Optimizing the Behaviour of Web Users Through Expectation Maximization Algorithm and Mixture of Normal Distributions

R. Sujatha[1]
Department of Mathematics
SSN College of Engineering, Chennai, India

P. Saravanan[3]
Research Scholar, Bharathiyar University
G B Pant Govt. Engineering College, New Delhi, India

D. Nagarajan[2]
Department of Mathematics
Hindustan Institute of Technology & Science, Chennai, India

J. Kavikumar[4]
Department of Mathematics and Statistics, Faculty of Applied Science and Technology
Universiti Tun Hussein Onm Malaysia, Malaysia

***Abstract*—The proposed work is to analyse the user's behaviour in web access. Worldwide, the web users are browsing through different websites every second. Aim of this paper is to identify the behaviour of user's in a time bound using an Expectation Maximization (EM) algorithm and the maximum likelihood estimates of the model parameters. A novel approach based on Mixture normal distribution is used to discuss the percentage of user along with web page frequency.**

***Keywords*—EM algorithm; maximum likelihood; mixture normal distribution; web page frequency**

I. Introduction

The number of accessible web pages grows significantly; it is becoming increasingly difficult for users to find documents that are relevant to their particular needs. Users must either browse through a large hierarchy of concepts to find information or submit a query to a widely available search engine [1]. Therefore, the process of understanding the user's navigation behaviour is challenging but fundamental in improving web query answering, link structure and in simplifying navigation through a large number of individual webpages. The web sites are making great effort to understand user's behaviour and make the web sites easy to access. To achieve this goal, researchers proposed lots of approaches to use web usage data.

Researchers studied this topic from different points of view. A broadly applicable algorithm for computing maximum likelihood estimates from incomplete data or hidden data is presented at various levels of generality [2], [3], [4] and [5]. EM algorithm to retrieve the complete scatterer trajectory matrix is discussed in [6].

Mixture distributions are extensively used to model a wide variety of empirical phenomena, in diverse fields such as biology, anthropology, psychology, economics, and marketing. Overviews of mixture distributions and many examples of their applications are given by [7]. Mixtures of t-distributions and their numerous variants are discussed by [8], [9], [10] and [11].

This work is part of the project SR/S4/MS:816/12, Science and Engineering Research Board, Department of Science and Technology, India.

EM algorithm and finite mixture model is discussed in [12]. The EM-GMM algorithm targets reconfigurable platforms, with five main contributions [13].

In this paper we have studied the web user's behaviour using EM algorithm. The web page access is predicted using mixture normal variate. The remaining of the paper is organized as follows. In section 2 we present the concept of EM algorithm. Section 3 gives the application of EM algorithm to the selected database. In Section 4 we deal with mixture normal variate and its application in predicting web page frequency and finally concluded in Section 5.

A. Data Base

The data is taken from the educational institute of Sri Sivasubramaniya Nadar College of Engineering (SSNCE), Chennai, Tamil Nadu, India.

II. Concept of EM Algorithm

The EM algorithm is a general method, to estimate the parameters using maximum-likelihood estimation.

EM algorithm is used when the data is incomplete, due to the limitations of the observation process. The algorithm consists of two steps. This is diagrammatically shown in Fig.1.

Given a set of parameter estimates the E-step calculates the conditional expectation of the complete-data log likelihood given the observed data and the parameter estimates. In this step, using conditional expectation, given the observed data and current estimate, the missing data is estimated. Given complete-data log likelihood, the M-step finds the parameter estimates in order to maximize the complete-data log likelihood from the E-step. These two steps are iterated until convergence. Convergence is assured since the algorithm is guaranteed to increase the likelihood at each iteration.

The sequences of web users randomly access the SSN educational institute website for various departments. Web user access 6 engineering departments with 4 independent various attributes in each department. For application of EM algorithm

the dataset corresponds to page views of a user. To predict the accessibility of various departments among users EM Algorithm is applied. From the Internet browsing logs, we could gather the following information about a web user the frequency.

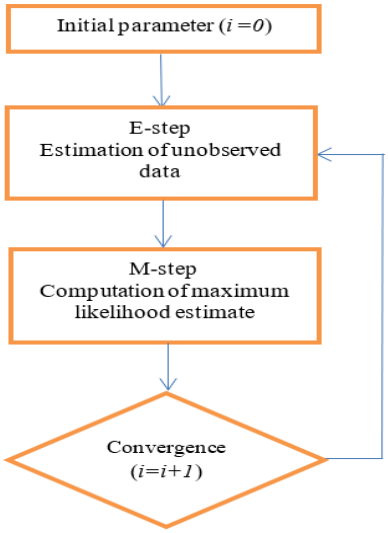

Fig. 1. EM Algorithm

III. Application of EM Algorithm

The sessions are grouped based on the user's profile. The sessions are grouped as various departments, namely, EEE,

