Analysis of Web User Clustering based on Users' Access Behavior

Theint Theint Shwe
University of Computer Studies, Mandalay
theint2shwe@gmail.com

Abstract

World Wide Web overwhelms us with the immense amounts of widely distributed interconnected, rich and dynamic information. Provision of services to users correctly according to their needs is one of the most important issues in Web. However, provision of services to individual users' need is time consuming and overburden for the web site developers or administrator. Not only for the developers but also for the users, group-based service provision can fulfill this situation at the same time. In this paper, clustering algorithms: Self Organizing Map (SOM) and K-Means are used to analyze the users' access behavior. The correctness of the clustering algorithms is tested with two external validation indexes. Our implementation results show that SOM gives better results than K-Means.

1. Introduction

Web Mining aims to discover useful information or knowledge from the World Wide Web. Generally, Web Mining tasks can be categorized into three main types: Web Structure Mining, Web Content Mining and Web Usage Mining.

Web Usage Mining refers to the discovery of user access pattern from the web access logs, which records every click made by each user. Web Usage Mining applies many data mining techniques to discover usage pattern from the web access logs.

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar amongst them and dissimilar compared to objects in other groups. The quality of cluster can be accessed based on a measure of dissimilarity of objects, which can be computed for various types of data, including interval-scaled variables, or combinations of these variable types [1].

Cluster analysis is the organization of a collection of patterns usually represented as a vector of measurements, or a point in a multidimensional space. There are many clustering algorithms to perform clustering.

In this paper, we would like to perform clustering based on Web server access logs using Self Organizing Map. The performance of the SOM is compared with K-means algorithm with the help of two validation indexes. The more correct the clustering results, the better provision of services to group of users with the most similar interests. The motivation of using SOM is that it can analyze users' behavior more correctly and because of its visualization facility.

The purpose of the paper is to analyze the final results (clusters) how much they perform precisely from the other clustering method and which results the higher quality for clusters.

The contribution of the paper is to correctly characterize the users' behaviors by using SOM.

Because of seeing the behavior of the users correctly, future access behavior of users can be predicted correctly.

2. Related Work

K. R. Suneetha et.al [2] described the in depth analysis of Web Log Data of NASA web site to find information about a web site, top errors, potential visitors of the site etc. They stated that the obtained results of their study will be used in the further development of the web site in order to increase its effectiveness.

T. Morzy et.al [3] presented a new clustering algorithm Pattern-Oriented Partial Clustering (POPC) based on the general idea of agglomerative hierarchical clustering. Their clustering algorithm starts with a number of small clusters and merges them together to reach the given number of resulting clusters. They applied the Jaccard Coefficient to merge the clusters iteratively. They said that their results may lead to an improved organization of the web documents for navigational convenience.

D. Qi et. al [4] presented web page clustering using Self Organizing Map together with association rules based on the users' browsing history. They used the classical k-means algorithm to classify the URLs into clusters based on users' browsing history. They demonstrated that their experimental results feasible and effective.

G. Pallis et. al [5] described a framework for model-based cluster analysis for web users' sessions. It deals with the problem of assessing the quality of user session clusters in order to make inferences regarding the users' navigation behavior. They validated their clusters by using statistical test, which measures the distance of clusters' distribution to infer their dissimilarity and distinguishing level. They applied two real data sets: msnbc and csd data sets.

F. M. Facca et. al [6] presented a survey of the recent developments in web usage mining area that is receiving increasing attention from the Data Mining Community. They made the survey based on more than 150 papers since 2000 in this topic.

3. Background Theory

Web usage mining refers to the automatic discovery and analysis of patterns in click stream. The main source of raw data is the web access Log. Based on the Web access Log data, preprocessing including data cleaning (removing irrelevant data), user identification and sessionization are performed that is ready for clustering of web usage data. In this paper, users' access behavior is analyzed with SOM and K-Means.

3.1. Self Organizing Map (SOM)

The self-organizing map (SOM) is an excellent tool in exploratory phase of data mining. It projects input space on prototypes of a low-dimensional regular grid that can be effectively utilized to visualize and explore properties of the data. When the number of SOM units is large, to facilitate quantitative analysis of the map and the data, similar units need to be grouped, i.e., clustered [7].

