User Behavior Modeling based on Adaptive Gaussian Mixture Model

M.Rekha Sundari
Asst. Professor, Dept. of CSE
GITAM UNIVERSITY
Visakhapatnam

Prasad Reddy PVGD
Professor, Dept. of CS&SE
ANDHRA UNIVERSITY
Visakhapatnam

Y.Srinivas
Professor, Dept. of IT
GITAM UNIVERSITY
Visakhapatnam

ABSTRACT
A remarkable technological development has been witnessed due to the recent advancements in the area of science and technology. This made the users to access the Internet and store the information retrieved in various databases at various servers across the globe, making World Wide Web as an information gateway. This technological growth has thrown a challenging situation to both the user and the business man. The various resources that are available through the Internet made the users to choose different alternatives which in turn makes it necessary for the businessman to bring out new strategies and alternatives so as to attract the users. In order to overcome these challenging situations, web usage mining brings out a solution to the business man by analyzing different user patterns that are available in the web. Many tools are available for this purpose but majority of the tools lag in including the complete details regarding the web log data .To overcome this disadvantage, in this paper a model based on Adaptive Gaussian Mixture Model, an extension of Gaussian Mixture Model (GMM) to interpret the user navigation behavior is brought out. The proposed model is applied on user traffic data taken from msnbc.com.

Keywords
Web usage mining, Gaussian Mixture Model, Adaptive Gaussian Mixture Model.

1. INTRODUCTION
Web usage mining is increasing enormously due to its ability to discover interesting user navigational patterns that can be applied to many real world problems such as improving Web sites/pages, making additional topic or product recommendations, user/customer behavior studies, etc. Web usage mining is generally a three level architecture, where in the first level the data accessed is preprocessed, in the second level the patterns discovered from the log files extracted from the Internet are analyzed, in the third level the identification of the user pattern is carried out. To identify the unstructured data and retrieve the exact information within a short duration of time is a challenging task. In order to overcome this difficulty it is needed to analyze the navigation pattern of the user on which some plausible comparisons and conclusions can be made. In some situations it is very much essential to correlate some of the patterns in order to find the exactness in the data. Several models have been proposed in the literature for the analysis of user navigational behavior; these models are based on the methodologies like Least Square Method [1], Factor Analysis [2], EM Algorithm [3], Variable Length Markovian Chains [4]. However these models are effective in the presence of data where the deeper knowledge about the data is not under consideration and these models have very less impact to identify the correlation between different patterns of a user which is very much essential in building a decision. Off late models based on association rules [5] are highlighted but these models discard the vital information regarding web log sessions. Hence to interpret and derive exact solutions, mixture models such as GMM [6] [7] have become popular. The main disadvantage with GMM is that large training data is necessary for decision making and another disadvantage is noise gets increased. It is necessary to utilize a model which minimizes the noise. Hence in this paper Adaptive Gaussian Mixture Model is used. The effectiveness of the model is that it includes a filter for minimizing the noise and it is inclusive of a covariance matrix which helps in identifying the correlation between the patterns. In order to have a comprehensive study of data, the data is to be clustered such that a structured information can be pooled among the clusters and based on these clusters of users behaviors, the data can be classified appropriately[8]. In order to cluster the data, partitioning based clustering method, K-means algorithm is used .The clustering is carried out based on the users and in order to initialize the value of K, sum of squared error criterion is used. To classify the data according to users' navigational pattern, Adaptive GMM is utilized. The developed model is tested using msnbc dataset. The rest of the paper is organized as follows. Section 2 of the paper highlights the Adaptive GMM. In section 3, K-means clustering algorithm is presented. The methodology along with experimentation results is presented in section 4 and finally section 5 of the paper summarizes the results with conclusion.

2. ADAPTIVE GAUSSIAN MIXTURE MODEL
In this paper, Adaptive GMM is utilized to classify the dynamic user behavior. This method is utilized since it identifies the users patterns based on the correlations. The probability density function (PDF) of Adaptive GMM is given by

$$f(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma}}\left(\frac{\mu}{N}\frac{e^{-(x-\mu)^{n+1}}}{2\sigma^2}+\frac{\mu}{N}\sum_{i=1}^{n}\left(\frac{x-\mu}{2\sigma^2}\right)^{n+1}\right)$$

Where N is total number of samples present in the data, n is number of samples in each cluster, x is a d-component feature vector, μ is the d-component vector containing the mean of each feature, σ is the d × d covariance matrix. It characterizes the dispersion of the data on the d-dimensions of the feature vector. The diagonal element σ_{ii} is the variance of x_i, and the non-diagonal elements are the covariances between the features.