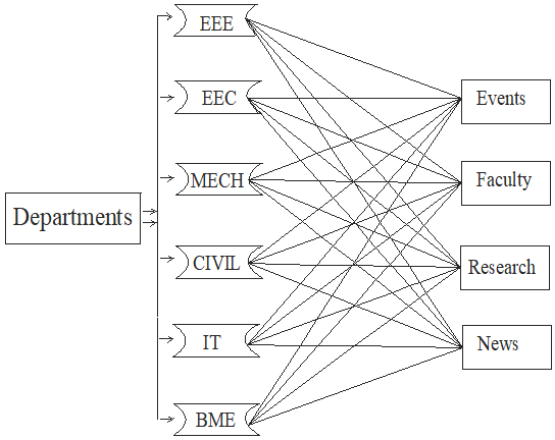

Fig. 2. Grouped webpages

ECE, MECH (MEC), CIVIL (CIV), IT and BME. The considered webpages are Events, Faculty (Fac), Research (Res) and News. The webpages considered is shown in Fig. 2. To determine the proportion of usage of various departments, we determine the likelihood of the webpage access and the sessions by using EM algorithm.

TABLE I. Initial and E-Step of the Web User

Dept.	EEE	ECE	MEC	CIV	IT	BME
Event	0.487	0.390	0.335	0.405	0.361	0.303
Fac	0.252	0.170	0.262	0.227	0.468	0.278
Res	0.162	0.317	0.012	0.025	0.117	0.177
News	0.121	0.121	0.387	0.341	0.053	0.240

TABLE II. Final and M-Step of the Web User

Dept.	EEE	ECE	MEC	CIV	IT	BME
Event	0.527	0.422	0.363	0.405	0.361	0.347
Fac	0.291	0.150	0.264	0.224	0.498	0.214
Res	0.201	0.345	0.114	0.051	0.196	0.200
News	0.147	0.151	0.314	0.342	0.053	0.296

The initial values *i.e.*, Expectation values are depicted in Table I as the values for E-step. By application of the Maximization step, the updated values are shown in Table II. Based on the calculations, the accessibility for each department can be determined and the results are depicted in Fig. 3.

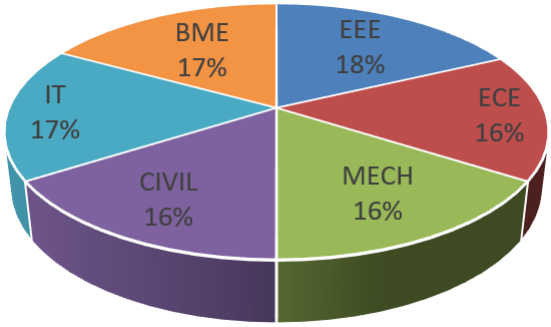

Fig. 3. Web User Activity

IV. Mixture of Normal Variates to Predict Percentage of User and Web Page Frequency

It is essential to predict the percentage of usage and web page frequency to understand the accessibility and popularity of the website among users. In this paper, we have used mixture of normal variates for this purpose.

Mixture of normal variates is used in statistical methods. Random vector x has a normal variate and it can be written as linear combination of variables from vectors x, all the samples of x variables from normal variates. It is independently distributed with zero covariance. The density function of a mixture of two univariate normal distributions is $f(y;w) = pf_1(y;w) + (1-p)f_2(y;w)$, where $f_j(y;w) = \frac{1}{\sigma_j}\phi\left(\frac{y-\mu_j}{\sigma_j}\right), j = 1$ and $\phi(.)$ is the standard normal distribution [14], [15], [16] and [17]. The interpretation of this system consists of mixture of two population and p lies between zero and one. The component of two mixture normal variates $\sum_j \sigma_j^2\, i$ where i is the unit matrix $N(x_n/X_n, \sigma_j^2 i) = \frac{1}{(2\pi)^{\frac{1}{2}}}\frac{1}{\sigma_j}$.

If $\sigma_j \to 0$ then the term goes to infinity. The variance of mixture components are finite and finite probability to all points. While other components can shrink onto the data point thus contributing the data point increasing additive value to the log likelihood. Two mixture A, B of normal distribution with mean μ_a, μ_b and standard deviations σ_a, σ_b to take mixture of distribution p and q where $0 \le p \le 1$, $q = 1-p$. Therefore the mixture of mean is $\mu_{ab} = (p \times \mu_a) + (q \times \mu_b)$ [18]. The mixture of the resulting normal curve is estimated using MATLAB and the results are shown in Fig. 4. From the graph, shown in Fig. 4 we observe that variance and mean are

different. It is an equally weighted average of the bell-shaped probability density function of the two normal distributions. The weights were not equal, the resulting distribution could still be bimodal but with peak of different height and split-up is a linear combination of two normal variates with means 11 and 18; variance 0, 1 and 4, given by *0.5N(11,1)+0.5N(18,1)* and *0.75(11,0)+0.25N(18,4)*.