The stages of the SOM algorithm can be summarized as follows:

1. ***Initialization*** – Choose random values for the initial weight vectors **w***j*.
2. ***Sampling*** – Draw a sample training input vector **x** from the input space.
3. ***Matching*** – Find the winning neuron *I*(**x**) with weight vector closest to input vector.
4. ***Updating*** – Apply the weight update equation

$$\Delta w_{ij} = \eta(t) T_{j.\tau(x)}(t)(x_i - w_{ji}) \quad (1)$$

5. ***Continuation*** – keep returning to step 2 until the feature map stops changing.

3.2. K-Means

The K-Means algorithm is one of the best known partitional clustering algorithms. It is also the most widely used among all clustering algorithms due to its simplicity and efficiency. Given a set of data points and the required number of k clusters (k is specified by the user), this algorithm iteratively partitions the data into k clusters based on a distance function. The processing steps of the K-Means algorithm are as follow:

1. Choose k data points as the initial centroid (cluster centers)
2. Repeat
3. For each data point $x \in D$ do
4. Compute the distance from x to each centroid
5. Assign x to the closest centroid
6. End for
7. Re-compute the centroid using the current cluster membership
8. Until the stopping criteria is met

3.3. External Validation Indexes

Clustering validation is a technique to find a set of clusters that best fit natural partitions (number of clusters) without any class information. In this paper, two types of external validation indexes are used to determine the correct number of groups from a dataset [8].

3.3.1. Entropy

Entropy measures the purity of the clusters' class labels. Therefore, if all clusters consist of objects with only a single class label, the entropy is zero. But, as the class labels of objects in a cluster become more varied, the entropy increases. To compute the entropy of the dataset, the class distribution of the objects in each cluster is needed to calculate as follows:

$$E_j = \sum_i p_{ij} \log(p_{ij}) \quad (2)$$

The sum is taken over all the classes. The total entropy for a set of clusters is calculated as the weighted sum of the entropies of all clusters as:

$$E = \sum_{j=1}^{m} \frac{n_j}{n} E_j \quad (3)$$

where n_j is the size of the cluster j, m is the number of clusters, and n is the total number of data points.

3.3.2. Purity

For each cluster, the purity is calculated as

$$P_j = \frac{1}{n_j} Max_i(n_j^i) \quad (4)$$

The overall purity of clustering solution can be obtained as a weighted sum of the individual cluster purities and it can be calculated as:

$$Purity = \sum_{j=1}^{m} \frac{n_j}{n} P_j \quad (5)$$

where n_j is the size of cluster j, m is the number of clusters, and n is the total number of objects.

3.4. System Architecture

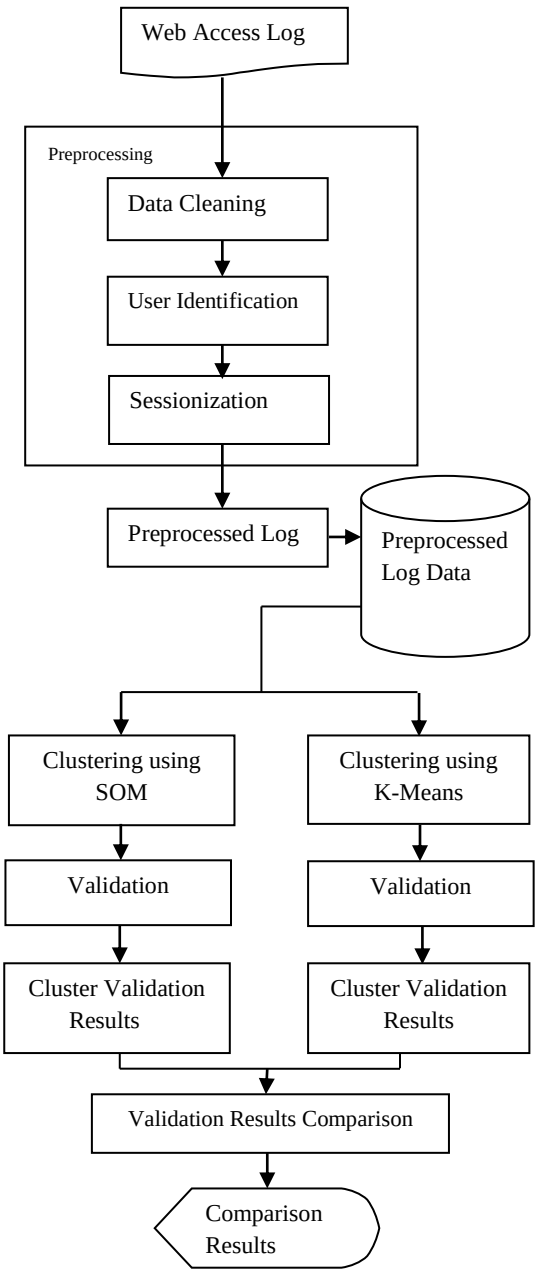

Figure1. Flow diagram of the proposed system

Web access log data set is used as the input data set. In the preprocessing step, data cleaning is performed firstly. The second step is user identification. In this system, the first attribute is defined as the user. The sessionization is third step. Maximum of thirty minute period is defined as one session for each user. After this stage, preprocessed log data are stored in the database.