We consider a set X of n observations of d features each, $X = [X_1, X_2, X_3\ldots, X_n]$. Each X_i is a d-dimensional feature vector, $X_i = [X_{i_1}, X_{i_2}, X_{i_3} \ldots, X_{i_d}]$. The covariance matrix of X is represented as

$$\begin{pmatrix} \text{cov}(x_{i_1}, x_{i_1}) & \text{cov}(x_{i_1}, x_{i_2}) \ldots. & \text{cov}(x_{i_1}, x_{i_d}) \\ \text{cov}(x_{i_2}, x_{i_1}) & \text{cov}(x_{i_2}, x_{i_2}) \ldots. & \text{cov}(x_{i_2}, x_{i_d}) \\ \text{cov}(x_{i_3}, x_{i_1}) & \text{cov}(x_{i_3}, x_{i_2}) \ldots. & \text{cov}(x_{i_3}, x_{i_d}) \\ \vdots & & \\ \vdots & & \\ \text{cov}(x_{i_d}, x_{i_1}) & \text{cov}(x_{i_d}, x_{i_2}) \ldots. & \text{cov}(x_{i_d}, x_{i_d}) \end{pmatrix}$$

The covariance is calculated using the following formula

$$\text{Cov}(x_i, y_i) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n}$$

3. K-MEANS ALGORITHM

K-means clustering is a partitional clustering approach where each cluster is associated with a centroid. Each point in the data is assigned to the cluster with closest centroid. Closeness is measured by Euclidean distance, Cosine similarity, Correlation etc. Number of clusters K must be predetermined and the algorithm referred is from [9] and is presented as follows.
Algorithm K-means (K, D)

a. Select K points as the initial centroids.
b. Repeat
c. Form K clusters by assigning all points to the closest centroids.
d. Recompute the centroid of each cluster.
e. Until the centroids don't change.

Initial centroids are often chosen randomly, the main problem with K-means algorithm is the choice of 'K', the initial clusters. Falsifying the value of K leads to bad clustering. To overcome this problem, Sum of Squared Error (SSE) is used as as a measure, for the goodness of a clustering structure. For each point, the error is the distance to the nearest cluster. To get SSE, these errors are squared and summed.

$$\text{SSE} = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2(m_i, x)$$

Where x is a data point in cluster C_i and m_i is the mean for cluster C_i. One easy way to reduce SSE is to increase K, the number of clusters. A good clustering with smaller K can have a lower SSE than a poor clustering with higher K.

4. METHODOLOGY & EXPERIMENTAL RESULTS

We applied the learning techniques described to a large Web navigation dataset. The data comes from Internet Information Server (IIS) logs for msnbc.com and news-related portions of msn.com [10] for the entire day of September, 28, 1999 (Pacific Standard Time). Each sequence in the dataset corresponds to page views of a user during that twenty-four hour period. Each event in the sequence corresponds to a user's request for a page. Requests are not recorded at the finest level of detail that is, at the level of URL, but rather, they are recorded at the level of page category (as determined by a site administrator). The categories are "front page", "news", "tech", "local", "opinion", "on-air", "misc", "weather", "health", "living", "business", "sports", "summary", "bbs" (bulletin board service), "travel", "msn-news", and "msn-sports". Any page requests served via a caching mechanism were not recorded in the server logs and, hence, not present in the data. The full data set consists of approximately one million sequences (users), with an average of 5.7 events per sequence. The data are available online at kdd.ics.uci.edu/databases/msnbc/msnbc.html. Table 1 contains an example of 6 user sequences.

Table 1: A sample of user sequences

User	Sequence
1	frontpage→frontpage
2	news
3	tech→news→local→news→tech→tech
4	news→ health→health→sports
5	frontpage→news→business→travel→weather
6	frontpage→sports→sports→weather

Each category is associated with an integer starting with "1". For example, "frontpage" is associated with 1, "news" with 2, and "tech" with 3. Each row in the column named "sequences" describes the hits of a single user. For example, the first user hits the "frontpage" category twice, and the second user hits the "news" category once. Table 2 illustrates the numerical representation of table 1. The data in this form is preprocessed and each entry in the processed data gives the count of user visits to that particular page. If the user doesn't visit that particular page that entry will be treated as 0. The sample of this form is illustrated in table 3.The total number of pages visited by each user for a sample of 350 users is graphically represented in fig1.

Table 2: Sample user sequences represented as integers

User	Sequence
1	1, 1
2	2
3	3, 2, 4, 2, 3, 3
4	2, 9, 9, 12
5	1, 2, 11, 15, 8
6	1, 12, 12, 8

Table 3: Sample user sequences represented by their Number of visits to a Page

User	frontpage	news	Tech	Local …
1	2	0	0	0
2	1	0	0	0
3	0	2	3	1
4	0	1	0	0
5	1	1	0	0
6	1	0	0	0

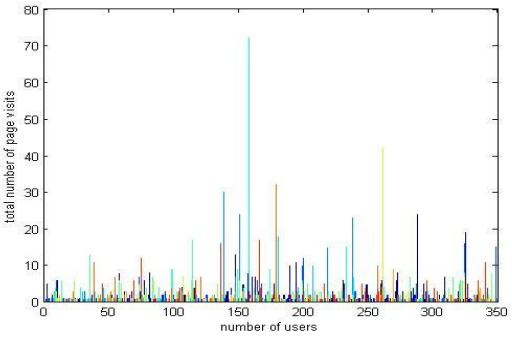

Fig 1: user's traffic flow in 24 hours

We learned models using a training set of 100,000 sequences sampled at random from the original one million. In order to model the users' behavior the data is first clustered into patterns based on user behavior using the K-means algorithm. Using this data, 100 clusters are generated as described in Section 3.For a sample of 350 users, the value of SSE varies with number of clusters as shown in fig 2. The value of SSE decreases as K increases but in order to achieve better clustered data a smaller K with lower SSE is selected. In this sample a K value of 8 is chosen.