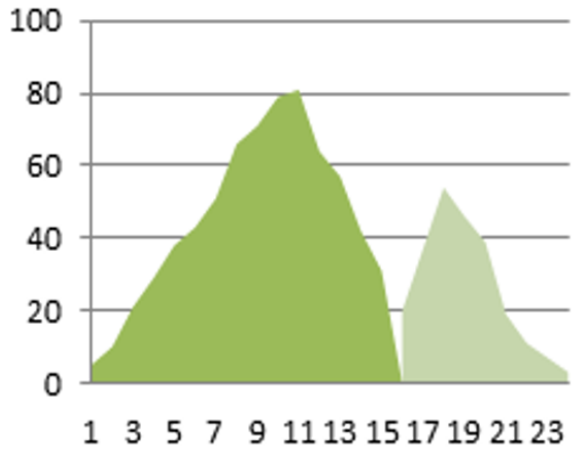

Fig. 4. Percentage user vs web page frequency

V. CONCLUSION

In this paper we proposed a method of using EM algorithm to predict the accessibility of webpages among users. We have used mixture distribution to identify web page frequency and percentage of users. Based on these the popularity of the web pages among users can be studied. The frequently accessed web pages can be updated. The study reveals that EEE department is popular among the users and is accessed much frequently when compared to the other departments. The study can be extended to centrality of networks.

REFERENCES

[1] G. Pallis, L. Angelis, A. Vakali and J. Pokorny, "A Probabilistic Validation Algorithm for Web Users Clusters", IEEE International Conference on Systems, Man and Cybernetics, pp. 4129–4134, 2004.

[2] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, pp.1–38, 1977.

[3] D. Chauveau, "A stochastic EM algorithm for mixtures with censored data", Journal of Statistical Planning and Inference, vol. 46, pp.1–25, 1995.

[4] R. A. Levine, and G. Casella, "Implementations of the Monte Carlo EM algorithm", Journal of Computational and Graphical Statistics, vol. 10(3), pp.422–439, 2001.

[5] S. Balakrishnan, M. J. Wainwright and B. Yu, "Statistical Guarantees For The EM Algorithm: From Population To Sample-Based Analysis", The Annals of Statistics, vol. 45(1), pp.77–120, 2017.

[6] L. Liu, F. Zhou, X. Bai, J. Paisley and H. Ji, "A Modified EM Algorithm for ISAR Scatterer Trajectory Matrix Completion", IEEE Transactions on Geoscience and Remote Sensing, vol. 56(7), pp. 3953-3962, 2018.

[7] G. J. McLachlan and D. Peel, Finite Mixture Models, Wiley Series in Probability and Statistics: Applied Probability and Statistics. John Wiley & Sons, 2000, New York.

[8] C. Archambeau and M. Verleysen, "Robust Bayesian clustering", Neural Networks, vol. 20(1), pp. 129–138, 2007.

[9] C. M. Bishop and M. Svensen, "Robust Bayesian mixture modelling", Neurocomputing, vol. 64, pp. 235–252, 2005.

[10] J. L. Andrews and P. D. McNicholas, "Model-based clustering, classification, and discriminant analysis via mixtures of multivariate tdistributions", Statistics and Computing, vol. 22(5), pp. 1021–1029, 2012.

[11] F. Forbes and D. Wraith, "A new family of multivariate heavy-tailed distributions with variable marginal amounts f tail weight: application to robust clustering", Statistics and Computing, vol. 24(6), pp. 971– 984, 2014.

[12] Y. Li and Y. Chen, "Research on Initialization on EM Algorithm Based on Gaussian Mixture Model", Journal of Applied Mathematics and Physics, vol. 6, pp. 11-17, 2018.

[13] C. He, H. Fu, C. Guo, W. Kuk and Guangwen, "A fully pipelined hardwared design for Gaussian mixture models", IEEE Transactions on Computers, vol. 66(11), pp. 1837–1850, 2017.

[14] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", Very Large Data Base Endowment, pp. 487-499, 1994.

[15] J. Xiao and Y. Zhang, "Clustering of Web Users Using Session-based Similarity Measures", International Conference on Computer Networks and Mobile Computing, pp. 223-228, 2001.

[16] S. Vishwakarma, S. Lade, M. K. Suman and D. Patel, "Web user prediction by integrating markov model with different features", International Journal of Engineering Research and Science & Technology, vol. 2(4), pp.74-83, 2013.

[17] M. Deshpande and G. Karypis, "Selective Markov Models for Predicting Web-Page Accesses", ACM Transactions on Internet Technology, vol. 4(2), pp.168-184, 2004.

[18] E. Meijer and J. Y. Ypma, "A Simple Identification Proof for a Mixture of Two Univariate Normal Distributions", Journal of Classification, vol. 25, pp.113-123, 2008.