After preprocessing, Self Organizing Map and K-Means are used for clustering. The same data set and the same amount of data are used for both algorithms. And the clustering accuracy of the algorithms is tested with entropy and purity measures. The analysis results of both algorithms are shown as comparison results.

4. Data Source for Clustering

Our proposed system uses the data source as web logs from NASA July 95 web access log dataset. A log file records activity information when web user submits a request to a Web Server.

4.1. Web Log Structure

Web Server logs are plain text (ASCII) files, that is independent from the server platform. The following is a fragment of the web access log format from NASA.

199.72.81.55 - - [01/Jul/1995:00:00:01 -0400] "GET /history/apollo/ HTTP/1.0" 200 6245

This reflects the information as follows:

- Remote IP address or domain name: An IP address is a 32 bit host address defined by the internet Protocol; a domain name is used to determine a unique internet address for any host on the internet. One IP address is usually defined for one domain name.
- Authuser: User name and password if the server requires user authentication
- Entering date and time
- Modes of request: GET, POST, or HEAD method of Common Gate Way Interface.

- Status: the HTTP status code returned to the client e.g., 200 is “ok” and 404 is “not found”
- Bytes: the content length of the document transferred.

4.2. Preprocessing of Web Access Logs

Web log data is usually diverse, voluminous and must be assemble into a consistent, integrated and comprehensive view in order to be used for pattern discovery. The following is the overview of data preprocessing for clustering.

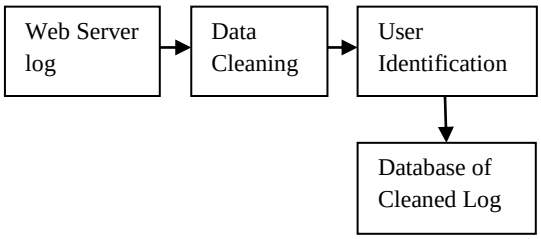

Figure2. Overview of Web Server Log Data Preprocessing

In data cleaning process, the entries that have status of “error” or “failure” should be removed, and then some access records generated by automated search engine agent should be identified and removed from the access log. This process removes requests concerning non-analyzed resources such as images, multimedia files, and page style files. By filtering out useless data, the log file size will be reduced to use less storage space and to facilitate upcoming tasks.

User Identification means identifying individual users by observing their IP address. If there is new IP address, then there is a new user. If the IP address is same but the operating system or browsing software is different, each different agent type for an IP address represents a different user.

A user session is defined as a sequence of requests made by a single user over a certain navigation period and a user may have a single (or multiple) session(s) during a period of time. If the time between page requests exceeds a certain limit, there is another user session even though IP address is the same. A Web user’s historical access pattern may have more than one session because the user may visit a Web site from time to time and spend arbitrary amount of time between consecutive visits.

NASA July 1995 log file of one hour is tested in this work and the results of preprocessed data is shown in table 1.

Table1. Results of preprocessed web server log data

Server log file	NASA July 1995
Duration	1 hour
Number of Entries	3570
Entries after preprocessing	3480
Number of Unique Users	346

After preprocessing, web user clustering is performed with SOM and K-Means to identify the similar web access behavior users. Then, the correctness of the clusters is measured with the two external validation indexes: entropy and purity. Implementation of the system is presented in the following section.

5. Web User Clustering System Implementation

Preprocessed web accessed logs are used as the input dataset in this system. We remove the users who use the only one web site at only one time because we assume that who doesn’t have special interest to use the internet and he may not be ordinary user. The users who access the internet two or more times to the same web site or two or more web sites at least one time are considered.

After preprocessing, the numbers of unique users who access the web are obtained. From

this, we remove the users who don't have special interest or regular access of web sites. Therefore we reduced the number of unique users from 346 to 312.

