www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 6 Issue 2 Feb. 2017, Page No. 20210-20214
Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i2.08

A Web Navigation Frame Work to Identify the Influence of Faculty on Students Using Datamining Tecniques

Rekha Sundari.M[1], Srininvas.Y[2], Prasad Reddy.PVGD[3]
[1]GITAM University, GITAM Institute of Technology,
Rushikonds, Visakhapatnam, Andhra Pradesh,India
mr.sundari@gmail.com.
[21]GITAM University, GITAM Institute of Technology,
Rushikonds, Visakhapatnam, Andhra Pradesh,India
ysrinivasit@rediffmail.com
[3] Andhra University, Department of CS&SE,
Visakhapatnam, Andhra Pradesh,India
Prasadreddy.vizag@gmail.com

Abstract:Analyzing student web browsing behavior is a challenging task. This paper mainly focuses on a methodology to identify the influencing factor that has driven a student towards navigating a particular web site. Most of the research in this direction is untouched to estimate the influence of faculty on the student's behavioral patterns. In this work we focus on a novel statistical approach based on adaptive Gaussian mixture model, where the data clustered is given as input to the model to classify the student navigating pattern. The concept of regression analysis is used to find the relationship between student's navigational behavior and faculty's experience and rating. This article considers a real time dataset of GITAM University for experimentation.

Keywords: Gaussian mixture model, Regression, Clustering, Classification;

1. Introduction

Present day students flock to the internet as the primary tool for researching any topic. In most of the cases, the students get influenced by different factors and these influences make them drive towards goal setting. This paper examines the student behavior on web and estimates the influence of a faculty / teaching on their behavior[3]. Perfect procedures are needed to find out this inclination, thus, faculty rating given by the students and experience of the faculty in a particular field of specialization were taken into consideration. In this work in-depth analysis of different kinds of students specifically related to the engineering group is concentrated[1][4].

Despite the fact that students join in an engineering college with the goal to receive degree, their future dreams are different[5][8]. This being the backdrop, we at GITAM University, have conducted a

Brain storming session of about one week to motivate the students regarding various opportunities that are ahead before them. To understand the impact of this session on the students, we have provided specific IP address and monitored their web navigation pattern thereof.

A statistical framework is developed for clustering the students into different groups basing on their navigation pattern. The objective of this prediction is to characterize the student behavior relating to a particular cluster. After classifying the students to one of the predefined groups, regression analysis is measured to find out the relationship between the students browsing pattern and the influence of an experienced teacher or teacher who has been rated good.

2. Dataset

This data set contains the sessionzed data for the gitam.edu web server (http://www.gitam.edu).This data is based on the students navigating pattern for a period of one week during which the motivating session is delivered. The following snap shot presents a view of the dataset before preprocessing.

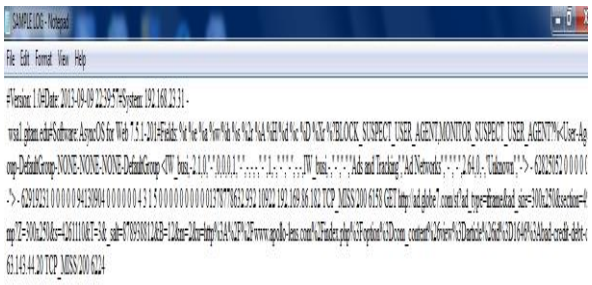

Figure 1: Data set before preprocessing

During preprocessing data is cleaned by removing whitespaces, images, audio and video files. The cleaned data is preprocessed for identifying sessions, different users; unique URL's or page views and time spent by them in each page view using different preprocessing techniques. As a first step in preprocessing all the Unique URL's in the dataset are identified and assigned unique identification numbers. In the second step, the dataset is presented with the student id, together with his access sequence. As sequence is not of priority in the proposed work, the third step is carried out, in which each entry in the dataset is redesigned such that if the student visits the page the entry is represented by 1 else 0.

The Tables below elucidate the outputs of the three preprocessing steps respectively.

Table 1 shows the list of distinct URL's U, with their corresponding id's.

ID	URL
1	www.goabroad.com
2	www.afsusa.org/study-abroad/faq/
3	www.uniguru.co.in/
4	www.usnews.com
5	www.i20fever.com
6	www.ustraveldocs.com
7	www.topuniversities.com/university-
8	rankings/world...rankings/2013
9	www.4icu.org › Asia
10	100bestschools.net/eng/schools/
14	www.cse.iitk.ac.in/acad/adm_mt.html
….	…….
60	http://testfunda.com/CAT

Table.1: Distinct URL's U

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 2: A sample of user sequences

User,Page	P1	P2	P3...	P60
1	1	0	0	0
2	1	0	0	0
3	0	1	1	1
4	0	1	0	0
5	1	1	0	0
6	1	0	0	0

Table 3: Sample user sequences represented by their visit to a Page

3. Clustering:

As the dataset under consideration consists different navigation patterns , to mine the relevant browsing patterns clustering is considered. In our work we use this technique to identify groups in students with browsing patterns[6].