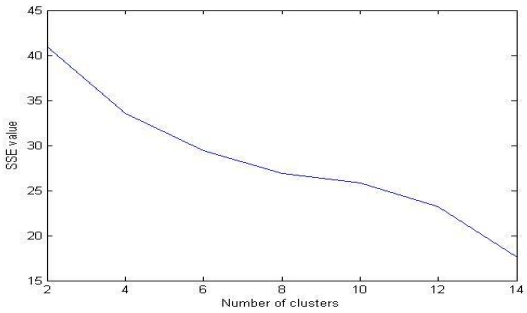

Fig 2: variation of SSE with K

4.1 Classification Using Adaptive GMM:

In order to classify an unknown sample X, $P(X/C_i)P(C_i)$ is evaluated for each C_i. Sample X is then assigned to the class C_i, if and only if $P(X/C_i)P(C_i) > P(X/C_i)P(Cj)$ for $1 \leq j \leq K$, $j \neq i$. In other words it is assigned to the class C_i for which $P(X/C_i)P(C_i)$ is maximum. Where $P(X/C_i)P(C_i)$ is calculated as below:

$$P(C_i) = \frac{\text{Number of samples in training set belonging to } C_i}{\text{Total Number of samples}}$$

$$P(X/C_i) = \prod_{k=1}^{n} P(x_k/C_i)$$

To classify the users the probability density function is extracted for each user and for each of the clusters derived a similar dimensionality PDF matrix is constructed. In order to classify the users' navigational behavior, a user is considered and his navigational pattern is analyzed based on the maximum likelihood estimate of the PDF.

5. CONCLUSION

In this paper a framework on Adaptive GMM is presented to interpret the user navigational pattern. In order to highlight the methodology an existing database msnbc.com is considered. The developed method as shown a drastic improvement in identifying the navigational pattern of user compared to GMM. In this paper the navigation pattern is modeled using the mixture model, however to obtain the robust performance, the initial estimates are to be modeled using EM algorithm to obtain the final updated parameters. Using these updated equations for μ, π and σ, we have to train the data to obtain the robust navigation patterns; this work is proposed as the future extension. The performance evaluation of the developed model is to be tested using evaluation metrics by taking into consideration of different data sets.

6. REFERENCES

[1] S.S.Patil "A Least Square Approach to Anlayze Usage Data for Effective Web Personalization" IJCTA | SPT-OCT 2011, Int. J. Comp. Tech. Appl., Vol 2 (5), 1192-1196.

[2] Yanzan Kevin Zhou,Bamshad Mobasher, "Web user segmentation based on a mixture of factor analyzers" EC-Web'06 Proceedings of the 7th international conference on E-Commerce and Web Technologies,pp.11-20,2006.

[3] Mustapha Norwati, Jalali Manijeh, Jalali Mehrdad," Expectation Maximization Clustering Algorithm for User Modeling in Web Usage Mining Systems", European Journal of Scientific Research,Jun2009, Vol. 32 Issue 4, p467.

[4] Jose Borges, Mark Levene, **"**Evaluating Variable-Length Markov Chain Models for Analysis of User Web Navigation Sessions", Journal IEEE Transactions on Knowledge and Data Engineering archive Volume 19 Issue 4, April 2007, Pages 441-452.

[5] A.Jebaraj Ratnakumar,"An Implementation Of Web Personalization Using Web Mining Techniques", Journal Of Theoretical And Applied Information Technology, 2005 - 2010 Jatit.

[6] Xiaofei He, Deng Cai, Yuanlong Shao, Hujun Bao, and Jiawei Han," Laplacian Regularized Gaussian Mixture Model for Data Clustering", IEEE Transactions On Knowledge And Data Engineering, Vol. 23, No. 9, September 2011.

[7] Rahul Telang, Peter Boatwright, And Tridas Mukhopadhyay" A Mixture Model for Internet Search-Engine Visits", Journal of Marketing Research,Vol. XLI (May 2004), 206–214.

[8] Kate A. Smith, Alan Ng " Web page clustering using a self-organizing map of user navigation patterns", Decision Support Systems 35 (2003) 245– 256.

[9] "Introduction to Data Mining", Pang-Ning Tan, Vipin Kumar, Michael Steinbach.

[10] Heckerman, David \MSNBC.com Anonymous Web Data Set", UCI KDD Archive, http://archive.ics.uci.edu/ml/ datasets/MSNBC.com+Anonymous+Web+Data