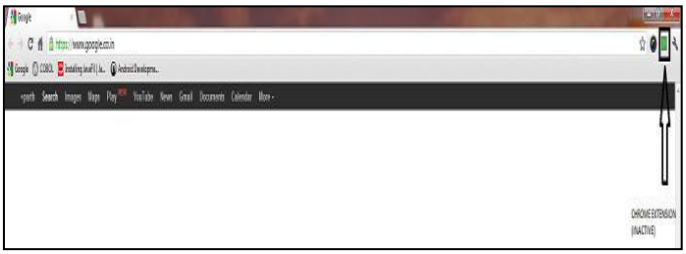

Figure 1. Google Chrome Extension when inactive

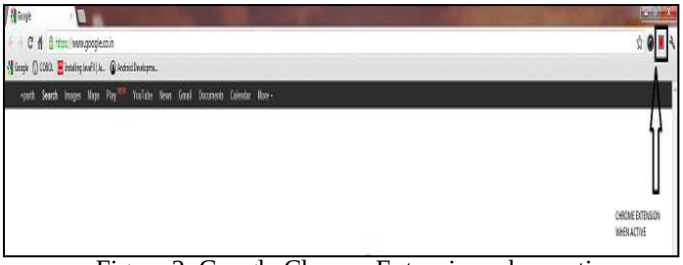

Figure 2. Google Chrome Extension when active

B. Downloading Log:

There are two ways to stored private information to make public. The one way is to downloading log module. Downloading log module send request to web mining service which is created on the server. Web mining service is a server side program to handle client request.

As shown in figure 3 the information downloaded on Downloading log module and at the same time user could do the analysis on them. For conceptual implementation, we have designed few graphs which provide information in graph form. Hit graph shows total number of time a web page visited. Time graph shows a time required to download a webpage.

It gives a lots of information to public like which page is more important to particular web site. This information will be helpful to modify their own web site web page or help to implement their own API [2].

We can apply clustering algorithm on collected public log file to extract related information. For example: The people interest according to region, age, sex, product, website etc.