3.1 Latent class Analysis:

Latent class analysis is considered for clustering the students into groups. This is a model based cluster analysis technique that uses mixture of probability distributions to assign a data point to a cluster.

The basic latent class cluster model is given by

$$P(y_n \mid \theta) = \sum_{1}^{s} \pi_i P_j (y_n / \theta_j) \qquad (1)$$

Where y_n is the nth observation of the manifest variables, S is the number of clusters and Π_j is the prior probability of membership in cluster j. P_j is the cluster specific probability of y_n given the cluster specific parameters Θ_j. For each data point LCA calculates the probability to the cluster membership. After the model is built data points are assigned to the clusters that have higher probability.

After performing clustering, we identified six different groups in students and when we analyzed the url's the groups have browsed we identified that the students of six different clusters concentrate on six different goals as their future endeavors.

Cluster 1 Research

Cluster 2 Placement

Cluster 3 GATE

Cluster 4 GRE

Cluster 5 CAT

Cluster 6 Entrepreneurship

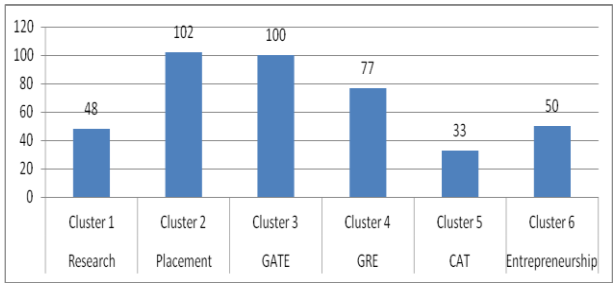

Figure 2: strength of all the clusters the students belong to

4. Classification:

The objective of classification in this context is to assign the student to a particular cluster that describes the student behavior more relatively with his similar group.

4.1 Adaptive Gaussian Mixture Model:

AGMM is an improvised version of GMM in which the probability density is a function of input vector x, mean μ,

standard deviation σ as equivalent to GMM[2] and with two additional parameters n and N. Where N is total number of samples present in the data and n is number of samples in each cluster.

The probability density function of Adaptive GMM is given by:

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

Assume that each sample x is a d dimensional vector. Let x = $[x_1, x_{2...} x_i..., x_d]$. As the features are independent, the mean and standard deviation are also calculated independently. For a cluster with n samples, the mean μ and standard deviation σ of each feature x_i is calculated by taking the x_i's of all the samples in that particular cluster. So the mean and the standard deviation are given by the equations 2 and 3

$$\mu=\sum_{j=1}^{n}\frac{x_{ij}}{n} \quad (3)$$

$$\sigma=\sqrt{\sum_{j=1}^{n}\frac{(x_{ij}-\mu)^2}{n-1}} \quad (4)$$

After classification the new student with access sequence is assigned to one of the above mentioned clusters.

5. Regression:

A data mining (machine learning) technique used to fit an equation to a dataset is called Regression. This is a statistical model for estimating relationships among variables. Regression analysis with a single explanatory variable is termed simple regression or linear regression Linear regression, uses the formula of a straight line (y = mx + b) and determines the appropriate values for m and b to predict the value of y based upon a given value of x. Multiple regression, allow the use of more than one input variable and allow for the fitting of more complex models, such as a quadratic equation[7].

In this article the concept of regression analysis is used to estimate the regression pattern of X(Independent variables, various navigation patterns of the student) and Y(Dependent variables rating and experience of the faculty).In our work basing on the regression analysis we made an attempt to identify the regression of X and Y and thereby identify the significance and inclination of a faculty experience and rating on student browsing behavior. Considering the multiple linear regression analysis by taking independent variables (1 to 48) and considering the Rating and Experience as dependent variables, we have the regression lines as