5.1. Clustering Using SOM

We perform the clustering using SOM with varied size of datasets. The clustering accuracy of the algorithm is validated with entropy and purity measures as follow:

Table2. Cluster validation results of SOM

Number of Users	Entropy	Purity
50	1.168	0.064
100	0.426	0.027
150	0.535	0.019
200	0.417	0.014
250	0.353	0.010
300	0.356	0.009
>300	0.354	0.009

5.2. Clustering Using K-Means

The same number of unique users is also clustered using K- Means. The results of the validation indexes of K-means algorithm is shown in table3 as:

Table3. Cluster validation results of K-means

Number of Users	Entropy	Purity
50	1.285	0.080
100	0.426	0.027
150	0.535	0.020
200	0.440	0.0153
250	0.353	0.012
300	0.378	0.01
>300	0.390	0.01

5.3. Comparison of SOM and K-Means

According to our implementation results, SOM can cluster more correctly than K-Means. SOM automatically produces the number of clusters in different number of users. Therefore, number of clusters equal to SOM is defined in K-Means to compare their respective results. The clustering results comparison of the two algorithms using entropy is shown in figure3 as follow:

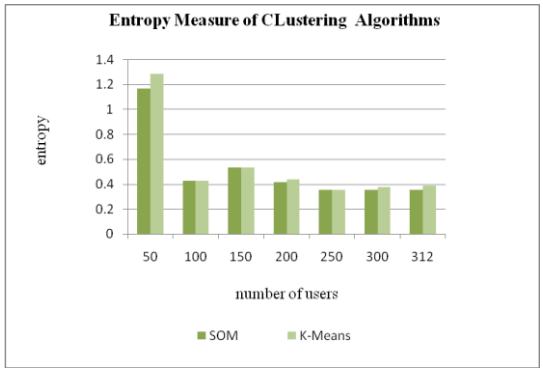

Figure3. Entropy measure of SOM versus K-Means

The cluster validation comparison results of SOM and K-Means measured by purity is presented in the following figure:

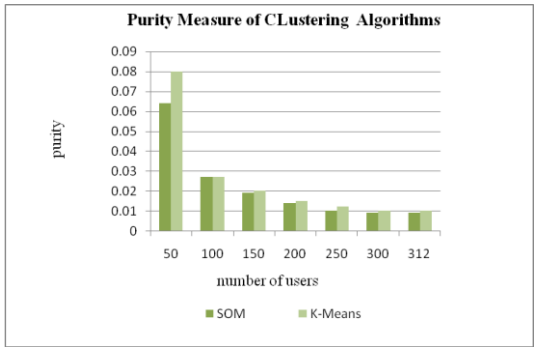

Figure4. Purity Measure of SOM and K-Means

6. Conclusion and Future Work

In order to facilitate data availability and accessibility, and at the same time to meet user preferences, clustering of user with similar browsing behavior is necessary. Because of the advantages of clustering, we can group the users with the most similar interests. Therefore,

provision of services to users with their needs can give the satisfaction of users. As a consequence of this, we can conclude that the more correct clustering results can give the more user satisfaction.

In the future, we would like to hybridize the SOM with Genetic Algorithm to get the optimal cluster results for the purpose of giving highest user satisfaction and accessibility. Then, Prediction of future access would be performed with Hidden Markov Model.

References

[1] B. Liu, Web Data Mining, Exploring Hyperlinks, Contents, and Usage Data, ISBN-10 3-540-37881 -2 Springer Berlin Heidelberg, New York.

[2] K. R. Suneetha, R. Krishnamoorthi,Identifying User Behavior by Analyzing Web Server Access Log File, IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.

[3] T. Morzy, M. Wojciechowski, M. Zakrzewicz, Web User Clustering, Poznan University, Institute of Computer Science.

[4] D. Qi,Clustering using Web Log, Computer Science Department and School of Information Technology, Lamar University and Illinois State University.

[5] G. Pallis, L. Angelis, A. Vakakli, Validation and Interpretation of Web Users' Sessions Clusters, Information Processing and Management Department of Informatics, Aristotle University of TGhessaloniki, Greece, 2006.

[6] F. M. Facca, P. L. Lanzi, Mining Interesting Knowledge from Web Logs: A Survey, Data and Knowledge Engineerin, 53 (2005) 225-241.

[7] S. Chakrabarti, Mining the Web, Discovering Knowledge from Hypertext Data, ISBN-13:978-1-55860-754-5, Indian Institute of Technology, Bombay.

[8] E. Rendon, I. M. Abundez, C. G. S.D. Zagal, A. Arizmendi, E. M. Quiroz, E. Arzateh, A Comparison of Internal and External Validation Indexes, Applications of Mathematics and Computer Engineering, ISBN:978-960-474-270-7.