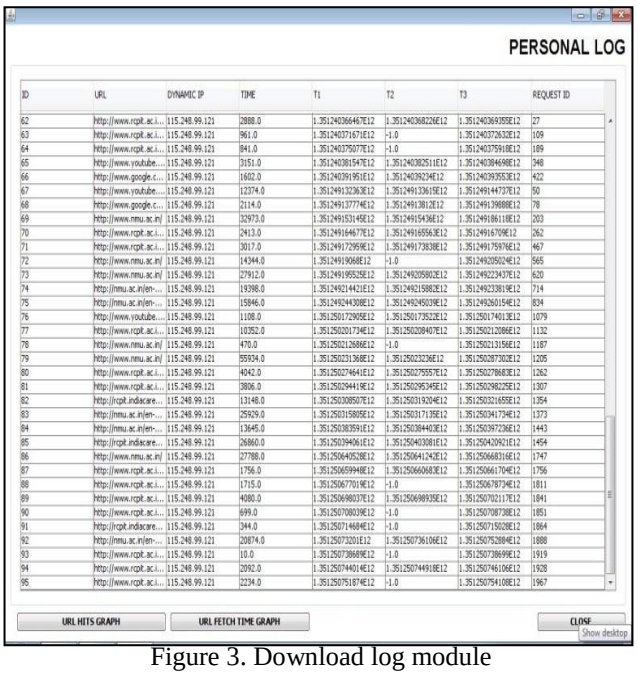

Figure 3. Download log module

III. RELATED WORK

Identifying Web browsing strategies is a crucial step in Website design and evaluation, and requires approaches that provide information on both the extent of any particular type of user behavior and the motivations for such behavior [5].Pattern discovery from web data is the key component of web mining and it converge algorithms and techniques from several research areas. Baraglia and Palmerini (2002) proposed a WUM system called SUGGEST that provide useful information to make easier the web user navigation and to optimize the web server performance. Liu and Keselj (2007) proposed the automatic classification of web user navigation patterns and proposed a novel approach to classifying user navigation patterns and predicting users' future requests and Mobasher (2003) presents a Web Personalizer system which provides dynamic recommendations, as a list of hypertext links, to users. Jespersen et al. (2002) [6] proposed a hybrid approach for analyzing the visitor click sequences. Jalali et al. (2008a [7] and 2008b [8]) proposed a system for discovering user navigation patterns using a graph partitioning model. An undirected graph based on connectivity between each pair of Web pages was considered and weights were assigning to edges of the graph. Dixit and Gadge (2010) [9] presented another user navigation pattern mining system based on the graph partitioning. An undirected graph based on connectivity between Referrer and URI pages was presented along with a preprocessing method to process unprocessed web log file and a formula for assigning weights to edges of the undirected graph. Ant-based clustering due to its flexibility and self-organization has been applied in a variety of areas from problems arising in e-commerce to circuit design, and text-mining to web-mining, etc (Jianbin et al., 2000. The various works proposed in this area with particular emphasize on web usage mining, clustering and

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

classification was provided in this section. In this present work, research work is one another attempt made to propose a hybrid system that uses clustering and classification methods to discover the user's navigation pattern and analyze them from the server's web log file.

IV. WEB CLUSTERING

Clustering is the process of assembling the data into classes or clusters so that objects within a cluster have high similarity in relationship to another, but are very dissimilar to objects in other clusters. Data clustering is under vigorous development and is applied to many application areas including business, biology, medicine, chemistry, etc. Owing to the huge amounts of data collected in databases, cluster analysis has recently become a highly active topic in data mining research [1][3]. For cluster analysis to work efficiently and effectively, as many literatures have presented, there are the following typical requirements of clustering in data mining

1. Scalability:
2. Ability to deal with different types of attributes:
3. Discovery of clusters with arbitrary shape.
4. Minimal requirements for domain knowledge to determine input parameters:
5. Ability to deal with noisy data:
6. Insensitivity to the order of input records:
7. High dimensionality:

The research is focused on finding user behavior by using efficient and effective cluster analysis.

A. Basic Clustering Step

- *Preprocessing and feature selection*

Most clustering models assume that n-dimensional feature vectors represent all data items. This step therefore involves choosing an appropriate feature, and doing appropriate preprocessing and feature extraction on data items to measure the values of the chosen feature set[5][8]. It will often be desirable to choose a subset of all the features available, to reduce the dimensionality of the problem space. This step often requires a good deal of domain knowledge and data analysis.

- *Similarity measure*

Similarity measure plays an important role in the process of clustering where a set of objects are grouped into several clusters, so that similar objects will be in the same cluster and dissimilar ones in different cluster. In clustering, its features represent an object and the similarity relationship between objects is measured by a similarity function. This is a function, which takes two sets of data items as input, and returns as output a similarity measure between them.

- *Clustering algorithm*

Clustering algorithms are general schemes, which use particular similarity measures as subroutines. The particular choice of clustering algorithms depends on the desired properties of the final clustering. Other considerations include the usual time and space complexity [8] [9]. A clustering algorithm attempts to find natural groups of components (or data) based on some similarity. The clustering algorithm also finds the centroid of a group of data sets. To determine cluster membership, most algorithms evaluate the distance between a point and the cluster centroid. The output from a clustering algorithm is basically a statistical description of the cluster centroid with the number of components in each cluster.

- *Result validation*

Do the results make sense? If not, we may want to iterate back to some prior stage. It may also be useful to do a test of clustering tendency, to try to guess if clusters are present at all; note that any clustering algorithm will produce some clusters regardless of whether or not natural clusters exist [10][11].

B. Clustering Algorithm:

- *Hierarchical algorithms:*

HA provide a hierarchical grouping of the objects. There exist two approaches, the bottom-up and the top-down approach [12]. In case of bottom-up approach, at the beginning of the algorithm each object represents a different cluster and at the end all objects belong to the same cluster. In case of top-down method at the start of the algorithm all objects belong to the same cluster which is split, until each object constitute a different cluster. A key aspect in these kinds of algorithms is the definition of the distance measurements between the objects and between the clusters. The advantage of the hierarchical algorithms is that the validation indices (correlation, inconsistency measure), which can be defined on the clusters, can be used for determining the number of the clusters.