y = 8.04 + 28.0 X2 + 193 X3 - 177 X4 - 318 X5 + 317 X6 - 17.0 X7 + 12.2 X9 - 12.3 X10 - 16.9 X11 + 2.1 X13 - 3.7 X15 - 114 X17 + 196 X18 + 40 X19 - 57 X20 - 81.2 X21 - 38.4 X23 + 6.6 X25 - 10.9 X26 + 0.1 X27 + 3.8 X28 + 9.7 X29 - 1.2 X31 - 65.1 X33 + 61.4 X34 + 41.4 X35 - 57.9 X36 + 5.7 X37 + 68.5 X39 + 294 X41 - 314 X42 - 301 X43 + 318 X44 - 11.3 X45 + 3.9 X47 +6.56X48

y = 1.75 + 1.57 X2 - 7.0 X3 + 8.8 X4 - 3.27 X5 + 1.63 X6 + 1.60 X7 - 17.6 X9+ 15.2 X10 - 4.18 X11 + 5.78 X13 - 12.0 X15 - 0.24 X17 + 1.29 X18 - 4.5 X19+ 7.0 X20 + 0.54 X21 + 1.94 X23 + 1.61 X25 + 0.42 X26 + 15.7 X27 - 15.1 X28 + 1.45 X29 + 0.08 X31 + 2.47 X33 + 0.06 X34 - 0.55 X35 + 3.36 X36 + 0.32X37+ 0.78 X39 + 6.53 X41 - 9.80 X42 - 9.77 X43 + 17.1 X44 - 4.03 X45 - 3.28X47-3.271X48.

The corresponding results are presented in Table-4 and Table-5

Predictor constant	Coef	SE Coef	T	P
X1	1.7468	0.9541	1.83	0.092
X2	1.572	1.777	0.88	0.394
X3	-6.95	13.29	-0.52	0.61
X4	8.84	13.43	0.66	0.523
X5	-3.271	8.157	-0.4	0.695
X6	1.629	2.086	0.78	0.45
X7	1.6	1.521	1.05	0.313
X9	-17.646	7.302	-2.42	0.033
X10	14.226	6.97	2.18	0.049
X11	-4.176	3.918	-1.07	0.308
X13	5.778	4.284	1.35	0.202
X15	-12.019	5.665	-2.12	0.055
X17	-0.244	2.74	-0.09	0.93
X18	1.289	2.479	0.52	0.612
X19	-4.46	13.08	-0.34	0.739
X20	6.96	12.79	0.54	0.596
X21	0.54	1.476	0.37	0.721
X23	1.944	1.761	1.1	0.291
X25	1.615	2.707	0.6	0.562
X26	0.417	2.663	0.16	0.878
X27	15.67	10.82	1.45	0.173
X28	-15.06	10.62	-1.42	0.182
X29	1.454	1.955	0.74	0.471
X31	0.085	1.801	0.05	0.963
X33	2.466	3.467	0.71	0.491
X34	0.06	4.193	0.01	0.989
X35	-0.545	4.858	-0.11	0.912
X36	3.363	3.666	0.92	0.377
X37	0.318	2.587	0.12	0.904
X39	0.784	2.751	0.29	0.78
X41	6.529	4.967	1.31	0.213
X42	-9.805	7.785	-1.26	0.232
X43	-9.77	7.994	-1.22	0.245
X44	17.09	12.05	1.42	0.182
X45	-4.032	3.453	-1.17	0.266
X47	-3.276	3.526	-0.93	0.371
X48	-3.271	8.157	-0.4	0.695

Table 4: Results obtained after considering rating as dependent Variable

Predictor constant	Coef	SE Coef	T	P
X1	8.036	2.347	3.42	0.005
X2	28.05	57.41	0.49	0.634
X3	193.4	147.4	1.31	0.214
X4	-176.8	141.8	-1.25	0.236
X5	-318.4	294.9	-1.08	0.302
X6	317.1	278.2	1.14	0.277
X7	-17.01	58.48	-0.29	0.776
X9	12.22	32.22	0.38	0.711
X10	-12.31	40.23	-0.31	0.765
X11	-16.88	23.31	-0.72	0.483
X13	2.12	23.75	0.09	0.93
X15	-3.73	27.18	-0.14	0.893
X17	-114.2	150.5	-0.76	0.462
X18	196.3	156.5	1.25	0.234
X19	40.1	159.7	0.25	0.806
X20	-56.5	168.9	-0.33	0.744
X21	-81.25	44.12	-1.84	0.09
X23	-38.44	63.45	-0.61	0.556
X25	6.56	22.45	0.29	0.775
X26	-10.87	22.49	-0.48	0.638
X27	0.08	28.59	0	0.998
X28	3.82	31.6	0.12	0.906
X29	9.68	18.36	0.53	0.608
X31	-1.23	14.74	-0.08	0.935
X33	-65.1	77.05	-0.84	0.415
X34	61.37	94.83	0.65	0.53
X35	41.39	71.57	0.58	0.574
X36	-57.88	78.75	-0.74	0.476
X37	5.75	38.86	0.15	0.885
X39	68.49	66.74	1.03	0.325
X41	293.6	268	1.1	0.295
X42	-314.3	267.3	-1.18	0.262
X43	-301.4	261	-1.15	0.271
X44	318.1	269.8	1.18	0.261
X45	-11.27	33.94	-0.33	0.746
X47	3.89	33.27	0.12	0.909
X48	6.56	22.45	0.29	0.775