- *Density-based algorithms:*

Start by searching for core objects, and they are growing the clusters based on these cores and by searching for objects that are in a neighborhood within a radius of a given object[8]. The advantage of these types of algorithms is that they can detect arbitrary form of clusters and it can filter out the noise.

5.2.3 Grid-based algorithms: GBA the grid-based algorithms

use a hierarchical grid structure to decompose the object space into finite number of cells [12] [8]. The advantage of this approach is the fast processing time that is in general independent of the number of data objects.

- *Ant-based Clustering*

Deneubourg et al. in [18] proposed ant-based clustering and sorting. In the case of ant-based clustering and sorting, two related types of natural ant behaviors are modeled. When clustering, ants gather items to form heaps. And when sorting, ants discriminate between different kinds of items and spatially arrange them according to their properties [19]. Lumer and Faieta in [20] proposed ant-based data clustering algorithm, which resembles the ant behavior described in [18].

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

- *Fuzzy Clustering Algorithm*

Fuzzy clustering algorithm is one of the approaches to derive user categories by capturing the similar user interests from web usage data available in log files. In particular CARD+ a fuzzy relational clustering algorithm that works on data quantifying similarity between user interest, with main two activities, and are the first method is to create the relation matrix containing the dissimilarity values among all pairs of users and the next approach is to categorize the user by grouping the similar user.

- *Graph*

Graph partitioning theoretic approach is presented by Perkowitz and Etzioni [21], who have developed a system that helps in making Web sites adaptive, i.e., automatically improving their Organization and presentation by mining usage logs. The core element of this system is a new Clustering method, called cluster mining, which is implemented in the Page Gather algorithm. Page Gather receives user sessions as input, represented as sets of pages that have been visited. Using these data, the algorithm creates a graph, as signing pages to nodes. An edge is added between two nodes if the corresponding pages co-occur in more than a certain number of sessions. Clusters are defined either in terms of cliques, or connected components. Clusters defined as cliques prove to be more coherent, while connected component clusters are larger, but faster to compute and easier to find. A new index page is created from each cluster with hyperlinks to all the pages in the cluster. The main advantage of Page Gather is that it creates overlapping clusters. Furthermore, in contrast to the other clustering methods, the clusters generated by this method group together characteristic features of the users directly. Thus, each cluster is a behavioral pattern, associating pages in a Web site. However, being a graph based algorithm, it is rather computationally expensive, especially in the case where cliques are computed.

- Page Cluster

In Page clustering algorithm page ratings are calculated, then the web pages with similar ratings still do not necessarily have similar contents or navigational functions. By taking into consideration the incoming links and the transition probabilities on them, to cluster Web pages having similar incoming links and ratings together to integrate with search results and give them more semantic meanings. We define incoming link similarity of two Web pages as the accumulated difference of transition probabilities on their incoming links. By setting a threshold, Web pages are clustered together based on both incoming links and ratings. The clustering algorithm reflects the observation that Web pages, that have links in a similar set of pages and receive a similar number of hits from these pages, tend to have similar contents or navigational functions. Each cluster of pages can be given a description based on concept learning.

CONCLUSION

In this paper we have extended our previous work, we have implemented a tool which share user log file with central database. The main goal of this paper is to analyzing hidden information from large amount of log data. This paper emphasizes on clustering among the different mining processes. We define various clustering algorithm for similar kind of web access pattern. These algorithms serve as foundation for the web usage clustering that were described and we conclude that web mining methods and clustering technique are used for self adaptive websites and intelligent websites to provide personalized service and performance optimization. The clustering algorithm is very helpful to extract useful information from the public log file. Through this mining project we want to show how these mining API and clustering algorithm can be useful to all people.