Table 5: Results obtained after considering experience as dependent variable

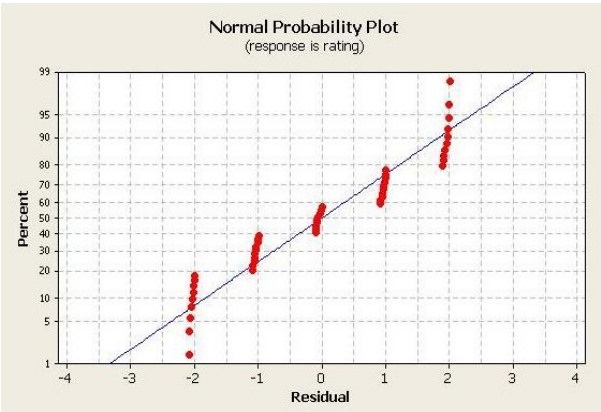

Figure 3: The results obtained from Table:4

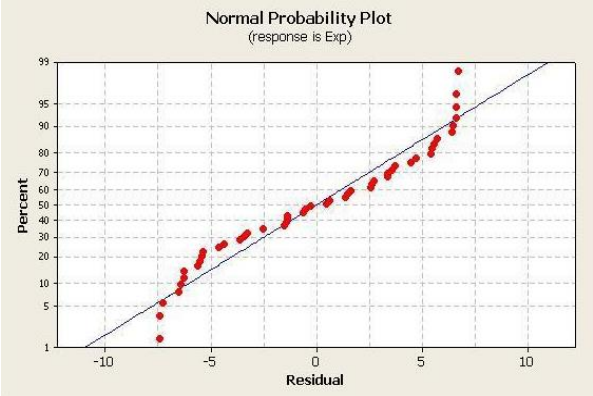

Figure 4: The results obtained from Table:4

From the above figures, it can be clearly seen that there is a considerable impact of experience on the students' behavioral pattern rather than the impact of the faculty having rated as good for a particular semester or a course.

6. CONCLUSION:

This paper presents the process of identifying different student clusters in a pool of students concentrating on different goals for their future, dynamically classifying a new student to one of the predefined clusters based on his behavioral pattern and then using regression to retrieve the effect of faculty experience and rating on students browsing behavior. Using regression analysis we concluded that the experience of the faculty impacts more on students behavior.

REFERENCES

[1] Alaa El-Halees, "Mining Students Data to Analyze Learning Behavior: A Case Study." Department of Computer Science, Islamic University of Gaza P.O.Box 108 Gaza, Palestine alhalees@iugaza.edu.ps .,2008;

[2] Daniel M. Steinberg"Towards autonomous habitat classification using Gaussian Mixture Models" The 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems october 18-22, 2010, Taipei, Taiwan.

[3] Eva M. Thury"Analysis of Student Web Browsing Behavior: Implications for Designing and Evaluating Web Sites" SIGDOC, page 265-270. 1998;

[4] Kunyanuth Kularbphettong, Cholticha Tongsiri"Mining Educational Data to Support Students' MajorSelection" World Academy of Science, Engineering and Technology,International Journal of Computer, Information, Systems and Control Engineering Vol:8 No:1, 2014;

[5] Luis Talavera and Elena Gaudioso "Mining Student Data To Characterize Similar Behavior Groups In Unstructured Collaboration Spaces" In Proc. European Conf. Artificial Intelligence 2004 ;

[6] O. A. Abbas, "Comparisons between Data Clustering Algorithms", International Arab Journal of Information Technology", Vol. 5(3), 2008, pp. 320-325.

[7] Prof.Dr.P.K.Srimani and Mrs. Malini M Patil"Linear Regression Model for Edu-mining in TES" International Journal of Conceptions on Electrical and Electronics Engineering Vol. 1, Issue 1, Oct 2013;

[8] Sergio Duarte Torres etal "Analysis of Search and Browsing Behavior of Young Users on the Web" ACM Transactions on Embedded Computing Systems, Vol. 0, No. 0, Article 0, Publication date: 2010.