REFERENCES

[1] U.M. Patil and S.N. Pardeshi. "A survey on user future request prediction: Web Usage Mining," (IJETAE) International Journal of Emerging Technology and Advanced Engineering, Vol. 2 (3), pp. 121-124, March 2012.
[2] U.M. Patil and S.N. Pardeshi. "Central Web Mining Services – Public and Free access log files," (WJST) World Journal Science of Technology, Vol. 2 (3), pp. 38-41, 21 April 2012.
[3] Rajni Pamnani, Pramila Chawan,"Web Usage Mining: A Research Area in Web Mining," in ISCET 2010, Punjab, India, 2010, pp.73-77.
[4] J.Han and M.Kamber by Data Mining Concepts and Techniques.
[5] Bradley P S, Fayyad U M. "Refining Initial Points for Kmeans,Clustering Advances in Knowledge Discovery and Data Mining", MIT Press.
[6] Ruoming Jin , Anjan Goswami and Gagan Agrawal. "Fast and exact out-of-core and distributed k-means clustering Knowledge and Information Systems", Volume 10, Number 1/July, 2006.
[7] Renata Ivancsy, Ferenc Kovacs "Clustering Techniques Utilized in Web Usage Mining" International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006 (pp237-242)
[8] M.N. Murty, A.K. Jain, P.J. Flynn, "Data clustering: a review",ACM Computer. Survey. 31 (3) (1999) 64– 323.
[9] V.V.R. Maheswara Rao , Dr. V. Valli Kumari , Dr. KVSVN Raju "Understanding User Behavior using Web Usage Mining" International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 7
[10] M.N. Murty, A.K. Jain, P.J. Flynn, "Data clustering: a review",ACM Computer. Survey. 31 (3) (1999) 64– 323.
[11] Ajith Abraham,"Business Intelligence from Web Usage Mining" Journal of Information & Knowledge Management, Vol. 2, No. 4 (2003) 375-390
[12] Hengshan Wang, Cheng Yang, Hua Zeng " Design and Implementation of a Web Usage Mining Model Based On Fpgrowth and Prefixspan, Communications of the IIMA, Volume 6 Issue 2
[13] Renata Ivancsy, Ferenc Kovacs "Clustering Techniques Utilized in Web Usage Mining" International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006 (pp237-242)
[14] M.N. Murty, A.K. Jain, P.J. Flynn, "Data clustering: a review",ACM Computer. Survey. 31 (3) (1999) 64– 323.
[15] Bradley P S, Fayyad U M. "Refining Initial Points for Kmeans,Clustering Advances in Knowledge Discovery and Data Mining", MIT Press.
[16] Ruoming Jin , Anjan Goswami and Gagan Agrawal. "Fast and exact out-of-core and distributed k-means clustering Knowledge and Information Systems", Volume 10, Number 1/July, 2006.
[17] Ji He,Man Lan, Chew-Lim Tan,Sam-Yuan Sung, Hwee-BoonLow, "Initialization of Cluster refinement algorithms: a review and comparative study", Proceeding of International Joint Conference on Neural Networks[C].Budapest,2004.
[18] J. Deneubourg -L., S. Goss, N. Franks, A. Sendova-Franks, C. Detrain, L. Chrétien, The dynamics of collective sorting: robot-like

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

ants and ant-like robots. Proceeding of the first international conference on simulation of adaptive behavior, pp. 356–365, MIT Press, 1991.

[19] J. Handl, B. Meyer, Ant-based and Swarm-based clustering, Swarm Intelligence, 1, pp. 95–113, 2007.

[20] E. Lumer, B. Faieta, Diversity and adaptation in populations of clustering ants. Proceeding of the third international conference on Simulation of adaptive behaviour, pp. 501–508, MIT Press, 1994.

[21] Ji He,Man Lan, Chew-Lim Tan,Sam-Yuan Sung, Hwee-BoonLow, "Initialization of Cluster refinement algorithms: a review and comparative study", Proceeding of International Joint Conference on

Neural Networks[C].Budapest,2004.